Leveraging Semantic Annotations for Event-focused Search & Summarization

Dissertation
submitted towards the degree Doctor Engineering (Dr.-Ing)
of the Faculty of Mathematics and Computer Science
of Saarland University

by

Arunav Mishra

Saarbrücken
September 2017

Day of Colloquium	12 / 03/ 2018
Dean of the Faculty	Univ.-Prof. Dr. Frank-Olaf Schreyer

Examination Board

Chair of the Committee	Univ.-Prof. Dr. Dietrich Klakow
First reviewer	Prof. Dr. Klaus Berberich
Second reviewer	Prof. Dr. Gerhard Weikum
Third reviewer	Prof. Dr. Claudia Hauff
Academic Assistant	Dr. Rishiraj Saha Roy

"Intelligence is not the ability to store information,
but to know where to find it."

-Albert Einstein

Dedicate to my wonderful teachers and loving family . . .

Acknowledgements

I would like to express my deepest gratitude to *Klaus Berberich* for giving me an opportunity to work under his guidance. This work is made possible with his unconditional support, expert scientific advice, and futuristic vision. However, the encouraging aspect of working under him was the exceptional freedom he granted to pursue challenging problems from various fields of information science (retrieval, summarization, and spatiotemporal text mining). In addition, our common interest in music that often triggered very interesting conversations made work even more enjoyable.

I am extremely thankful to *Gerhard Weikum* for supporting me throughout my Master's and Ph.D studies. His high standards of conducting research constantly inspired and trained me to become a better researcher. I also thank the additional reviewers and examiners, *Dietrich Klakow* and *Claudia Hauff* for providing valuable feedback for further improvements of this work.

I acknowledge that this work would have not been possible without the influence, teachings, and guidance of several people. Firstly, I am extremely grateful to *Martin Theobald* for first introducing me to the department group and presenting me an opportunity to work on a Master's thesis under his guidance. His teachings motivated and prepared me for my Ph.D studies. Secondly, I offer a special thanks to *Avishek Anand* and *Sarath Kumar Kondreddi* for their constant support and guidance.

The institute not only presented a high class academic but also a very lively and social environment which made this journey very enjoyable. It was made possible by the wonderful friends and colleagues that I made and met here, especially - *Shubhabrata, Mittul, Kai, Amy, Yusra, Kaustubh, Saurabh, Kashyap, Johannes, Luciano, Dragan, Kiril,* and *the list goes on.* I also thank *Petra Schaaf* and all other staff members within the Max Planck Institute for being very supportive towards the students and abstracting them from all fine-grained administrative issues.

Last but not least, I am extremely grateful to my parents, *Ramakanta Mishra* and *Saswati Mishra,* and my brother *Anubhav Mishra* for always supporting me in all my endeavors. Finally, I believe the biggest contributor in the success of this work is my wife, *Bhawna Kalra,* and my son *Neer Mishra,* who have always been by my side.

Abstract

Today in this Big Data era, overwhelming amounts of textual information across different sources with a high degree of redundancy has made it hard for a consumer to retrospect on past events. A plausible solution is to link semantically similar information contained across the different sources to enforce a structure thereby providing multiple access paths to relevant information. Keeping this larger goal in view, this work uses Wikipedia and online news articles as two prominent yet disparate information sources to address the following three problems:

- We address a linking problem to connect Wikipedia excerpts to news articles by casting it into an IR task. Our novel approach integrates time, geolocations, and entities with text to identify relevant documents that can be linked to a given excerpt.
- We address an unsupervised extractive multi-document summarization task to generate a fixed-length event digest that facilitates efficient consumption of information contained within a large set of documents. Our novel approach proposes an ILP for global inference across text, time, geolocations, and entities associated with the event.
- To estimate temporal focus of short event descriptions, we present a semi-supervised approach that leverages redundancy within a longitudinal news collection to estimate accurate probabilistic time models.

Extensive experimental evaluations demonstrate the effectiveness and viability of our proposed approaches towards achieving the larger goal.

Kurzfassung

Im heutigen Big Data Zeitalters existieren überwältigende Mengen an Textinformationen, die über mehrere Quellen verteilt sind und ein hohes Maß an Redundanz haben. Durch diese Gegebenheiten ist eine Retroperspektive auf vergangene Ereignisse für Konsumenten nur schwer möglich. Eine plausible Lösung ist die Verknüpfung semantisch ähnlicher aber über mehrere Quellen verteilter Informationen, um dadurch eine Struktur zu erzwingen, die mehrere Zugriffspfade auf relevante Informationen, bietet. Vor diesem Hintergrund benutzt diese Dissertation Wikipedia und Onlinenachrichten als zwei prominente aber dennoch grundverschiedene Informationsquellen um die folgenden drei Probleme anzusprechen:

- Wir adressieren ein Verknüpfungsproblem, um Wikipedia-Auszüge mit Nachrichtenartikeln zu verbinden und das Problem in eine Information-Retrieval-Aufgabe umzuwandeln. Unser neuartiger Ansatz integriert Zeit- und Geobezüge sowie Entitäten mit Text, um relevante Dokumente, die mit einem gegebenen Auszug verknüpft werden können, zu identifizieren.
- Wir befassen uns mit einer unüberwachten Extraktionsmethode zur automatischen Zusammenfassung von Texten aus mehreren Dokumenten um Ereigniszusammenfassungen mit fester Länge zu generieren, was eine effiziente Aufnahme von Informationen aus großen Dokumentenmassen ermöglicht. Unser neuartiger Ansatz schlägt eine ganzzahlige lineare Optimierungslösung vor, die globale Inferenzen über Text, Zeit, Geolokationen und mit Ereignis-verbundenen Entitäten zieht.
- Um den zeitlichen Fokus kurzer Ereignisbeschreibungen abzuschätzen, stellen wir einen semi-überwachten Ansatz vor, der die Redundanz innerhalb einer langzeitigen Dokumentensammlung ausnutzt, um genaue probabilistische Zeitmodelle abzuschätzen.

Umfangreiche experimentelle Auswertungen zeigen die Wirksamkeit und Tragfähigkeit unserer vorgeschlagenen Ansätze zur Erreichung des größeren Ziels.

Contents

Chapter 1

Introduction

1.1 Motivation

Today in this digital age, the global news industry is going through a drastic shift as the broadband penetration exponentially rises and new devices for delivering digital content become affordable. This is due to a substantial increase in online news consumption as the general public gains access to sources, such as news portals and archives, made available over the World Wide Web. For example, a user can easily access a local news archive by simply connecting through broadband on a mobile device. However, this ease of access to overwhelming amounts of information from diverse sources (typically large unstructured textual document collections) has made it difficult for a user to retrospect on past events that have spanned over considerable amount of time and received large media coverage. This is referred to as *information overloading* or having too much information. One remedy to this information overloading problem is to come up with a better structure and organization of the information across different sources so as to aid effective exploration and retrospection of past events.

One traditionally considered plausible solution to organize information across (and within) large document is to automatically establish links between documents with semantically similar information [7, 40, 57, 59, 178, 207]. Such a linking introduces structure on large (Web) collections with unstructured data and provides efficient access to information in multiple ways. Moreover, with the links between documents in place, a user can use an information retrieval (IR) system or *search engine* to identify initial documents with relevant information, and then navigate to directly associated or linked documents to efficiently satisfy her information needs [7]. Further, automatic linking systems can leverage IR techniques to identify similar documents and in turn IR systems can leverage the links to further improve their quality. However, standard IR techniques based on syntactic matching have been found to be insufficient for achieving high linking quality. What is missing in such techniques is the semantic component to improve the document content representation [178].

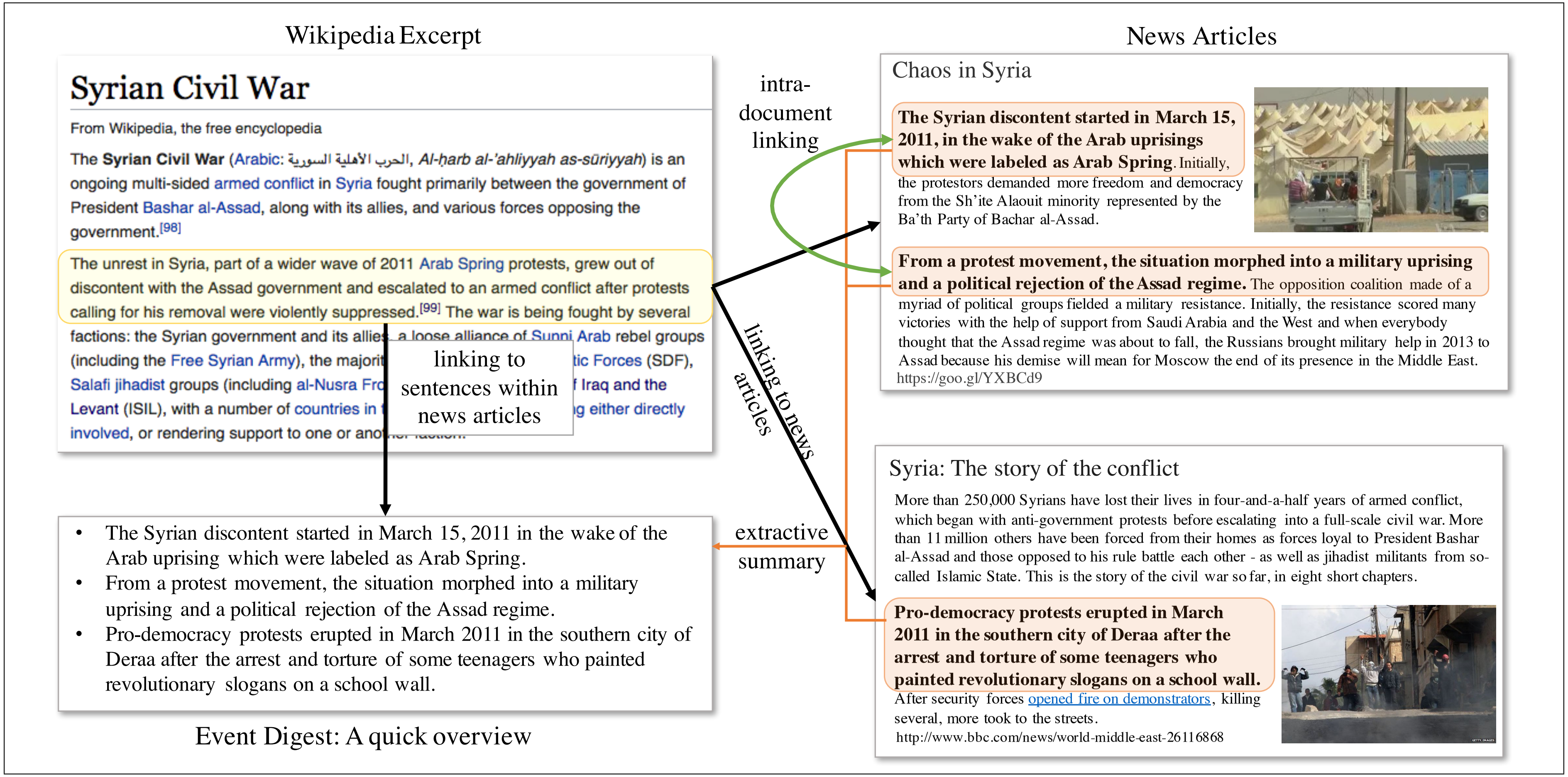


Figure 1.1 Linking a Wikipedia excerpt describing an aspect of "Syrian Civil War" to news articles at different granularities

In a longitudinal news collection, such as the New York Times [4], there often exist many articles that report on different aspects of an important event with long ramifications. Consider "European Debt Crisis" and "Syrian Civil War" as examples of such events that have been extensively covered by the media. A standard search engine (e.g. Google News) will retrieve a large number of relevant news articles for such events (issued as queries). For example, Google News retrieves 8,810 relevant New York Times articles[1] for the query "syrian civil war". In such a case, a user who is retrospecting on the event has to still sift through the large number of documents so as to develop a full understanding. An important power tool that has been effective in dealing with this information overloading problem is an automatic text summarization system [34, 63, 83, 116, 120, 137, 172, 221]. When dealing with large amounts of information spread across multiple documents, short informative summaries giving an overview can facilitate efficient consumption of the information. In this context, a query-focused extractive summarization [83] system can additionally accept a query (e.g. an event description) and generate a summary by selecting relevant and informative text units (typically sentences) from a set of documents (e.g. retrieved by a search engine).

From the perspective of linking information on past events, a query-focused extractive summary generated from a set of *target* documents, for a text unit taken from a *source* document (by treating it as a query) can be considered as linking the source and target documents at a finer textual granularity. Further, text units selected from within a single document can be connected thereby representing intra-document linking. Similar ideas have been leveraged in the past [181] where intra-document links are generated to improve summarization task. We motivate that an extractive summarizer can spot relevant and non-redundant textual units which can improve the linking quality.

As a primary source of information, news articles on the Web provide contemporary reports on events of global or local importance as they happen. For a long time, news articles have been archived for future generations as a part of our cultural heritage by media houses themselves (e.g., the archives of The New York Times go back until 1851[2]), national libraries, or efforts such as the Internet Archive[3]). Other examples large news collections are the English Gigaword [5] and New York Times Annotated corpus [4] are two examples of public news archives, while ClueWeb09 [2] and ClueWeb12 [3] are web crawls including articles that report on newsworthy events.

As another orthogonal source of information, the free encyclopedia Wikipedia has emerged as one of the largest reference websites with the English Wikipedia attracting 7.8 billion monthly average page views as of July 2017[4]. Wikipedia articles often summarize events, seminal for the central entity (topic), as passages or sentences by abstracting from the fine-grained details that mattered when the events took place. Figure 1.1

[1] accessed on 21/08/2017 22:00 hr
[2] http://www.nytimes.com/ref/membercenter/nytarchive.html
[3] https://archive.org/
[4] https://tools.wmflabs.org/siteviews

illustrates such a sentence from the article central to the "Syrian Civil War" that abstractly describes the beginning of the event. Moreover, since they are edited collaboratively and continuously, its account on past events can be seen as a reflection of their *collective memory* [210] and how it changes over time.

Wikipedia and news articles as prominent sources of information, individually lack in providing complete clarity on multi-faceted events. On one hand, it becomes difficult for readers of Wikipedia to dig deeper into the details of an event, which are missing from encyclopedia articles. For example, the detail about "killing of teenagers" highlighted from the second news article illustrated in Figure 1.1 is abstracted from the Wikipedia article. On the other hand, when reading a past news article, the implication or background of what is reported is not always apparent. What is missing are connections between Wikipedia and past news articles. With Wikipedia-news connections in place, readers of Wikipedia can refer to news for fine-grained details and readers of news may resort to Wikipedia to get a better understanding of the big picture. Figure 1.1 illustrates the above ideas for connecting Wikipedia and news with an example excerpt taken from the Wikipedia article central to the "Syrian Civil War".

Events in general have been defined in different ways at different textual granularities. For example, events can be defined as a single or cluster of news articles [134, 183], or entities from knowledge bases [62, 104]. A more generic definition was adopted in the task of Topic Detection and Tracking (TDT) [9] as something that happens at a particular time at a particular place. However as described by Alan et al. [10], this definition is preliminary and partial as events may be associated with multiple time periods and geographical locations. Consider our previous example of the "European Debt Crisis" that lasted several years since 2009 and affected multiple countries in Europe. Similarly, "Syrian Civil War" broke in 2011 and has continued till date. Further, we motivate that events will often have multiple seminal named entities (people and organizations) associated with them. For example, the "Syrian Civil War" involves entities such as *Syrian Democratic Forces* (SDF), *Salafi jihadist, al-Nusra Front, Islamic State of Iraq and the Levant* (ISIL), etc. Thus in this work, we define a newsworthy event as follows:

Definition 1.1: *An event covered by news media is a significant occurrence of something with unavoidable ramification that is associated with a spatio-temporal scope and involving seminal entities.*

The above definition is more generic than the one used in TDT as it allows an event to be associated with multiple entities, geolocations, and time periods that define its scope.

Short (single sentence or passage) event descriptions are prevalent across different online information sources, for example, within news articles to refer to related past events. As another source, Wikipedia systematically lists events in portals such as the Wikipedia Current Events portal [1] and *Year* pages (e.g., http://en.wikipedia.org/wiki/1987) that come with a short description and a specific date indicating their occurrence.

Additionally, short abstract textual descriptions often occur within regular Wikipedia articles that refer to past events. An example of a short description on "Arab Spring" in the "Syrian Civil War" Wikipedia article is illustrated in Figure 1.1.

In this work, we consider Wikipedia and online news articles as two prominent and disparate (e.g., authoring style, language structure, etc.) examples of information sources on past events. Additionally, by leveraging more generic event definition, this work addresses the following three problems with the goal of achieving the larger objective of better organizing event-focused information:

(I) Constantly evolving Wikipedia articles tend to summarize past event (and stories describing event aspects) into short passages (or sentences) by abstracting fine-grained details that mattered when the event happened. On the other hand, contemporary news articles provide details of events as they had happened. However, as motivated before, both Wikipedia and news articles individually become insufficient to present a holistic view on past events. By connecting Wikipedia excerpts from general articles that describe events and news articles, a quick navigation between the two information sources in facilitated. We motivate that this can help a general user to quickly develop a better understanding on the events. An example of a linking established between a Wikipedia excerpt and news articles is illustrated in Figure 1.1. We note that excerpts in Wikipedia articles that describe events often mention time, geolocations, and named entities in their textual descriptions. These semantic annotations can act as a strong indicator to identified highly relevant news articles that can be connected.

Connecting excerpts from Wikipedia describing an event to past news articles providing contemporary accounts is the first problem that is addressed.

(II) State-of-the-art vertical news search engines, like Google News[1], are among the first choices of a general user when seeking information on past events. However, these search engines are keyword-based and retrieve a large ranked list of news articles, all of which are temporally biased to the query issuing time. It is hard for a user to sift through all retrieved news articles so as to get a holistic view on a past event. For such an information need, it would be useful if a system could automatically generate an *event digest* by extracting text from retrieved news articles. In context of a linking task defined to connect Wikipedia events to news (as in our first problem), a Wikipedia event with long ramifications gets linked with many news articles that give information on different aspects. In such a case, an event digest that presents a *holistic view* on the event represents an intermediate level of linking, i.e., connecting the Wikipedia event to smaller textual units within documents. An example is illustrated in Figure 1.1. This is because, the Wikipedia event descriptions as an *abstract view*, is connected to

[1] https://news.google.com

excerpts in the digest which are in-turn connected to news articles that give a *detailed view* on the event.

Leveraging semantic annotations to generate an event digest from news articles to presents a holistic view on a given event is the second problem that is addressed.

(III) Time is considered an important dimension for understanding and retrospecting on past events. Temporal information associated with a past event not only indicates its occurrence period but also helps to understand its causality, evolution, and ramifications. Thus, the time dimension has been leveraged in several applications such as temporal information extraction [204], temporal information retrieval [13], and event-focused multi-document summarization [11]. Temporal information extraction [204] focuses on recognizing and normalizing temporal expressions embedded in text into precise time points or intervals. In temporal information retrieval [15, 26, 93, 162], the goal is to rank documents based on their temporal relevance in addition to textual relevance. For this, approaches often leverage temporal expressions that come with the documents as meta-data or embedded in their content to estimate their temporal relevance to a given query. Time has also been leveraged to order sentences in automatically generated textual summaries [21, 28, 157]. For event digest generation (second problem described before), our goal is to explicitly leverage temporal expressions in news article excerpts along with other semantic annotations. However, all the above mentioned tasks suffer due to sparsity of temporal expressions especially at finer textual granularities such as passages and sentences within larger documents like news articles. Thus, it becomes important to design method for better estimation of temporal scopes for events with short textual descriptions.

Estimating an accurate time models to capture the temporal scope of a given news article excerpt is the third problem addressed in this work.

1.2 Contributions

This work makes the following key contributions while presenting approaches towards addressing the three important problems defined before:

(I) We propose a novel linking problem with the goal of connecting textual excerpts extracted from Wikipedia articles, coined *Wikiexcerpts*, to news or web articles in a large archive. For this, we cast the linking problem into an information retrieval task where the goal is to retrieve a ranked list of documents for a given Wikiexcerpt thus treating it as a user query. We propose a framework to leverage a combination of text, time, geolocations, and named entities to identify relevant documents that could be linked to a given excerpt. In detail, we present:

(a) Several time-aware language models capturing different temporal intents, and a two-stage cascade approach that estimates a probabilistic text and time model representing the temporal scope of the query from pseudo-relevant documents so as to retrieve textually and temporally relevant news articles.

(b) Novel query modeling techniques, and a KL-divergence [214] based retrieval framework to incorporate text, time, geolocations, and named entities to retrieve relevant documents that can be linked to a given excerpt. To the best of our knowledge, we are the first to present a *unified method* that explicitly combines text, time, geolocations, and entities for identifying relevant news articles for a given short event description as a user query.

(c) A prototype time-aware exploratory search system for exploring events that were seminal in the past. We implement our different time-aware retrieval models as modes of exploration available to a user as options.

(II) We propose a new problem of event digest generation as a special case of the unsupervised extractive multi-document text summarization with the goal of providing an effective holistic view on a past event that is considered as an input. For this, we motivate that an digest should present diverse information across text, time, geolocation, and entity event dimensions. We propose an Integer Linear Programming (ILP) based approach for global inference [20] along the event dimensions. In detail, we present:

(a) A novel method that uses a *divergence-based framework* and formulates the problem as ILP for the event digest creation. To the best of our knowledge, we are the first to present a *unified method* to explicitly diversify across text, time, geolocations, and entities using query modeling approaches.

(b) An experimental evaluation on three real-world datasets by treating Wikipedia articles central to an event query as a gold standard. For this, we present novel variants of Rouge metrics that aims to measure how holistic is an automatically generated summary.

(c) A crowdsourcing-based study to understand the effect of the summary structure on its readability. We release a corpus with four variants of 10-sentence summaries for 100 Wikipedia events along with pair-wise human preference judgments on their readability and coherence quality.

(III) To estimate probabilistic time models that more accurately capture the temporal scope of news article excerpts, we present a distribution propagation framework that leverages redundancy in a large longitudinal document collection. Probabilistic time models that are estimated for excerpts that come with temporal expressions are propagated to those that give similar information but do not come with any temporal expression to estimate their time models.

1.3 Publications

The results presented in this work have appeared in several publications. Next, we briefly summarize these publications and point out their connection to the subsequent chapters.

Connecting Wikipedia Events to News Articles

Connecting Wikipedia excerpts to news articles can be done in two directions. In the first direction, we investigate connecting a short event description from Wikipedia to news articles. In the second direction, first events are detected as clusters of news articles, then connected to the relevant Wikipedia current events portal page. In this thesis, we describe the first direction in detail. However, in this section, we briefly summarize the publications that addresses the second linking direction as a reference for the readers.

- [141] Arunav Mishra. *Linking Today's Wikipedia and News from the Past.* PIKM 2014.

In [141], we begin by defining the linking problem, identifying the challenges, and propose a preliminary approach before proceeding in the following two directions:

Wikipedia to News: In [147], we address a linking task where our goal was to connect Wikipedia event descriptions in special Year pages (that list all seminal events occurred in a specific year) to news articles. We cast the problem into an information retrieval task by considering a given Wikipedia event description as a user query. We present several time-aware language modeling approaches, and a two-stage cascade approach that estimates pseudo-relevance feedback to estimate query time and text models.

- [147] Arunav Mishra, Dragan Milchevski, and Klaus Berberich. *Linking wikipedia events to past news.* TAIA 2014.

After investigating the simpler problem, in [145] we extend our approach to accommodate additional signals associated with a given event, namely named entities and geolocations in addition to text and time. Moreover, we address the problem of connecting arbitrary Wikipedia excerpts extracted from regular articles to news articles. To address this linking problem, we cast it into an information retrieval task by treating a given excerpt as a user query with the goal to retrieve a ranked list of relevant news articles. Our retrieval model leverages text and the semantic annotations as different dimensions of an event by estimating independent query models to rank documents.

- [145] Arunav Mishra and Klaus Berberich. *Leveraging Semantic Annotations to Link Wikipedia and News Archives.* ECIR 2016.

In [142], we introduce EXPOSÉ exploratory search system that explicitly uses temporal information associated with events to link different kinds of information sources for effective exploration of past events. Our demo includes several time-aware retrieval approaches developed in [147] that a user can employ for retrieving relevant news articles, as well as a timeline tool for temporal analysis and entity-based facets for filtering.

- [142] Arunav Mishra and Klaus Berberich. *EXPOSÉ: EXploring Past news fOr Seminal Events*. WWW 2015.

News to Wikipedia: In [183], our main contribution is to provide effective models for improved news event ranking. We propose novel event mining and feature generation approaches for improving estimates of event importance. Extensive evaluation of our approaches on two large real-world news corpora each of which spans for more than a year with a large volume of up to tens of thousands of daily news articles demonstrates the effectiveness of our approach in comparison to several baselines.

- [183] Vinay Setty, Abhijit Anand, Arunav Mishra, and Avishek Anand. *Modeling Event Importance for Ranking Daily News Events*. WSDM 2017.

Based on the work done in [183], we demonstrate BioNex, a system to mine, rank and visualize biomedical news events. BioNex visualizes the retrieved event clusters to highlight the top news events and corresponding news articles for the given query. Further, the visualization provides the context for news events using (1) a chain of historically relevant news event clusters, and (2) other non-biomedical events from the same day.

- [52] Patrick Ernst, Arunav Mishra, Avishek Anand, and Vinay Setty. *BioNex: A System For Biomedical News Event Exploration*. SIGIR 2017.

Summarizing Wikipedia Events with News Excerpts

In our previous work in [145], we found that events with larger ramifications get linked to a large number of relevant news articles. In [143] we define the problem of automatic event digest generation to aid effective and efficient retrospection. For this, in addition to text, a digest should maximize the reportage of time, geolocations, and entities to present a holistic view on the past event of interest. In our approach, we propose a novel divergence-based framework that selects excerpts from an initial set of pseudo-relevant documents, such that the overall relevance is maximized, while avoiding redundancy across the event dimensions. Our method formulates the problem as an Integer Linear Program (ILP) for global inference to diversify across the event dimensions. Using Wikipedia articles as gold standard summaries, we compare all methods using standard Rouge-1, -2, and -SU4 along with Rouge-NP, and a novel weighted variant of Rouge.

- [143] Arunav Mishra and Klaus Berberich. *Event Digest: A Holistic View on Past Events.* SIGIR 2016.

In [146], we conduct an empirical study on a crowdsourcing platform to get insights into regularities that make a text summary coherent and readable. For this, we generate four variants of human-written text summaries with 10 sentences for 100 seminal events and conduct three experiments. Experiment 1 and 2 focus on analyzing the impact of sentence ordering and proximity between originally occurring adjacent sentences, respectively. Experiment 3 analyzes the feasibility of conducting such a study on a crowdsourcing platform.

- [146] Arunav Mishra and Klaus Berberich. *How Do Order and Proximity Impact the Readability of Event Summaries?* ECIR 2017.

Estimating Time Models for Short News Excerpts

It is often difficult to ground text to precise time intervals due to the inherent uncertainty arising from either missing or multiple expressions at year, month, and day time granularities. In [144], we address the problem of estimating an *excerpt-time model* capturing the temporal scope of a given news article excerpt as a probability distribution over *chronons.* For this, we propose a semi-supervised *distribution propagation* framework that leverages redundancy in the data to improve the quality of estimated time models. In our experiments, we first generate a test query set by randomly sampling 100 Wikipedia events as queries. For each query, making use of a standard text retrieval model we retrieve top-10 documents with an average of 150 excerpts. From these, each temporally annotated excerpt is considered as gold standard. The evaluation measures are first computed for each gold standard excerpt for a single query by comparing the estimated model with our method to the empirical model from the original expressions. Experiments on the English Gigaword corpus [5] show that our method estimates significantly better time models than several baselines motivated from the literature.

- [144] Arunav Mishra and Klaus Berberich. *Estimating Time Models for News Article Excerpts.* CIKM 2016.

Short textual event descriptions (e.g. single sentences) prevalent in Web documents (also considered as inputs in the above applications) often lack explicit temporal expressions for grounding them to a precise time period. Thus, in [45] we address the problem of estimating *event focus time* defined as a time interval with maximum association thereby indicating its occurrence period. We propose several estimators that leverage distributional event and time representations learned from large document collections by adapting the word2vec.

- [45] Supratim Das, Arunav Mishra, Vinay Setty, and Klaus Berberich. *Estimating Event Focus Time Using Neural Word Embeddings.* CIKM 2017.

1.4 Outline

The remainder of the thesis is organized as follows. Chapter 2 recapitulates fundamental techniques from different areas of Computer Science with the goal of giving sufficient background required for understanding subsequent chapters. In Chapter 3, we address a linking task defined to connect excerpts from Wikipedia articles to news articles. In this chapter, we also present an exploratory search system that builds upon several approaches designed by us. Chapter 4 describes our approach for the event digest generation problem. We additionally present our findings from a user study done on a crowdsourcing platform with the goal of finding regularities that make a short summary coherent and readable. In Chapter 5, we present our approach to estimate accurate probabilistic time models for news article excerpts by leveraging redundancy in a longitudinal data collection. Finally, we conclude this work in Chapter 6.

Chapter 2

Foundations & Background

The contributions described in the subsequent chapters build upon techniques from different areas of Computer Science. In this chapter, we recapitulate fundamental techniques and also review state-of-the-art approaches that are essential for this work.

2.1 Information Retrieval

The rise of the World Wide Web (WWW) has led to an ever increasing amount of information that is made available online. For a general user with a specific information need, this has led to the problem of *information overloading* (too much information). Web search engines have been considered as the most successful power tools when it comes to dealing with the information overloading problem by providing effective and efficient access to relevant information. Search engines such as Google and Bing have managed to influence every sector of society and have become an integral part of day-to-day life. In this context, information retrieval (IR) can be concisely defined as the science of search engines [214]. Generally, IR is also often connected to several other tasks such as text categorization, text clustering, text summarization, question answering, and information filtering. Since, we focus on the task of searching textual documents from large collection, we adopt the more focused definition of IR given by Manning et al. [130]:

> **Definition 2.1:** *Information Retrieval (IR) is finding unstructured textual documents from within a large collections that satisfy a user's information need expressed as a query.*

Such a task is also often referred to as the *ad-hoc retrieval task*. Given the above definition, information retrieval task is formally set up as follows:

Let us assume that a given *document collection* or *corpus* c is represented as a set of independent *documents* d_i, i.e., $c = \{d_1, d_2, ...d_N\}$ where N is the total size of the collection. For a given *query q*, the goal is to retrieve a *ranked list* of documents such that the top-ranked document is considered most *relevant* to q. Next, we look into following aspects of the problem definition that need more explanation:

Query: A user's information need is typically assumed to be expressed as *query q*. Here, an information need can be understood as a topic that the user wants to know more about. For this, a user can select the most representative words or *keywords* that describe the information need. Such a query is also known as a *keyword query*.

Document and Ranking: A document is usually defined as the single unit of text that a retrieval system must consider while retrieving relevant information for a user query. The goal of an IR system is to automatically estimate relevance of a document to a given query. Typically, the output of a retrieval system is a ranked (or relevance ordered) list where the top document is speculatively most relevant. For this, IR systems implement what is referred to as a *ranking function* $r(q,d)$ that assigns scores to documents against a query. It is desired that scores model the notion of relevance, i.e., a document with larger score is more relevant.

Relevance and Effectiveness: Any document is considered *relevant* if it contains information that satisfies the user's need. Often this is decided following the notion of *binary relevance*, i.e., either a document is relevant or irrelevant to a given query. Relevance of a document can also be represented at a finer granularity as *graded relevance* on a discrete scale. For example, on a scale of three, a document can be *highly relevant*, *marginally relevant*, or *irrelevant* to a query. A ranked list presented by an ideal IR system would exactly coincide with the user's perception thus satisfying information need. Thus, for empirical evaluation of retrieval systems, manual binary or graded *relevance judgments* are gathered for each query-document pair.

Vocabulary: Typically, textual queries and documents are composed of words or terms w that are assumed to be sampled from a fixed vocabulary V. Here, the vocabulary V as a set is assumed to comprise all the words that can be used to generate text.

Bag-of-Words Model: Further, it is assumed that the words are independently sampled from a vocabulary while generating a document and query. This is referred to as the *bag-of-words* model of a query and document. Though this may be considered as an oversimplification due to the disregard for word order and proximity which may be essential to capture text semantics, the bag-of-words model enables designing efficient scalable ranking functions which have also been empirically shown to be effective.

Empirical studies of ranking models proceeded in two directions resulting in two categories of approaches. They can broadly be referred to as *deterministic approaches* that include variants of the *vector-space models* [39, 73, 84, 148, 159, 160, 177, 179, 186, 188], and *probabilistic approaches* [106, 165, 173, 175, 203, 203] that include *binary independence model* and *statistical language and inference model*. While the deterministic approaches rely on heuristics such as *tf-idf*, probabilistic language modeling approaches are based on the Probability Ranking Principle (PRP) [174] and develop upon strong mathematical foundations. Due to this, many variations of the language modeling approaches have found application in several tasks such as, cross-lingual information retrieval [111], distributed information retrieval [187], expert finding [53], passage retrieval [121], web search [102, 135, 155], topic tracking [90, 110], and sub-topic information retrieval [218].

Okapi BM25 Ranking Function

Among the probabilistic retrieval approaches, one of the most popular ranking function is the Okapi BM25 presented by Spärk Jones et al. [92, 94]. The BM25 ranking function is very similar to the *tf-id* based vector space retrieval function but motivated from the 2-Poisson probabilistic retrieval model with heuristic approximations. As an extension of the classical binary independence model, BM25 is considered to be one of the most effective and robust retrieval function in practice. Primarily, the ranking function can be explained as a combination of three parts: *term frequency* (tf), *inverse document frequency* (idf), and *document length normalization*.

Formally, the BM25 ranking function is defined as,

$$r(q,d) = \sum_{w \in q} \omega_{idf}(w) \cdot \omega_{tf}(w,d)\,. \tag{2.1}$$

The first factor is the logarithmically dampened word inverse document frequency, which is computed as the inverse of the count of documents containing w.

$$\omega_{idf}(w) = \log \frac{N - df_w + 0.5}{df_w + 0.5}\,, \tag{2.2}$$

where df_w is the document frequency of a word w. The second factor in Equation 2.1 represents the term frequency of a word w with the length normalization component. Formally,

$$\omega_{tf}(w,d) = \frac{(k_1+1)\cdot tf_{w,d}}{k_1 \cdot ((1-b) + b \cdot (|d|)/avdl)) + tf_{w,d}}\,, \tag{2.3}$$

where k_1 and b are tunable parameters. For a more detailed discussion, we point readers to the work done by Spärk Jones et al. [92, 94].

2.1.1 Statistical Language Models

Broadly, a statistical language model can be understood as a function that puts a probability distribution over words sequences [130]. Here the words are assumed to be taken from a fixed vocabulary V for a specific language. Let us assume we have a language model θ for an English text corpus containing mostly sports related documents. Given θ, the following word sequences can be sampled with different probabilities:

$$P(football\ player|\theta) = 0.001$$
$$P(movie\ actor|\theta) = 0.00001$$

Given such a language model θ, the sampling probabilities are also referred to as *generative probabilities* where it is assumed that a user used θ to generate text in the documents. Thus, the language models are also often called *generative models* for text. It is not hard to see that such a model presents an opportunity to principally quantify the uncertainty of occurrence of a piece of information in text by looking at word generative probabilities. For example, we can probabilistically determine that a document in our corpus more likely contains information about a *football player* than a *movie actor*.

Estimating a language model that encapsulates all possible word sequence becomes infeasible due to the explosion of parameters though theoretically possible. Here, the word probabilities are referred to a parameters of a language model. It thus becomes essential for practicality to consider only n-grams (extracted from the target corpus c) with small values of n. Most commonly, in an IR task defined over huge document collections (resulting in a large vocabulary) unigram language models are generated and also have shown to be effective in addition to being efficient. Mathematically, this is done with an independence assumption between the words occurring in text (analogous to the bag-of-words model described earlier). Probabilities of word sequences can be estimated by multiplying the generative probabilities of independent words in the sequence. Formally,

$$P(w_1, w_2, ... w_n) = \prod_{i=1}^{n} P(w_i|\theta) \, .$$

For example, $P(football\ player|\theta) = P(football|\theta) \cdot P(player|\theta)$. However, with the simplistic (arguably unrealistic [214]) independence assumption results in the loss of word-order information in text.

Parameter estimation for generating a language model can be done with different statistical methods such as using maximum a posteriori probability (MAP), maximum likelihood estimation (MLE), and full Bayesian estimation. Assuming a multinomial word distribution, the parameters, i.e., generative probability of a word, can be estimated from the empirical words occurring in the documents. Such a method estimates a language model that best describes the data. For this, one popular method is using the

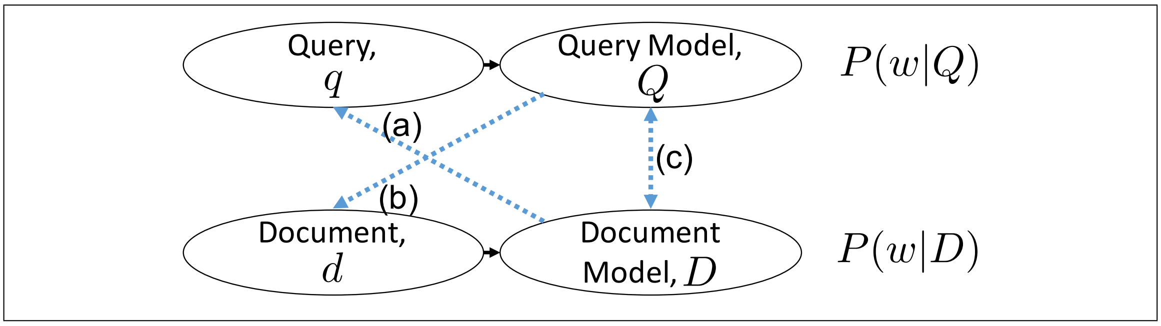


Figure 2.1 The three ways to design ranking function: (a) query likelihood, (b) document likelihood, and (c) model comparison.

maximum likelihood estimator that is computed as,

$$P(w|\theta) = \frac{count(w,d)}{|d|}, \tag{2.4}$$

where $count()$ function computes the frequency of w in d. In this formulation, the parameters are estimated from a single document d. It is easy to see that different documents will result in different language models. A language model that is specific to a document is referred to as a document language model and we represent this as D. Following this notion, a language model can be estimated from the entire corpus by concatenating the documents. This is referred to as the collection language models and is represented as C.

Next, we describe how the language models can be leveraged in the context of IR to estimate the relevance of documents to a given query for generating a ranked list. We identify that there are three ways to design a ranking function as illustrated in Figure 2.1.

2.1.2 Query Likelihood Retrieval Model

The query likelihood retrieval model was first introduced by Ponte and Croft [165]. The basic assumption behind the query likelihood retrieval model is that a user generates a query by sampling words from a document language model. With this assumption, a document is scored against a given query with the following two-step process: first, a language model is estimated for a given document. Second, the likelihood of generating the query from the document language model is estimated. The final ranked list is obtained by ordering the documents in descending order of their likelihood scores.

Formally, let q be a given query, and D be the document model estimated for the document d. The similarity score of d to a given query q is defined as the probability $P(d|q)$. Using Bayes' rule we have,

$$score(q,d) = P(d|q) = \frac{P(q|D) \cdot P(d)}{P(q)}. \tag{2.5}$$

Often in the above equation, an *uninformed document prior* is assumed by setting $P(d)$ to be uniform across all documents. Further, the denominator $P(q)$ is the same for all the documents and does not affect the ranking. Thus, the $score(q,d)$ depends only on the first factor, which is the likelihood of generating the query from the document language model. Further, assuming a multinomial language model with the independence assumption, the generative probability of a query from a document, $P(q|D)$ is estimated as,

$$P(q|D) = \prod_{w \in V} P(w|D)^{count(w,q)}, \tag{2.6}$$

where the $count(w,q)$ returns the count of a word w in query q. With this formulation, the problem is reduced to estimating $P(w_i|D)$, and can be done as described in Equation 2.4.

Though a multinomial language model is often used for estimating document language models, other distributions have also been considered in the past. Ponte and Croft [165] assume a multiple Bernoulli distribution. In another study, Mei et al. [139] evaluate the effectiveness of multiple Possion distribution. However, the multinomial model has been empirically found to be more effective [189].

2.1.3 Kullback-Leibler Divergence Retrieval Model

In the context of ranking models, one important signal that has been shown to improve retrieval quality is pseudo-relevance feedback (explicit and implicit) [175, 176]. One drawback of the query-likelihood retrieval model (described before) is that it is not straightforward to accommodate relevance feedback into the ranking function [214]. This is primarily due the underlying assumption of the model, i.e., the query is generated by sampling from independent document models. To address this issue, Lafferty and Zhai [216] introduced the Kullback-Leibler (KL) divergence retrieval model. Additionally, the KL-divergence retrieval model can also be motivated as *a risk minimization model* that reduces the risk of retrieving irrelevant documents to a given query by computing how close they are based on their language models.

The distance-based KL-divergence retrieval model introduced by Lafferty and Zhai [216], addresses this issue by separating the query and document representations from the ranking function. In this retrieval model, as the first step, the two independent language models are estimated. The first model, analogous to the query-likelihood model, is estimated for a document such that it captures the information (topic) described in its content. The second language model is independently estimated for a given query such that it captures the information intent of the user specifying the query. Given the two models, the relevance score of the document to the query can be computed based on computing the KL-divergence between the two language models. In Figure 2.1, this is illustrated as the model comparison-based approach.

Formally, let Q be the query model estimated for a given query q and D is the document language model estimated for d. The relevance score of d to q is computed as the KL-divergence $KLD(Q||D)$ as,

$$score(q,d) = -KLD(Q||D) = -\sum_{w \in V} P(w|Q) \log \frac{P(w|Q)}{P(w|D)} . \tag{2.7}$$

In the above equation, the negative sign implies that lower divergence indicates higher relevance. With a simple derivation, it can be shown that the ranking is based on computing the cross entropy of a query and document models.

Generalization of Query Likelihood Model

It can be shown that the KL-divergence model is a generic model and the query likelihood can be derived as a special case when the query model is generated only from the terms occurring in the original query. We refer to such a query model that is estimated based on only the original query terms as an *empirical query model*. Given the query model Q, the generative probability of a word w is estimated as,

$$P(w|Q) = \sum_{w \in V} \frac{count(w,q)}{|q|} . \tag{2.8}$$

With the empirical query model estimated from the original terms in the query, it is easy to show that the KL-divergence retrieval model essentially scores a document d against a query q based on the cross-entropy between their language models. Starting with this, we can make the following derivation:

$$\begin{aligned} score(q,d) &\equiv \sum_{w \in V} P(w|Q) \cdot \log P(w|D) \\ &= \sum_{w \in V} \frac{count(w,q)}{|q|} \cdot \log P(w|D) \\ &\approx \sum_{w \in V} count(w,q) \cdot \log P(w|D) \\ &= \log P(q|D) \end{aligned}$$

Finally, in the derivation, $\log P(q|D)$ is the log likelihood of generating the query words from the document language model.

The main motivation of using the KL-divergence model is to improve the query model with the addition of relevance feedback. However, in practical scenario, a query model estimated for a short query would often result in only a few terms receiving high probabilities while most terms receive zero probability. Thus as a common practice, the query models are generated by keeping only the high probability terms [216]. Such a truncated query model results in improved scoring efficiency [214].

Interpretation and Other Divergence Functions

The KL-divergence retrieval model was proposed as a distance-based retrieval model which computes a (negative) distance between a document and query. The first interpretation of this retrieval is consistent with the query likelihood model (as derived before) i.e., a user uses a the document model estimated from few words to generate the query. Thus, the divergence score indicates how close this document model is to a query model that capture the "true" intent of the user. The second interpretation comes from the comparison with the vector-space models. With the KL-divergence retrieval model, the query and document models are decoupled. Thus additional techniques (like smoothing) can be designed to improve the document models. Thus a document model captures the "real" information mentioned in the content than simply estimated from a few empirical terms. In this case, the KL-divergence scores indicate how close is the user's intent to the content of a document.

The distance-based retrieval framework can be applied by using other divergence measures [118] such as K-divergence, J-divergence, I-divergence, and JS-divergence. Among the divergence measures, the Jensen-Shannon (JS) divergence has been extensively used is several applications to compute similarity (distance) between probability distributions such as clustering since it is a metric. The JS-divergence (JSD) is the symmetric variant of the KL-divergence and is bounded between $[0,1]$. Formally, JS-divergence between a query model Q and a document model D is defined as,

$$JSD(Q||D) = \frac{1}{2}\sum_{w\in V} P(w|Q)\log\frac{P(w|Q)}{P(w|M)} + \frac{1}{2}\sum_{w\in V} P(w|D)\log\frac{P(w|D)}{P(w|M)}\,. \tag{2.9}$$

where M is the average model computed with an interpolation parameter $\pi = [0,1]$ as,

$$P(w|M) = \pi P(w|Q) + (1-\pi)P(w|D)$$

Though JS-divergence is a distance metric, in the context of IR, [54] found the KL-divergence to be the most effective divergence measure while JS-divergence is shown to dissatisfy length normalization constraint.

2.1.4 Estimation of Query Models

By separating the representation of the queries and documents in the KL-divergence retrieval model, it is possible to re-estimate the query term probabilities with feedback documents, thereby presenting a natural way of incorporating relevance feedback into language models. In this context, the feedback documents can either be explicitly selected by a user, or obtained with an initial round of retrieval with the empirical query model. These documents are then treated as pseudo-relevant.

Let the set R contain the feedback documents, the general framework is to first estimate a feedback language model F' from the documents contained in R. It is then combined with the empirical query model Q. Formally, the generative probability of a word w from the updated query model Q' is computed as,

$$P(w|Q') = (1-\alpha)P(w|Q) + \alpha P(w|F') \,, \tag{2.10}$$

where $\alpha \in [0,1]$ is an interpolation parameter that controls the influence of the feedback documents. Usually, this parameter can be set empirically by training on a certain corpus. In a study, Zhai et al. [216] empirically test the effect of α on different TREC corpus. Next, we discuss popular methods to estimate the feedback model F'.

Mixture Model Feedback

One approach to estimate F' is to use a mixture model with two components where one is a background language model C estimated from the entire corpus c, and the other is the empirical feedback model F that is estimated from the words occurring in documents d_i in set R. Such a model will put more stress on the discriminating terms in the feedback documents while factoring out undiscriminating terms that are more frequent in the entire corpus. Formally, the probability of generating a word w is estimated as,

$$P(w|F') = (1-\lambda)P(w|F) + (\lambda)P(w|C) \,.$$

In the above equation, a word w is generated from the empirical feedback model F estimated from the documents in R with a $()1-\lambda)$, and from the entire corpus with a probability of λ . Here, λ is an interpolation parameter which is set between $[0,1]$. This parameter can be tuned using an expectation maximization (EM) algorithm as shown by Zhai et al. [216].

Robust Mixture Model Feedback

As an extension of the mixture model, Tao et al. [200] presented the robust mixture model by introducing discriminative priors for documents in the set R. More specifically, this model additionally incorporates the following three ideas: **1)** each feedback document has an independent λ_d in Equation 2.1.4 thereby allowing modeling different noise level within the individual documents; **2)** the empirical query is treated as a prior for the feedback documents using the Bayesian estimation framework, thereby putting more stress on the query-specific terms; **3)** the EM algorithm is regularized which prevents topical drifts while estimating feedback models.

For our approaches designed in this thesis, we adopt the simpler mixture feedback model to capture information along time, geolocations, and named entities along with text. Experiments with more advanced feedback models remain as a future work.

2.1.5 Document Models and General Smoothing Methods

So far in both the query likelihood and KL-divergence based retrieval model, we have leveraged a language model for a given document. For a given document, this model can be understood as a probability distribution over words that is used to generate its content. In other words, the textual document content is assumed to be sampled from its document language model.

Formally, for a given document d, the generative probability of a word w from its document language model D can be estimated as,

$$P(w|D) = \frac{count(w,d)}{|d|} . \tag{2.11}$$

In both query likelihood and the KL-divergence retrieval framework, it becomes integral to smooth an estimated document model with a more general and a fixed background model. When comparing statistical language model based approaches to the *tf-idf* heuristics (used in deterministic approaches), smoothing language models (either for documents or queries) can be observed to play the following three important roles: first, smoothing document models have the idf-like effect which has been shown as an important heuristic to factor out undiscriminating terms in a document [214]. Undiscriminating terms can be defined as those that are frequent in the corpus and likely to occur in a large number of documents.

Second, smoothing enables reducing the sparsity in the language models. That is in a language model estimated from a short document, many terms receive a zero probability. Smoothing with a fixed general model, like the corpus language model, assigns a small probabilities to these terms. This is important to both the retrieval models discussed in this section. In Equation 2.6 that describes the query likelihood retrieval model, an unseen term, i.e., term in a query that is missing in a document, without smoothing receives zero probability and would result in the overall likelihood to be zero. This is commonly referred to as the *zero probability issue*. In Equation 2.7 that describes the KL-divergence retrieval model, due to the zero probability issue for unseen terms, the divergence becomes unbounded due to division by zero.

Finally, smoothing with appropriate background models and result in generation of better document and query models. This becomes an advantage specially in the KL-divergence retrieval model that allows incorporating feedback documents as described before. Next we look into the two most popular smoothing methods proposed in the literature that have repeatedly proven to be most effective.

Jelinek-Mercer Smoothing (Fixed Coefficient Interpolation)

A simple method to smooth a document language model is to combine it with the collection language model via linear interpolation with a fixed coefficient. This is referred to as Jelinek-Mercer (JM) smoothing. Formally, let D be the document language for d and C be the collection model estimated from the entire corpus c, the updated document language model D' after smoothing is computed as,

$$P(w|D') = (1-\lambda)\cdot P(w|D) + \lambda \cdot P(w|C)\,. \tag{2.12}$$

The interpolation parameter λ controls the amount of smoothing. If $\lambda = 1$ then the document model will be same as the collection model. By setting $\lambda = 0$, the smoothing can be turned off entirely.

Dirichlet Prior Smoothing (Bayesian Interpolation)

In JM-smoothing, all the documents are treated equally, i.e., they are assumed to contain an equal amount of noise. However, the Dirichlet prior smoothing takes advantage of Bayesian estimation [214] to impose a document specific prior. This allows to apply different amounts of smoothing on independent documents, i.e., longer documents are smoothed relatively less.

Formally, given a document language model D, and the collection language model C, the updated document language model D' with Dirichlet smoothing is computed as,

$$P(w|D') = \frac{count(w,d) + \mu \cdot P(w|C)}{|D| + \mu}\,, \tag{2.13}$$

where $count(w,d)$ is the count of w in d and μ is a parameter that can be safely set to a number close to or larger than the average document length [217] in the corpus.

Zhai et al. [215] systematically studied the effects of different smoothing methods. Their study shows that Dirichlet smoothing works better than JM-smoothing for estimating accurate document models while the latter proves to be better for estimating query models with feedback.

2.2 Specialized Information Retrieval

So far we have looked into approaches that address the information retrieval task of retrieving relevant textual documents for a given keyword query. However, there have been efforts to consider additional signals from documents to further improve retrieval quality. For example, computing document relevance by explicitly considering the temporal information has led to the research direction of temporal information retrieval. Similarly, in geographic information retrieval, methods leverage geolocations, and in entity-based information retrieval methods leverage named entity mentions in documents to estimate relevance by combining the additional signals with text. Next, we look into the individual research directions in detail as specialized branches of the larger information retrieval problem.

2.2.1 Temporal Information Retrieval

Time has been considered as an important indicator when it comes to retrieving temporal documents from a longitudinal corpus, for example a news archive like the New York Times Annotated corpus [4]. Queries in a large Web search log have been considered to exhibit certain temporal patterns as discussed by Kanhabua et al. [98]. For example, queries that intent *breaking news*, *short-span events*, and information about *celebrities*, have *sporadic* distribution over time. Similarly, queries relating to *seasonal events* such as annual sports series and TV series exhibit a *periodicity* in their distribution. Finally, queries that seek information on past events with longer ramifications may exhibit a smooth decay in their distribution over time. Another class of temporal queries are those that do not explicitly exhibit patterns over time. Such queries can be categorized as two types. First, those that come with explicit temporal criteria [26], and second those that do not come with explicit temporal criteria [32, 74]. Several studies have reported that significant fraction of web search queries have temporal intent. For example, Zhang et al. [219] showed that 13.8% contain explicit time and 17.1% of queries have implicit temporal intent, i.e., they target documents that give information on a specific time period which can be inferred for the query.

For a given temporal query, a time-aware retrieval system should consider the temporal relevance of the documents in addition to its textual relevance. For this, it is desired to estimate the temporal scopes of the target documents so as to compute their relevance. For this work we adopt the definition as given by Campos et al. [33]:

Definition 2.2: *Temporal information retrieval aims to satisfy search needs by combining the traditional notion of document relevance with temporal relevance.*

Significant efforts have been put in several research directions to understand the temporal dynamics associated with text. Next, we discuss several research directions that are defined to leverage the temporal dynamics in context of temporal IR.

Temporal Information Extraction

Temporal information extraction is a task of extracting textual phrases or *temporal expressions* which can be normalized into precise time intervals [204] on a time line. As concrete example, Figure 2.2 illustrates a sample news article with highlighted temporal expressions that are normalized into time intervals described in the TIMEX3 format which is a part of the TIMEML language [167]. Temporal expressions have been categorized into the following categories: *explicit*, *implicit*, *relative*, and *free text*. For example, “January 1, 2015” is an explicit expression while “yesterday” is a relative expression which requires an additional reference time point for normalization. Similarly, “New Year’s Day” is an implicit expression that points to “January 1, 2015”. The free text temporal expressions refer to larger textual units that convey information which can be grounded

WASHINGTON — President Trump embraced a proposal on Wednesday to slash legal immigration to the United States in half within a decade by sharply curtailing the ability of American citizens and legal residents to bring family members into the country. The plan would enact the most far-reaching changes to the system of legal immigration in decades and represents the president's latest effort to stem the flow of newcomers to the United States. Since taking office, he has barred many visitors from select Muslim-majority countries, limited the influx of refugees, increased immigration arrests and pressed to build a wall along the southern border. In asking Congress to curb legal immigration, Mr. Trump intensified a debate about national identity, economic growth, worker fairness and American values that animated his campaign last year. Critics said the proposal would undercut the fundamental vision of the United States as a haven for the poor and huddled masses, while the president and his allies said the country had taken in too many low-skilled immigrants for too long to the detriment of American workers. "This legislation will not only restore our competitive edge in the 21st century, but it will restore the sacred bonds of trust between America and its citizens," Mr. Trump said at a White House event alongside two Republican senators sponsoring the bill. "This legislation demonstrates our compassion for struggling American families who deserve an immigration system that puts their needs first and that puts America first." In throwing his weight behind a bill, Mr. Trump added one more long-odds priority to a legislative agenda already packed with them in the wake of the defeat of legislation to repeal and replace President Barack Obama's health care program. The president has already vowed to overhaul the tax code and rebuild the nation's roads, airports and other infrastructure. Continue reading the main story But by endorsing legal immigration cuts, a move he has long supported, Mr. Trump returned to a theme that has defined his short political career and excites his conservative base at a time when his poll numbers continue to sink. Just 33 percent of Americans approved of his performance in the latest Quinnipiac University survey, the lowest rating of his presidency, and down from 40 percent a month ago. Democrats and some Republicans quickly criticized the move. "Instead of catching criminals, Trump wants to tear apart communities and punish immigrant families that are making valuable contributions to our economy," said Tom Perez, the chairman of the Democratic National Committee. "That's not what America stands for." The bill, sponsored by Senators Tom Cotton of Arkansas and David Perdue of Georgia, would institute a merit-based system to determine who is admitted to the country and granted legal residency green cards, favoring applicants based on skills, education and language ability rather than relations with people already here. The proposal revives an idea included in broader immigration legislation supported by President George W. Bush that died in 2007. More than one million people are granted legal residency each year, and the proposal would reduce that by 41 percent in its first year and 50 percent by its 10th year, according to projections cited by its sponsors. The reductions would come largely from those brought in through family connections. The number of immigrants granted legal residency on the basis of job skills, about 140,000, would remain roughly the same. Under the current system, most legal immigrants are admitted to the United States based on family ties. American citizens can sponsor spouses, parents and minor children for an unrestricted number of visas, while siblings and adult children are given preferences for a limited number of visas available to them. Legal permanent residents holding green cards can also sponsor spouses and children. In 2014, 64 percent of immigrants admitted with legal residency were immediate relatives of American citizens or sponsored by family members. Just 15 percent entered through employment-based preferences, according to the Migration Policy Institute, an independent research organization. But that does not mean that those who came in on family ties were necessarily low skilled or uneducated. The legislation would award points based on education, ability to speak English, high-paying job offers, age, record of achievement and entrepreneurial initiative. But while it would still allow spouses and minor children of Americans and legal residents to come in, it would eliminate preferences for other relatives, like siblings and adult children. The bill would create a renewable temporary visa for older-adult parents who come for caretaking purposes.

Text	Value	Timex3 Tag
Wednesday	2017-08-02	<TIMEX3 tid="t1" type="DATE" value="2017-08-02">Wednesday</TIMEX3>
a decade	P10Y	<TIMEX3 tid="t2" type="DURATION" value="P10Y">a decade</TIMEX3>
decades	PXY	<TIMEX3 tid="t3" type="DURATION" value="PXY">decades</TIMEX3>
last year	2016	<TIMEX3 tid="t4" type="DATE" value="2016">last year</TIMEX3>
the 21st century	20XX	<TIMEX3 tid="t5" type="DATE" value="20XX">the 21st century</TIMEX3>
a month ago	2017-07-03	<TIMEX3 tid="t6" type="DATE" value="2017-07-03">a month ago</TIMEX3>
2007	2007	<TIMEX3 tid="t7" type="DATE" value="2007">2007</TIMEX3>
each year	P1Y	<TIMEX3 periodicity="P1Y" quant="each" tid="t8" type="SET" value="P1Y">each year</TIMEX3>
first year		<TIMEX3 alt_value="P1Y-#1" tid="t9" type="DATE">first year</TIMEX3>
10th year	2026	<TIMEX3 tid="t10" type="DATE" value="2026">10th year</TIMEX3>
current	PRESENT_REF	<TIMEX3 tid="t11" type="DATE" value="PRESENT_REF">current</TIMEX3>
2014	2014	<TIMEX3 tid="t12" type="DATE" value="2014">2014</TIMEX3>

Figure 2.2 Illustration of extraction and normalization of temporal expression in a sample news article with the SUTime toolkit.

to a precise time interval. This category becomes most difficult to deal with and there seems to be few works in this context. In this context, recently Kuzey et al. [105] propose to automatically detect and normalize free-text temporal expressions or coined *temponyms* such as "Obama's presidency" to a time interval (from 2008 to 2016). The main goal of their work was to curate temporal facts for a knowledge base, however, currently available temporal taggers [35, 193, 205] are unable to deal with this class of expressions.

Recently, there have been developments in direction of temporal information extraction. TempEval [204] competitions are held with a specific goal to evaluate temporal taggers on different document collections. Through the community effort there are several open-source temporal taggers that are available today. Some existing systems are HeidelTime [193], SUTime [35], and Tarsqi toolkit [205], although all these systems mainly target explicit and relative temporal expressions. The focus of our work in this thesis is on temporal ranking, thus we refrain from getting into the details of these systems. However, in our work we make use of the open-source temporal taggers to annotate our document collections and query benchmarks.

Document Dating

Typically, documents that punctually report on newsworthy events, such as news articles, often come with a date as a meta data indicating its creation or publication date. For such documents, the creation date becomes a strong indicator of the temporal scope of its content. Additionally, the document creation time is also often considered as a good reference time point to resolve relative expressions mentioned in their content. However, in collections such as Web crawls [2, 3], documents do not come with an explicit creation date instead a date indicating when they were crawled. Such a date in meta data may not be related to the temporal scope of the information in their content. Naïvely relying on such meta data may severely affect result quality for tasks such as temporal information retrieval and Topic Detection and Tracking [10] that require estimation of the temporal scope of the documents. Thus, it becomes important to infer the document creation dates when it is unavailable.

Document dating is defined as the task of automatically determining the creation time of non-timestamped documents. In this realm, the methods proposed to address the task can be categorized into two classes: content-based and non-content-based. Content-based methods [46, 64, 96, 101] consider only the textual content of a document to determine its creation dates. Here the dating process is performed by finding content-wise most similar time-stamped documents to the document to be dated. For example, Jong et al. [46] present an approach that estimates temporal language models for documents. The temporal language model of a document D_T is generated by employing word usage statistics over time. For this, they first assume a corpus as input that exhibits the following characteristics: **1)** it is sufficiently large; **2)** has a balanced distribution over time; **3)** covers the topical domain of the document to be dated; **4)** finally, covers the time period of the document to be dated. Then the corpus is partitioned based on the publication, i.e., $C_P = \{p_1, p_2, ...p_n\}$. Given the partitioned corpus, the date of a non-timestamped document d is determined by computing the normalized log likelihood ratio $NLLR(d, p_i)$ between the partition p_i. Formally, this is computed as,

$$NLLR(d, p_i) = \sum_{w \in d} P(w|d) \cdot \log \frac{P(w|p_i)}{P(w|C_P)} \,. \tag{2.14}$$

This method was extended by Kanhabua et al. [96] with an additional temporal entropy $TE(w_i)$ based weighting which is computed as,

$$TE(w_i) = 1 + \frac{1}{\log |C_P|} \sum_{p \in C_P} P(p|w_i) \cdot \log P(p|w_i) \,, \tag{2.15}$$

where $|C_P|$ is the number of partitions of the input collection.

Document Focus Time Estimation

Focus time estimation is defined as the task of mapping the content of a document to a time period indicating the scope of the information given by the document [87]. This problem is different from the task of estimating document creation time discussed before and applies to document with temporal content.

Most approaches [87, 194] designed to address this task, first extract temporal expressions from a large corpus, then cast the document focus time estimation problem into the problem of ranking the temporal expressions for a given textual document.

Here it is assumed that the top ranked temporal expressions represent the focus time. Jatowt et al. [87] present a generic graph-based approach where words and temporal expressions are taken from a large longitudinal document collection. Edge weights are computed based on temporal entropy as described in Equation 2.15.

Time-Aware Ranking Models

In temporal information retrieval, a time-aware ranking model takes into account the relevance of documents to a given query in the time dimension in addition to text. Several ranking models proposed in the literature capture different temporal intents of the queries. As one of the early works, Li and Croft [115] propose recency-aware ranking. In their approach, they introduce a time-dependent exponential document prior $P(d|t_d)$in the query-likelihood retrieval model as described in Equation 2.5. That is the score of a document d for a given a query q is computed as,

$$score(q,d) = P(q|D) \cdot P(d|t_d)\,, \tag{2.16}$$

where t_d is the publication date of the document d. For a query where recency is a major concern, the document prior is estimated by leveraging an exponential decay function as,

$$P(d|t_d) = \lambda \exp(\frac{(t_c - t_d)}{\mu})\,, \tag{2.17}$$

where t_c is the most recent date in the document collection, λ is a tuning parameter representing decay rate that is set to $[0,1]$, and μ is set to the unit of time distance, i.e., the temporal granularity that is set in prior.

Motivated from [115], Peetz et al. [161] design several temporal document priors motivated from cognitive science. The document priors are modeled using retention functions represent the human memory retention behavior. Among the different retention function $f()$, Peetz et al. [161] find that the Weibull function proves to be the most effective which is computed as,

$$f_{Weibull}(d,q,g) = b + (1-b)\mu \exp(-\frac{a\delta g(d,q)^s}{s})\,. \tag{2.18}$$

We point the readers to [161] for a more detailed description of this function.

The recency-based ranking models discussed so far, rely on the publication dates of the documents to determine their temporal scopes. However, often documents come with additional temporal expressions that are mentioned in their content. Leveraging, these temporal expressions can lead to better temporal scope estimation for the documents. Berberich et al. [26] present an approach that extends the query-likelihood approach to estimate temporal relevance of the documents to a given query. However, in their method, they assume that a query explicitly comes with a time interval q_{time} expressing the temporal information need of an user.

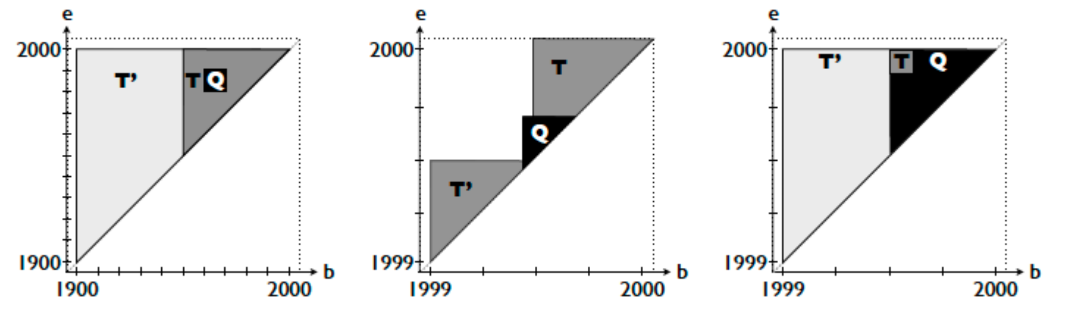


Figure 2.3 Three requirements for a time generative model. Figure adapted from [26].

Berberich et al. [26] represent a temporal expression t in a discrete time domain as a quadruple,

$$t = (tb_l, tb_u, te_l, te_u) ,$$

where tb_l and tb_u give the lower and the upper bounds for the begin time tb of the interval, thus indicating the earliest and the latest possible time points for tb. Analogously, te_l and te_u give the bounds for the end time te of the interval. The quadruple representation of a time interval enables capturing uncertainty in temporal expressions for which the time boundaries cannot be exactly determined. As a concrete example, *"from summer of 2000 to winter of 2001"* will be resolved to;

$$t = (2000/05/01, 2000/07/31, 2001/11/01, 2001/12/31);$$

and *"in May 2000"* will be resolved to the quadruple

$$t = (2000/05/01, 2000/05/31, 2000/05/01, 2000/05/31).$$

In their ranking model, Berberich et al. [26] assume independence between the text and time to extend the query likelihood retrieval model. With the independence assumption, the query likelihood ranking model is formally defined as,

$$P(q|d) = P(q_{text}|d_{text}) \cdot P(q_{time}|d_{time}) , \tag{2.19}$$

where $P(q_{text}|d_{text})$ is the generative probability of the query text from the document text, and $P(q_{time}|d_{time})$ is the generative probability of the query time from the temporal expression in the document. For computing the second factor, following requirements are identified: specificity, coverage, and maximality. Adopting the two-dimensional representation of time, the conditions are illustrated in Figure 2.3.

Capturing the necessary requirements, Berberich et al. [26] present two models, namely, uncertainty-ignorant and aware model, where the later has been found to be more effective [97]. The uncertainty-aware model estimates the generative probability of the query time part q_{time} from the time part of a document d_{time} where both are represented as a bag of temporal expressions. Finally, the generative probability of a temporal expression $Q \in q_{time}$ from a temporal expression $T \in d_{time}$ is estimated by computing their overlap. Formally,

$$P(Q|T) = \frac{1}{|Q|} \sum_{[qb,\, qe] \in Q} \frac{1}{|T|} \mathbb{1}([qb,\, qe] \in T)\,, \tag{2.20}$$

which can be simplified as,

$$P(Q|T) = \frac{|T \cap Q|}{|T| \cdot |Q|}\,. \tag{2.21}$$

For any given $t = (tb_l, tb_u, te_l, te_u)$, $|t|$ can be computed as ,

$$|t| = \left\{ \begin{array}{ll} (tb_u - tb_l + 1) \cdot (te_u - te_l + 1) & \text{if } tb_u \leq te_l \\ (tb_u - tb_l + 1) \cdot (te_u - te_l + 1) + & \\ (tb_u - te_l) \cdot (te_u - te_l + 1) - & \\ 0.5 \cdot (tb_u - te_l) \cdot (tb_u - te_l + 1) & \text{if } tb_u \leq te_l \end{array} \right\}.$$

Given two temporal expressions t and q, their intersection $|t \cap q|$ can be computed as,

$$max(tb_l, qb_l), min(tb_u, qb_u), max(te_l, qe_l), min(te_u, qe_u)\,.$$

2.2.2 Geographic Information Retrieval

In the context of information retrieval, geolocations mentioned in text have been considered as addition signals to further improve results. Such signals have been leveraged to provide various services with unique user interfaces, such as map-based exploration (such as Google maps) and route planning in hotel reservation. Also, it has been found that a large fraction of queries that are issued to standard search engines require processing and reasoning of geographic locations [125]. While, such services may use different visualization tools to organize information that are salient to specific geographic locations, it is important to integrate such signals into ranking functions of retrieval systems to make them geographically-aware. A concrete example of query according to [125] is "riots in Paris and their consequences". We adopt the definition of geographic information retrieval as considered in the GeoCLEF conference [65, 66, 125]:

Definition 2.3: *Geographic Information Retrieval (GIR) concerns the retrieval of information involving some kind of spatial awareness.*

GeoCLEF was an evaluation platform that investigated effectiveness of geographically-aware retrieval systems. It was introduced as a pilot track and then later ran as a regular track for the consequent three years. As a more recent initiative, NTCIR-GeoTime workshop [67, 68] ran for two years with the goal of evaluating the effectiveness of combing geographic and temporal signals to improve retrieval quality. Next, we look into some approaches that were presented in the GeoCLEF and NTCIR-GeoTime workshops.

MIRACLE Toolbox

As one of the participants of GeoCLEF 2005 and 2006, the MIRACLE team that was composed of researchers from three institutes tested the effectiveness of combining textual (keyword based) and geographical entities. In their first approach [107], they focused on gazetteer creation for geographic entities tagging and recognition, processing spatial queries, and document-topic expansion. In their first approach, they focused on monolingual collection.

For lexicon creation, the MIRACLE team coalesced two existing gazetteers: the Geographic Names Information System (GNIS) gazetteer of the U.S. Geographic Survey[1] and the Geonet Names Server (GNS) gazetteer of the National Geospatial Intelligence Agency (NGA)[2]. The combined gazetteer is leveraged in the entity recognition and tagging modules.

In their topic expansion module, MIRACLE leveraged various spatial relations to expand the geographical scopes of the queries. For a given geographic expression that is normalized to a geolocation, its scope is determined by a bounding box placed on the centroid (in terms of latitude and longitude). Here, the dimensions of the box are obtained by examining various hand crafted spatial relations proposed by them. Some concrete examples of spatial relations are *IN*, *NEAR*, *NORTH*, *SOUTH*, *EAST*, and *WEST*. The original topic (query) is expanded with the geographic entities with the box.

The MIRACLE system implements a novel trie [70] data structure to index the textual and the geospatial data. Finally, the expanded topic is treated as a standard keyword query. In their later work, the MIRACLE team extended their approach to cater to multilingual documents~[71] in English, German, Portuguese, and Spanish.

Hariharan's Approach

Hariharan et al. [75] present a general system to address several GIR applications. However, the main focus of this work was on the efficient processing of geographic queries. For this, they develop a novel KR^*-tree data structure that captures the joint distribution of keywords that represented geolocations and normal text.

An overview of the GIR system is illustrated in Figure 2.4. The major components are GIR database, indexer, ranker, and interface. The GIR database provides effective storage of the geospatial data. To generate the database from unstructured textual documents, the GIR system mainly performs geographic entity extraction, matching, and propagation steps (described in detail by Markowetz et al. [133]). An object in

[1] http://www.usgs.gov
[2] http://www.nga.mil

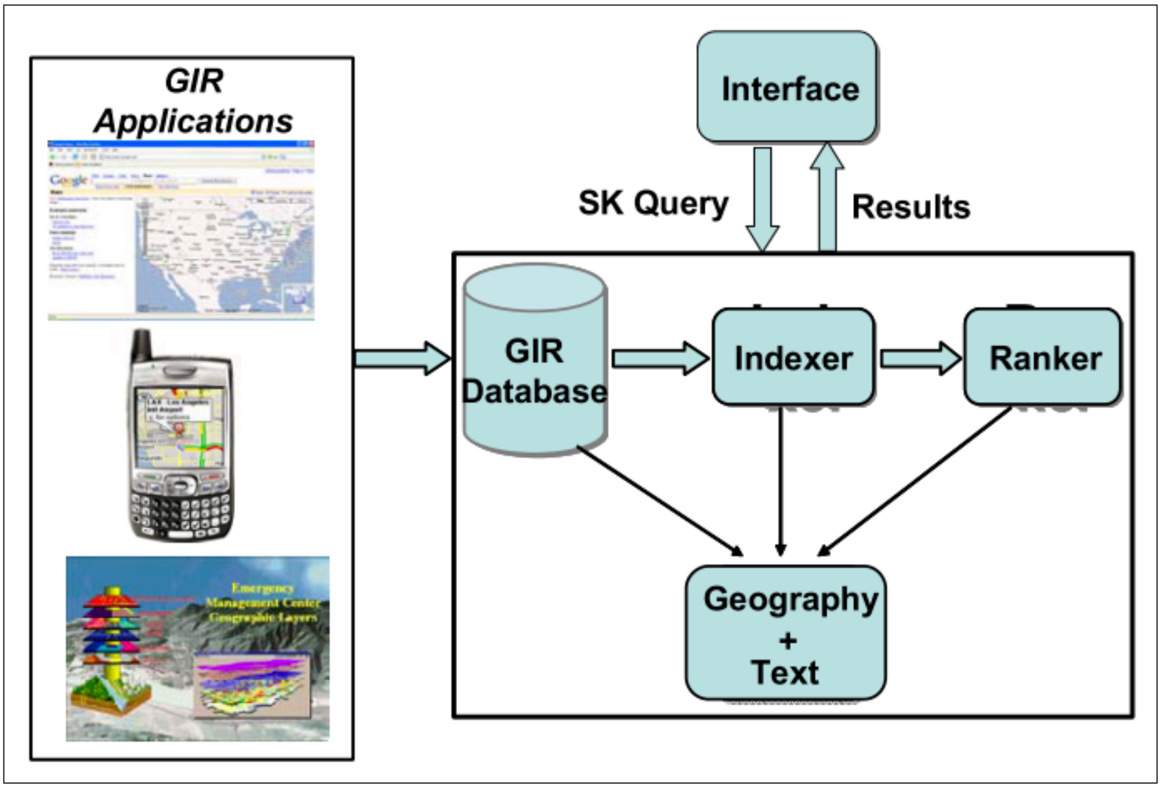


Figure 2.4 Overview of the GIR system. Figure adapted from [75].

the database is described as a triple $< o_{id}, o_r, o_t >$ where o_r is a minimum bounding rectangle (MBR) and o_t is a set of textual terms. With the database generated, the indexer module facilitates efficient processing of the coined SK queries. A SK query Q is defined as a combination of two parts, q_r and q_t. First, the spatial part q_r is represented as a minimum bounding rectangle (MBR) that is generated from the geolocations mentioned in the query. Second, textual part q_t is set of keywords. The ranker module assigns scores to the objects from the GIR database by comparing against the query. For a given SK query $Q = q_r, q_t$, score of an object O is formally computed as,

$$F_{sk} = \alpha_1 F_r(q_r, o_r) + \alpha_2 F_t(q_t, o_t)$$

where $F_r()$ is a geographic ranking function and $F_t()$ is a keyword-based text ranking function. α_1 and α_2 are weights assigned to each function such that $\alpha_1 + \alpha_2 = 1$.

As the main contribution, Hariharan et al. [75] propose the KR^*-tree data structure that stands for keyword R^* -tree. The KR^*-tree data structure is an extension of R^*-tree-Inverted File proposed by Zhou et al. [223]. Most importantly, the KR^*-tree differs from the standard R^*-tree in two aspects, firstly, it facilitates pruning text and geolocations in a single step. All internal and leaf nodes are augmented with a set of keywords. Secondly, the KR^*-tree captures the joint distribution of keywords which enhances the performance for multi-word queries. We point the readers to [75] for a detailed description of the KR^*-tree data structure.

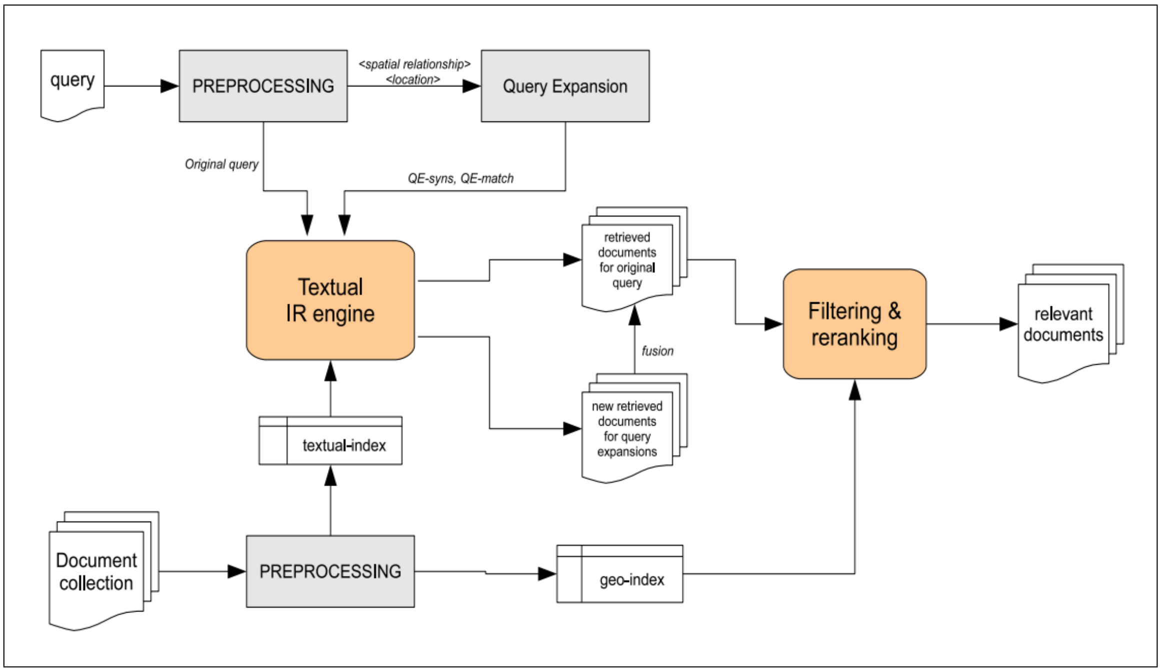


Figure 2.5 Overview of the SINAI-GIR system [163].

SINAI-GIR System

The SINAI-GIR was proposed by Perea-Ortega et al. [163] as a geographically-aware information retrieval system. In their approach, they present NLP techniques to expand a given textual query to incorporate its geographical aspect.

A given query is treated as a triplet of <*theme*> <*spatial relationship*> <*location*>. Here, the SINAI-GIR system treats the <*theme*> as the main subject of a given query, the <*location*> as the geographical aspect, and <*spatial relationship*> as the relationship between the geographical scope and the main subject. As a concrete example, the query "plane crashes close to Russian cities" is resolved to <*airplane crashes*> <*close to*> <*Russian cities*>. With such a query representation, the geographic terms identified from a given textual queries are expanded based on pseudo-relevant documents.

The system overview of the SINAI-GIR system is illustrated in Figure 2.5. SINAI-GIR is composed of mainly three stages: query preprocessing, indexing, and reranking of pseudo-relevant documents initially retrieved with the original query by additionally comparing against the geographical aspect. For query preprocessing, SINAI-GIR leverages Geo-NER [164] tool kit to recognize spatial entities.

In their reranking module, the initial similarity scores of documents retrieved against the textual part of the query are linearly combined with the similarity scores obtained for the geographic part of the query. Formally, the final similarity $sim(Q,D)$ between a query Q and document D is estimated as,

$$sim(Q,D) = \alpha sim_{text}(Q,D) + (1-\alpha) sim_{geo}(Q,D)$$

where α is an interpolation that is set to 0.5. The $sim_{geo}(Q,D)$ is computed as,

$$sim_{geo}(Q,D) = \frac{\sum_{i \in geoEntites(D)} match(i,GS,SR) \cdot freq(i,D)}{|geoEntites(D)|}$$

where the function $match(i,GS,SR)$ returns 1 if the geographic entity i satisfies the geographic scope GS for the spatial relationship SR. The function $freq(i,D)$ returns the frequency of an entity i in document D while $geoEntities(D)$ returns the unique entities in document D.

The query expansion of the geographic part is achieved with two methods, *QE-syn* and *QE-match*. The first method is based on expanding the geographic part based on the synonyms of the original expressions. The second method is based on expanding the geographic part using locations that match with the geographic scope. Experiments on four years of GeoCLEF data show that the method that leverages the expansion and the reranking techniques with the combination of text and geographic query shows significant improvement over textual query.

2.2.3 Entity-based Information Retrieval

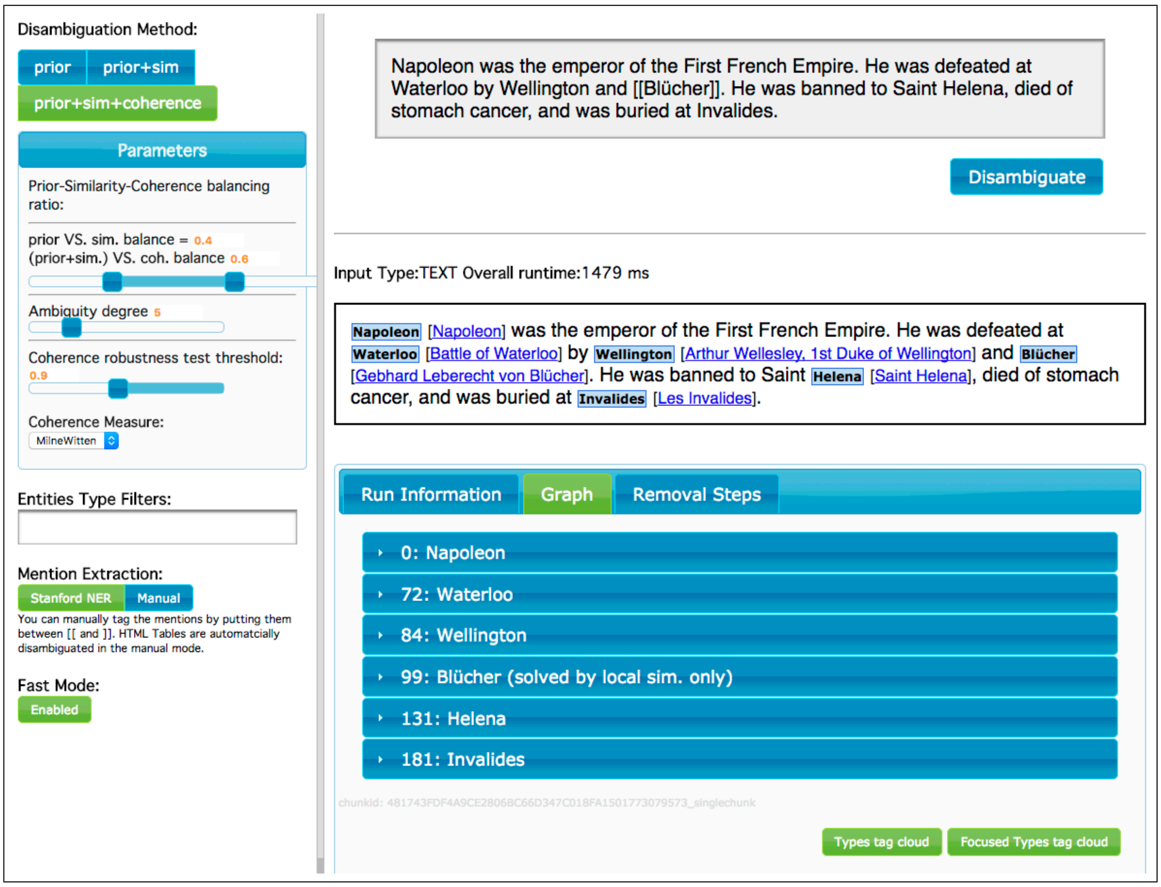


Figure 2.6 Entity disambiguation example by AIDA system [81]

The improvements in the entity linking systems have led to the development of several entity recognition and disambiguation tools such as AIDA [81] and Tagme [58]. An example of entity disambiguation with the AIDA system is illustrated in Figure 2.6. Construction of high quality knowledge bases such as YAGO [82] and DBpedia [16] have propelled high precisions entity linkers. It has been acknowledged that defining entities in Web data has been difficult [18], however, often entities are categorized as people, organizations, and locations. In the context of information retrieval, such entities tagged in textual documents can be leveraged as additional signals to improve result quality.

Several analytics tasks that involve entities have been studied recently. Some examples of the tasks are entity ranking [166] that was investigated as an INEX track [47], related entity finding and entity list completion were also a part of the TREC 2010 Entity Track [18]. However, explicitly combining annotated entities in documents to improve document retrieval quality is an open problem. For this thesis, we define the entity-based information retrieval as,

> **Definition 2.4:** *Entity-based information retrieval is finding text documents, annotated with disambiguated named entity mentions, that satisfy a user's intent that is expressed as a query which can be represented as combination of textual keywords and entities.*

One of the bottlenecks has been a lack of large annotated document collections. However, recently Google released the FACC1 dataset [61] with entity mentions linked to the Freebase knowledge base [27] for the TREC ClueWeb09 [2] and ClueWeb12 [3] collections. As the first attempt to leverage entities for document retrieval Dalton et al. [44] perform experiments on the TREC and the Robust04 collection [8]. For their study, they use the TREC Web track queries with explicit manual entity annotations. The focus of their work was twofold. First, to design appropriate representations of queries and documents using the annotated entities so as to improve retrieval effectiveness. Second, inferring latent entities that are salient to a query.

In their study, Dalton et al. [44] introduce a new query expansion technique, an entity modeling technique for capturing query-specific context, and a ranking function that leverages a combination of text and entities. In their experiments, they found that leveraging named entity mentions in textual documents can significantly improve retrieval quality.

2.3 Text Summarization

The primary consumption of news is now increasing online and consequently, the global news industry has witnessed a drastic shift of its focus from traditional print media to publishing digital content. The vast amount of online information being generated from various news agencies, independent providers, and sometimes the end users themselves, has made it difficult to retrospect and develop a holistic understanding of daily news events. Thus, it has become integral to design tools that facilitate efficient and effective consumption of information in addition to tools that provide efficient access to the information contained over the World Wide Web or in large document collections such as news and Internet archives.

Text summarization [55] has been considered as an important tool to tackle the information overloading problem. Automatically producing summaries from large sources of text is one of the oldest studied problems in both IR and NLP. One of the first studies by Luhn [124] on automatic summarization can be traced back to 1958. A summary is defined by Radev et al. [168] as a text that is produced from one or more texts and contains a significant portion of the information in the original text(s), and that is no longer than half of the original text(s). With this definition of a summary, Mani et al. [127] define text summarization as follows:

Definition 2.5: *Text summarization is the process of distilling the most important information from a source (or sources) to produce an abridged version for a particular user (or user) and task (or tasks).*

This definition captures the following three important aspects that characterize research directions on automatic summarization [168]:

- Summaries may be produced from a single document or multiple documents.
- Summaries convey important and diverse topics in the source documents.
- Summaries should be short (a length budget is specified before summarization).

One can note that the above definition of text summarization is fairly broad and prior works can be categorized based on the different problem definitions. Problem definitions that consider a single document as input are categorized as *single-document summarization* tasks. On the other hand, those that consider a cluster of documents as input are categorized as *multi-document summarization* tasks [20–22, 34].

To ensure summaries convey important and diverse topics, systems try to optimize mainly three properties of the summaries [137]: *relevance* to the user needs often expressed as a query is maximized; *redundancy* of information is avoided; and *length* of the summary is bounded by a given budget often expressed in terms of *words*. This can be achieved by two means. First, selecting appropriate text (sentences and paragraphs) from the source documents, and second, generating text in a new way. This categorizes summarization tasks into *abstractive* and *extractive* [172]. While the abstractive summarization focuses on generating more coherent text with deeper understanding of the text semantics, the extractive summarization [34, 83, 116, 120, 137, 172, 221] mainly consists of selecting most informative and non-redundant sentences from the input set of documents to constitute their summary.

Finally, text summarization can be posed either as a supervised or an unsupervised task [126]. In a supervised setting [56, 83, 116, 221], the task additionally comes with a set of example summaries which can be used to learn summarization models. Learning techniques can be designed by leveraging sufficient labeled data, where sentences that end up in a summary become positive samples. In this context, approaches cast the summarization problem into sentence classification and leverage SVM-based and neural network based classifiers. On the other hand, in an unsupervised setting [34, 120, 137, 172] , the task does not consider any training examples as input. An unsupervised summarization system has access only to the corpus and relies on document and corpus level statistics while generating a summary.

The problem addressed in this thesis falls into the category of unsupervised extractive multi-document summarization. Thus, in the rest of the section, we focus on prior works in this research direction.

2.4 Extractive Multi-Document Summarization

The primary goal of extractive text summarization can concisely be defined as extracting textual units which convey information about most important concepts and topic in input text (s). This problem definition encapsulates various challenges. Firstly, given a text as input, it must be broken into smaller units (sentences or paragraphs) such that each unit provides a complete piece of information. A naïve method, such as a simple sentence splitter based on period, may result with sentences with unresolved anaphoric mentions which can lead to complete misinterpretation of information in a summary. Secondly, the importance of the concepts and information in the different textual units must be determined for a given information need. Thirdly, the textual units must be chosen by keeping other units in view to avoid redundancy in a generated summary. Finally, methods have to often perform the above steps with the length constraint imposed. We next describe the steps in a little detail.

2.4.1 Global Inference Problem

Extractive summarization methods attempt to optimize the following three properties:

1. *Relevance*: Summaries should contain relevant and informative textual units. Let function $Rel(i)$ return a relevance score of a textual unit t_i.

2. *Redundancy*: Summaries should not contain multiple textual units that convey the same information. Let function $Red(i, j)$ return a score indicating the information redundancy between text units t_i and t_j.

3. *Length*: Summaries are bounded in length. Let function $l(i)$ return the length of textual unit t_i (usually in terms of number of words).

The joint optimization of all the three properties is referred to as the *global inference problem.* Formally, McDonald [137] defines the global inference problem as,

$$S = \arg\max_{S \subset R} \sum_{t_i \in S} Rel(i) - \sum_{t_i, t_j \in S, i<j} Red(i,j) \qquad (2.22)$$
$$\text{subject to: } \sum_{t_i \in S} l(i) \leq L$$

Here, R is a set of input text units and L the total length budget that is considered as an additional input. According, to the above definition, the goal is to select a set of text units t_i to compose a summary S such that the total length (for example in terms of word count) is under a specified budget L by maximizing the relevance of the summary to the user's intent specified as a query while avoiding redundancy.

Important textual units are identified via a pair-wise comparison of all textual units. Alternatively, in case a query is given (e.g. for query-focused summarization), relevance can be computed by comparing the textual units against the query. In multi-document summarization, high degree of redundancy among extracted textual units from different documents may result in uninformative summaries. Finally, the length budget constraint enables fair empirical comparison [117] of summaries that are automatically and human written. Furthermore, it represents an important real-world scenario where summaries are generated in order to be displayed on small screens, like smart phones.

It can be shown that a global summarization system is NP-hard through the reduction from 3-D matching (3DM). For the reduction and proof, we refer to McDonald et al. [137]. Most commonly, the global inference problem is solved in two ways. The first approach is to optimize the three criteria in tandem. Carbonell and Goldstein [34] present the maximal marginal rank criterion as one of the first global models where the sentence scoring is achieved by linearly combining relevance and redundancy estimates (we discuss the details of this method later in the section). The second approach is to optimize the relevance and redundancy separately [138]. In this context, often the textual units are first clustered based on topical or thematic similarity. Then, redundancy is reduced by selecting units from different clusters.

In this thesis, we propose a global inference model that jointly optimizes all the necessary conditions for event-focused summarization. Thus in rest of the section, we review some popular approaches to global inference problem.

2.4.2 Maximal Marginal Relevance

In the context of unsupervised extractive multi-document summarization, one of the most widely known frameworks is the maximal marginal relevance (MMR) that is proposed by Carbonell and Goldstein [34]. The algorithm iteratively selects sentences from a given input set that is most relevant to a user query Q and least redundant to the set S of already selected sentences D in the summary. Thus, the objective is modeled

Algorithm 1 Greedy Approach for Global Inference.

Input: Set of sentences $R = \{t_1, ... t_n\}$, and length budge L

Output: Summary S

1: sort R so that $Rel(i) > Rel(i+1) \quad \forall\, i$
2: $S = \{t_1\}$
3: **while** $\sum_{t_i \in S} l(i) < L$ **do**
4: $\quad t_j = \arg\max_{t_j \in R \setminus S} score(S \cup t_j)$
5: **end while**
6: return S

as a linear combination of a relevance function and the *novelty* of a sentence that is considered to be selected. Here, the novelty is estimated based on the most redundant or least similar sentence in S and used as a penalization component in the MMR criterion. The relevance score of a sentence against a query that is penalized with its novelty score computed against already-selected sentences into the summary is referred to as the *marginal relevance*.

Formally, let c be the document collection, q is a user query, and L is a budget, the set R contains top-k sentences (or documents) in c ranked by an IR system. The MMR criterion is defined as,

$$MMR = \arg \max_{d_i \in R \setminus S} \left[\lambda(Sim_1(d_i, q)) - (1-\lambda) \max_{d_j \in S} Sim_2(d_i, d_j) \right] \tag{2.23}$$

In this criterion, the function $Sim()$ computes the similarity of a sentence to the query. Further, set difference $R \setminus S$ represents the sentences that are not already selected into the summary. Finally, the parameter λ balances relevance and redundancy.

2.4.3 Greedy Approach

As described before, the global inference problem that is a reduction of the 3DM problem is NP-hard. Thus, greedy approximations with MMR style algorithms have widely been used in prior works. The main advantage is that such algorithms can easily be implemented and are computationally efficient.

The greedy algorithm to the global inference problem with MMR style criterion proposed by McDonald [137] is defined in Algorithm 1. The greedy algorithm starts by selecting most relevant sentence to query. Then the algorithm proceeds by greedily selecting the next sentence t_j that maximizes the objective $score(S \cup t_j)$ and computes a sentence score analogous to Equation 2.22. The runtime complexity of the algorithm is $O(n \log n + Kn)$ when calculating $O(S)$ is considered to be $O(1)$.

Algorithm 2 Knapsack Algorithm for Global Inference.

Input: Set of sentences $R = \{t_1, \dots t_n\}$, and length budge L

Output: Summary S

Initialize: $S[i][0] = \{\} \quad \forall\, 1 \leq i \leq n$
for $i : 1 \to n$ **do**
  for $k : 1 \to L$ **do**
    $S' = S[i-1][k]$
    $S'' = S[i-1][k - l(i)] \cup \{t_i\}$
    if $s(S') > s(S'')$ **then**
      $S[i][k] = S'$
    else
      $S[i][k] = S''$
    end if
  end for
end for
return $\arg\max_{s[n][k], k \leq L} score(S[n][k])$

2.4.4 Dynamic Programming Approach

An alternative solution to the global inference problem is to cast it into a 0-1 knapsack problem. The problem is to fill a knapsack of K capacity with a set of items each having a value and weight. In the context of summarization, the value is the score of a sentence and weight is its length. The goal is to fill a knapsack of total capacity L which is the length budget. Without the redundancy component in the global inference as shown in Equation 2.22, the problem will be identical to the 0-1 knapsack problem. To incorporate the redundancy component in Equation 2.22, after each step the novelty score of the sentences have to be recomputed. For detailed explanation of the algorithm, we refer to the description given by McDonald [137].

The dynamic programming algorithm to the global inference problem with MMR style criterion proposed by McDonald [137] is defined in Algorithm 2. In comparison to the greedy approach described before, the dynamic programming algorithm avoids selecting longer and noisy sentences. Greedily selecting longer sentences limits the space (due to the length constraint) for future inclusion of other informative sentence, thereby reducing the summary quality. The runtime of this algorithm is $O(n \log n + Ln)$, where L is a fixed lower bound on the run time.

2.4.5 Integer Liner Programming-based Approaches

In the greedy and the dynamic programming algorithm, discussed before, gives an approximated solution to the global inference problem. An alternative method to

Algorithm 3 McDonald's ILP to generate an event digest.

Maximize:

$$\sum_i [\lambda \cdot Rel_i S_i - (1-\lambda) \cdot \sum_{j \neq i} Red_{ij} S_{ij}]$$

Subject to:

1: $S_{ij} \leq S_i \quad \forall\ i$
2: $S_{ij} \leq S_j \quad \forall\ i$
3: $S_j + S_j \leq 1 + S_{ij} \quad \forall\ j$
4: $\sum_i l_i S_i \leq L$

generate an exact solution is by the use of Integer Linear Programming (ILP). An ILP is used to solve constrained problems where the objective and constraints are linear and in a set of integer variables. ILPs are well studied and are known to provide optimal solutions with branch and bound algorithms [112]. Further, modern ILP solvers such as GNU Linear Programming Kit[1] and Gurobi[2] can efficiently solve large ILP problems. Next, we look into some of the popular ILP formulations to address the global inference to address the summarization task.

McDonald's Approach

The global inference problem that is defined as the MMR criterion can be modeled also as ILP. The first attempt to design an ILP-based global inference approach that had a MMR-style objective was made by McDonald [137]. For an optimal solution to the original MMR criterion as presented by Carbonell and Goldstein [34] (described as Equation 2.22) requires solving 0-1 quadratic problem which includes a $max()$ function, thus making the overall problem non-linear. To make it linear, McDonald [137] proposed to change the $max()$ function to summation with additional constraints to ensure validity of the solution. Formally, the ILP is described as Algorithm 3.

In the objective, S_{ij} is a binary variable that indicates the addition of sentence i and j into the summary. Further, the $max()$ function is replaced by summation thereby penalizing a sentence with the average redundancy to the other sentences in summary.

Riedhammer's Approach

The formulation proposed by McDonald [137] is close to the original MMR, however the linear approximation used for redundancy part makes it slightly different. As a more recent approach Riedhammer et al. [172] proposed an ILP formulation that expresses the global MMR criterion more accurately. They extend McDonald's formulation by adding another additional binary variable M_{ij} that keeps track of the most redundant sentence

[1] http://www.gnu.org/software/glpk/
[2] http://www.gurobi.com/

Algorithm 4 Riedhammer's ILP to generate an event digest.

Maximize:

$$\sum_i \left[\lambda \cdot rel_i S_i - (1-\lambda) \cdot \sum_{j \neq i} red_{ij} M_{ij}\right]$$

Subject to:

1: $\sum_j M_{ij} = S_i \qquad \forall\ i$
2: $M_{ij} \leq S_i \qquad \forall\ i$
3: $M_{ij} \leq S_j \qquad \forall\ j$
4: $M_{ik} \geq S_k - (1 - S_i) - \sum_{j:red_{ij} \geq red_{ik}} S_j \qquad \forall i \neq k$
5: $\sum_i l_i S_i \leq L$

Algorithm 5 Gillick's approach to generate an event digest.

Maximize:

$$\sum_i w_i \cdot C_i$$

Subject to:

1: $O_{ij} S_j \leq C_i \qquad \forall\ i, j$
2: $\sum_j O_{ij} S_j \geq C_i \qquad \forall\ i$
3: $\sum_i l_i S_i \leq L$

in a summary to a given sentence that is being considered. This makes their formulation identical to the MMR criterion. Formally, the ILP is described as Algorithm 4.

The main idea behind the algorithm is to explicitly select the sentence that is most redundant to a sentence in the summary. For each sentence, the summary sentences are ordered based on their respective redundancy. Then, for a given sentence i, the most redundant sentence k is selected and the indicator variable M_{ik} is set to 1. This formulation has more constraint as compared to Algorithm 3, however it gets rid of the linear approximation and results in the optimal solution to the MMR criterion.

Gillick's Approach

In the context of global inference for summarization, another popular approach is presented by Gillick et al. [69] as the coverage-based framework based on ILP. As a contribution, they propose a concept-based summarization approach where the redundancy is avoided implicitly by maximizing over important *concepts* extracted from a given set of documents to be summarized. In their work, they define a concept as bigrams extracted from the documents. Formally, the ILP is described as Algorithm 5.

In the objective, the indicator variable C_i denotes if a concept i is included in the summary. The variable S_j indicates the presence of sentence j in the summary, Finally, the indicator variable O_{ij} indicates if a sentence j contains the concept i or not. Each concept is associated with an importance weight w_i. The score of the summary is determined based on the sum of the weights of the concepts that are selected. The first constraint ensures that the summary is less than the given budget L. The second constraint ensures that every concept that appears in the summary is also incorporated in the objective. The third constraint ensures that if a concept i is in the summary, then at least one sentence contains it.

2.5 Summary

In this chapter, we discussed technical foundations and looked into popular prior approaches in information retrieval and text summarization. We deem the fundamental techniques discussed in this chapter to be essential for understanding the work done in this thesis. In the upcoming chapters we either directly use or built upon the techniques discussed in this chapter. In the context of information retrieval, we specifically looked into the branch of techniques that deal with statistical and probabilistic approaches. We then discuss the sub-directions of information retrieval that look into incorporating additional signals like time, geolocations, and named entity mentions.

We acknowledge that text summarization is one of the oldest problems studied by the NLP and IR community. This also implies that there is a wide variety of approaches studied in the past. However, in this thesis we aim to design an extractive unsupervised summarization system that focuses on event queries. Further, we desire to design Integer Linear Programming-based methods that perform global inference. Thus, we described the popular and state-of the-art methods as technical background.

Chapter 3

Connecting Wikipedia Events to News Articles

3.1 Motivation & Problem Statement

The incomprehensible amount of information available online has made it difficult to retrospect on past events. As a remedy to this overloading problem, it is required to come up with a better structure and organization of the information across different sources so as to aid retrospection on past events. As an instance of such structure and organization, linking the different sources past events to provide multiple access paths to relevant information can be considered as a viable solution.

Online news articles are published contemporary to the events and report fine-grained details by covering all angles. These articles have been preserved for a long time as part of our cultural heritage through initiatives taken by media houses, national libraries, or efforts such as the Internet Archive. As a concrete example of such an effort, the archives of The New York Times go back until 1851.

The collaboratively authored free Wikipedia encyclopedia has emerged as a prominent source of information on past events. As a whole, the Wikipedia articles give contextual information on a central event (or an entity) and can help to build a good understanding on their evolution. However, often the articles summarize different aspects or other related events as short passages by abstracting from fine-grained details. Such events in the short descriptions do not always end up having their independent Wikipedia article giving additional information.

Individually, both Wikipedia and news articles are ineffective in providing complete clarity on multi-faceted events. On one hand, brief summaries in Wikipedia that abstract from the fine-grained details, make it difficult to understand all dimensions of an event. On the other hand, news articles that report a single story from a larger event do not make its background and implications apparent. What is badly missing are the

Table 3.1 Examples of Wikiexcerpts.

	No.	Wikiexcerpt
Year Article	1	**January 3, 1987:** Aretha Franklin becomes the first woman inducted into the Rock and Roll Hall of Fame
	2	**October 11, 1987:** The first National Coming Out Day is held in celebration of the second National March on Washington for Lesbian and Gay Rights.
General Article	3	**Jaber Al-Ahmad Al-Sabah:** After much discussion of a border dispute between **Kuwait** and **Iraq**, **Iraq** invaded its smaller neighbor on *August 2, 1990* with the stated intent of annexing it. Apparently, task of the invading **Iraqi** army was to capture or kill Sheikh Jaber.
	4	**Static cling:** In advertising Advertisers in urban areas, eager to use guerilla marketing techniques, have turned to static cling as a distribution medium. In an advertising campaign for Microsoft's MSN 8 Internet service, on *October 24, 2002*, hundreds of decals of the MSN butterfly logo were affixed to surfaces in **New York City** and held there with static cling.

connections between excerpts from Wikipedia articles summarizing events and news articles. With these connections in place, a Wikipedia reader can jump to news articles to get the missing details.

The idea of organizing vast document collections such as archives or newswire services by automatically constructing hypertext links has been studied extensively in the past [7, 40, 57, 178, 207]. Connections between documents provide structure to a large collection thereby enabling access to information in multiple ways. With the connections in place, and an additional search engine over the collection, a user can satisfy her information need from directly associated documents based on the hyperlinks, and by retrieving content with the search engine [7]. Traditional information retrieval techniques based on syntactic matching have been considered to be insufficient for achieving high linking quality in automatic hypertext systems. What is missing in such techniques is the semantic component to improve the document content representation [178]. In our context of linking information across different sources on past events, the document content can be better represented based upon time, geolocations, and entity mentions that come as additional semantics.

In this chapter, to establish connections between Wikipedia and news articles as two sources of information on past events, we define the following linking problem:

> **Problem Statement:** *Given an excerpt from Wikipedia, coined Wikiexcerpt, as a source summarizing an event, automatically identify past news articles as targets providing contemporary accounts.*

We cast the linking problem into information retrieval tasks thereby considering a given Wikipedia excerpt as a user query. Our goal is to retrieve a ranked list of news articles from a target corpus which can be used as references for the event described by a given Wikiexcerpt. The larger problem is broken into two subproblems that independently consider two types of Wikiexcerpts as input. First, those that occur in special *Year* articles (e.g., http://en.wikipedia.org/wiki/1987) that list seminal newsworthy events that have happened in a specific year. The first two examples illustrated in Table 3.1 represent such excerpts. Second, we consider excerpts that occur as arbitrary passages in general Wikipedia articles. The last two examples in Table 3.1 represent such excerpts.

Standard document retrieval models for keyword queries rely on syntactic matching and are ineffective for our task. Due to the verbosity of Wikiexcerpts, they are prone to topic drift and result in lower retrieval quality. The Wikiexcerpts also contain additional semantics like temporal expressions, geolocations, and named entities which can be leveraged to identify relevant documents. Making standard document retrieval models aware of these semantic annotations so as to identify contemporary and relevant documents is not straightforward.

Approach Overview

We start with the simpler problem by considering the Year article events as queries, and news articles in news archives as targets of the linking task. This problem is relatively simpler as the input events in the year pages have short (usually single sentence) descriptions. In addition, they come with an explicit date indicating the occurrence period of the event. We motivate that leveraging time in combination with the text may lead to improved linking quality. To address this task, we develop several methods that integrate publication dates and temporal expressions (phrases like “in 2000” representing time) into statistical language models. Additionally, we propose a two-stage cascade approach that leverages text and time as event dimensions to retrieve relevant news articles from a large archive. This method, in the first stage, estimates query-text and -time models from textual terms and temporal expressions respectively, from a set of initially retrieved pseudo-relevant news articles. Then in the second stage, it re-ranks the initially retrieved articles by comparing them against the query models. Our experimental evaluation on 50 randomly sampled Wikipedia events with crowd-sourced relevance assessments shows that the two-stage cascade approach is found to be the most effective among the methods under comparison that models time differently.

To demonstrate the practicality for the retrieval models designed to connect Wikipedia events to news articles, we introduce *EXPOSÉ*, a time-aware exploratory search system that explicitly uses temporal information associated with events to link different kinds of information sources for effective exploration of past events. In this demonstration, we use Wikipedia and news articles as two orthogonal sources where Wikipedia is viewed as an event directory that systematically lists seminal events in a year, and news articles as a source of detailed information on each event. To this end, our demo exhibits implementations of different time-aware retrieval models which a user could choose while retrieving relevant news articles, a timeline tool for temporal analysis, and entity-based facets for filtering results so as to pin point documents that satisfy information needs.

Gathering insights from our study to address the simpler linking problem, we address the more difficult problem of connecting arbitrary passages extracted from general Wikipedia articles to news articles. The difficulty increases due to the higher degree of verbosity. Additionally, we find that excerpts describing events that are taken from general Wikipedia articles, often come with additional semantics in their textual descriptions representing the time, geolocations, and named entities involved in the event. Our method leverages text and semantic annotations as different dimensions of an event by estimating independent query models to rank documents. In our experiments on two datasets and with 150 randomly sampled excerpts as test queries, we compare methods that consider different combinations of dimensions. We find that the method that combines all dimensions is the most effective to address our problem in comparison to other combinations of the dimensions.

Other Potential Applications

We motivate the linking task as an effort to better structure vast amounts of information contained in different information sources to aid retrospection on past events. However, our proposed generic methods can be potentially applied to several downstream applications. The retrieval model proposed to address the linking task can be used to identify informative documents, within large collections, in the context of a specific event. These documents can then be considered as input for further tasks like event information extraction [37], event summarization (described in Chapter 4), and exploratory search systems [30]. Other more focused potential applications where the methods can be applied with minor customization are: enriching short social network post with news articles [202]; connecting news articles to online news broadcast [79]; inter-connecting news articles within an archive to increase their consumption by end users [14]; and automatically suggesting relevant news articles to content providers as an authoring tool [50]. Such tools are developed to assist users to efficiently author text . We motivate these tools can leverage the linking systems to identify high quality reference documents from large unstructured textual document collections.

Contributions

We make the following key contributions in this chapter:

- We propose a novel linking problem with the goal of connecting Wikipedia excerpts describing events to news or web articles in a large archive.
- We design time-aware language models, and a two-stage cascade model that estimates query-time model from pseudo-relevant documents to retrieve textually and temporally relevant news articles.

- To leverage additional semantic annotations, namely time, geolocations, and named entities, along with the textual descriptions of Wikiexcerpts, we design novel query modeling techniques to estimate independent models, thus treating them as dimensions of an event.
- Leveraging the independent query models, we present a KL-divergence-based framework to generate a ranked list of news articles by comparing them to a given query along the event dimensions.
- Finally, we present a prototype time-aware exploratory search system for exploring events that were seminal in the past. We implement our different time-aware retrieval models as modes of exploration available to a user as options.

Organization

The rest of the chapter is organized as follows. In Section 3.2, we put our work in context with prior work. We focus on leveraging a combination of text and time that comes with event descriptions, and describe our methods in Section 3.3. We present the *EXPOSÉ* exploratory search system in Section 3.4 that builds on our time-aware methods, and illustrate the practicality of our retrieval framework. Methods designed to leverage additional semantics like time, geolocations, and named entities to improve the linking task are described in Section 3.5. Finally, we conclude this chapter and point out the future directions in Section 3.6.

3.2 Related Work

We review the following five lines of prior research that are related to our work: **1)** first we look into the efforts made in the past to automatically link different document archives. **2)** Next, we look into the different approaches to address temporal information retrieval that combines text and time for document ranking. **3)** We also look into prior efforts made to address geographical information retrieval. **4)** As the fourth research line, look into methods that additionally leverage disambiguated named entity mentions for document ranking. **5)** Finally, our as the last line of prior works, we review approaches that leverage statistical language models to address inflammation retrieval.

Linking Document Archives

As the first line of prior work, we look into efforts to link different document collections. Henzinger et al. [78] automatically suggested news article links for an ongoing TV news broadcast by treating the embedded closed captions as the viewer's information need. More recent studies have found that linking similar (related) items across multiple archives can supplement background information to improve exploration [30, 76]. Jiyin et al. [76] link narrative accounts of a radiologist's findings, diagnoses, and recommendations to Wikipedia articles as a source of background information. Born et al. [30] link event descriptions in textually rich news archives to a sparsely annotated video library containing TV news. Another direction of linking efforts goes towards enriching social media posts with links to news articles. Tasgkias et al. [202] aims at identifying news articles that are implicitly referred to by social media posts. Recently, Arapakis et al. [14] propose an automatic linking system between news articles describing similar newsworthy events. The focus of all these work has been on coping with disparate quantities of text and bridging language gaps between the source and the target. As a difference, we aim at connecting excerpts from Wikipedia articles to news articles by leveraging semantic annotations.

Temporal Information Retrieval

As the second line, we look into prior works that make use of temporal information such as publication dates, and temporal expressions to improve effectiveness in ad-hoc retrieval. Li and Croft [115] introduce a time-dependent exponential document prior in a language modeling approach. Other document priors motivated by findings from cognitive science have been considered more recently by Peetz et al. [161]. Time has also been considered as a feature for query profiling and classification [103]. Berberich et al. [26] made use of temporal expressions in documents to better deal with explicitly temporal queries. Other efforts such as Jones and Diaz [93] and Peetz et al. [162] have used temporal information for query modeling in temporal query classification and ad-hoc retrieval, respectively. Efron et al. [51] present a kernel density estimation method to temporally match relevant tweets. All of the approaches use only system-generated timestamps such as publication dates to filter documents [33]. In contrast, our approach exploits the publication date and the temporal expressions that are mentioned in pseudo-relevant documents to design a ranking model.

Geographical Information Retrieval

There have been many prior initiatives to evaluate approaches in the realm of geographical information retrieval. The GeoCLEF search task examined geographic search in text corpus [65, 66, 125]. More recent initiatives, like the NTCIR-GeoTime task [67] evaluated adhoc retrieval with geographic and temporal constraints. Among early work, Lana et al. [107] presented the MIRACLE system that performs expansion of the user's query by identifying geographic entities and spatial relationships between them. More recently, Perea-Ortega et al. [163] propose NLP techniques to expand geographic queries considering both textual and geographic aspects. Studies [42, 75] have also found that combining geographic and textual ranking functions shows overall improvements in retrieval quality. Hariharan et al. [75] propose a combined ranking function that linearly interpolates two functions assigning textual and geographical relevance scores to documents. Geolocation-based ranking has also been studied for post retrieval in social media. As a recent work, [42] ranks social media posts by combining geographical, temporal, and textual ranking functions that take popularity of user posts into consideration. In this chapter, we target a different task. In our approach, we use the geographic information in queries to build a query-space model. This is then combined with other query models for other dimensions to rank documents based on their divergence.

Entity Retrieval

We look into prior research efforts that use entities for information retrieval. Earlier initiatives like the INEX entity ranking track [47], and the TREC entity track [18] focus on retrieving relevant entities for a given topic. More recently, the INEX Linked Data track [24] aimed at evaluating approaches that additionally use text for entity ranking. Other works like Yahya et al. [211] exploit a combination of text and entities for question answering. As the most recent work, Dalton et al. [44] exploits entities for document retrieval. In our approach, we treat entities as an additional dimension of the query. We build a query-entity model using the entity mentions in the query. This is then combined with query models for other dimensions.

Text Retrieval

Divergence retrieval models for text have been well studied in the past. The main motivation for these models is that they can easily accommodate feedback to improve the query models. Zhai et al. [216] combine a generative feedback model to the query model through linear interpolation. In their study, Zai et al. [217] compare techniques of combining background models to query and document models. To further improve the query model estimation, Shen et al. [185] exploit contextual information like query history and click-through history. Bai et al. [17] present a general framework to combine query models estimated from multiple contextual factors. In our approach, we use a KL-divergence retrieval model to rank documents. We find prior studies investigating methods that combine query with background models helpful to develop our approach. We design event-specific background models from a proxy corpus.

3.3 Leveraging Time + Text

Descriptions of events are abundant in Wikipedia, and are systematically curated in Year, Decade, and Century articles. Moreover, initiatives like the Wikipedia Current Events[1] portal aim at organizing seminal events as timelines. Generally, the events listed in these special Wikipedia articles come with short (usually single sentence)

[1] https://en.wikipedia.org/wiki/Portal:Current_events

descriptions. Additionally, they come with an explicit date indicating their occurrence period. However, even though the special articles represent elaborate lists, the events often are missing independent Wikipedia pages centrally describing them. Thus, it becomes difficult for a user browsing through such lists to get additional information on the events from Wikipedia.

News articles act as a secondary source of information on such seminal events as they were likely to receive vast media coverage. However, for a user, it becomes difficult to sift through large number of news articles with a high degree of redundancy to gather a holistic understanding on the events. This causes the information overloading problem as described in Section 3.1. In this context, with connections between the Wikipedia events and news articles in place, a user browsing through the lists can navigate between them to quickly gather more detailed information.

As illustrated in Figure 3.1, we address the problem of automatically linking seminal Wikipedia events that are listed in special Wikipedia Year articles to news articles from the past. We cast this problem into the following information retrieval task:

Problem Statement: *Given a Wikipedia event as a user query consisting of a short textual description and a date, automatically retrieve a ranked list of news articles providing details on the event which could be used as a background reference.*

As input, we consider short single sentence textual event descriptions that additionally come with a date indicating the occurrence period. Consider the following as a concrete example of such a Wikipedia event description or Wikiexcerpt (also illustrated in Table 3.1):

Example 3.1: January 3, 1987: Aretha Franklin becomes the first woman inducted into the Rock and Roll Hall of Fame.

As output, our goal is to retrieve a ranked list of news articles that give additional information on the event descriptions for a given target collection.

Challenges

Addressing the above problem includes the following key challenges:

- The Wikiexcerpts come with a date indicating the occurrence period of the described event. Leveraging time in addition to text may lead to improved result quality. Thus designing methods that leverage a combination of text and time becomes a challenge.

- The textual descriptions of Wikipedia events as Wikiexcerpts are typically verbose. As a well-known challenge, standard retrieval models are prone to topic drift.

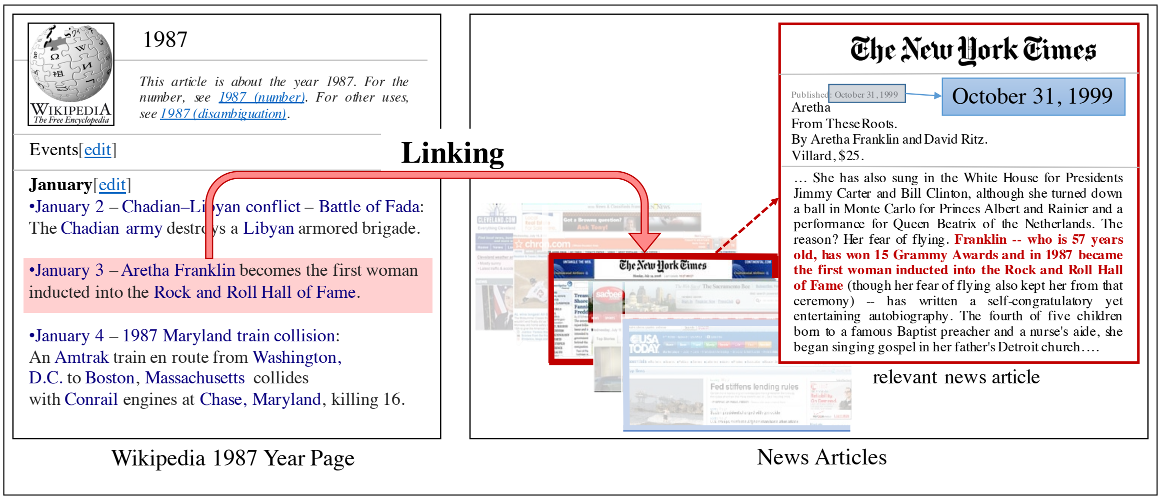


Figure 3.1 Connecting Wikiexcerpt in Example 3.1 from year "1987" to news articles.

In our approach to address the above described challenges, we design several time-aware methods that integrate time into language modeling techniques. We note that the input date along with a Wikiexcerpt denotes a single day. To deal with specificity of the date, we propose a two-stage cascade approach that estimates query-time model from pseudo-relevant temporal expressions. However, we first introduce the notation and representations used to design our methods.

3.3.1 Notation & Representations

We begin by describing the notation and representations used to design our methods.

Time Domain & Temporal Expression Model

Our time domain $T \times T$ is modeled as a discrete probability distribution over integers Z representing time units or *chronons* that are assumed to reflect the time passed (to pass) since (until) a reference date such as the UNIX epoch.

Any temporal expression mentioned in text is modeled as a time interval $[b, e] \in T \times T$ and assume $b \leq e$. For time intervals $[b, e]$ with coinciding boundaries (i.e., $b = e$) we use $[b]$ as a shorthand. When mapping temporal expressions contained in a document to this representation, we map onto the largest plausible time interval. Thus, "in May 2014", as a concrete example would be mapped onto [2014.05.01, 2014.05.31]. Likewise, "in the evening of March 31st 2011" would be mapped onto [2011.03.31].

Document Collection

Let D denote our document collection. A document $d^t \in D$ comes with a publication date $t \in T$, where T is our time domain. The document d^t consists of a textual part d_{text} and a temporal part d_{time}. The textual part d_{text} is a bag of words drawn from a fixed vocabulary V. The temporal part d_{time} is a bag of temporal expressions that includes its publication date.

Sometimes it will be convenient to treat the entire document collection as a single (coalesced) document. We use D_{text} and D_{time} to refer to the corresponding textual part and temporal part, respectively.

Query

Given a Wikipedia event like the ones illustrated in Example 3.1, we derive a query q from it as follows. The textual part q_{text} is obtained directly from the description of the event. The temporal part q_{time} is obtained from the indicated date. Note that, for the scope of this work, we only consider Wikipedia events that indicate a specific day as their date. Hence while $q_{time} \in T \times T$, we refer to q_{time} as a single time when convenient.

3.3.2 Time-Aware Language Models

We propose methods that leverage publication dates of the documents, and temporal expressions encoded in content to estimate temporal relevance of the documents to a given query. Additionally, we propose a novel two-stage cascade approach that builds a query model from the temporal expressions in pseudo-relevant documents. All our time-aware models are built upon the standard unigram language model with Dirichlet smoothing [214] as a foundation.

Text Only Approach

As the simplest method, the text-only method leverages standard query-likelihood retrieval method to rank documents. When considering only the textual parts of the query q_{text} and a document d_{text}, the relevance of a document d to the query q is estimated based on the probability of generating the document from the query that is computed as

$$P(q \mid d^t) = \prod_{w \in q_{text}} \frac{P(w \mid d_{text}) + \mu \cdot P(w \mid D_{text})}{|d_{text}| + \mu} \,. \tag{3.1}$$

Here, $P(w|d_{text})$ and $P(w|D_{text})$ are maximum-likelihood estimates of generating a word w from a single document d and the entire corpus D, respectively. Further, parameter μ is the parameter used in the Dirichlet smoothing and is set to the average document length [214].

Publication Date-based Approach

Intuitively, documents published around the time when the query event happened are more likely to discuss the event in detail. Our second method implements this idea by relating a document's publication date to the temporal part of the query q_{time}. Relevance of a document d to q is estimated as the generative probability,

$$P(q \mid d^t) = P(q_{text} \mid d_{text}) \cdot P(q_{time} \mid t)\,. \tag{3.2}$$

Here, the first factor is estimated according to Equation (3.1), and the second factor is defined as,

$$P(q_{time} \mid t) = \frac{1}{1 + e^{r|q_{time} - t|}}\,. \tag{3.3}$$

With this function, we thus favor documents published shortly before or after the time q_{time} when the query event at hand took place. Thus assuming recency as user's intent. This is illustrated in Figure 3.2.

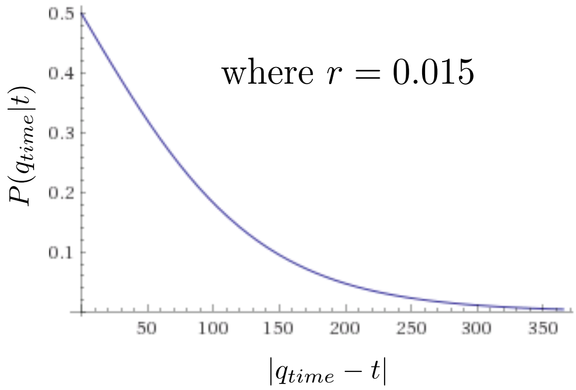


Figure 3.2 Plot illustrating the effect on $P(q_{time} \mid t)$ w.r.t. $|q_{time} - t|$ with $r = 0.015$.

Temporal Expression-based Approach

Temporal expressions can be another strong indicator for whether a document discusses the query event at hand. Intuitively, if many of a document's temporal expressions refer to the time period when the event happened, chances are that the document discusses the event. Our third method leverages this idea and estimates relevance of a document d to a query q based on the following generative probability,

$$P(q \mid d^t) = P(q_{text} \mid d_{text}) \cdot P(q_{time} \mid d_{time})\,. \tag{3.4}$$

Analogous to the publication date-based method, the first factor follows Equation (3.1), and the second is estimated based on a simplified variant of [26]. Formally, this is computed as,

$$P(q_{time} \mid d_{time}) = \frac{1}{|d_{time}|} \sum_{[b,e] \in d_{time}} \frac{\mathbb{1}(q_{time} \in [b,e])}{e - b + 1}\,. \tag{3.5}$$

To avoid the issue of zero probabilities [214], if none of a document's temporal expressions includes the query time q_{time}, we smooth this estimate by interpolating with a small constant. According to the above, a document yields a high probability $P(q_{time} \mid d_{time})$ if many of its temporal expressions, or more precisely the corresponding time intervals, are at a fine temporal granularity and include the query time q_{time}.

Publication Date + Temporal Expression-based Approach

As the method that gives importance to the publication dates and temporal expressions, we combine the above approaches, and rank documents according to the generative probability of a query q from a document d which is computed as follows,

$$P(q \mid d^t) = P(q_{text} \mid d_{text}) \cdot P(q_{time} \mid t) \cdot P(q_{time} \mid d_{time}) \tag{3.6}$$

with factors as defined in Equations (3.1), (3.3), and (3.5) respectively.

Two-Stage Cascade Approach

We now describe our *two-stage cascade approach*. As a key difference to the previous methods, the two-stage cascade method performs an initial round of retrieval using the unigram language model with Dirichlet smoothing described in Equation (3.1), and estimates a *query-time model* from the top-k documents retrieved, thus treating them as pseudo-relevant. It then re-ranks the top-K documents (with $k \le K$) from the initial round of retrieval by taking into account their divergence from the temporal query model, their publication date, and their fit to the textual description of the query event.

Intuitively, by estimating a query-time model from pseudo-relevant documents, we cope with an overly specific query time q_{time} and instead consider salient temporal expressions related to the query event. This is expected to be particularly helpful for events that did not receive extensive coverage when they happened or whose ramifications linger on for long.

Stage 1: Let $R_k \subseteq D$ denote the set of top-k documents retrieved based on $P(q_{text} | d_{text})$. From those pseudo-relevant documents, we estimate the query-time model Q_{time} as,

$$P(\tau \mid Q_{time}) = \sum_{d \in R_k} \frac{P(q_{text} \mid d_{text})}{\sum_{d' \in R_k} P(q_{text} \mid d'_{text})} \cdot P(\tau \mid d_{time})\,. \tag{3.7}$$

Akin to Equation (3.5), the second factor is estimated as,

$$P(\tau \mid d_{time}) = \frac{1}{|d_{time}|} \sum_{[b,e] \in d_{time}} \frac{1}{|\tau|} \cdot \frac{\mathbb{1}(\tau \in [b,e])}{e-b+1}. \tag{3.8}$$

It can be noted that this factor computes the probability of generating a time unit τ from the time part of a document. This can be understood as the document-time model of d.

Equations (3.7) and (3.8) describe probability distributions over times $\tau \in T$. The document-time model assigns higher probability to times that are mentioned more often or through more specific temporal expressions in the document. Our query-time model then aggregates these probabilities from the pseudo-relevant documents in R_k, taking their initial query likelihoods into account. As a result, our query-time model assigns high probability to times mentioned often in textually relevant documents through temporal expressions.

Stage 2: In the second stage of our cascade approach, documents from R_K, the set of top-K documents retrieved initially, are re-ranked. As a first factor for re-ranking, we take the Kullback-Leibler (KL) divergence between the temporal query model estimated in Stage 1 and a smoothed temporal document model into account. Formally, this is computed as,

$$D(Q_{time} \parallel d_{time}) = \sum_{\tau \in Q_{time}} P(\tau \mid Q_{time}) \cdot \log \frac{P(\tau \mid Q_{time})}{P'(\tau \mid d_{time})} . \tag{3.9}$$

Analogous to our previous methods, we tackle the zero-probability issue which arises due to the sparsity in the documents by additionally smoothing with the time part of the entire corpus D_{time}. Thus we treat the entire corpus as a background model and linearly interpolate with the document models as,

$$P'(\tau \mid d_{time}) = \lambda \cdot P(\tau \mid d_{time}) + (1 - \lambda) \cdot P(\tau \mid D_{time}) . \tag{3.10}$$

As a second factor, we consider the publication dates of documents. To this end, we determine the KL-divergence between q_{time} and the publication date t, in analogy to Equation (3.3), which in our specific setting, where q_{time} refers to a single time, simplifies to

$$KL(q_{time} \parallel t) = \log(1 + e^{r|q_{time}-t|}) . \tag{3.11}$$

Finally as a third factor, we also take the textual parts of the query and the document into account. To this end, we consider the KL-divergence $KL(q_{text} \parallel d_{text})$ between their Dirichlet smoothed unigram language models estimated in analogy to Equation (3.1). Putting it together, we re-rank documents according to the score computed as,

$$score(q,d) = \alpha \cdot KL(Q_{time} \| d_{time}) + \beta \cdot KL(q_{time} \| t) + \gamma \cdot KL(q_{text} \| d_{text}) \tag{3.12}$$

where α, β, and γ are tunable parameters.

3.3.3 Experimental Evaluation

Next, we provide details on the setup and results of our experimental evaluation.

Document Collection

We use The New York Times Annotated Corpus [4] which contains about 2 million documents published between 1987 and 2007. The detailed statistics of the dataset is described in Table 3.2.

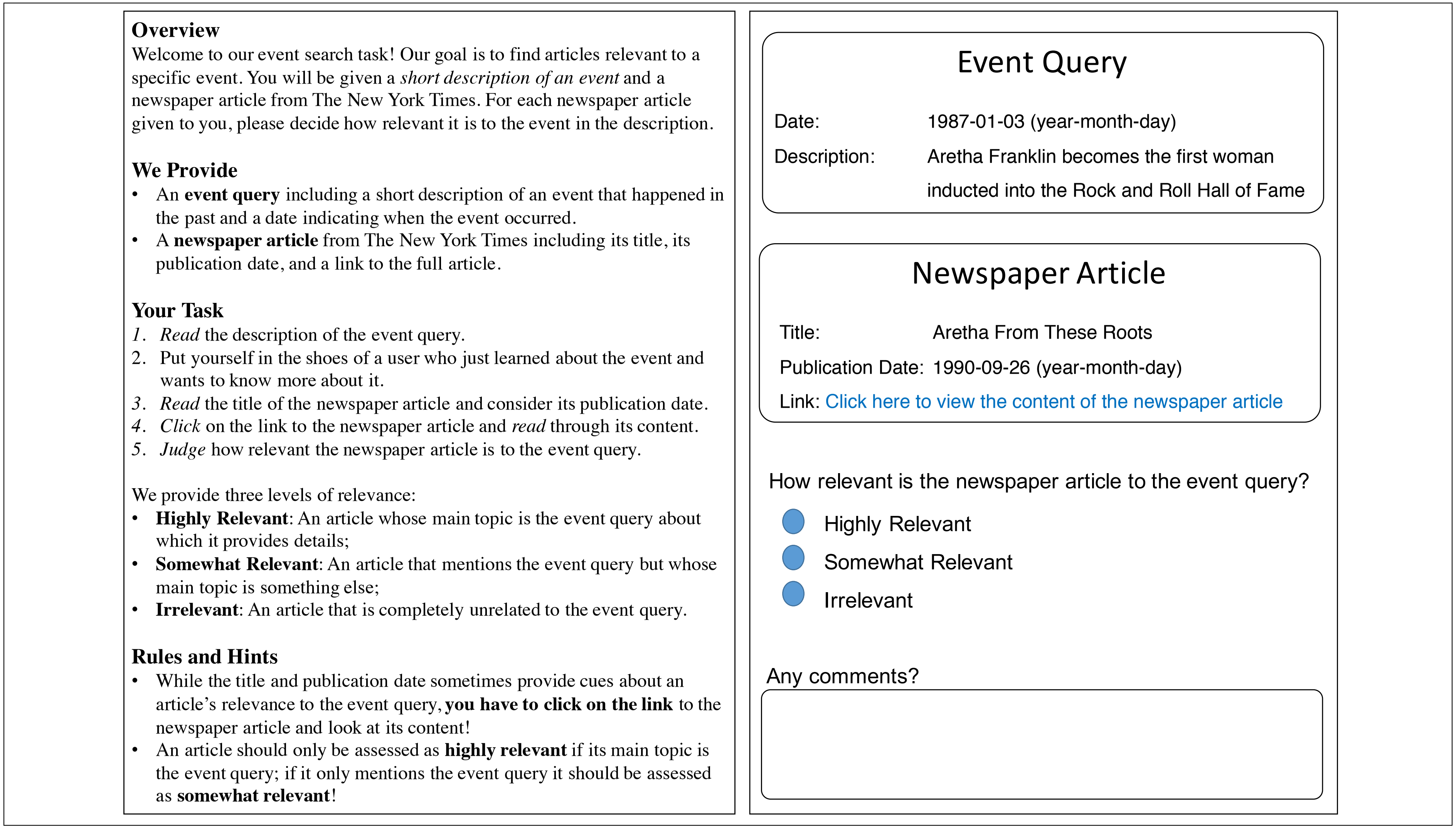


Figure 3.3 Crowdflower instructions and interface shown to a contributor.

Table 3.2 Document collection statistics

Number of documents	1,855,656
Mean document length in words	691.79
Standard deviation of document length in words	722.88
Number of temporal expressions per document	6.35
Standard deviation of temporal expressions per document	5.86

Test Queries

We use *The English Wikipedia* dump released on February 3rd 2014 and randomly sample 50 Wikipedia events as queries from the articles for the years 1987 to 2007. We include a full list of the test queries for experiments in Appendix A.1.

Relevance Assessments

We collected relevance judgments using the Crowdflower platform. We pooled top-10 results for the methods under comparison, which resulted in 1,297 unique query-document pairs. We asked contributors to judge whether the document was **(0)** irrelevant, **(1)** somewhat relevant, or **(2)** highly relevant to the given query. Our instructions said that a document can only be considered highly relevant if its main topic was the event given as a query. The full instructions specified and the Crowdflower interface shown to a contributor is illustrated in Figure 3.3. For quality control, we additionally judged 10% of the units ourselves which were then set as test questions. To further ensure good quality results, we only allowed "highest quality" contributors to work on our task. These contributors are described as the smallest group of the most experienced and the highest accuracy contributors. We set the language requirement of the contributors to "English". Each query-document pair was judged by at least three contributors. We paid $0.03 per batch of five query-document pairs. The total agreement (computed by Crowdflower) between the contributors was found to be 66.57% which indicates that the judgments may be subjective. Thus final judgments were obtained with majority voting however, the votes were weighted based on a trust score (based on prior jobs computed by Crowdflower) of a contributor. The entire job ran for a total of 152 hours and its total cost came to be 57$ including the additional fees charged by Crowdflower.

Methods & Parameters

In our experiments, we compare the following methods:

- *LM* as a unigram language model with Dirichlet smoothening according to Equation (3.1) using $\mu = 1{,}000$ (this parameter can be safely set to a value larger than the average document length [214]).

	LM	*LM+P*	*LM+T*	*LM+PT*	*CM*
MAP	0.35	0.43	0.40	0.42	**0.45**
P@5	0.55	0.63	0.61	0.61	**0.66**
P@10	0.48	0.57	0.54	0.54	**0.58**
NDCG@5	0.53	0.58	0.58	0.60	**0.60**
NDCG@10	0.54	0.62	0.60	0.62	**0.63**

Table 3.3 Comparison of effectiveness of the different methods.

- *LM+P* as the method integrating documents' publication date according to Equation (3.3) using $r = 0.015$ [162].
- *LM+T* as the method integrating temporal expressions alongside unigrams according to Equation (3.5).
- *LM+PT* as the approach considering both publication dates and temporal expressions according to Equation (3.6).
- *CM* as our two-stage cascade approach. Here, we found that building the query model from the top-10 initially retrieved documents and using it to re-rank the top-30 initially retrieved documents gives the best result. We employ an exhaustive grid search to find most suitable mixing parameter values of α, β, and γ for our data. For this, we test one parameter at a time while keeping the others fixed and optimize for MAP values across all the test queries. We find that the values $\alpha = 0.20$, $\beta = 0.60$, and $\gamma = 0.20$. The smoothing parameter is borrowed from the literature [214] and set as $\lambda = 0.85$.

Experimental Results

We measure retrieval effectiveness using Mean Average Precision (MAP) as well as Precision (P) and Normalized Discounted Cumulative Gain (NDCG) at cut-offs 5 and 10. For MAP and P we consider a document relevant to a query, if the majority of assessors judged it with label **(1)** or **(2)**. For NDCG we plug in the mean label assigned by assessors.

Table 3.3 lists retrieval effectiveness measures for our five methods under comparison. It can be seen that *CM* consistently outperforms the baselines across all effectiveness measures, proving to be the most effective approach for the linking task. Comparing the baselines, we see that both *LM+P* and *LM+T* methods yield improvements over the text-only approach as in the *LM* method. Thus, considering either kind of temporal information helps with our linking task. Unfortunately, their effects are not additive, as can be seen from the performance of *LM+PT*, which does not yield a consistent improvement over the two and sometimes even performs worse than *LM+P*.

Gains & Loss Analysis

To get some insight into where *CM*'s improvements come from, we perform a gain/loss analysis based on P@10 by ranking queries according to the difference between *CM* and the best performing baseline. The biggest gain of +0.2 (with *LM+P* and *LM+TP* as the best performing competitors) is observed for the following query:

Example 3.2: March 19 2002: US war in Afghanistan: Operation Anaconda ends after killing 500 Taliban and Al-Qaeda fighters, with 11 allied troop fatalities.

For this relatively verbose query, the two-stage cascade approach yields a P@10 of 0.8. Here, the temporal query model shifts focus from the specific day to March 2002, which for this query (related to a relatively short operation) turns out beneficial.

We observe the biggest loss of −0.3 for the following query:

Example 3.3: February 27 1991: President Bush declares victory over Iraq and orders a cease-fire.

Interestingly, the best performing baseline in this case is *LM+PT*, which achieves a perfect P@10 of 1.0. In contrast to that, the *LM* achieves only a P@10 of 0.2. Here, the text-only baseline suffers from the ambiguity of the query (i.e., multiple presidents called Bush and multiple wars in Iraq) and is unable to focus its results on the right time period. As a building block, it also negatively affects the performance of our two-stage cascade approach *CM*, which through its re-ranking still achieves an P@10 of 0.7.

Easy & Hard Query Events

Finally, we identify easy and hard query events in our testbed. The easiest one, having the highest minimum P@10 across all methods, is the following query:

Example 3.4: August 4 1993: A federal judge sentences Los Angeles Police Department officers Stacey Koon and Laurence Powell to 30 months in prison for violating motorist Rodney King's civil rights.

The event description contains several discriminative terms that are part of an entity surface forms such as " Los Angeles Police Department", "Stacey Koon", and "Laurence Powell". Thus, even a simple *LM* methods, as the worst performing method, achieves a high P@10 score of 0.8 for this query event.

We identify the following query as the hardest query event in our testbed:

Example 3.5: May 3 1989: Cold War - Perestroika - The first McDonald's restaurant in the USSR begins construction in Moscow. It will open on 31 January 1990.

For the above query, upon a closer examination, we found that our target NYT corpus does not contain any relevant news article. Thus, all methods receive a P@10 score of 0.

3.4 EXPOSÉ: Exploring Past News for Seminal Events

To design an effective exploratory search system for seminal past events, one could think of connecting the Wikipedia events to relevant news articles. With these badly-missing connections in place, a user with an unclear information need can start by browsing through the Wikipedia lists organized collectively as a *directory of events* and jump to connected news articles for more fine-grained details. In addition, flexibility to submit typed description and date extends the scope of exploration to unlisted events or enables the user to simply refine a listed event. Further, visualization tools like facets and timelines offer valuable learning assistance while exploring past events.

Time becomes an interesting dimension of past events and can help to effectively explore their ramifications [192]. However, incorporating time into exploration is not straight-forward. General users may refer to a specific event date to explore news articles by having different temporal intent. For example, a user could choose to see news articles that were published close to the indicated date, or be interested in news articles that discuss the date in their content, or a combination of both. Decoupling text and time for exploration of temporal events and giving explicit handles for each would result in more effective and efficient exploration.

Consider the following use cases for further motivation:

- Melita Garner, a news curator, researching on *1987 Maryland train collision* wants to better understand how the story presented by the media unfolded as the event had happened. For this, she wants to explore news articles that are published around the event time period and sets her preference accordingly.

- Tony Stark, a computer scientist, wants to understand the ramifications of the *Dot-com bubble burst in January 2000*. For this, he would intend to explore all articles that are published around the event time period or discuss the time period in their content. In this regard, he sets his preference to view relevant news for exploration.

Both users performing exploratory search over news have different temporal intents. Melita Garner desires to explore a punctual event that has a short scope. For this, she expects relevant documents discussing different aspects such as causality, casualty, rescue operations, etc. to be published with a short time frame after the happening of the event. On the other hand, Tony Stark explores an event that has long ramifications. Thus, he expects relevant documents, irrespective of their publication date, should refer to the occurrence period of the event.

Motivated by the above scenarios, EXPOSÉ offers five modes of temporal exploration:

1. *relevant news* – automatically finds interesting time periods associated with an event as its temporal scope to identify relevant and contemporary news articles that discuss date in their content;

2. *news published around the event date* – uses the event date to identify relevant and contemporary news articles;
3. *news that mention the event date* – identifies relevant news articles that discuss the event date in their content;
4. *news published around the event date or mention the event date* – identifies relevant and contemporary news articles that also discuss the event date in their content;
5. *news by a standard text search* – identifies relevant news articles with the given textual description and completely ignores the event date.

Depending on information need, a user may choose to use one of the above modes to explore past events. Each mode uses a different temporal retrieval method in the background that we discuss in the next subsection.

State-of-the-Art Systems

State-of-the-art vertical search engines like Google news that provide only lookup services, do not help in exploration [132]. To satisfy temporal needs, they provide simple filtering on publication dates and do not explicitly regard the temporal content of news articles. In addition, an event description issued as a keyword query to such systems suffers from verbosity which leads to topic drifts.

For time-aware exploratory search, Odijk et al. [153, 154] design interfaces to visualize document temporal distribution and word clouds. These visualizations are useful for analyzing word meanings and system usage-history over time but do not seem to work for event exploration. As an extension, Reinanda et al. [171] presented a system to explore entity associations in longitudinal document collections. Though interesting, this system becomes difficult to extend to events.

3.4.1 Algorithmic Building Blocks

In this section, we describe five approaches with example queries for which they become most effective. For full details on individual approaches and an evaluation of their performances, we refer to the previous Section 3.3. Our EXPOSÉ system encodes each of the approaches as a search mode to allow exploration with different temporal intents.

To develop our time-aware approaches, we adopt a simplified representation for temporal expression as given in Berberich et al. [26]. Temporal expressions are modeled simply as time intervals $[b, e] \in T \times T$ with begin time point b and end time point e and with the assumption $b \leq e$. Any given event, like the one in our previous example, is treated as a query with two parts, q_{time} that is generated from the date; and q_{text} that is generated from textual description. Analogously, we represent a document (news article) d with publication date t as a combination of bag of temporal expressions d_{time} and bag of textual terms d_{text}.

Text-Only Approach

As a simple approach, we compare the query description q_{text} to the textual terms in a document d_{text}. For this, we use a query-likelihood approach with Dirichlet smoothing that ranks a document as per Equation 3.1.

Intuitively, the text-only approach is effective for short event descriptions with discriminative terms. Further, this approach is also useful when a user does not have any information on the event time period. Consider the following example where this approach works well:

Example 3.6: **December 24 1992:** President George H. W. Bush pardons 6 national security officials implicated in the Iran-Contra affair, including Caspar Weinberger.

The example above refers to a specific event about the Iran-Contra scandal (of 1986) that happened during the final days of presidency of George H. W. Bush. The relatively short textual description contains many terms as surface forms of entities that makes it selective. This example retrieves relevant articles in top-10 with our simple text-only approach that employs the query likelihood-based retrieval model.

Publication-Date Approach

Newsworthy events receive quick media attention with articles reporting on the details as the events happen. Intuitively, the news articles that are published around the event date are more likely to discuss the event in more detail. This approach ranks documents by relating their publication dates to the event date by assuming recency as the user's temporal information intent. That is, a user is satisfied if the publication date of a news article is close to the event date and quickly becomes dissatisfied as the temporal distance between the two dates grows.

To rank documents, the publication-date approach first assumes independence between the temporal and the textual part of a given query. It then estimates the likelihood of generating the query from the document as per Equation 3.3.This approach is most effective for events with short ramifications. As an example query for which this approach becomes effective, consider the following query:

Example 3.7: **September 16 2000:** Winnie Mandela is convicted of kidnapping. On May 14, she is sentenced to 6 years in prison.

This example can be considered as a facet of a larger event that spans from the kidnapping to trial. In this example, the date in the query becomes a strong signal to identify relevant news articles by comparing to their publication dates.

Temporal-Expression Approach

News articles that are not contemporary to an event may still be relevant if they provide information pertaining to the event. Such articles often refer to the time period of the event in their content. Motivated by this, the temporal-expressions approach ranks an article higher if many temporal expressions in its content refer to the time period when the event happened.

To rank documents, our approach first assumes independence between the textual and temporal part of a query and ranks documents according to Equation 3.5. This approach is most effective for events that did not receive adequate media coverage and have few relevant contemporary articles. Consider the following example query:

Example 3.8: **September 16 2000:** Ukrainian journalist Georgiy Gongadze is last seen alive; this day is taken as the commemoration date of his death.

In this example, though journalist Georgiy Gongadze was last seen alive on the given date, his death was confirmed months later. Thus, we find that news articles that elaborate on his death are published much later to the event date, however they mention this time period in their content.

Publication-Date + Temporal-Expression Approach.

This approach combines the publication-date and temporal-expression approaches to retrieve relevant news articles as Equation 3.6.

Intuitively, this approach performs best for events that have larger ramifications or are discussed in the media at diverse time points. This approach is effective for ambiguous event descriptions. Consider the following example query:

Example 3.9: **February 27 1991:** President Bush declares victory over Iraq and orders a cease-fire.

The description shown above contains a reference to President Bush and war in Iraq which are ambiguous since there are multiple presidents named Bush and multiple wars in Iraq. For this, any news article that is retrieved by syntactic matching, and has publication date close to the event date or mentions the event time period turns out to be relevant.

Two-Stage Cascade Approach.

The temporal part of a given query refers to a single day and becomes overly specific to indicate happening of events. This is often seen for events that spanned over a larger time period. The two-stage cascade approach aims at automatically identifying additional time points (days) that are seminal to the event. As a key difference to other

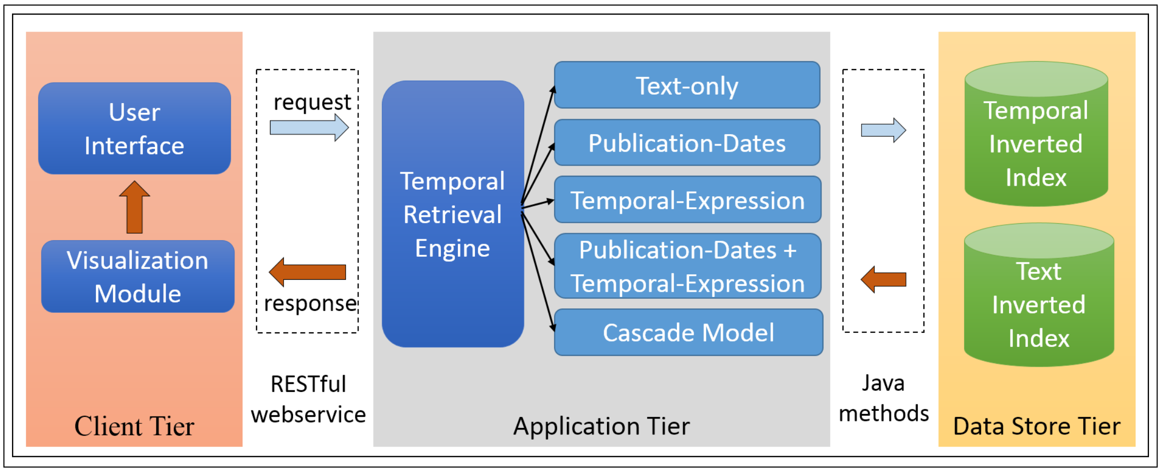


Figure 3.4 System architecture of EXPOSÉ

approaches, the two-stage cascade approach uses relevance feedback to expand the q_{time} and shift the focus to a larger time period.

In *Stage 1*, it performs an initial round of retrieval with the text-only approach, and estimates a temporal query model using the temporal expressions in the retrieved articles, thus treating them as pseudo-relevant. In *Stage 2*, our approach re-ranks the initially retrieved candidate articles by taking into account the Kullback-Leibler (KL) divergence between the query and document temporal models. To generate the final document score we additionally combine the divergence between the originally given event date and publication date of the document, and the text parts of the query and document to preserve the textual relevance. The details of this method are described in Section 3.3.2. This approach is most effective for events that did not receive extensive coverage when they happened or that spanned over a long time period. Consider the following example query:

Example 3.10: March 19 2002: US war in Afghanistan: Operation Anaconda ends after killing 500 Taliban and Al-Qaeda fighters, with 11 allied troop fatalities.

The example above refers to the end of Operation Anaconda which itself lasted for a longer time period. For this, building a temporal model with feedback expressions as salient time points shifts the focus from the overly specific event date to a larger time period.

3.4.2 Demonstration Setup

Implementation

The responsive web interface is implemented using the Twitter Bootstrap toolkit. The server side is implemented in Java and deployed in an Apache Tomcat server. We use Stanford CoreNLP [131] to tokenize documents, extract entities, and annotate temporal expressions in their content.

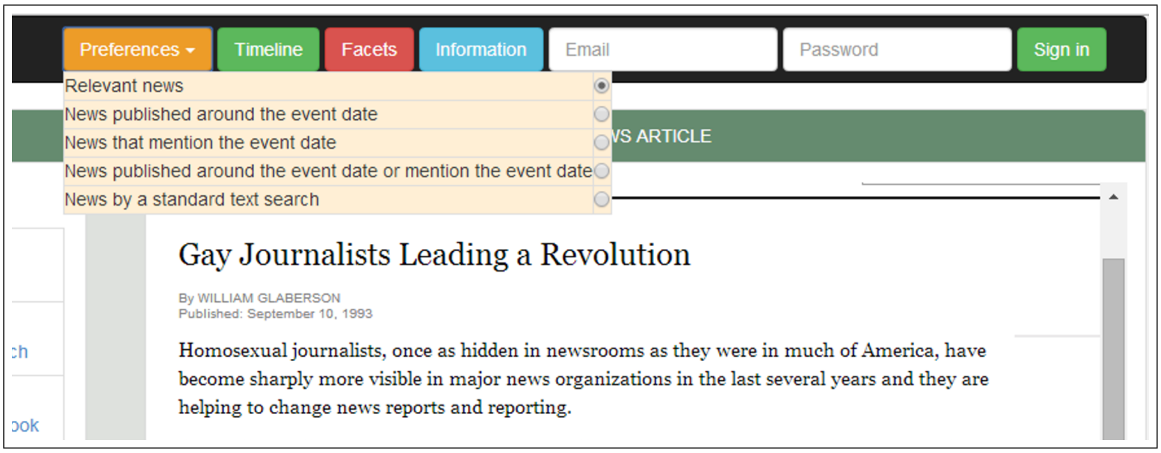


Figure 3.5 Different temporal search preference options

News Collection

For the prototype, we use The New York Times Annotated Corpus [4], which contains about 2 million documents published between 1987 and 2007.

Directory Events

We use The English Wikipedia dump released on February 3^{rd} 2014 and generate 3,076 Wikipedia events as queries from the articles for the years 1987 to 2007. These events are provided in the prototype as a directory.

System Architecture

The architecture of EXPOSÉ is illustrated in Figure 3.4. A user submits a preference option, an event description and a date as a RESTful web-service request to the temporal retrieval engine. This engine implements five retrieval methods, one of which is selected by the user. The retrieval engine uses inverted text and temporal indexes to efficiently retrieve and rank relevant news articles. The visualization module receives a ranked list of top-10 news articles as a response. Finally, the visualization module performs facet extraction and updates the user interface.

3.4.3 Exploratory Interface

The Web interface exhibits a three column layout that we describe next. Illustration of the full user interface is shown in Figure 3.6.

The left-most holds two components: first, *Search Box* panel where a user can type in an event description and a date; and second, *Event Directory* panel that contains 3,076 clickable events extracted from Wikipedia. To select a particular mode for time-aware exploration, a user can choose one the following options in *Preference* tab as illustrated in Figure 3.5: **1)** *Relevant news* – is the default option and uses the two-stage cascade approach; **2)** *News published around the event date* – regards event date and uses publication-date approach; **3)** *News that mention the event date* – considers the temporal expressions in article content and uses the temporal-expression approach; **4)** *News published around the event date or mention the event date* – regards both temporal expressions and the publication date, and uses the publication-date + temporal-expression approach; **5)** *News by a standard text search* – activates the text-only approach.

The middle column contains the *Search Results* panel that displays the relevant news articles after processing the user input. For each identified news article, its title and the publication date are displayed as a ranked list.

A user can then click on a title to read its content which opens in the *News Article* panel aligned to the right-most column.

The *Timeline* panel visualizes the publication dates distribution of the retrieved news articles for a given event. The timeline plots titles of the news articles which when clicked opens corresponding content in the *News Article* panel. Assuming that the distribution represents salient time points for the event, this visualization for publication date-based methods helps to explore the temporal evolution of the event. For methods exploiting temporal expressions, this visualization is useful to identify related events whose corresponding news articles refer to the current event in their content. This feature can be activated or deactivated in the timeline panel by toggling the *Timeline* button which is shown in Figure 3.5.

The *Facet Panel* displays three sets of entity labels semantically representing person, location, and organization that are extracted from relevant news articles. The aggregated term frequency of each label is also displayed which gives an understanding of their prominence. The three sets are organized into sub panels and can be selected in a desired combination to dynamically filter news articles. This feature can be activated or deactivated in the facet panel by toggling the *Facet* button as shown in Figure 3.5.

3.4.4 Demonstration Scenario

Let us consider the example event in Section 3.4. To explore, a user begins by looking for this event in the directory using a directory search (or simply types in the search box). By setting the preference to *News published around the event date* option, she quickly identifies relevant contemporary news articles that elaborate on the event. Further, the user uses a timeline tool in the system to visualize and learn salient time periods when this event was discussed in the media, like *November 1987* and *April 1988.* Entities like *Pope John Paul* (of type person), *United States of America* (of type geographic location) that are mentioned in the identified articles are extracted as facets for further filtering. For further exploration, the user also has the flexibility to submit a typed description of a related event that is not listed in the directory, or is simply a refinement for an existing description.

A video showcasing EXPOSÉ is available at: http://youtu.be/t4frzvvonqE

Figure 3.6 Illustration of the EXPOSÉ user interface.

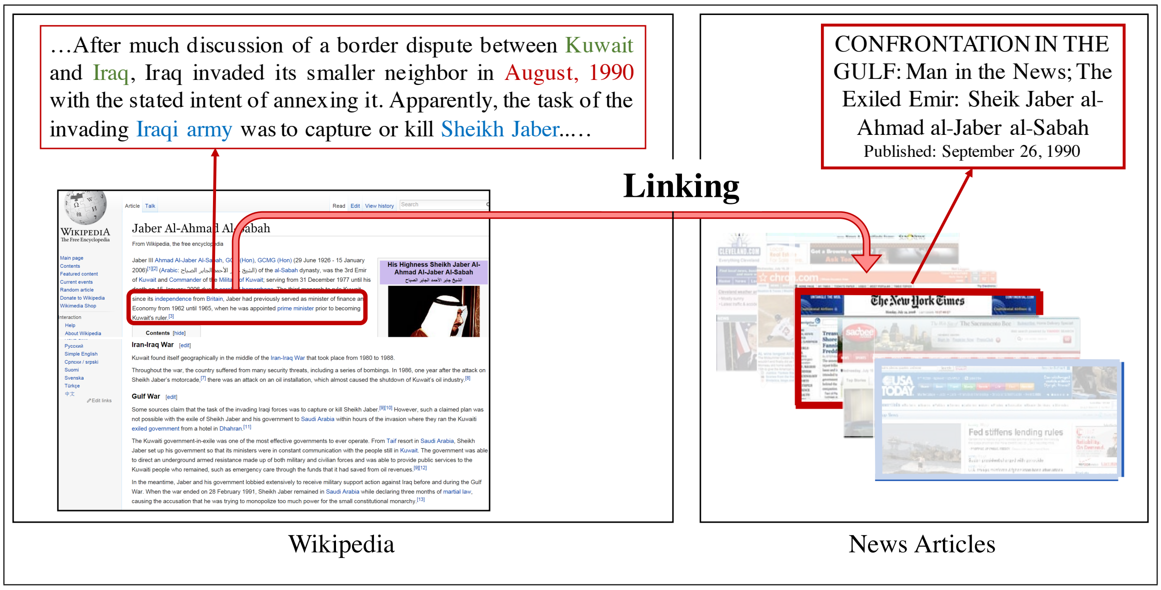


Figure 3.7 Connecting Wikiexcerpt from Wikipedia Jaber Al-Ahmad Al-Sabah to news articles.

3.5 Leveraging Text + Time + Geolocation + Entity

Articles in Wikipedia are often collectively authored. As one perspective, they can be considered as a collective memory on past events. Since they are constantly evolving, Wikipedia articles tend to summarize past events by abstracting from fine-grained details that mattered when the event happened. As a general observation, Wikipedia articles describe relevant events related to the central entity (or topic) as sentences or paragraphs. Often such fine-grained events described are less likely to have an independent central Wikipedia article. Thus it becomes difficult for a user to get additional information on such events by tedious reading through large Wikipedia articles to look for bits of relevant information scattered throughout. It can be noted that often such passages additionally come with few (usually single) citations. However, their main purpose is to validate the information in contrast to providing additional details.

A secondary source of information on past events is online news articles, and general web pages reporting on them. Such articles are often published contemporary to the events and report on all fine-grained that mattered when the events had happened. Since Wikipedia articles tend to describe events (or aspects of a larger event) of local and global importance (a side effect of collective authoring) such events can be assumed to be covered in news. However, for a general user, it becomes difficult to sift through large amounts of news or general web articles with high degree of redundancy to develop a holistic understanding on such fine-grained events. This gives rise to the information overloading problem as described in Section 3.1.

Consider the following two scenarios for further motivation:

- Laura Lane, a journalist, researching on the Iraq-Kuwait war in 1990, comes across the following Wikipedia passage:

Example 3.11: Jaber Al-Ahmad Al-Sabah: After much discussion of a border dispute between **Kuwait** and **Iraq**, **Iraq** invaded its smaller neighbor on *August 2, 1990* with the stated intent of annexing it. Apparently, task of the invading **Iraqi** army was to capture or kill Sheikh Jaber.

The above passage states that the intention of Iraqi army behind invading Kuwait in August 1990, was to kill or capture Sheik Jaber Al-Ahmad Al-Sabah. After an exhaustive search in Wikipedia, she fails to find any central article describing the specific event. However, after issuing a series of keyword queries to a vertical news search engine, she finds that a lot of relevant and redundant news articles are retrieved in context of the war. She then has to tediously sift through the large number of articles to find more information on the specific event.

- John Green, an advertising agent, comes across the following Wikipedia passage:

Example 3.12: Static cling: In advertising Advertisers in urban areas, eager to use guerilla marketing techniques, have turned to static cling as a distribution medium. In an advertising campaign for Microsoft's MSN 8 Internet service, on *October 24, 2002*, hundreds of decals of the MSN butterfly logo were affixed to surfaces in **New York City** and held there with static cling.

It mentioning that in an advertising campaign for Microsoft's MSN 8 Internet service, on October 24, 2002, hundreds of decals of the MSN butterfly logo were affixed to surfaces in New York city. With lack of further information on the event in Wikipedia, and the need to dig deeper, John switches to a Web search engine to find web articles describing the event. He finds a lot of relevant web articles relevant to Microsoft's MSN 8 Internet service. However, after shifting through the results, he pinpoints few local news articles relevant to the event.

In both aforementioned scenarios, what is missing are connections from such Wikipedia passages describing events to relevant news articles giving additional information. With connections between passages and news articles in place, if a user while reading a Wikipedia article comes across such a passage describing an event, she can then jump to the connected news articles to quickly gather more detailed information. Motivated from the above scenarios, and as illustrated in Figure 3.7, we address the following linking problem to connect excerpts taken from Wikipedia articles, to news articles:

Problem Statement: *Given an excerpt extracted from a Wikipedia article, coined Wikiexcerpt, summarizing an event as a user query, automatically retrieve a ranked list of news articles providing details on the event which could be provided as a background reference.*

As input, we consider a Wikiexcerpt that is either a single sentence or a passage describing event. Two concrete examples of such Wikipedia event descriptions are illustrated as Example 3.11 and Example 3.12. As output, our goal is to retrieve a ranked list of news articles that give detailed information on the event descriptions.

Challenges

To address the linking problem, we are faced with the following key challenges:

- The Wikiexcerpts also contain additional semantics like temporal expressions, geolocations, and named entities which can be leveraged to identify relevant documents. Making the retrieval model aware of these semantic annotations so as to identify contemporary and relevant documents is not straightforward.
- Wikiexcerpts extracted from general articles are typically verbose. Due to the verbosity of Wikiexcerpts, they are prone to topic drift and result in lower retrieval quality.

In our approach, we design a *two-stage* cascade retrieval model. In the first stage, our approach performs an initial round of retrieval with the text part of the query to retrieve a set of top-K documents. It then treats these documents as pseudo-relevant, and expands the *temporal*, *geospatial*, and *entity* parts of the query. Then, in the second stage, our approach builds independent query models using the expanded query parts, and re-ranks the initially retrieved documents based on their overall divergence from the final integrated query model. However, before going into detail, we begin by describing our notation and representations.

3.5.1 Notation & Representations

We use the following notation and representations to design our methods.

Document Collection

Each document d in our document collection C consists of a textual part d_{text}, a temporal part d_{time}, a geospatial part d_{space}, and an entity part d_{entity}. As a bag of words, d_{text} is drawn from a fixed vocabulary V derived from C. Similarly, d_{time}, d_{space}, and d_{entity} are bags of temporal expressions, geolocations, and named-entity mentions respectively. We sometimes treat the entire collection C as a single coalesced document and refer to its corresponding parts as C_{text}, C_{time}, C_{space}, and C_{entity} respectively. In our approach, we use the Wikipedia Current Events Portal[1] to distinguish event-specific terms by coalescing into a single document d_{event}.

[1]http://en.wikipedia.org/wiki/Portal:Currentevents

Temporal Expressions

Time unit or *chronon* τ indicates the time passed (to pass) since (until) a reference date such as the UNIX epoch. A temporal expression t is an interval $[tb, te] \in T \times T$, in time domain T, with begin time tb and end time te. Moreover, a temporal expression t is described as a quadruple $[tb_l, tb_u, te_l, te_u]$ [26] where tb_l and tb_u gives the plausible bounds for begin time tb, and te_l and te_u give the bounds for end time te.

Geographical Locations

A geospatial unit l refers to a geographic point that is represented in the geodetic system in terms of latitude (lat) and longitude $(long)$. A geolocation s is represented by its minimum bounding rectangle (MBR) and is described as a quadruple $[tp, lt, bt, rt]$. The first point $[tp, lt]$ specifies the top-left corner, and the second point $[bt, rt]$ specifies the bottom-right corner of the MBR. We fix the smallest MBR by setting the resolution $[resol_{lat} \times resol_{long}]$ of space.

Named Entity Mentions

A named entity e refers to a location, person, or organization from the YAGO [82] knowledge base. We use YAGO URIs to uniquely identify each entity in our approach.

Query

A query q is derived from a given Wikiexcerpt in the following way: the text part q_{text} is the full text, the temporal part q_{time} contains explicit temporal expressions that are normalized to time intervals, the geospatial part q_{space} contains the geolocations, and the entity part q_{entity} contains the named entities mentioned. To distinguish contextual terms, we use the textual content of the source Wikipedia article of a given Wikiexcerpt and refer to it as d_{wiki}.

3.5.2 IR Framework

In our approach, we design a *two-stage* cascade retrieval model. In the first stage, our approach performs an initial round of retrieval with the text part of the query to retrieve top-K documents. It then treats these documents as pseudo-relevant and expands the *temporal*, *geospatial*, and *entity* parts of the query. Then, in the second stage, our approach builds independent query models using the expanded query parts and re-ranks the initially retrieved K documents based on their divergence from the final integrated query model. As output it then returns top-k documents $(k < K)$. Intuitively, by using pseudo-relevance feedback to expand query parts, we cope with overly specific (and sparse) annotations in the original query and instead consider those that are salient to the query event for estimating the query models.

For our linking task, we extend the KL-divergence framework [216] to the text, time, geolocation, and entity dimensions of the query and compute an overall divergence score. This is done in two steps: First, we independently estimate a query model for each of the dimensions. Let Q_{text} be the unigram *query-text model*, Q_{time} be the *query-time model*, Q_{space} be the *query-space model*, and Q_{entity} be the *query-entity model*. Second, we represent the overall query model Q as a joint distribution over the dimensions and exploit the additive property of the KL-divergence (KLD) to compute the overall divergence $KLD(Q\,||\,D)$between the query Q and document model$D$ as,

$$KLD(Q_{text}||D_{text})+KLD(Q_{time}||D_{time})+KLD(Q_{space}||D_{space})+KLD(Q_{entity}||D_{entity}).$$

According to additive property, the overall KL-divergence score is a simple sum of divergence scores across independent dimensions. In the above equation, analogous to the query, the overall document model D is also represented as the joint distribution over D_{text}, D_{time}, D_{space}, and D_{entity} which are the independent document models for the dimensions.

The KL-divergence framework with the independence assumption gives us the flexibility of treating each dimension in isolation while estimating query models. This would include using different background models, expansion techniques with pseudo-relevance feedback, and smoothing. The problem thus reduces to estimating query models for each of the dimensions.

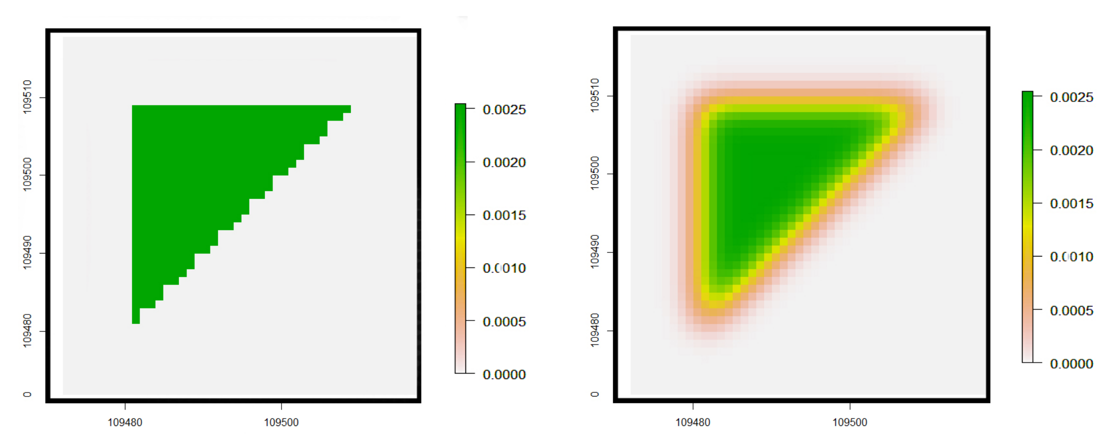
3.5.3 Query Models

We describe the query models estimated independently for the event dimensions.

Query-Text Model

Standard likelihood-based query modeling methods that rely on the empirical terms become ineffective for our task. As an illustration, consider Example 3.11. A likelihood-based model would put more stress on {*Iraq*} that has the highest frequency, and suffer from topical drift due to the terms like {*discussion*, *border*, *dispute*, *Iraq*}. It is hence necessary to make use of a background model that emphasizes event-specific terms.

We observe that a given q_{text} contains two factors, first, terms that give background information, and second, terms that describe the event. To stress on the latter, we combine a query-text model with a background model estimated from: **1)** the textual content of the source Wikipedia article d_{wiki}; and **2)** the textual descriptions of events listed in the Wikipedia Current Events portal, d_{event}. The d_{wiki} background model puts emphasis on the contextual terms that are discriminative for the event, like {*Kuwait*, *Iraq*, *Sheikh*, *Jaber*}. On the other hand, the background model d_{event} puts emphasis on event-



(a) before **(b)** after

Figure 3.8 Effect of applying Gaussian smoothing on the 2D representation of *August 1990*.

specific terms like {*capture, kill, invading*}. Similar approaches that combine multiple contextual models have shown significant improvement in result quality [185, 197].

We combine the query model with a background model by linear interpolation [217]. The probability of a word w from the Q_{text} is estimated as,

$$P(w \mid Q_{text}) = (1-\lambda) \cdot P(w \mid q_{text}) + \lambda \cdot \left[\beta \cdot P(w \mid d_{event}) + (1-\beta) \cdot P(w \mid d_{wiki})\right]. \quad (3.13)$$

A term w is generated from the background model with probability λ and from the original query with probability $1-\lambda$. Since we use a subset of the available terms, we finally normalize the query model as in [162]. The updated generative probability $\hat{P}(w \mid Q_{text})$ is computed as,

$$\hat{P}(w \mid Q_{text}) = \frac{P(w \mid Q_{text})}{\sum_{w' \in V} P(w' \mid Q_{text})}. \quad (3.14)$$

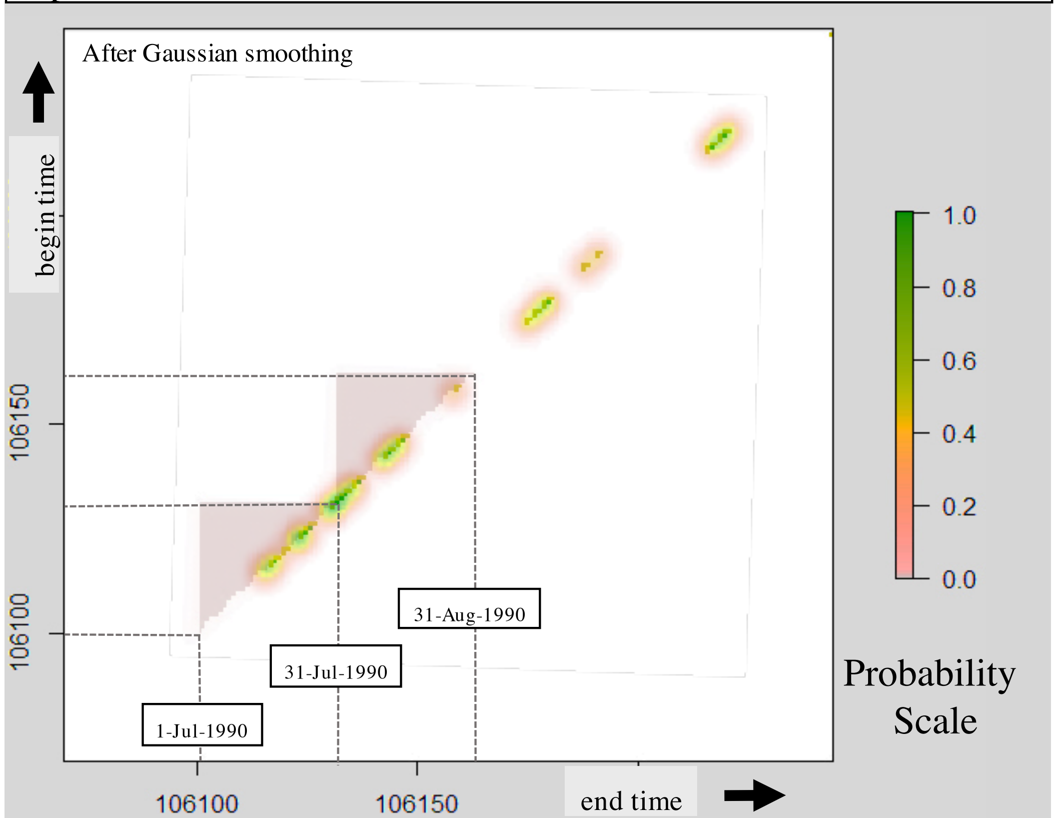
Query-Time Model

We assume that a temporal expression $t \in q_{time}$ is sampled from the query-time model Q_{time} that captures the salient periods for an event in a given q. The generative probability of any time unit τ from the temporal query model Q_{time} is estimated by iterating over all the temporal expressions $t = [tb_l, tb_u, te_l, te_u]$ in q_{time} as,

$$P(\tau \mid Q_{time}) = \sum_{[tb_l, tb_u, te_l, te_u] \in q_{time}} \frac{\mathbb{1}(\tau \in [tb_l, tb_u, te_l, te_u])}{|[tb_l, tb_u, te_l, te_u]|} \quad (3.15)$$

where the $\mathbb{1}(\cdot)$ indicator function indicates containment of a time unit t within an interval that is represented as $[tb_l, tb_u, te_l, te_u]$, i.e., does the point t lie within the interval. The denominator computes the area of the temporal expression in $T \times T$.

…After much discussion of a border dispute between Kuwait and Iraq, Iraq invaded its smaller neighbor in **August, 1990** with the stated intent of annexing it. Apparently, the task of the invading Iraqi army was to capture or kill Sheikh Jaber…



	Begin Date	End Date
Year	1-Jan-1990	31-Dec-1990
	1-Jun-1990	12-Sep-1990
	1-Sep-1990	1-Dec-1990
Month	1-Jul-1990	31-Jul-1990
	1-Aug-1990	31-Aug-1990
Day	20-Jul-1990	21-Jul-1990
	24-Jul-1990	25-Jul-1990
	25-Jul-1990	26-Jul-1990
	26-Jul-1990	27-Jul-1990
	31-Jul-1990	1-Aug-1990
	1-Aug-1990	2-Aug-1990
	2-Aug-1990	3-Aug-1990
	3-Aug-1990	4-Aug-1990
	5-Aug-1990	6-Aug-1990
	6-Aug-1990	7-Aug-1990
	12-Aug-1990	13-Aug-1990
	13-Aug-1990	14-Aug-1990
	15-Aug-1990	16-Aug-1990
	16-Aug-1990	17-Aug-1990
	28-Aug-1990	29-Aug-1990
	13-Sep-1990	14-Sep-1990
	15-Sep-1990	16-Sep-1990

Figure 3.9 Query-time model estimated with selected pseudo-relevant temporal expressions.

For any given temporal expression, we can compute its area and its intersection with other expressions as described in [26]. Intuitively, the above equation assigns higher probability to time units that overlap with a larger number of specific (smaller area) intervals in q_{time}.

The query-time model estimated so far has hard temporal boundaries and suffers from the issue of *near-misses*. For example, if the end boundary of the query-time model is "10 January 2014" then the expression "11 January 2014" in a document will be disregarded. To address this issue, we perform an additional *Gaussian smoothing*. The new probability is estimated as,

$$\hat{P}(\tau \mid Q_{time}) = \sum_{t \in T \times T} G_\sigma(t) \cdot P(\tau \mid Q_{time}) \tag{3.16}$$

where G_σ denotes a Gaussian kernel that is defined as,

$$G_\sigma(i) = \frac{1}{2\pi\sigma^2} \cdot exp\left(-\frac{(tb_l, tb_u)^2 + (te_l, te_u)^2}{2\sigma^2}\right). \tag{3.17}$$

Gaussian smoothing computes a weighted average of adjacent units where the weight decreases with the increase in spatial distance to the center position τ in the space domain. The σ in the kernel defines the neighborhood size and can be empirically set. As a result of the Gaussian smoothing, the temporal boundaries are blurred, spilling some probability mass to adjacent time units. Figure 3.8 illustrates the effect of Gaussian smoothing on a two dimensional illustration of the month "August 1990".

Finally, since we use only a subset of temporal expressions we normalize similar to Equation 3.14. This step is motivated in [162].

An illustration of query-time model estimated for Example 3.11 along with the Gaussian smoothing is shown in Figure 3.9. We gather temporal expressions at the year, month, and day granularities as mentioned in the pseudo-relevant documents. In Figure 3.9, larger triangular shapes represent the temporal expression at the month granularity. Expressions representing a single day occur on the diagonal. Note that the days receive larger probability mass as compared to the longer month intervals.

Query-Space Model

We assume that a user samples a geolocation s from query-space model Q_{space} to generate q_{space}. The query-space model captures salient geolocations for the event in a given Wikiexcerpt. The generative probability of any spatial unit l from the query-space model Q_{space}, by iterating over all geolocations $[tp, lt, bt, rt] \in q_{space}$, is estimated as,

$$P(l \mid Q_{space}) = \sum_{[tp,lt,bt,rt] \in q_{space}} \frac{\mathbb{1}(l \in [tp, lt, bt, rt])}{|[tp, lt, bt, rt]|}. \tag{3.18}$$

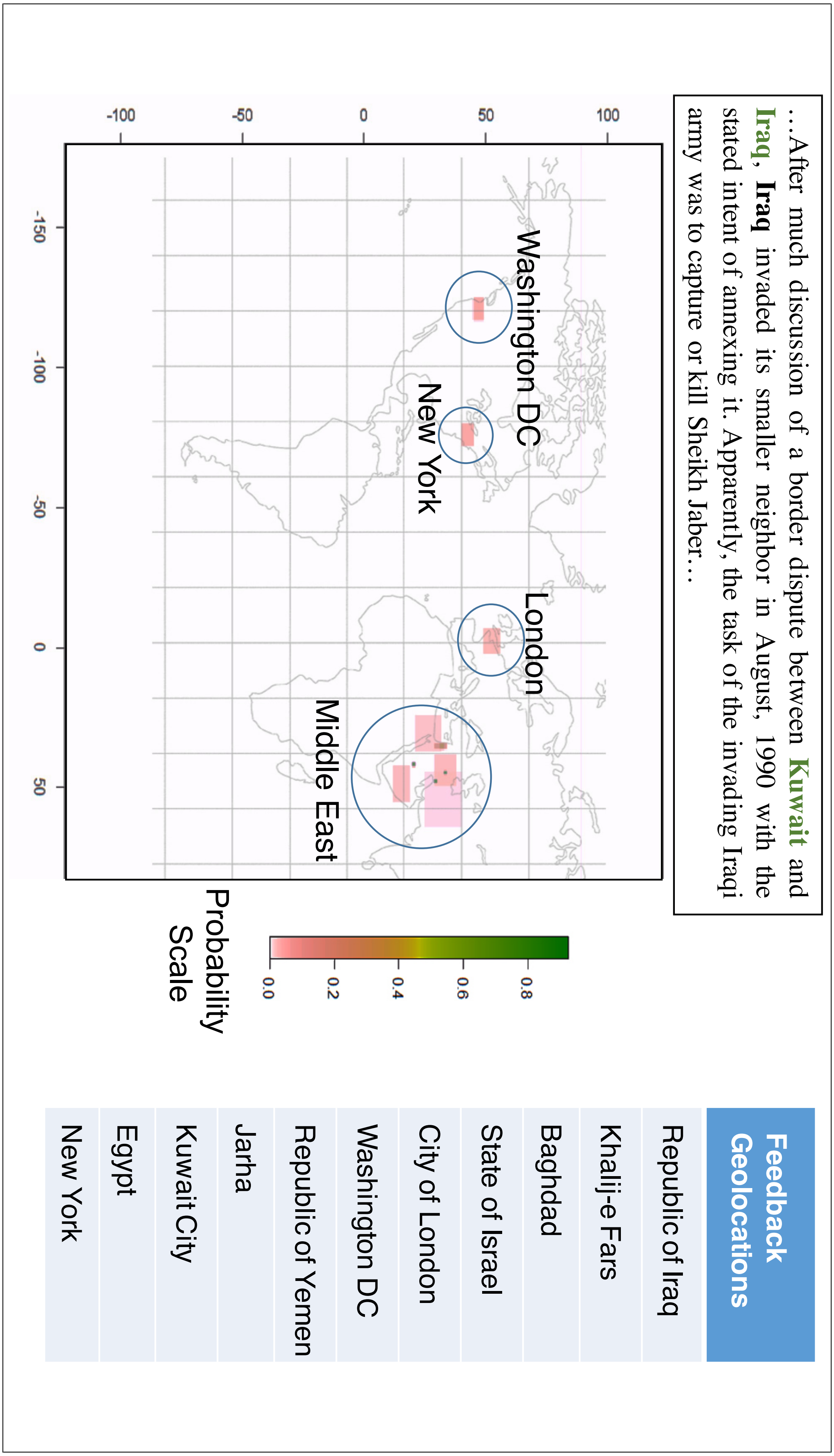
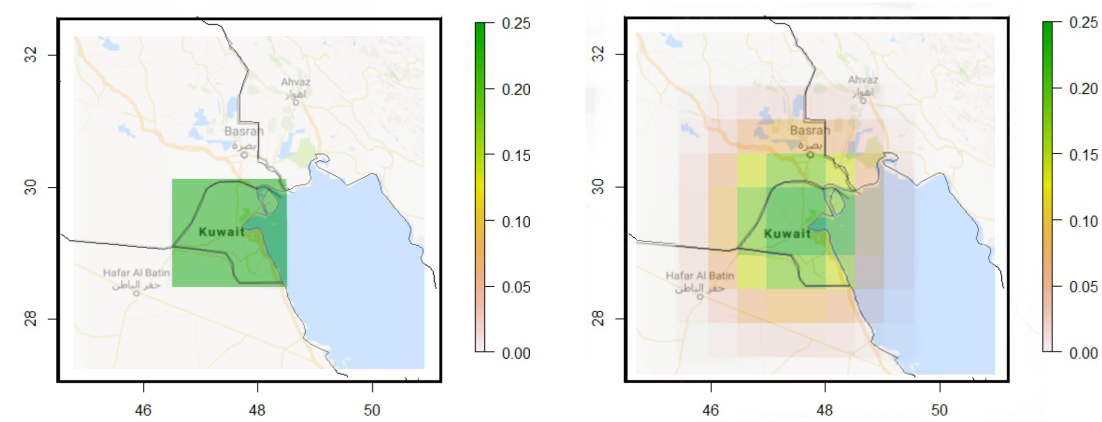


Figure 3.10 Query-space model estimated with selected pseudo-relevant geolocations.



(a) before **(b)** after

Figure 3.11 Effect of applying Gaussian smoothing on the 2D MBR representation of *Kuwait* on a topological map.

Analogous to the Equation 3.15, the $\mathbb{1}(\cdot)$ indicator function indicates containment, of a space unit g within a MBR $[tp, lt, bt, rt]$, i.e., does the point l lie within the MBR. Intuitively, our query-space model assigns higher probability to l if it overlaps with a larger number of more specific (MBR with smaller area) geolocations in q_{space}. As the denominator, it is easy to compute $|[tp, lt, bt, rt]|$ as $|s| = (rt - lt + resol_{lat}) * (tp - bt + resol_{long})$. The addition of the small constant ensures that for all s, $|s| > 0$. Similar to the query-time model, to address the issue of near misses we estimate $\hat{P}(l \,|\, Q_{Space})$ that additionally smooths $P(l \,|\, Q_{space})$ using a Gaussian kernel as described in Equation 3.16 and also normalize as per Equation 3.14. Figure 3.11 illustrates the effect of Gaussian smoothing on the MBR-based representation.

An illustration of query-space model estimated for Example 3.11 along with the Gaussian smoothing is shown in Figure 3.10. We gather the geolocations mentioned in the pseudo-relevant documents as described before. Figure 3.9 overlays the estimated model on a world map for illustration. Note that based on the probability scale for this query, most of the probability mass is concentrated around middle east indicating the conflict geographical region.

Query-Entity Model

The query-entity model Q_{entity} captures the entities that are salient to an event and builds a probability distribution over an entity space. To estimate Q_{entity} we make use of the initially retrieved pseudo-relevant documents to construct a background model that assigns higher probability to entities that are often associated with an event. Let D_R be the set of pseudo-relevant documents. The generative probability of entity e is estimated as,

$$P(e \mid Q_{entity}) = (1-\lambda) \cdot P(e \mid q_{entity}) + \lambda \cdot \sum_{d \in D_R} P(e \mid d_{entity}) \tag{3.19}$$

where $P(e \mid q_{entity})$ and $P(e \mid d_{entity})$ are the likelihoods of generating the entity from the original query and document $d \in D_R$ respectively.

3.5.4 Document Models

To estimate the document models for each dimension, we follow the same methodology as for the query with an additional step of Dirichlet smoothing [217]. This has two effects: First, it prevents undefined KL-divergence scores. Second, it achieves an IDF-like effect by smoothing the probabilities of expressions that occur frequently in the C. The generative probability of a term w from document-text model D_{text} is estimated as,

$$P(w \mid D_{text}) = \frac{\hat{P}(w \mid D_{text}) + \mu P(w \mid C_{text})}{|D_{text}| + \mu} \tag{3.20}$$

where $\hat{P}(w \mid D_{text})$ is computed according to Equation 3.14 and μ is set as the average document length of our collection [217]. Similarly, we estimate D_{time}, D_{space}, and D_{entity} with C_{time}, C_{space}, and C_{entity} as background models respectively to tackle the above mentioned issues. To estimate D_{time} and D_{space}, we follow methods similar to Equations 3.15 and 3.18. However, we do not apply the Gaussian smoothing (as described in Equation 3.16) as it tends to artificially introduce temporal and spatial information into the document content.

3.5.5 Experimental Evaluation

Next, we describe our experiments to study the impact of the different components of our approach. We make our experimental data publicly available[1].

Document Collections

We consider the following document collection for the experimental evaluations. Additional details are illustrated in Table 3.4.

- **New York Times Annotated Corpus (NYT):** As the first dataset, we use The New York Times[2] Annotated Corpus which contains daily news articles published between 1987 and 2007.

[1] http://resources.mpi-inf.mpg.de/d5/linkingWiki2News/
[2] http://corpus.nytimes.com

Table 3.4 Document collection statistics

	NYT	CW12
Indexed documents	1,855,656	46,797,647
Average document length in words	691.79	740.04
Average of temporal expressions per document	8.61	5.98
Average of entity mentions per documents	6.3	6.6

- **ClueWeb12-B13 corpus (CW12):** As our second dataset, we use the ClueWeb12-B13 (CW12) corpus[1] with 50 million web pages crawled in 2012.

Test Queries

We use the English Wikipedia dump released on February 3^{rd} 2014 to generate two independent sets of test queries: **1)** *NYT-Queries* containing 150 randomly sampled Wikiexcerpts targeting documents from the NYT corpus; **2)** *CW-Queries* containing 150 randomly sampled Wikiexcerpts targeting web pages from CW12 corpus. *NYT-Queries* have 104 queries out of 150, that come with at least one temporal expression, geolocation, and named-entity mention. In the remaining 46 test queries, 17 do not have any temporal expressions, 28 do not have any geolocations, and 27 do not have any entity mentions. We have 4 test queries where our taggers fail to identify any additional semantics. *CW-Queries* have 119 queries out of 150, that come with at least one temporal expression, geolocation, and entity mention. 19 queries do not mention any geolocation, and 26 do not have entity mentions. We include a full list of the NYT-Queries queries in Appendix A.2 and CW-Queries in Appendix A.3.

Effectiveness Measures

As a strict effectiveness measure, we compare our methods based on Mean Reciprocal Rank (MRR). We also compare our methods using Normalized Discounted Cumulative Gain (NDCG) and Precision (P) at cut-off levels 5 and 10. We also report the Mean Average Precision (MAP) across all queries. For MAP and P, we consider a document relevant to a query if the majority of assessors judged it with label **(1)** or **(2)**. For NDCG, we plug in the mean label assigned by assessors.

Relevance Assessments

Manual relevance judgments were collected using the Crowdflower platform[2]. We pooled top-10 results for the methods under comparison and asked contributors to judge a document as **(0)** irrelevant, **(1)** somewhat relevant, or **(2)** highly relevant to a

[1] http://www.lemurproject.org/clueweb12.php/
[2] http://www.crowdflower.com/

query. Our instructions said that a document can only be considered highly relevant if its main topic was the event given as the query.

We ran two independent Crowdflower jobs to get judgments on NYT documents and the CW12 web pages. This resulted in only a slight variation in the instructions given to the contributors and user interface. For NYT documents, the instruction set was kept identical to our previous setting as described before in Figure 3.3. An illustration of the full instruction set for getting judgments on the CW12 documents is illustrated in Figure 3.12. For the NYT job, we used the URL of documents that comes with the corpus to render the content of the documents in the user interface. However, we found that almost 40% of the URLs were no longer working. Thus we hosted a subset of the corpus with the documents retrieve in our runs and provided a link in the user interface. Since the web pages in the CW12 corpus retain their html structure, we could render the documents fairly close to the original web pages. Both experiments resulted in 1,778 and 1,961 unique query-document pairs respectively. Each query-document pair was judged by three assessors. The final relevance label was obtained using the Crowdflower aggregation tool that based on majority voting, however, giving more weight to trusted contributors based on the entire prior job history. We paid \$0.03 per batch of five query-document pairs for a single assessor. The total cost, including additional charges by Crowdflower, of the two experiments was \$118.34 and took a total of 182 hours to complete.

Methods under Comparison

We compare the following methods:

- *txt* considers only the query-text model that uses the background models estimated from the current events portal and the source Wikipedia article (Equation 3.13).
- *txtT* uses the query-text and query-time model (Equation 3.15).
- *txtS* uses the query-text and query-space model (Equation 3.18).
- *txtE* uses the query-text and query-entity model (Equation 3.19).
- *txtST* uses the query-text, query-time, and query-space model.
- *txtEST* uses all four query models to rank documents.

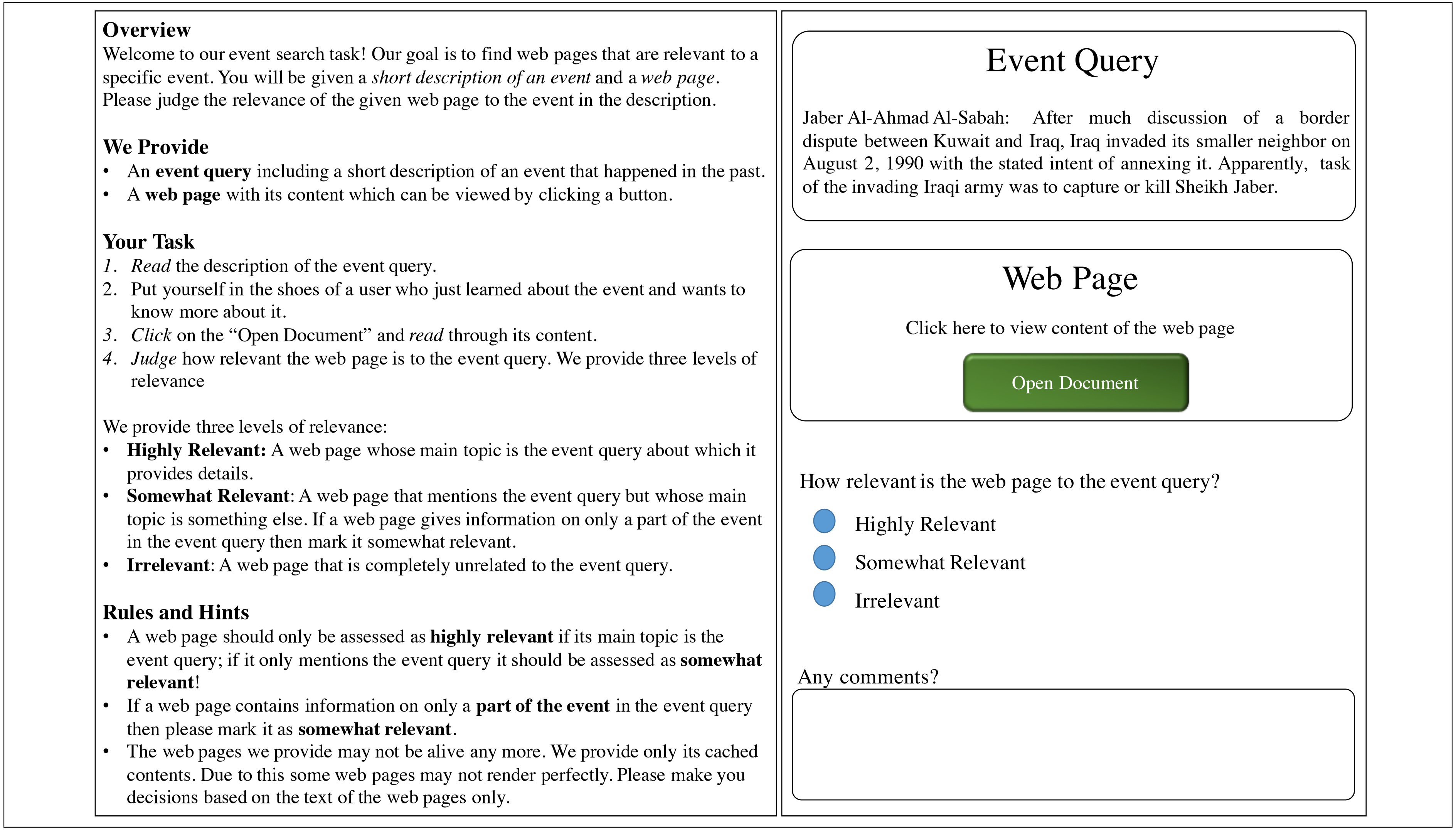


Figure 3.12 Crowdflower interface and the instructions given to contributors.

Parameters

We set the values for the different parameters in query and document models for all the methods by following [216]. For the NYT corpus, we treat top-100 documents retrieved in the first stage as pseudo-relevant. For CW12 corpus with general web pages, we set this to top-500. The larger number of top-K documents for the CW12 corpus is due to the fact that web pages come with fewer annotations than news articles. In Equation 3.13, for estimating the Q_{text}, we set $\beta = 0.5$, thus giving equal weights to the background models. We borrow standard settings for the smoothing parameters from the literature [214] and set $\lambda = 0.85$ in Equations 3.13 and 3.19. For the Gaussian smoothing in Equation 3.17 we set $\sigma = 1$. The smallest possible MBRs in Equation 3.18 is empirically set as $resol_{lat} \times resol_{long} = 0.1 \times 0.1$. Setting this to a resolution did not have significant effect on the quality, however, makes estimating space models computationally expensive.

Implementation

All methods have been implemented in Java. To annotate named entities in the test queries and documents from the NYT corpus, we use the AIDA [81] system. For the CW12 corpus, we use the annotations released as Freebase Annotations of the ClueWeb Corpora[1]. To annotate geolocations in the query and NYT corpus, we use an open-source gazetteer-based tool[2] that extracts locations and maps them to the GeoNames[3] knowledge base. To get geolocations for CW12 corpus, we filter entities by mapping them from Freebase to GeoNames ids. Finally, we run Stanford Core NLP[4]on the test queries, NYT corpus, and CW12 corpus to get the temporal annotations.

Experimental Results

Table 3.5 and Table 3.7 compare the different methods on our two datasets. Both tables have two parts: **(a)** results on the entire query set; and **(b)** results on a subset of queries with at least one temporal expression, geolocation, and entity mention. To denote the significance of the observed improvements to the *txt* method, we perform one-sided paired student's T test at two alpha-levels: 0.05 (‡), and 0.10 (†), on the MAP, P@5, and P@10 scores [43]. We find that the *txtEST* method is most effective for the linking task.

In Table 3.5 we report results for the NYT-Queries. We find that the *txtEST* method that combines information in all the dimensions achieves the best result across all metrics except P@5. The *txt* method that uses only the text already gets a high MRR score. The *txtS* method that adds geolocations to text is able to add minor improvements in NDCG@10 over the *txt* method. The *txtT* method achieves a considerable improvement

[1]http://lemurproject.org/clueweb12/FACC1/
[2]https://github.com/geoparser/geolocator
[3]http://www.geonames.org/
[4]http://nlp.stanford.edu/software/corenlp.shtml

Table 3.5 Results for NYT-Queries 150 queries with significance against the *txt* method.

Measures	*txt*	*txtT*	*txtS*	*txtE*	*txtST*	*txtEST*
MRR	0.898	0.897	0.898	0.898	0.898	**0.902**
P@5	0.711	0.716	0.709	0.716 ‡	**0.719**	0.717
P@10	0.670	0.679	0.669	0.671	0.679	**0.682** †
MAP	0.687	0.700	0.687	0.688	0.701	**0.704** †
NDCG@5	0.683	0.696	0.682	0.685	0.697	**0.697**
NDCG@10	0.797	0.813	0.798	0.796	0.814	**0.815**

Table 3.6 Results for NYT-Queries 104 queries containing at least one annotation in each dimension with significance against the *txt* method.

Measures	*txt*	*txtT*	*txtS*	*txtE*	*txtST*	*txtEST*
MRR	0.921	0.936	0.921	0.921	0.936	**0.942**
P@5	0.715	0.740 ‡	0.715	0.723 ‡	**0.742** ‡	0.740 ‡
P@10	0.682	0.692	0.681	0.684	0.692	**0.696** †
MAP	0.679	0.702 †	0.679	0.682	0.703 †	**0.708** ‡
NDCG@5	0.686	0.721	0.686	0.689	0.721	**0.723**
NDCG@10	0.794	0.823	0.795	0.795	0.823	**0.825**

over *txt*. This is consistent for both NYT-Queries (a) and NYT-Queries (b). The *txtE* method that uses named entities along with text shows significant improvement in P@5 and marginal improvements across other metrics. The *txtST* method that combines time and geolocations achieves significant improvements over *txt*. Finally, the *txtEST* method proves to be the best and shows significant improvements over the *txt*.

In Table 3.7, we report results for the CW-Queries. We find that the *txtEST* method outperforms other methods across all the metrics. Similar to previous results, we find that the *txt* method already achieves high MRR score. However, in contrast, the *txtT* approach shows improvements in terms of P@5 and NDCG@5, with a marginal drop in P@10 and MAP. The geolocations improve the quality of the results in terms of MAP and significantly improve P@10. Though individually time and geolocations show only marginal improvements, their combination as the *txtST* method shows significant increase in MAP. We find that the *txtE* method performs better than other dimensions with a significant improvement over *txt* across all metrics. Finally, the best performing method is *txtEST* as it shows the highest improvement in the result quality.

Discussion

As a general conclusion of our experiments we find that leveraging semantic annotations like time, geolocations, and named entities along with text, improves the effectiveness of the linking task. Because all our methods that utilize semantic annotations (*txtS*, *txtT*, *txtE*, *txtST*, and *txtEST*) perform better than the text-only (*txt*) method. However,

Table 3.7 Results for CW12-Queries 150 Queries with significance against the *txt* method.

Measures	*txt*	*txtT*	*txtS*	*txtE*	*txtST*	*txtEST*
MRR	0.824	0.834	0.831	0.827	0.833	**0.836**
P@5	0.448	0.460	0.451	0.456 †	0.467 ‡	**0.475** ‡
P@10	0.366	0.349	0.366	0.375 ‡	0.367	**0.375** †
MAP	0.622	0.616	0.628	0.640 ‡	0.640 †	**0.653** ‡
NDCG@5	0.644	0.657	0.651	0.654	0.666	**0.673**
NDCG@10	0.729	0.723	0.734	0.746	0.744	**0.755**

Table 3.8 Results for CW12-Queries 119 Queries containing at least on annotation in each dimension with significance against the *txt* method.

Measures	*txt*	*txtT*	*txtS*	*txtE*	*txtST*	*txtEST*
MRR	0.837	0.855	0.846	0.842	0.854	**0.855**
P@5	0.456	0.468	0.459 †	0.466 ‡	0.478 ‡	**0.488** ‡
P@10	0.377	0.358	0.378	0.390 ‡	0.377	**0.386** ‡
MAP	0.623	0.616	0.631	0.647 ‡	0.642 †	**0.661** ‡
NDCG@5	0.655	0.675	0.664	0.669	0.684	**0.695**
NDCG@10	0.736	0.736	0.744	0.759	0.755	**0.769**

the simple *txt* method already achieves a decent MRR score in both experiments. This highlights the effectiveness of the event-specific background model in tackling the verbosity of the Wikiexcerpts. Time becomes an important indicator to identify relevant news articles but it is not very helpful when it comes to general web pages. This is because the temporal expressions in the news articles often describe the event time period accurately thus giving a good match to the queries while this is not seen with web pages. We find that geolocations and time together can identify relevant documents better when combined with text. Named entities in the queries are not always salient to the event but may represent the context of the event. For complex queries, it is hard to distinguish salient entities which reduce the overall performance due to topical drifts on a news corpus. However, they prove to be effective in identifying relevant web pages which can contain more general information thus also mentioning the contextual entities. The improvement of our method over a simple text-based method is more pronounced for the ClueWeb corpus than the news corpus because of mainly two reasons: firstly, the news corpus is too narrow with much smaller number of articles; and secondly, it is slightly easier to retrieve relatively short, focused, and high quality news articles. This is supported by the fact that all methods achieve much higher MRR scores for the NYT-Queries.

Gain/Loss analysis

To get some insights into where the improvements for the *txtEST* method comes from, we perform a gain/loss analysis based on NDCG@5. The *txtEST* method shows biggest gain (+0.13) in NDCG@5 for the following query in *NYT-Queries*:

Example 3.13: West Windsor Township, New Jersey: The West Windsor post office was found to be infected with anthrax during the anthrax terrorism scare back in 2001-2002.

The single temporal expression *2001-2002* refers to a time period when there were multiple anthrax attacks in New Jersey through the postal facilities. Due to the ambiguity, the *txtT* and *txtS* methods become ineffective for this query. Their combination, however, as the *txtST* method becomes the second best method achieving NDCG@5 of 0.7227. The *txtEST* combines the entity *Anthrax* and becomes the best method by achieving NDCG@5 of 0.8539. This method suffers worst in terms of NDCG@5 (−0.464) for the following query in *CW-Queries*:

Example 3.14: Human Rights Party Malaysia: The Human Rights Party Malaysia is a Malaysian human rights-based political party founded on 19 July 2009, led by human rights activist P. Uthayakumar.

The two entities, *Human Rights Party Malaysia* and *P. Uthayakumar*, and one geolocation *Malaysia*, do prove to be discriminative for the event. Time becomes an important indicator to identify relevant documents as *txtT* becomes most effective by achieving NDCG@5 of 0.9003. However, a combination of text, time, geolocations, and named entities as leveraged by *txtEST* achieves a lower NDCG@5 of 0.4704.

Easy & Hard Query Events

Finally, we identify the easiest and the hardest query events across both our testbeds. We find that the following query, in the *CW-Queries*, gets the highest minimum P@10 across all methods:

Example 3.15: Primal therapy: In 1989, Arthur Janov established the Janov Primal Center in Venice (later relocated to Santa Monica) with his second wife, France.

For this query, even the simple *txt* method gets a perfect P@10 score of 1.0. Terms *Janov*, *Primal*, and *Center* retrieve documents that are pages from the center's website, and are marked relevant by the assessors. Likewise, we identify the hardest query as the following one from the *NYT-Queries* set:

Example 3.16: Police aviation in the United Kingdom: In 1921, the British airship R33 was able to help the police in traffic control around the Epsom and Ascot horse-racing events.

For this query none of the methods were able to identify any relevant documents thus all getting a P@10 score equal to 0. This is simply because this is relatively an old event and is not covered in the NYT corpus.

3.6 Summary & Future Directions

In this chapter we have addressed the novel linking task of connecting Wikipedia past event descriptions to online news articles. As two instances of such event descriptions, we consider those that are listed in special Wikipedia year pages, and those that occur as arbitrary passages within general articles.

To connect Wikipedia year page events, we cast the problem into an information retrieval task by considering an event description as a user query. To address this task, we designed several time-aware language models that leveraged the publication dates of the target news articles and additional temporal expression encoded in text. Finally, we presented a two-stage cascaded approach that first estimated independent query models for text and time from a set of pseudo relevant documents retrieve with a standard text retrieval framework. Then in the second stage, our approach re-ranked the initially retrieved documents by comparing them to the text and time query models. To illustrate a practical application of our time-aware retrieval models, we presented EXPOSÉ, a time-aware exploratory search system for past events.

To connect Wikipedia excerpts extracted from general pages, analogous to before, we cast the linking problem into an information retrieval task thereby considering the excerpt as a user query. In our approach to address the information retrieval task, we presented a framework that leverages additional semantics, namely time, geolocations, and named entities, that come with a given Wikiexcerpt. Under this framework, our method estimates independent query-text, -time, -space, and -entity models by considering them as event dimensions. Finally, documents are ranked by comparing them to the query across all the dimensions. Comprehensive experiments on two large datasets show that our approach that combines all the event dimensions to address the linking task is found to be most effective.

Future Directions

The problems defined in this chapter are novel. In our first of its kind approach, we study the effectiveness of combining additional semantics that are associated with events to identify relevant documents that give details. We speculate on how plausible result improvements can be achieved and plan to design such methods as future work.

- Even though in this work we focus on Wikipedia as a source of event descriptions, our approaches designed to connect them to news articles are fairly generic. Given the specific use case, one could leverage additional signals from Wikipedia to further improve the estimated query and document models in our approach. For example, Wikipedia revision history and page-view statistics can be leveraged to estimate pre-retrieval results thereby enabling treating difficult events as queries differently. As a future work, we plan to design better collection-specific features [59] to improve linking quality.

- In case of entity models, one can think of leveraging rich taxonomies that clearly categorize the entities to determine their salience. Leveraging such additional signals may lead to better entity models thus improving the result quality. As future work, we plan to design methods to estimate more accurate entity models leveraging additional information from knowledge bases such as YAGO [82].

- In our approach, we use a simple minimum bounding rectangles (MBRs) to model the scope of the geolocations associated with the events. Due to arbitrary shapes of geographical areas at coarser granularities like states or countries, such a representation may lead to false positives. A better representation such as convex hulls may be adopted to improve the estimation of the space models in our approach. It can also be noted that sometimes the geographical locations come with spatial relations like, "north of", and "south of" [107]. As future work, we plan to design methods to incorporate the additional information to estimate more accurate space models.

- We adopt a simplified representation for time in all our approaches. This enables us to leverage the KL-divergence-based framework to estimate query-document similarity. However, more complicated representations that capture the notion uncertainty [26] of coarse-grained temporal expressions like "decade" and "century" may be used to estimate better time models thus improving the result quality further. However, this opens a new future research direction.

Chapter 4

Summarizing Wikipedia Events with News Excerpts

4.1 Motivation & Problem Statement

Today, in this era of digitization, the World Wide Web plays an integral role as an effective and efficient digital medium for providing information on events of global as well as local importance. Large volumes of online news data are generated by media houses and other independent providers as they report eagerly on current events or those that have happened in the past. Contributing to the volume, variety, and velocity of the data, social media is also proving to be a new popular medium of news propagation across the globe. On one hand, this change from traditional print media to publishing online news has given rise to less polarizing and more democratic journalism. On the other hand, from the perspective of a general user, vast amounts of information contained within numerous online news articles with a high degree of redundancy have made it difficult to connect the dots to get a holistic understanding on past events.

Information retrieval systems or search engines have today proven to be one of the preferred power tools when seeking information from within large document collections such as news archives. However, state-of-the-art vertical news search engines (e.g., Google News[1]) are keyword-query-driven and often retrieve a large temporally-biased ranked list of news articles from various sources. For a (professional and general) user retrospecting on a past event, it becomes hard to sift through all retrieved news articles so as to get a holistic view on a past event. One could motivate that for such an information need, a concise *event digest* that is automatically generated by extracting text form the large number of retrieved news articles can prove to be useful. With such a digest, the user can first get a quick access to consolidated information presenting a broader view on the event. Then, if desired, the user may refer to individual documents

[1] https://news.google.com

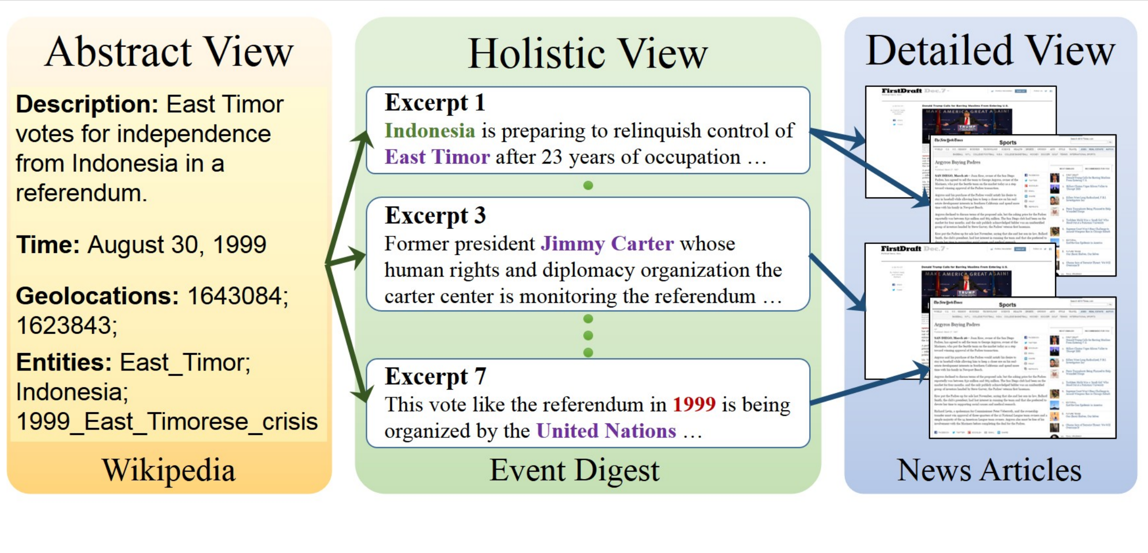


Figure 4.1 Different views on a past event.

for necessary details. Thus an event digest can facilitate more efficient consumption of information and there by aid in retrospection. A concrete example of an event digest is given in Table 4.2. To further motivation this problem consider the following real-world scenario:

- A journalist, Laura Lang, wants to quickly get a holistic view on the event of East Timor's independence, illustrated in Table 4.2. She uses Google News and issues the keyword query {*East, Timor, votes, independence, Indonesia, referendum*}. Not to her surprise, she finds that the system retrieves numerous (more than one thousand) news articles published by different news agencies. To get a good understanding of the event, she tediously sifts through many articles, most of which contain redundant information. However, with a concise digest given, she can first get an overview of the event, and then jump into news articles connected to *excerpts* in the digest to get necessary details.

In Chapter 3, we motivated that one plausible solution to the overloading problem arising due to information spread across a large number of documents is to better structure the information available online through linking orthogonal sources of information on past events (also motivated by [184, 202]). With a similar goal, we investigated a linking task that leveraged semantic annotations to identify relevant news articles that can be linked to events from Wikipedia articles. Two concrete examples of such events are illustrated in Table 4.1. We motivate that Wikipedia articles summarize past events by often abstracting from fine-grained details, and on the other hand, online news is published as the events happen and covers all angles with necessary details. Individually, they both fall short in providing a full picture due to context missing from news articles, and fine-grained details omitted from Wikipedia articles. However, connections can facilitate navigation between them and help in getting a larger picture. One drawback

Table 4.1 Examples of Wikipedia events.

1	**February - March 2002:** Riots and mass killings in the Indian state of Gujarat.
2	**November 22, 1995:** Rosemary West is sentenced to life for killing 10 women and girls, including her daughter and stepdaughter, after the jury returns a guilty verdict at Winchester Crown Court. The trial judge recommends that she should never be released from prison, making her only the second woman in British legal history to be subjected to a whole life tariff (the other is Myra Hindley).

of such a linking task is that for an excerpt from Wikipedia that represents an event with long ramifications, like the one in our example, many news articles get linked to it. An event digest in such a case becomes an intermediate level of linking that presents a *holistic view*. Excerpts from Wikipedia, an *abstract view*, are connected to excerpts in the digest which are in-turn connected to news articles that give a *detailed view* as illustrated in Figure 4.1. As a different perspective, the event digests can be additionally used to paraphrase abstract event descriptions in Wikipedia.

We address the following problem with the goal of connecting Wikipedia events to excerpts from news articles:

Problem Statement: *Given an event description from Wikipedia as a source, automatically connect it to excerpts from within news articles as targets such that together they present a holistic view on the event.*

Keeping this larger problem in view, we address the following two important subproblems: As the first subproblem to identify information excerpts from news articles for a Wikipedia event, we address a query-focused multi-document summarization task where the Wikipedia event description, coined Wikiexcerpt, is considered as a user query. The generated fixed-length event-focused summary of the news articles is referred to as *event digest*. We note that the Wikiexcerpts that describe events usually come with additional semantics namely, temporal expressions indicating time period, geolocations indicating the place, and named entities indicating the actors of the event. Intuitively, these semantic annotations can help to identify informative news excerpts that contain information on the Wikiexcerpt in question. Moreover, the additional semantics can be leveraged to avoid redundancy within selected news excerpts in an event digest so that they present a holistic view on the event.

As the second subproblem, we look into the evaluation of the quality of an event digest. We note that standard metrics such as ROUGE [117], and Pyramid [151], evaluate the content quality of generated multi-document summary. However, evaluation of coherence of text is often based on human judgments. In context of events, evaluation of coherence becomes much more difficult due to the large variance in the discourse

Event Description: East Timor votes for independence from Indonesia in a referendum
Time: August 30, 1999
Geolocations: 1643084; 1623843; 7289708;
Entities: East_Timor; Indonesia; 1999_East_Timorese_crisis;

Event Digest (with chronological ordering on publication dates)

• **Publication Date:** *July 20, 1999* **Source Link:** *http://goo.gl/rJYDiZ*
(1) **Indonesia** is preparing to relinquish control of **East Timor** after 23 years of occupation and it believes that independence advocates are highly likely to win a referendum **next month** says an authentic internal government report that has been made available to reporters by advocates of independence. **(2)** **Late next month** estimated 400 000 **East Timorese** are to choose between broad autonomy within **Indonesia** option 1 or independence option 2.

• **Publication Date:** *August 29, 1999* **Source Link:** *http://goo.gl/Cz6Jkk*
(3) Former president **Jimmy Carter** whose human rights and diplomacy organization the **Carter Center** is monitoring the referendum here said this **this month** some top representatives of the government of **Indonesia** have failed to fulfill their main obligations with regard to public order and security.

• **Publication Date:** *November 21, 1999* **Source Link:** *http://goo.gl/hdqYm8*
(4) The last time it was **East Timor** which voted for independence from **Indonesia** in **August** only to be plunged into a spasm of violence that required an **Australian** led international military force to quell it. **(5)** **Acehs** latest push for independence began with the fall of President **Suharto** in **May 1998** and accelerated after the **East Timor** referendum.

• **Publication Date:** *September 24, 2000* **Source Link:** *http://goo.gl/AijWVY*
(6) **East Timor** has been under a transitional **United Nations** administration since the **Aug. 30** independence vote **last year**. **(7)** The groups pillaged **East Timor** after **last year's** independence vote which freed the territory from military control.

• **Publication Date:** *August 24, 2001* **Source Link:** *http://goo.gl/EAGBxC*
(8) This vote like the referendum in **1999** is being organized by the **United Nations** which has continued to administer **East Timor** a former **Portuguese colony** annexed by **Indonesia** as it struggles to its feet economically and politically.

Figure 4.2 Example of an event digest.

structure [150] of text conventionally followed across the event. Thus, we look into the problem of designing a test collection comprising of large number of manually judged event summaries pairs exhibiting different structures which can be in future leveraged to design automatic evaluation frameworks for event-focused text summary coherence.

Approach Overview

To address the first subproblem, we motivate that in addition to text, a digest should also maximize the coverage of time, geolocations, and entities to present a holistic view on the past event of interest. Intuitively, while selecting excerpts from the input documents, if coverage of time period, geolocations, and named entities associated with the input event is maximized then a holistic view can be developed. To address the problem of automatic event digest generation, we propose a novel divergence-based framework that selects excerpts from an initial set of pseudo-relevant documents, such that the overall relevance is maximized, while avoiding redundancy in text, time, geolocations, and named entities, by treating them as independent dimensions of an event. Our method formulates the problem as an Integer Linear Program (ILP) for global inference to diversify across the event dimensions. Relevance and redundancy measures are defined based on Jensen–Shannon (JS) divergence between independent query and excerpt models estimated for each event dimension. Elaborate experiments on three real-world datasets are conducted to compare our methods against the state-of-the-art from the literature. Using Wikipedia articles as gold standard summaries in our evaluation, we find that the most holistic digest of an event is generated with our method that integrates all the event dimensions. We compare all methods using standard Rouge-1, -2, and -SU4 along with Rouge-NP, and a weighted variant of Rouge.

To design a test collection of human-written event-focused summaries as our second subproblem, we conduct an empirical study on a crowdsourcing platform to get insights into regularities that make a text summary coherent and readable. For this, we conduct three experiments by generating four variants of human-written text summaries with 10 sentences for 100 seminal events. Experiments 1 and 2 focus on analyzing the impact of sentence ordering and proximity between originally occurring adjacent sentences respectively. Experiment 3 analyzes the feasibility of conducting such a study on a crowdsourcing platform. As a contribution, we release our data to facilitate designing measures to evaluate summary coherence in future.

Other Potential Applications

In this chapter, we motivate the event digest generation as an approach to connect Wikipedia events to excerpts from news articles. However, the event digest defined as a more event-focused multi-document summary, can be used in a few downstream applications. Automatically generated event digests with large word lengths can be used to aid retrospection on past events, thus becoming a power tool for professionals such as journalists, blog writers, and news curators, who desire to get a quick overview on an event. Examples of commercial text extraction and text summarization real-world systems that provide such functionalities are Ultimate Research Assistant [1] and iResearch Reporter[2]. The Ultimate Research Assistant performs text mining on Internet search results to help summarize and organize them to make it easier for conducting online research on events. On the other hand, iResearch Reporter accepts user-entered query, passes it on to Google search engine, retrieves multiple relevant documents, and pro-

[1]http://ultimate-research-assistant.com/
[2]http://www.iresearch-reporter.com

duces easily readable natural language summary reports covering multiple documents (in the retrieved set) with links to original documents on the Web. It additionally provides post-processing functionalities like entity extraction, event relationship extraction, text clustering, etc. As another application, event digests can be used as an authoring tool for Wikipedia article population. Wikipedia articles evolve over time. Short and newly created articles are referred to as stubs [19]. Articles, more specifically those that are central to events, usually grow as more information is available in authoritative sources. Here, news become an important source that is used as a reference to edit such articles. In this context, an event digest generated in an update summarization setting can be used by Wikipedia editors to update the articles. Recent efforts [48, 182, 208, 213] have been made to crisply define problems in the direction of automatic Wikipedia article creation.

Contributions

We make the following key contributions in this chapter:

- We propose the new problem of event digest creation with the goal of providing an effective holistic view on a past event that is considered as an input.
- In our approach, we formulate the problem as an Integer Linear Program (ILP) to perform global inference for the event digest creation. To the best of our knowledge, we are the first to present a *unified method* to explicitly diversify across text, time, geolocation, and entity event dimensions.
- We present an experimental evaluation on three real-world datasets by treating Wikipedia articles central to an event query as a gold standard.
- We present a crowdsourcing-based study to understand the effect of the summary structure on its readability.
- We release a corpus with four variants of 10-sentence summaries for 100 Wikipedia events along with pair-wise human preference judgments on their readability.

Organization

The rest of the chapter is organized as follows. In Section 4.2, we put our work in context of prior work. In Section 4.3, we present our approach for event digest generation. We view this as a representation of intermediate linking between Wikipedia events and news articles at a finer granularity. Section 4.4 presents our approach to design a test collection of human-written summaries that are event-focused. Here we describe the different Crowdflower-based experiments conducted while developing the test set. Finally, we conclude in Section 4.5 and point out future directions for further improvements.

4.2 Related Work

We identify related prior works along the following eight lines: **1)** we motivate that the event digest generation problem is a special variant of unsupervised query-focused extractive text summarization. Thus as the first line of research, we look into prior works that address the extractive text summarization task. **2)** Next we look into approaches that address the query-focused summarization task. **3)** As the third research line, we look into the task of generating a timeline of events that are associated with entities. Next we look into approaches to address: **4)** passage retrieval and **5)** diversify search results since we aim to address similar challenges for event digest generation. In this work, we also look into the problem of identifying regular patterns in coherent texts to design better evaluation measures as future work. As the next lines of research, we look into: **6)** different evaluation techniques to estimate the quality, and **7)** the coherence of an automatically generated summary. **7)** Finally, since we use crowdsourcing platform for conducting our studies, we review prior works that present appropriate techniques to conduct such studies.

Extractive Summarization

Extractive summarization focuses on selecting sentences from a single or multiple input documents to create a summary. This line of research has received much attention in the past [34, 83, 116, 120, 137, 172, 221]. It was also investigated at the Document Understanding (DUC) and Text Analysis Conferences (TAC). From various subclasses of extractive summarization, we identify three that seem to be most related to our task: *multi-document summarization*, *query focused multi-document summarization*, and *timeline generation*. In the realm of unsupervised summarization techniques, MMR [34] stands as the most popular approach that defines an objective function rewarding relevance and penalizing redundancy. McDonald et al. [137] proposed an ILP formulation with a slight change to the original MMR objective function. Litvak and Last [120], and Riedhammer et al. [172] proposed to use key phrases to summarize news articles and meetings. Gillick et al. [69] maximized the coverage of the salient terms in input documents to generate summaries. However, in the problems investigated by works mentioned above, there is no notion of a user query. This stands as a difference to our problem where we have to generate a digest for a given event query. Further, we incorporate additional semantics to identify informational excerpts for the digest. However, in our approach, we incorporate the formulation given by Riedhammer et al. [172] to develop our text-only method.

Query-Focused Multi-Document Summarization

Query-focused multi-document summarization takes into consideration a topic that is input as a user query to generate a topic-focused summary. For this task, supervised approaches have recently proved to be effective [83, 116, 221]. However, they require labeled data for training. Firstly, these approaches focus on short queries, like TREC adhoc topics used in TAC, whereas in our problem, event queries are verbose textual descriptions of events. Secondly, we present an unsupervised approach that formulates an ILP for event digest generation.

Timeline Generation

Timeline generation as a subclass of query-focused summarization focuses on events, has also received attention [11, 37, 212]. Here, the main goal is to generate a time-stamped list of updates as sentences, key phrases, etc., covering different facets of an event. As an early approach, Allan et al. [11] proposed clustering-based approaches on entities and noun phrases to generate a timeline for a given event. Chieu et al. [37] leveraged burstiness as a ranking metric to identify sentences to be included into a timeline. McCreadie et al. [136] proposed an incremental update summarization task and presented supervised methods to address it. Recently, in a different direction, Shahaf et al. [184] addressed the information overloading problem by presenting a map of connected news articles that captures the story development of a given event. Timeline generation and incremental update summarization tasks aim at presenting a concise ordered summary of events. This is different from our task in two ways: firstly, we do not focus on the ordering of the excerpts in a digest; and secondly, we focus on generating a holistic view by explicitly diversifying across different dimensions of a past event to aid retrospection.

Passage Retrieval

Passage retrieval tasks have been well studied in the past. Systems retrieving passages have been proven to be effective for IR tasks when the documents are long or contain diverse topics. One popular way to define a passage is based on the document structure [31, 77, 180]. Another example of passages are windows consisting of a fixed number of words. These can be further classified into overlapping [31] or non-overlapping windows [100]. The traditional passage retrieval tasks do not take diversity of the passages into consideration. However, we find the definitions of the passages to be complementary to our excerpts.

Search Result Diversification

Search result diversification problem originally aimed at identifying documents from a relevant set that catered to different information needs of a user query. Further, we look into prior novelty-based strategies to diversify search results. MRR [34] is among the first formulations that penalized documents based on redundancy. This was extended by Zhai et al. [218] for language models and they proposed a risk minimization framework to diversify search results. Wang et al. [206] proposed a mean-variance analysis (MVA) diversification objective. A recent work that becomes interesting is presented by Dou et al. [49] and Leelanupab et al. [114] as they attempt to diversify across multiple implicit subtopics by treating them as dimensions of the query. All the methods above cater only to text, and extending them to time, geolocations, and entity dimensions is not straightforward.

Text Summarization Evaluation

Evaluation of automatically system-generated summaries is a hard task. This is primarily due to the absence of an *"ideal"* summary that can be leveraged as a ground truth. Commonly, the automatically generated summaries are compared with human-written or so-called *reference* summaries [119]. However, such an evaluation setting also poses several drawbacks as pointed out by Nenkova et al. [151]. The most prominent of them is low agreement in the reference summaries, that is, different sentences are selected by different humans while generating a summary. Thus, often multiple reference summaries are used for evaluating the content of a system-generated summary.

In intrinsic evaluations [191], there exist several metrics in the literature to measure the goodness of a generated summary. Largely, they can be categorized into text, co-selection, and content quality measures [169]. To evaluate content quality of a summary, measures like ROUGE [117], Pyramid [151], and longest common subsequence [169] are used. The co-selection and content quality measures can automatically be computed from gold-standard reference summaries. However, estimating the text quality of a summary, like coherence and readability often require human judgments

Summary Coherence Evaluation

For event-related text or news article summarization tasks, the structure of a generated summary becomes important for its readability. However, lack of an ideal (correct) ordering for a given set of sentences in a summary makes the evaluation of summary coherence (defined in terms of sentence ordering) a challenge. In single document summarization, one possible ordering is provided by the source document itself. However, Jing [91] observed that extracted sentences may not retain the ordering of the documents. An alternative evaluation method can be to compare the prevalence of the discourse structure of the source documents represented by rhetorical [80] and coherence relations [85]. However, Ono et al. [158] discovered significant differences in accuracy when building a discourse representation from technical tutorial and newspaper texts. Few prior works [88, 151] have leveraged the small number of available reference summaries to evaluate a summary structure. Other approaches [20, 22, 109, 191] have resorted to human preference judgments for system-generated summaries which proves to be expensive for a large scale evaluation.

Crowd-based Text Summarization Evaluation

Crowdsourcing services have been successfully used in various natural language processing (NLP) [190], information retrieval (IR) [12], and text summarization tasks [21, 22, 95, 123]. Primarily, they have been leveraged to obtain human annotations for generating ground truths that are then used to evaluate automatic systems. In the context of evaluating short-fixed length summaries, Barzilay et. al. [21] conduct a study to create a collection of multiple orderings generated by humans for 9 sentences extracted from news articles. In another study, Kaisser et. al. [95] look into the effect of changing the summary length. Recently, Lloret et al. [123] study the viability of using a crowdsourcing service to generate reference summaries.

We leverage a crowdsourcing service to gather human assessments for four possible orderings of 10 sentences for 100 events from Wikipedia. This is different from the work done by Barzilay et. al. [21] as they generate a corpus of 10 orderings from sentences that are selected with their MultiGen system, from only 10 sets of news articles. We start with human-written summaries that are accepted as coherent, and use a crowdsourcing service to get insights into impact of altering the ordering and proximity between originally occurring adjacent sentences on the coherence and readability of text. For this, pairs of summaries with different sentence ordering are manually compared on the Crowdflower platform and preference judgments are gathered based on text coherence.

4.3 Event Digest Generation

We look into the problem of automatic event digest generation to aid effective and efficient retrospection. We motivate that to generate a holistic view on a past event, a digest should maximize the reportage of time, geolocations, and entities in addition to salient textual terms to present a holistic view on the past event of interest. Thus, we propose to address the following problem:

Problem Statement: *Given a Wikiexcerpt describing an event, along with a time interval indicating its occurrence period, automatically generate a fixed length digest comprising of news excerpts that presents a holistic view on the event.*

As input, we consider: **1)** an *event query* derived from a Wikiexcerpt with textual description, and time interval; **2)** a set of textually *pseudo-relevant documents* retrieved using a standard retrieval model with a keyword query generated from the event description; and **3)** a length budget for the digest in terms of number of words. As output, our goal is to return a diverse set of excerpts from news articles to compose an *event digest* with its total length under a given length budget such that it presents a holistic view on the event in the query. We interchangeably refer to a Wikiexcerpt as an event query or simply query; and a news excerpt as simply an excerpt.

Challenges

Addressing the above defined problem includes the following challenges:

- Leveraging semantic annotations, namely text, time, geolocations, and named entities, associated with a given event from Wikipedia can lead to better capturing its information content. Thus designing appropriate event-query models for the semantic annotations, thereby treating them as event dimensions is a challenge.
- We note that the Wikipedia events come with few semantic annotations. Thus it becomes a challenge to deal with the sparsity.
- Analogous to the Wikipedia event, it becomes important to capture the information content of the News excerpts by leveraging the semantic annotations associated with them. Thus designing appropriate excerpt models is a challenge.
- Since we treat the event digest generation problem as unsupervised event-focused multi-document summarization, it becomes a challenge to design a framework to address the global inference problem to jointly maximize relevance to the query while avoiding redundancy across the event dimensions.
- Since there are no existing benchmarks with human-written gold-standard summaries to evaluate the quality of a system-generated event digest, it becomes a challenge to design an appropriate evaluation framework for our task.
- Finally, it is a challenge to design measures which evaluate how holistic is the information presented by the event digest that was generated by our methods.

State-of-the-Art Systems

Traditional multi-document extractive summarization tasks [34, 83, 116, 120, 137, 172, 221] focus on generating textual summaries from filtered relevant documents such that they are as close as possible to a manually created summary. Unsupervised methods in this realm consider only text to maximize relevance and reduce redundancy in the generated summaries. However, we define an event to be a joint distribution over independent *text*, *time*, *geolocation*, and *entity* dimensions, indicating the time period, geographic locations, and entities affected by its ramifications. To present a holistic view on an event, we motivate that relevant information along all four dimensions has to be diversified. For the example in Figure 4.2, diversifying across time will result in information on *causality* (excerpt 1 and 2), *effects during happening* (excerpts 3 to 5), and *after-math* (excerpts 6 to 8) of the event. Similarly, diversifying across geolocations will give information on the entire geographical scope, and diversifying across entities will give information on all persons, places, and organizations involved. We refer to such a view as an *event digest* that contrasts from a traditional notion of a summary.

4.3.1 Notation & Representations

We begin by defining our notation and representations.

Event Query

An event query q is derived from a given Wikiexcerpt that comes with a short textual description and a time interval indicating its occurrence period. We assume that an event is a joint distribution over four independent dimensions: text, time, geolocation, and entity. Thus, from a given query we derive the following four parts from the textual description: query-text part q_{text} as a bag of textual words; query-time part q_{time} as a bag of explicit temporal expressions; query-space part q_{space} as a bag of geolocations; and query-entity part q_{entity} as a bag of entity mentions.

News excerpts

A news excerpt ε is a single unit of an input document that gives information on an event. In this work, we fix an excerpt as a single sentence, however other definitions may be adopted depending on the application. Analogous to the query, each excerpt has four parts: text ε_{text}, time ε_{time}, geolocation ε_{space}, and entity ε_{entity} part.

In our method, we sometimes use the entire collection as a single coalesced document and refer to its corresponding parts as C_{text}, C_{time}, C_{space}, and C_{entity}.

Time

Our time domain $T \times T$ is modeled as a two-dimensional distribution over integers or *chronons* that represent the time passed (to pass) since (until) a reference date such as the UNIX epoch. We adopt a simplified representation of the time domain as proposed by Berberich et al. [26]. We normalize a temporal expression to an interval $[tb, te]$ with begin time tb and end time te. Further, each interval is described as a quadruple $[tb_l, tb_u, te_l, te_u]$ where tb_l gives the lower bound, and tb_u gives the upper bound for the begin time tb of the interval. Analogously, te_l and te_u give the bounds for the end time te. In our work, we fix the granularity of a time unit to a single day, however, this can be varied based on the application.

Geolocations

Geolocations mentioned in text are modeled using the geodetic system in terms of $latitude \times longitude$. A geolocation s is represented by its Minimum Bounding Rectangle (MBR). Each MBR is described as a quadruple $[tp, lt, bt, rt]$, where the point (tp, lt) is the top-left corner and (bt, rt) is the bottom-right corner of the MBR. A geolocation unit g refers to a geographical point in our two-dimensional grid. Further, we empirically set a minimum resolution $resol_{lat} \times resol_{long}$ of the grid to smooth out noisy annotations at a very high granularity (like streets and avenues). A further filtering of geolocations can be achieved by considering populated places with a population over a certain threshold (for example 1000).

Named Entities

In our model, an entity e refers to a location, person, or organization. Our entity dimension represents all entities in the YAGO2 [82] knowledge base. We use the YAGO URI of an entity as its unique identifier while estimating query- and excerpt-entity models. Geographic locations are redundantly considered as entities in our approach. This is essential to deal with annotation errors from either the geolocation tagger or entity disambiguation tool. For example, *Washington* can sometimes be used in text to refer to the state or the US government.

4.3.2 Event -Digest Generation Framework

We present our approach to create a concise digest for a given event that presents a holistic view by describing as many aspects as possible. Intuitively, while selecting excerpts from the input documents, if the coverage of time, geolocations, and entities associated with the input event is maximized then a holistic view can be developed. We propose a novel *divergence-based framework* for event digest creation. Under this framework, our method estimates independent query and excerpt models, and maximizes the relevance while avoiding inter-excerpt redundancy based on the KL-divergence between the models. We define an event as a joint distribution over text, time, geolocation, and entity dimensions. Our method extends the divergence-based retrieval framework, and formulates a single *unified* linear problem to perform global inference across the event dimensions by formulating an ILP.

4.3.3 Query and Excerpt Models

The divergence-based framework with the independence assumption between the dimensions allows us to estimate the corresponding query models uniquely. For this, we first expand the original parts of a query (other than text) with the given input set of the documents, thus treating them as pseudo-relevant. Intuitively, by expanding the query parts we cope with overly specific annotations in the original query. We refer to the previous Chapter 3 for more details.

Query-Text Model

Query modeling for text is a well-studied problem. The main intuition is that the query-text model should capture the true intent of the query in the text dimension. In our approach, we treat the set of excerpts R in the input documents as pseudo-relevant, and consider their excerpt text model $\mathscr{E}_{text}$ to estimate a feedback model. We then combine the feedback model with the empirical query model, estimated from q_{text}, to boost salient words for the event in the query. Since the best way of combining is through linear interpolation [217], we define the generative probability of a word w as,

$$P(w \mid Q_{text}) = (1-\theta) \cdot P(w|q_{text}) + \theta \cdot \sum_{\mathscr{E}_{text} \in R} P(w|\mathscr{E}_{text}). \tag{4.1}$$

A word w is generated from the feedback model with θ probability and from the original query with $(1-\theta)$ probability. Since we use a subset of the available words, we finally re-normalize as,

$$\hat{P}(w \mid Q_{text}) = \frac{P(w \mid Q_{text})}{\sum_{w' \in V} P(w' \mid Q_{text})}. \tag{4.2}$$

Query-Time Model

Query-time model Q_{time} can be understood as a probability distribution in our time domain $T \times T$ that captures the true temporal scope of an input event. We assume that the time part q_{time} of a query is sampled from Q_{time}. The generative probability of a time unit t from the query-time model Q_{time} is estimated by iterating over all the time intervals $[tb, te] \in q_{time}$ as,

$$P(t \mid Q_{time}) = \sum_{[tb_l, tb_u, te_l, te_u] \in q_{time}} \frac{\mathbb{1}(t \in [tb_l, tb_u, te_l, te_u])}{|[tb_l, tb_u, te_l, te_u]|} \tag{4.3}$$

where $\mathbb{1}(\cdot)$ indicator function indicates containment of a time unit t within an interval that is represented as $[tb_l, tb_u, te_l, te_u]$, i.e., does the point t lie within the interval. If a time unit overlaps with a time interval then we add a probability mass proportional to the inverse of the interval's area in the space. Intuitively, this assigns higher probability to time units that overlap with a layer of specific (smaller area) intervals in q_{time}. For computation of areas and intersections of temporal intervals we refer to [26]. To handle *near misses* (described in Chapter 3) we perform an additional two-dimensional Gaussian smoothing that blurs the boundaries of Q_{time} by spilling some probability mass to adjacent time units. With this, the new generative probability is estimated as,

$$\hat{P}(t \mid Q_{time}) = \sum_{t \in T \times T} G_\sigma(t) \cdot P(t \mid Q_{time}) \tag{4.4}$$

where G_σ denotes a 2-D Gaussian kernel that is defined as,

$$G_\sigma(t) = \frac{1}{2\pi\sigma^2} \cdot exp\left(-\frac{(tb_l, tb_u)^2 + (te_l, te_u)^2}{2\sigma^2}\right).$$

Finally, we normalize analogous to Equation 4.2.

Query-Space Model

Analogous to time, the query-space model is a probability distribution from which the geolocation part of a given query is sampled. It captures the true geographical scope of the event described in the query. The generative probability of a geolocation unit g from the query-space model Q_{space} by iterating over all $[tp, lt, bt, rt] \in q_{space}$ is estimated as,

$$P(g \mid Q_{space}) = \sum_{[tp,lt,bt,rt] \in q_{space}} \frac{\mathbb{1}(g \in [tp, lt, bt, rt])}{|[tp, lt, bt, rt]|}. \tag{4.5}$$

The $\mathbb{1}(\cdot)$ indicator function indicates containment, of a space unit g within a MBR $[tp, lt, bt, rt]$, i.e., does the point g lie within the MBR. Since we normalize with the area of the MBR, a geolocation unit gets higher probability if it overlaps with many specific geolocations (MBR with smaller area) in q_{space}. Area of a MBR, $|[tp, lt, bt, rt]|$ can easily be computed as $(rt - lt + resol_{lat}) * (tp - bt + resol_{long})$. To avoid the issue of near misses, we estimate $\hat{P}(l|Q_{Space})$ with additional Gaussian smoothing as described in Equation 4.4, and finally re-normalize as per Equation 4.2.

Query-Entity Model

The query-entity model Q_{entity} is a probability distribution over our entity space and captures the entities that are salient to the event in a given query. To estimate Q_{entity} from q_{entity}, we follow a similar process as described for the query-text model, by combining the empirical entity model with a feedback model estimated from the entity models $\mathscr{E}_{entity}$ pseudo-relevant excerpt set R. The generative probability of an entity e is estimated analogous to the query-text model as,

$$P(e \mid Q_{entity}) = (1-\theta) \cdot P(e \mid q_{entity}) + \theta \cdot \sum_{\mathscr{E}_{entity} \in R} P(e \mid \mathscr{E}_{entity}) \tag{4.6}$$

where $P(e \mid q_{entity})$ and $P(e \mid \mathscr{E}_{entity})$ are the likelihoods of generating e from the query and pseudo-relevant excerpts $\varepsilon \in R$ respectively. We finally normalize as in Equation 4.2.

Excerpt Model

An excerpt model for each dimension is estimated by following a similar methodology as for the query modeling. However, we additionally add Dirichlet smoothing [217] to the excerpt models with the collection C as a background model. For the text dimension, the excerpt-text model $\mathscr{E}_{text}$ is formally estimated as,

$$P(w \mid \mathscr{E}_{text}) = \frac{\hat{P}(w \mid \mathscr{E}_{text}) + \mu \cdot P(w \mid C_{text})}{|\mathscr{E}_{text}| + \mu} \tag{4.7}$$

where $\hat{P}(w|\mathscr{E}_{text})$ is computed according to Equations 4.1 and μ is set as the average excerpt length of our collection [217]. Similarly, for time, geolocations, and entity models we follow Equation 4.3, 4.5, and 4.6 respectively. However, for estimating $\mathscr{E}_{time}$ and $\mathscr{E}_{space}$, we do not employ the Gaussian smoothing (Equation 4.4) as this tends to introduce additional information into the excerpts artificially.

4.3.4 ILP Formulation

In the realm of unsupervised extractive multi-document summarization, the primary goal is to jointly optimize the three properties of a summary . One of the first global models was presented by Carbonell and Goldstein [34] that used the maximum marginal relevance (MMR) criteria to score sentences in documents with a linear combination of relevance and redundancy between already selected sentences in the summary. While it has become standard to use greedy approximations of MMR, with the recent advent of more powerful optimizers, such as the Gurobi optimizer [6], it has become feasible to find exact solutions to the global inference problem using integer linear programming (ILP) [69, 137, 172].

After estimating necessary query and excerpt models, we next describe our ILP designed for the event digest generation. With our assumptions in mind, we first specify our exact requirements. A digest should portray the following characteristics:

1. A digest should be comprised of relevant excerpts to a given event query.
2. Information redundancy between the excerpts should be avoided.
3. Coverage of salient time, geolocations, and entities should be maximized.
4. The event digest length in words should not be more than a given budget L.

To design an ILP, we define the following binary indicator variables: **1)** S_i indicates if a candidate excerpt ε_i is selected into the digest. **2)** For a given excerpt ε_i, M_{ij} indicates the single most redundant excerpt ε_j that is already selected into the digest. **3)** T_{it} indicates if there is an overlap with $t \in Q_{time}$. **4)** G_{ig} indicates if there is an overlap with $g \in Q_{space}$. **5)** E_{ie} indicates if there is an overlap with $e \in Q_{entity}$. With the indicator variables defined, we can now precisely formulate our ILP. The objective function and the constraints are illustrated in Algorithm 6. We next describe the formulation in detail.

Algorithm 6 ILP to generate an event digest.

Maximize:

$$\sum_i \Big[\alpha \Big(\lambda rel_i S_i - (1-\lambda) \sum_{j \neq i} red_{ij} M_{ij} \Big) + \frac{\beta}{N_t} \sum_{t \in Q_{time}} \omega_{it} T_{it} + \frac{\gamma}{N_g} \sum_{g \in Q_{space}} \omega_{ig} G_{ig} + \frac{\psi}{N_e} \sum_{e \in Q_{entity}} \omega_{ie} E_{ie} \Big]$$

Subject to:

Constraints on text:

1) $\sum_j M_{ij} = S_i \qquad \forall\, i$

2) $M_{ij} \leq S_i \qquad \forall\, i$

3) $M_{ij} \leq S_j \qquad \forall\, j$

4) $M_{ik} \geq S_k - (1 - S_i) - \sum_{j:red_{ij} \geq red_{ik}} S_j \qquad \forall i \neq k$

Constraints on time:

5) $\sum_i T_{it} \geq 1 \qquad \forall\, t \in Q_{time}$

6) $T_{it} \geq S_i \cdot O_{it} \qquad \forall\, i,\, t \in Q_{time}$

7) $T_{it} \leq S_i \qquad \forall\, i,\, t \in Q_{time}$

Constraints on geolocation:

8) $\sum_i G_{ig} \geq 1 \qquad \forall\, g \in Q_{space}$

9) $G_{ig} \geq S_i \cdot O_{ig} \qquad \forall\, i,\, g \in Q_{space}$

10) $G_{ig} \leq S_i \qquad \forall\, i,\, g \in Q_{space}$

Constraints on entity:

11) $\sum_i E_{ie} \geq 1 \qquad \forall\, e \in Q_{entity}$

12) $E_{ie} \geq S_i \cdot O_{ie} \qquad \forall\, i,\, e \in Q_{entity}$

13) $E_{ie} \leq S_i \qquad \forall\, i,\, e \in Q_{entity}$

Objective Function

In the ILP in Algorithm 6, the objective function can be explained as four parts: text, time, geolocation, and entity. In the text part, rel_i function computes the relevance between an excerpt ε_i and query q. Each excerpt is penalized with the maximum textual redundancy score red_{ij} with the already-selected excerpts into the digest. The parameter λ balances textual relevance and redundancy estimates.

To explain the rest of the formulation, let us consider the time part in isolation. For each excerpt, the following steps are followed: first, identify the time units that overlap with the given Q_{time}. Second, weights ω_{it} of the time units are summed up and assigned as temporal scores. The rest of the parts, i.e., space and entity, are handled similar to time. To specify the global importance of the dimensions, four parameters, α, β, γ, and ψ are introduced into the objective function. Finally, we normalize the time, geolocation, and entity parts with the size of their corresponding query models denoted as N_t, N_g, and N_e respectively.

Constraints

The constraints defined for our ILP in Algorithm 6 are categorized into four parts corresponding to the objective function. Constraints on text are defined on the binary indicator variable M_{ij}. Constraints 1-3 enforce that exactly one excerpt in the digest is selected for consideration as most redundant. Constraint 4 is to ensure that if $M_{ik} = 1$, then ε_j is most redundant to ε_i and they both are selected into the final digest [172]. Constraints 5, 8, and 11 ensure that each unit in the query model is covered by at least one excerpt in the digest, thus maximizing total coverage of the query units in the digest. Constraints 6, 9, and 12 specify that if an excerpt is selected then overlapping units across the dimensions are covered. For these constraints, we introduce an additional binary variable O_{ik} that indicates if there is an overlap between a k^{th} unit in query with excerpt ε_i model. Finally, constraints 7, 10, and 13 are required as sanity check that a unit can be covered only if an overlapping excerpt is selected.

Textual Relevance

In our ILP in Algorithm 6, the textual relevance rel between a given query q and excerpt ε in the objective function is estimated by computing the KL-divergence KLD between their language models, denoted as Q_{text} and $\mathscr{E}_{text}$ respectively. Formally, the relevance score of an excerpt ε_i is computed as,

$$rel_i = -KLD(Q_{text} \| \mathscr{E}_{i\ text})\,. \tag{4.8}$$

Textual Redundancy

In our ILP in Algorithm 6, the textual redundancy red between any two excerpts can simply be interpreted as the similarity between them. For this, we compute the Jensen-Shannon divergence JSD which is the symmetric variant of the KLD and a popularly used distance metric. In this case, lower divergence indicates higher redundancy between the excerpts. Formally, this is defined as,

$$red_{ij} = -JSD(\mathcal{E}_{i\ text} || \mathcal{E}_{j\ text}) \tag{4.9}$$

Weights

The weights ω_{it}, ω_{ig}, and ω_{ie} specify the importance of the time t, geolocation g, and entity unit e respectively for an excerpt ε_i. Under our divergence-based framework, we define the weights as the negative KL-divergence between the generative probability of a unit from the query and excerpt models. Formally, we define weights for all dimensions analogous to time as,

$$\omega_{it} = -KLD\big(P(t|Q_{time}) || P(t|\mathcal{E}_{i\ time})\big) . \tag{4.10}$$

For a single dimension considered in isolation, summing over the weights gives the overall divergence between excerpt and query models in that dimension. For example, in the time dimension the divergence $KLD(Q_{time} || \mathcal{E}_{i\ time})$ can be computed by summing over query time units as $\sum_{t \in Q_{time}} \omega_{it} T_{it}$. Intuitively, objective is maximized by globally minimizing the overall divergence of a query with the entire digest. However, since the KL-divergence scores are not bounded, we normalize the weights across all excerpts as,

$$\frac{KLD - KLD_{min}}{KLD_{max} - KLD_{min}} .$$

In our divergence framework, the redundancies in time, geolocations, and entity dimensions are minimized implicitly by maximizing the coverage of the units in query models. However, at the same time, the relevance of excerpts in these dimensions are also considered. The ILP solver first selects excerpts that cover most important units (receiving high probability in query model) with the lowest divergence scores as indicated by their weights.

4.3.5 Experimental Evaluation

Next, we give details on the conducted experiments. For reproducibility, we make the experimental data publicly available[1]. We begin by describing our test collections, query set, gold standards, and measures used in the experimental setup.

[1] http://resources.mpi-inf.mpg.de/d5/eventDigest/

Table 4.2 Document collection statistics

	NYT	Giga	CW12
Indexed documents	1,855,656	4,937,785	46,797,647
Average document length in words	691.79	423.61	740.04
Average of temporal expressions per document	8.61	8.03	5.98
Average of entity mentions per documents	6.3	4.12	6.6

Document Collections

We consider the following document collection for the experimental evaluations. Additional details are illustrated in Table 4.2.

- **The New York Times Annotated Corpus (NYT):** As our first corpus, we use the The New York Times Annotated Corpus (NYT) [4] with about 2 million news articles published between 1987 and 2007.
- **English Gigaword Corpus (Giga):** We use the English Gigaword [5] corpus with about 9 million news articles published between 1994 and 2010 as our second corpus. This collection contains news articles from five different news sources.
- **ClueWeb12-B13 (CW12):** Finally, as our third corpus we consider ClueWeb12-B13 [3] with about 50 million web pages crawled in 2012. We process the queries in our test set with a standard query likelihood document retrieval model.

For each query in the test set, Top-10 retrieved documents from each dataset are considered pseudo-relevant and input into our methods. Note that our first two collections are news archives while the third collection contains general Web pages. The NYT collection containing well-authored news articles from a single source comprises of least redundancy among its documents. The Giga collection is also a collection of well-authored news articles. However, since it is comprised of articles from five different sources reporting on a given event, there exists a higher degree of redundancy across the documents. Finally, the CW12 with Web articles represent a corpus with wide variety of documents for example representing blog posts, news, or general web pages.

Test Queries

The test queries are generated from the *timeline of modern history*[1] in Wikipedia that contains the most prominent news events in the past. We randomly sample 100 events occurring between 1987 and 2007 as test queries. Each query comes with a short textual description and a time interval indicating when the event happened. Further, each query is automatically annotated with time, geolocations, and entities. We include a full list of the test queries for experiments in Appendix A.4.

[1] https://en.wikipedia.org/wiki/Timeline_of_modern_history

Gold Standard

We consider a Wikipedia article that describes the specific event in a query as its human-generated or *gold standard digest* created by Wikipedians. Since these articles on past events are elaborate and cover most of the important aspects, they are apt for evaluating our task. We manually identify Wikipedia articles that are central to an event query.

Effectiveness Measures

We compare the use the following measures:

- ***Rouge-1***, ***Rouge-2***, and ***Rouge-SU4*****:** The Rouge-based measures [117] are well established for evaluating method-generated against gold standard (human-generated) summaries.

- ***Rouge-NP*****:** We introduce a new measure that takes into consideration only the noun phrases overlap. Generally, noun phrases represent the key concepts in a gold standard, and a larger overlap indicates better information coverage in the digest. This is further motivated by Taneva et al. [198] in their evaluation.

- ***Weighted-Rouge (w-Rouge)*****:** The above Rouge measures evaluate how close a method-generated digest is to the gold standard. However, due to the disparate quantity of text between a method-generated digest and the gold standard, these measures are not indicative of the diversity of excerpts in a digest. Thus, we introduce w-Rouge that computes Rouge-1 score of a method-generated digest S with each paragraph p of the corresponding gold standard GS. The individual Rouge-1 scores are weighted with the normalized length of each paragraph $\frac{|p|}{|GS|}$. To get the final score for a method, we average over all queries q in query set QS. Formally,

$$w\text{-}Rouge = \frac{1}{|QS|} \sum_{q \in QS} \sum_{p \in GS} \frac{|p|}{|GS|} \text{Rouge-1}(S, p) \tag{4.11}$$

  Additionally, we also report the mean variance ($MVar$) of the w-Rouge across all the queries. Formally, this is given as,

$$MVar = \frac{1}{|QS|} \sum_{q \in QS} \frac{1}{N} \sum_{p \in GS} \left[w\text{-}Rouge(S, p) - w\text{-}Rouge(S, GS) \right]^2$$

  where N is the number of paragraphs in GS. We assume that in a long Wikipedia article, each paragraph describes an aspect of the central event. Thus, a method-generated summary that gives diverse information should show overlap with a large number of paragraphs in the gold standard Wikipedia article. A method that generates a digest that is closer to the gold standard by covering more aspects of the given event should have a higher mean F1 score and larger mean variance of w-Rouge. Thus, we look at the scores together to identify a more effective method.

Implementation

All the methods are implemented in Java. For the temporal annotation, we use Stanford CoreNLP toolkit [1]. To annotate geolocations, we use an open-source gazetteer-based tool[2] that extracts locations and maps them to the GeoNames[3] knowledge base. For entity annotations, we use the AIDA [81] system. We use the Gurobi ILP solver[4] to solve our ILPs, and the Galago system[5] for implementing the pseudo-relevance feedback.

Methods under Comparison

We next describe the different methods that are compared in our experiments. We distinguish three frameworks that use integer linear programs for global inference: **1)** maximum marginal relevance [34, 137, 172], **2)** coverage-based [60, 69, 209], and **3)** divergence-based methods. While the first two are derived from the literature as state-of-the-art frameworks for unsupervised methods, the third divergence-based framework is proposed in this work. We extend the frameworks to incorporate time, geolocations, and entities, and design methods that leverage their different combinations under each framework. All our methods formulate the event digest problem as ILPs.

Maximum Marginal Relevance [34]: The maximal marginal relevance (MMR) is arguably considered as the most popular unsupervised framework for generating document summaries. We compare the following two methods under this framework:

- ***Mcd*:** As first method, we consider the summarizer presented by McDonald et al. [137] that uses an ILP for global inference in summarization. Though they follow the MMR [34] style formulation, they make a slight change to the global objective function by introducing a linear approximation. This results in candidates being penalized with the average redundancy to the already selected excerpts. Their objective is defined as,

$$Maximize: \sum_{i} \left[\lambda \cdot rel_i S_i - (1-\lambda) \cdot \sum_{j \neq i} red_{ij} S_{ij} \right]. \tag{4.12}$$

  We refer to [137] for the full set of constraints. The generalized ILP framework allows us to define the rel and red functions using language modeling methods as described in Equation 4.8 and 4.9.

- ***Rdh*:** More recently, Riedhammer et al. [172] propose an ILP formulation that got rid of the linear approximation in the global objective function of *Mcd*, thus giving

[1] http://stanfordnlp.github.io/CoreNLP/
[2] https://github.com/geoparser/geolocator
[3] http://www.geonames.org/
[4] http://www.gurobi.com
[5] https://www.lemurproject.org/galago.php

an optimal solution. In their formulation, they introduce an additional binary variable M_{ij} to indicate the maximum redundancy of an excerpt to the already selected excerpts in the digest. Further, they have additional constraints that are defined on this variable which leads to efficient convergence to the optimal value. Their global objective function is defined as,

$$Maximize: \sum_i \left[\lambda \cdot rel_i S_i - (1-\lambda) \cdot \sum_{j \neq i} red_{ij} M_{ij}\right]. \tag{4.13}$$

We refer to [172] for full set of constraints. Similar to *Mcd*, we use the definitions for rel and red as given in Equations 4.8 and 4.9.

Coverage-Based Framework [69]: The coverage-based framework is also popular in the summarization community as an unsupervised global inference method. It follows the idea of implicitly reducing the redundancy in the final summary by maximizing the coverage of textual units. Prior works [60, 209] propose various definitions for such units in the context of different tasks. This framework remains state-of-the-art, and approaches have shown to work well in comparison to other unsupervised global inference methods. At large, the framework is general enough and can easily be extended to our event dimensions. Their global objective function is defined as,

$$Maximize: \sum_i w_i \cdot C_i \,, \tag{4.14}$$

where w_i is defined as $P(c \mid Q_{text})$ probability of generating a term from query-text model, and C_i is a binary indicator variable that marks the occurrence of a term c in an excerpt ε_i. We design the following methods:

- ***Cov-txtEM*** and ***Cov-txtQM*:** As text-only methods, *Cov-txt* maximizes the coverage of the salient terms associated with an event. In their original work, Gillick et al. [69] relied heavily on preprocessing the documents to be summarized including key-phrase extraction. For this work, we do not do any preprocessing. In contrast, we make use of a query-text model Q_{text} to capture salient textual terms for the event in the query. We motivate that this makes their method more query-focused and stronger as a baseline. To demonstrate the advantage of incorporating a query model, we compare two methods: *Cov-txtEM* that uses only the empirical terms (after stopword removal), and *Cov-txtQM* that estimates a query model that re-weights the terms using pseudo-relevant documents as shown in Equation 4.1.

- ***Cov-T***, ***Cov-S***, ***Cov-E***, ***Cov-ST***, and ***Cov-EST*:** In principle, the coverage-based method can easily be extended to time, geolocations, and entity dimensions. In the time dimension, we adapt the objective function in Equation 4.14, such that it selects excerpts by maximizing the global coverage of all time units $\tau \in Q_{time}$.

We label this method as *Cov-T*. Similarly, *Cov-S* and *Cov-E* maximize the coverage of geolocation units $l \in Q_{space}$ and entities $e \in Q_{entity}$ respectively. We motivate that *Cov-E* method is similar to the original approach proposed by Gillick et al. that maximizes concepts, which in our case are entities. It is not hard to think of methods that maximize a combination of the dimensions in their objective functions. *Cov-ST* maximizes the coverage of geolocations and time, and *Cov-EST* additionally combines entities.

In all the above methods, weights w_i as in Equation 4.14, are generative probabilities from query models described in Section 4.3.3.

Divergence-Based Framework: Our divergence-based framework that is discussed in Section 4.3.2, takes into consideration the divergence between query and excerpt models in all the dimensions. We note that the text-only method under this framework is equivalent to *Rdh* that defines its rel and red functions in Equation 4.13 based on the KL-divergence between corresponding text models. We design the following methods:

- ***Div-T***, ***Div-S***, ***Div-E***, ***Div-ST***, ***Div-txtST***, ***Div-EST***, and ***Div-txtEST*:** Section 4.3.2 presents a unified divergence-based framework that maximizes the textual relevance and minimizes the redundancy across text, time, geolocations, and entities. We label this as the *Div-txtEST* method. However, one can think of methods that consider only a subset of the dimension. We thus design *Div-T*, *Div-S*, *Div-E*, *Div-ST*, *Div-txtST*, and *Div-EST* methods that leverage a combination of text (*txt*), time (*T*), space (*S*), and entities (*E*) as indicated by the suffixes in their labels.

Rand: This final method selects excerpts at uniform random from the input top-10 pseudo-relevant documents until the length constraint is satisfied.

Parameter Setting

We have two groups of parameters, first denoted by λ that balances the relevance rel with redundancy red. We vary λ from $[0,1]$ with fixed increments of 0.05. For NYT, Gigaword, and CW12 datasets, we set λ to 0.85, 0.90, and 0.95 respectively. The second group specifies the importance of text, time, geolocations, and entities with parameters, α, β, γ, and ψ respectively. Each parameter has a range $[0,1]$ and we fix the increments to 0.05 while maximizing Rouge-2 scores. We tune β, γ, and ψ while fixing $\alpha = 1-(\beta+\gamma+\psi)$ and empirically observe that setting the three parameters too high leads to a deterioration in the results. For NYT, Gigaword, and CW12, we set $\beta = [0.10, 0.10, 0.10]$, $\gamma = [0.05, 0.01, 0.01]$, and $\psi = [0.01, 0.30, 0.01]$. We adopt an exhaustive search strategy to tune the parameters while more advanced learning methods are out-of-scope of this work. Finally, for the query models described in Equation 4.1, 4.3, 4.5, and 4.6, we use settings described in Section 3.5.5.

Table 4.3 Results on The New York Times dataset.

Methods	Rouge 1			Rouge 2			Rouge NP			Rouge SU4		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
Rand	0.618	0.073	0.113	0.155	0.014	0.024	0.084	0.007	0.012	0.209	0.025	0.039
Mcd	0.652	0.078	0.121	0.192	0.019	0.030	0.086	0.008	0.013	0.205	0.026	0.039
Rdh	0.662	0.078	0.121	0.203	0.019	0.031	0.088	0.008	0.013	0.210	0.026	0.040
Cov-txtEM	0.652	0.051	0.085	0.180	0.011	0.020	0.091	0.005	0.009	0.219	0.017	0.029
Cov-txtQM	0.646	0.079	0.122	0.190	0.018	0.030	0.086	0.009	0.014	0.204	0.026	0.039
Cov-T	0.544	0.023	0.041	0.188	0.007	0.012	0.079	0.003	0.006	0.182	0.008	0.014
Cov-S	0.464	0.027	0.043	0.154	0.006	0.011	0.058	0.003	0.005	0.149	0.009	0.014
Cov-E	0.666	0.069	0.110	0.215	0.018	0.031	0.076	0.007	0.011	0.213	0.023	0.036
Cov-ST	0.647	0.038	0.062	0.214	0.010	0.017	0.087	0.005	0.008	0.212	0.013	0.021
Cov-EST	0.666	0.073	0.115	0.214	0.019	0.032	0.078	0.008	0.012	0.214	0.024	0.038
Div-T	0.647	0.068	0.110	0.195	0.017	0.028	0.088	0.008	0.013	0.211	0.023	0.037
Div-S	0.653	0.072	0.113	0.199	0.018	0.030	0.090	0.008	0.013	0.212	0.025	0.038
Div-E	0.652	0.077	0.120	0.195	0.019	0.031	0.091	0.008	0.014	0.213	0.026	0.041
Div-ST	0.662	0.081	0.124	0.210	0.020	0.033	0.090	0.008	0.014	0.215	0.027	0.041
Div-txtST	0.667	0.082	0.125	0.214	0.021	0.034	0.090	0.009	0.014	0.214	0.027	0.041
Div-EST	0.649	0.080	0.122	0.196	0.020	0.031	0.087	0.008	0.013	0.211	0.027	0.041
Div-txtEST	0.675	0.084	**0.127**	0.219	0.022	**0.035**	0.089	0.010	**0.016**	0.219	0.028	**0.042**

Table 4.4 Results on the Gigaword dataset.

Methods	Rouge 1			Rouge 2			Rouge NP			Rouge SU4		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Rand	0.643	0.069	0.109	0.191	0.016	0.027	0.163	0.017	0.028	0.242	0.026	0.041
Mcd	0.663	0.073	0.113	0.204	0.019	0.030	0.191	0.020	0.032	0.255	0.028	0.043
Rdh	0.654	0.075	0.119	0.201	0.019	0.031	0.190	0.021	0.034	0.247	0.028	0.045
Cov-txtEM	0.676	0.052	0.086	0.204	0.013	0.022	0.184	0.012	0.021	0.257	0.020	0.033
Cov-txtQM	0.652	0.078	0.122	0.197	0.019	0.031	0.178	0.021	0.033	0.242	0.029	0.045
Cov-T	0.372	0.015	0.027	0.120	0.004	0.007	0.127	0.005	0.008	0.150	0.006	0.010
Cov-S	0.509	0.021	0.036	0.167	0.005	0.008	0.164	0.006	0.011	0.206	0.008	0.013
Cov-E	0.664	0.070	0.111	0.212	0.018	0.030	0.173	0.017	0.028	0.255	0.027	0.042
Cov-ST	0.553	0.028	0.048	0.169	0.007	0.012	0.180	0.008	0.014	0.214	0.011	0.018
Cov-EST	0.665	0.071	0.112	0.214	0.018	0.030	0.174	0.017	0.028	0.257	0.027	0.043
Div-T	0.650	0.077	0.120	0.198	0.019	0.031	0.191	0.022	0.035	0.244	0.029	0.045
Div-S	0.647	0.071	0.112	0.201	0.017	0.028	0.191	0.020	0.033	0.245	0.026	0.042
Div-E	0.646	0.077	0.121	0.190	0.018	0.029	0.189	0.022	0.035	0.241	0.029	0.045
Div-ST	0.668	0.045	0.074	0.212	0.011	0.019	0.201	0.013	0.021	0.260	0.017	0.028
Div-txtST	0.655	0.081	0.125	0.201	0.020	0.033	0.192	0.023	0.036	0.248	0.031	0.047
Div-EST	0.619	0.074	0.116	0.189	0.018	0.030	0.182	0.022	0.034	0.233	0.028	0.044
Div-txtEST	0.655	0.081	**0.126**	0.205	0.021	**0.034**	0.192	0.022	**0.036**	0.248	0.031	**0.048**

Table 4.5 Results on the ClueWeb12-B13 dataset.

Methods	Rouge 1			Rouge 2			Rouge NP			Rouge SU4		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
Rand	0.604	0.058	0.095	0.173	0.013	0.022	0.135	0.011	0.018	0.248	0.021	0.035
Mcd	0.642	0.058	0.095	0.197	0.014	0.024	0.157	0.011	0.019	0.269	0.021	0.036
Rdh	0.630	0.063	0.101	0.193	0.015	0.026	0.156	0.012	0.020	0.261	0.023	0.038
Cov-txtEM	0.649	0.052	0.084	0.209	0.012	0.022	0.167	0.011	0.018	0.278	0.019	0.032
Cov-txtQM	0.650	0.065	0.105	0.206	0.016	0.028	0.164	0.014	0.022	0.277	0.024	0.040
Cov-T	0.410	0.011	0.020	0.155	0.004	0.007	0.140	0.003	0.006	0.186	0.005	0.009
Cov-S	0.375	0.014	0.026	0.131	0.004	0.008	0.112	0.004	0.007	0.168	0.006	0.010
Cov-E	0.628	0.060	0.097	0.199	0.015	0.026	0.140	0.012	0.020	0.260	0.022	0.036
Cov-ST	0.532	0.019	0.036	0.191	0.006	0.011	0.170	0.005	0.010	0.237	0.008	0.015
Cov-EST	0.634	0.060	0.097	0.203	0.015	0.026	0.146	0.012	0.020	0.265	0.022	0.037
Div-T	0.608	0.070	0.110	0.173	0.016	0.026	0.124	0.011	0.018	0.248	0.024	0.039
Div-S	0.568	0.055	0.092	0.163	0.013	0.022	0.117	0.009	0.015	0.235	0.019	0.033
Div-E	0.617	0.073	0.114	0.174	0.016	0.026	0.119	0.011	0.017	0.253	0.025	0.040
Div-ST	0.654	0.045	0.069	0.215	0.010	0.017	0.171	0.008	0.013	0.284	0.016	0.025
Div-txtST	0.614	0.072	0.113	0.172	0.017	0.027	0.122	0.011	0.019	0.249	0.025	0.041
Div-EST	0.613	0.073	0.113	0.170	0.016	0.026	0.115	0.011	0.017	0.248	0.025	0.041
Div-txtEST	0.632	0.075	**0.117**	0.192	0.018	**0.030**	0.152	0.014	**0.023**	0.264	0.027	**0.044**

Table 4.6 Statistical significance test with two tailed paired t test at three α levels: 0.01 (▲/▼), 0.05 (△/▽), and 0.10(˄/˅); of all against the text-only methods. Three symbols in each cell denote increase (up), decrease (down), or no change (–) in Rouge-1, -2, and -SU4, respectively.

Datasets	Methods	Mcd	Rdh	Cov–txtQM	Cov–T	Cov–S	Cov–E	Cov–ST	Cov–EST	Div–T	Div–S	Div–E	Div–ST	Div–txtST	Div–EST	Div–txtEST
NYT	Mcd	– – –	▲ △ △	– – –	▼ ▼ ▼	▼ ▼ –	▼ – ▽	▼ ▼ –	▼ ˄ –	▼ ▼ –	▼ – –	– – ˄	– ▲ –	▲ ▲ ▲	– – ▲	▲ ▲ ▲
	Rdh	▼ ▽ ▽	– – –	▲ ▽ –	▼ ▼ ▼	▼ ▼ ▼	▼ – ▽	▼ ▼ ▼	▼ – ▽	▼ ▼ ▽	▼ ˅ ▽	▼ ˅ ▲	△ – ▲	– ▲ ˄	▲ – ▲	– ▲ ▲
	Cov–txtQM	– – –	▼ △ –	– – –	▼ ▼ ˅	▼ ▼ –	▼ – ˅	▼ ▼ –	▼ △ –	▼ ▽ –	˅ – –	– – △	– ▲ ˄	▲ ▲ ▲	– – △	▲ ▲ ▲
Giga	Mcd	– – –	▲ ▲ △	▲ △ △	▼ ▼ ▼	▼ ▼ ▼	– – –	▼ ▼ ▼	– – –	▲ – –	– – –	▲ – –	▼ ▼ ▼	▲ ▲ ▲	▲ – –	▲ ▲ ▲
	Rdh	▼ ▼ ▽	– – –	– – –	▼ ▼ ▼	▼ ▼ ▼	▼ – ▼	▼ ▼ ▼	▼ – ▼	– – –	– – –	– – –	▼ ▼ ▼	▲ – △	– – –	▲ ▲ ▲
	Cov–txtQM	▼ ▽ ▽	– – –	– – –	▼ ▼ ▼	▼ ▼ ▼	▼ ▽ ▼	▼ ▼ ▼	▼ – ▼	– – –	▽ ˅ –	– ▽ –	▼ ▼ ▼	– – △	– – –	– ˄ ˄
CW12	Mcd	– – –	▲ ▲ –	▲ ▲ ˄	▼ ▼ ▼	▼ ▼ ▼	– – –	▼ ▼ ▼	– ˄ –	▲ ˄ ▲	˅ – –	▲ ▲ ▲	▼ ▼ ▼	▲ △ ▲	▲ △ ▲	▲ ▲ ▲
	Rdh	▼ ▼ –	– – –	– – –	▼ ▼ ▼	▼ ▼ ▼	▽ – –	▼ ▼ ▼	˅ – –	– – –	– – –	△ – –	▼ ▼ ▼	△ – –	△ – –	▲ ▲ –
	Cov–txtQM	▼ ▼ ˅	– – –	– – –	▼ ▼ ▼	▼ ▼ ▼	˅ ˅ –	▼ ▼ ▼	˅ – –	△ – –	– ˅ –	▲ – –	▼ ▼ ▼	▲ – –	▲ – –	▲ △ ˄

4.3.6 Results & Analysis

We compare the quality of event digests generated from the three different datasets. We also compare the variance of weighted-Rouge measure that we propose to highlight the diversification effect of each method.

Rouge score analysis

Table 4.3, Table 4.4, and Table 4.5 show the results of generating an event digest of length 250 words from NYT, Gigaword, CW12 documents respectively. We compare the digest generated by different methods against Wikipedia articles as gold standards, and report Rouge-1, Rouge-2, Rouge-NP, and Rouge-SU4 scores. We find that across all the three datasets the best quality digest is generated by the *Div-txtEST* method.

Firstly, we note that the random method *Rand* already achieves a decent F1 score. Across the three datasets, selecting excerpts randomly from the top-10 input pseudo-relevant NYT and Gigaword news articles generates better digests as compared to CW12 web pages. The text-only methods, *Mcd*, *Rdh*, and *Cov-txtQM* perform significantly better than *Rand*, as expected. Among the text-only methods, *Rdh* and *Cov-txtQM* show significant improvements over *Mcd* in terms of Rouge scores. At the same time, *Rdh* and *Cov-txtQM* follow different frameworks and prove to have overall similar performance. However, *Cov-txtQM* method proves to be better for Gigaword, and CW12 datasets while for NYT, *Cov-txtQM* gets significantly lower Rouge-2 F1 score. *Cov-txtEM* that uses only the empirical query terms proves to be the worst text-only method, thus highlighting the advantage of incorporating our query-text model. Next, we look at the methods that extend the coverage based framework into the different dimensions. *Cov-T* and *Cov-S* do not prove to be effective in any dataset. The *Cov-E* gets significantly higher score from its contemporary methods considering only time and geolocations. This method gets the highest gain over its contemporaries in the CW12 dataset. However, it always performs significantly worse than the text-only methods as shown in Table 4.6. The coverage-based methods in the time, geolocations, and entity dimensions get worse scores than the *Rand* due to relatively shorter digest generated. Later in the section, we discuss this in detail. Our proposed divergence-based methods perform better that the coverage-based methods across all three datasets. This is because *Div-T*, *Div-S*, *Div-E*, *Div-ST*, and *Div-STE* always perform better than the *Cov-T*, *Cov-S*, *Cov-E*, *Cov-ST*, and *Cov-STE* methods. Next we analyze the different combinations of dimensions in the divergence framework. The text-only method under this framework is equivalent to *Rdh* that performs significantly better than *Div-T*, *Div-S*, and *Div-E*. However, different combinations of the time, geolocations, and entities as *Div-ST* and *Div-STE* perform better than the text-only method. Finally, the *Div-txtSTE* with highest score proves to be the best method for our task.

Variance Analysis

From our experiments so far it is clear that event digests generated by the *Div-txtEST* method are closest to the gold standard Wikipedia articles. However, we perform an extended evaluation using the proposed w-Rouge measure that computes the mean Rouge-1 F1 score with respect to individual paragraphs in the Wikipedia articles. Firstly, we assume that each paragraph describes some aspect of the event. Thus, a larger mean with high variance in the weighted F1 score indicates higher coverage of the Wikipedia paragraphs, and hence better diversity in the generated digest. As shown in Table 4.7, all methods get higher mean and variance from *Rand* across all three datasets. We find that the *Div-txtEST* method proves to be the method that generates most diversified digest by achieving the highest mean and MVar scores. The next best method is *Div-txtST*.

Varying Digest Length

Next we analyze the effect on the quality of the digest by varying the digest *length budget*. Table 4.8 compares the *Div-txtEST* method with the text-only methods, *Cov-txtQM* and *Rdh*, in terms of Rouge-2. What we find is that for smaller length budget, *Cov-txtQM* performs better than other methods. We also find that both *Cov-txtQM* and *Div-txtEST* perform better than the *Rdh* across all the length budgets in Gigaword and CW12 datasets. The poor performance of *Rdh* as compared to *Cov-txtQM* can easily be understood by analyzing their formulation. For a very small budget of only 50 words, on average only two excerpts are selected into the digest by all the methods. *Rdh* first selects the most relevant excerpt, and then as the next it selects the one that is least redundant from the first. This causes the Precision to fall and an overall decrease in the F1 score. On the other hand, *Cov-txtQM* attempts to cover as many important terms with high probability in the Q_{text} to maximize the coverage. This generates a better digest for the smaller length budget. However, in the NYT dataset, for larger length budgets, *Cov-txtQM* suffers due to the lower redundancy in the news articles. Thus, as it tries to maximize term coverage, the Precision falls resulting in lower F1 scores as compared to *Rdh*. The effect of diversifying across time, geolocations, and entities in the *Div-txtEST* method is evident when the length budget is larger than 100.

Table 4.7 Comparison of methods in Weighted-Rouge1.

	NYT			GIGA			CW12		
	F1	**MVar**	**MSD**	**F1**	**MVar**	**MSD**	**F1**	**MVar**	**MSD**
Rand	0.022	1.87E-03	0.032	0.031	1.90E-03	0.030	0.025	1.51E-03	0.026
Mcd	0.023	2.10E-03	0.034	0.033	2.27E-03	0.031	0.026	1.47E-03	0.026
Rdh	0.023	2.07E-03	0.034	0.032	2.34E-03	0.033	0.026	1.58E-03	0.027
Cov-txtQM	0.023	2.19E-03	0.034	0.033	2.30E-03	0.033	0.029	1.72E-03	0.028
Cov-T	0.009	5.55E-04	0.012	0.008	3.95E-04	0.008	0.005	1.93E-04	0.005
Cov-S	0.010	7.29E-04	0.013	0.011	5.52E-04	0.011	0.006	2.59E-04	0.007
Cov-E	0.022	1.93E-03	0.032	0.031	1.97E-03	0.030	0.026	1.48E-03	0.026
Cov-ST	0.013	8.95E-04	0.018	0.014	8.28E-04	0.014	0.008	3.90E-04	0.010
Cov-EST	0.023	2.06E-03	0.033	0.031	2.00E-03	0.030	0.027	1.44E-03	0.026
Div-T	0.021	1.89E-03	0.032	0.033	2.47E-03	0.033	0.029	1.93E-03	0.030
Div-S	0.022	1.92E-03	0.032	0.029	2.28E-03	0.031	0.020	1.33E-03	0.024
Div-E	0.023	2.05E-03	0.034	0.033	2.36E-03	0.032	0.030	2.00E-03	0.030
Div-ST	0.023	2.22E-03	0.035	0.025	1.39E-03	0.021	0.022	1.28E-03	0.019
Div-txtST	0.023	2.16E-03	0.035	0.034	2.47E-03	0.033	0.030	1.99E-03	0.030
Div-EST	0.023	2.09E-03	0.034	0.031	2.40E-03	0.031	0.030	1.99E-03	0.030
Div-txtEST	**0.024**	**2.30E-03**	**0.036**	**0.034**	**2.49E-03**	**0.033**	**0.031**	**2.07E-03**	**0.031**

Table 4.8 Varying the length budget of methods.

Datasets	Length	Rdh			Cov-txtQM			Div-txtEST		
		P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
NYT	50	0.166	0.003	0.006	0.231	0.005	**0.009**	0.236	0.004	0.008
	100	0.222	0.009	**0.016**	0.211	0.008	0.015	0.223	0.008	0.015
	200	0.208	0.016	0.027	0.194	0.015	0.026	0.213	0.016	**0.027**
	300	0.199	0.022	0.036	0.184	0.021	0.034	0.205	0.023	**0.036**
	400	0.194	0.029	0.044	0.176	0.026	0.040	0.199	0.029	**0.044**
	500	0.189	0.034	0.050	0.169	0.032	0.046	0.194	0.035	**0.051**
Giga	50	0.146	0.003	0.005	0.268	0.006	**0.010**	0.242	0.005	0.009
	100	0.225	0.009	0.016	0.236	0.010	**0.018**	0.227	0.009	0.017
	200	0.209	0.016	0.027	0.206	0.016	0.027	0.209	0.017	**0.028**
	300	0.196	0.022	0.036	0.193	0.022	0.035	0.200	0.024	**0.037**
	400	0.189	0.028	0.043	0.186	0.028	0.043	0.191	0.030	**0.044**
	500	0.184	0.033	0.049	0.182	0.034	0.049	0.186	0.035	**0.051**
CW12	50	0.192	0.004	0.007	0.265	0.006	**0.010**	0.198	0.005	0.009
	100	0.189	0.007	0.013	0.223	0.009	**0.016**	0.187	0.009	0.016
	200	0.184	0.012	0.021	0.191	0.016	0.026	0.185	0.016	**0.027**
	300	0.187	0.017	0.028	0.176	0.021	0.032	0.176	0.022	**0.035**
	400	0.185	0.021	0.032	0.168	0.026	0.039	0.172	0.028	**0.041**
	500	0.183	0.023	0.035	0.164	0.030	0.044	0.169	0.033	**0.046**

Discussion

As the first point, we discuss the poor performance of the coverage framework in time, geolocation, and entity dimensions. The coverage-based framework selects excerpts such that the maximum number of *units* (of time, geolocations, or entities) in a given query are covered. The associated weights, as described in Equation 4.14, denote the importance of the units for a given query, thus forcing the ILP solver to cover more important units first. As a drawback, this framework does not take into consideration the importance of the units for the excerpts. This causes selection of excerpts that may not be relevant. Moreover, in the dimensions other than text, excerpts that do not come with explicit annotations are automatically disregarded. Since a single temporal or geographical expression can represent a large time interval (e.g., a century) or geographic area (e.g., a continent) respectively, the coverage of query units are easily maximized by selecting few excerpts. For example, the entire temporal scope of the query in Figure 4.2 is covered by excerpt 8. This causes the *Cov-T* and *Cov-S* methods to generate digests with fewer excerpts. Hence, they receive a worse Rouge score than the *Rand* which simply benefits from generating a longer digest. On the other hand, the divergence-based framework additionally regards the importance (higher generative probability) of a unit for the individual excerpts. While generating the digest, the ILP solver first selects excerpts which cover important query units with higher probability, thus lowering the overall divergence of the digest to the query. Moreover, since each excerpt is associated with an independent smoothed model for each dimension, no excerpt is disregarded.

Next, we discuss the diversification of excerpts in the digest achieved by each method. We note that Wikipedia articles as gold standard digests are textually larger than the system generated digests. We assume that Wikipedia articles cover most aspects of a given event as a query. To get a better insight into the diversification of the excerpts, we compare the methods using w-Rouge, proposed by us. The *Div-txtEST* gets the highest mean and variance scores, and proves that it achieves the best diversification.

We next discuss the individual dimensions. Text is clearly the most important dimension. However, we find that the text-only methods heavily rely on the query modeling techniques. Using only the empirical query terms leads to worse performance of *Cov-txtEM*. The *Cov-E* method uses only entities instead of all terms, as motivated by Gillick et al. [69], and is not able to beat text-only methods in terms of Rouge-2 scores. Time and geolocations are important indicators of identifying event-related excerpts and work well as a combination. Individually, we run into sparsity problems with very few annotations in excerpts. This is more pronounced in the CW12 dataset. The *Div-txtST* proves to be the second best method in the news datasets where we get comparatively more annotations. This is because every excerpt from a news article is also annotated with the publication date of the source article. Entities are more important for the CW12 documents and help to reinforce the text model. However, across all the datasets, combination of all four dimensions in our *Div-txtEST* proves to be the most effective.

Gain/Loss Analysis

The Rouge measures assume that higher n-gram overlap with gold standard implies more relevant excerpts in the digest. Thus, to get insights into the overall quality of digests, we manually identify relevant excerpts (highlighted in green) by referring to their source documents. We look at queries for which *Div-txtEST* shows the highest gain and loss in w-Rouge scores, when compared to the best among the text-only methods.

It achieves the highest gain in w-Rouge from the best text-only method for the query:

Example 4.1: January 26, 2001: An earthquake hits Gujarat, India, killing almost 20,000.

For this query, the best text-only method proves to be *Cov-txtQM*. Let us compare the digests generated by both the methods:

Event Digest by the *Div-txtEST* method: • In 1988 a devastating earthquake struck northern Armenia, killing 25,000 people, and in 1990 50,000 were killed by an earthquake that struck Rasht, Iran. • AP, 10405 2005 Oct 8, A 7.6-magnitude earthquake hit Kashmir near the Pakistan-India border reaching to Afghanistan. • Reuters, 725052005 Jul 24, A 7.2 earthquake hit Indias southern Andaman and Nicobar Islands and part of Indonesia. • **Shobha De, The Week, February 18, 2001 The story of the devastating earthquake in Gujarat is the story of women.** • **Scientists are already working to prepare earthquake probability map.** • **Skyscrapers need special construction to make them earthquake resistant.** • **This earthquake was not felt in Amreli, Junagarh or Porbander districts.** • **Leaders called for greater cooperation within the region to deal with the aftermath of disasters like the Kashmir earthquake and last year's devastating tsunami.** • **Additional Info The earthquake was centred 4.5 kms E of Gandhidham Gujarat, India.** • The Herald,India – was affected by the December 26, 2004 earthquake and subsequent tsunamis. • **Exactly a month ago, 18,122 people were killed in a deadly earthquake in the Kutch district.** • **The earthquake was felt strongly in parts of east-central Kachchh near the towns of Bachau and Vondh.** • Quito, now the capital of Ecuador, was shaken by an earthquake in 1797, and more than 40,000 people died. • **Almost four weeks after the earthquake, Gujarat is still coming to terms with what was and what is.** • The newly reopened Peace Bridge linking the Indian and Pakistani portions of disputed Kashmir nearly collapsed during the earthquake.

Event Digest by the *Cov-txtQM* method: • 543 AD - Disastrous earthquakes shaked much of the world; 803 AD - Fierce storms lashed the west coast of Ireland, killing close to 1,000; 851 AD - Rome had a violent earthquake that damaged Pope Leos 4-year-old Leonine Wall and further destroyed the Colosseum; 856 AD - an earthquake at Corinth killed an estimated 45,000 Greeks; 856 AD - an earthquake at Damghan, Iran killed an estimated 200,000; 893 AD - an earthquake at Ardabil, Iran killed about 150,000 people1138 AD - an earthquake at Aleppo, Syria claimed lives of aprox 230,000 people; 1290 AD - an earthquake at Chihli, China killed about 100,000 people; 1319 AD - an Armenian earthquake shatters the city of Ani. • By Dear Anonymous As far as policymakers go, it is an ardous task to make them understand the dangers of exploiting all our...By Scientists suggest local-level mapping as India upgrades its seismic map. • The earthquake that hit Sikkim on September 18, killing some 150 people and devastating the Himalayan state, was unexpected. • 1976 AD - an earthquake and tidal wave hit Mindanao, Philippines; 1976 AD - an earthquake hit Guatemala; 1978 AD - an earthquake destroyed Tabas a city in eastern Iran; 1985 AD - a magnitude 8.1 earthquake devastated part of Mexico City and three coastal states1; 985 AD - Nevada del Ruiz errupted, 85 mi northwest of Bogot. • In North America, the San Francisco earthquake of 1906 caused extensive damage and claimed about 700 lives. • **Almost four weeks after the earthquake, Gujarat is still coming to terms with what was and what is.**

We note that the *Div-txtEST* method selects more relevant excerpts. The *Cov-txtQM* method maximizes the coverage of the textual terms in the query-text model which leads to selection of irrelevant excerpts that enlist past earthquakes. These are however not selected by *Div-txtEST* due to their large divergence with query-time and -space models. Thus, due to combined diversification across time, geolocations, and entities, *Div-txtEST* generates a better quality digest with more relevant excerpts.

The *Div-txtEST* method gets the worst loss in w-Rouge against *Cov-txtQM*, for query:

Example 4.2: November 22, 1995: Rosemary West is sentenced to life for killing 10 women and girls, including her daughter and stepdaughter, after the jury returns a guilty verdict at Winchester Crown Court. The trial judge recommends that she should never be released from prison, making her only the second woman in British legal history to be subjected to a whole life tariff (the other is Myra Hindley).

Let us compare the digest generated by both the methods:

Event Digest by the *Div-txtEST* method: • An Orange County jury found Miller guilty in November, and on Nov. 27 see voted 11-1 to recommend the death sentence. • A Pasco County jury found Partin guilty of 1st-degree murder in March. • In July, a jury found Ballard, 67, white, guilty of the 1st-degree murder of Traub. • The same jury found Abdool guilty of first-degree murder last week for the Feb. 25, 2006 death of 17-year-old Amelia Sookdeo. • Last month, on March 5 see, a jury voted 7-5 to recommend a death sentence for Smith. • The sentence followed an 11-1 jury recommendation that Wade receive the death penalty. • The judge may override the jury decision. • Albright Gregory Murphy, 45, was convicted of the 1st-degree premeditated killing in a March jury trial. • The points in Welchs appeal, he challenged his sentence on the basis of jury selection. • The jury voted 7 to 5 to recommend the killer receive a death sentence. • The jury voted 10 to 2. • **She was found guilty of murder and sentenced to life imprisonment.** • The jury recommended by a 9-3 vote that Ballard be executed. • Judges typically follow jury recommendations, so Teppers ruling was somewhat surprising. • On Sept. 26 the same jury that found him guilty, unanimously recommended the death penalty. • After a bench trial, JudgeBurge found Kovarbasich guilty of voluntary manslaughter on April 29, 2010. • A life sentence was recommended by Florida prosecutors. • Watson vacated the sentence in 2005. • The sentence is determined by a jury.

Event Digest by the *Cov-txtQM* method: • An Orange County jury found Miller guilty in November, and on Nov. 27 see voted 11-1 to recommend the death sentence. • It is a very difficult case, but it is the verdict of the jury, he said. • The sentence is determined by a jury. • Albright Gregory Murphy, 45, was convicted of the 1st-degree premeditated killing in a March jury trial. • As he had during the verdict, Riley showed little emotion as his sentence was imposed. • On Sept. 26 the same jury that found him guilty, unanimously recommended the death penalty. • But next week vital and previously withheld testimony from one of her children could overturn the sentence By Tracy McVeigh • **A battered wife serving life imprisonment for killing her husband may soon be freed following evidence from her traumatize daughter which was held back from the original jury.** • **The jury was unanimous West was guilty of ten murders, and the judge, Mr Justice Mantell, sentenced her to life imprisonment with a recommendation that she should serve at least 25 years.** • The jury voted 10 to 2. • Miscarriage of justice and womens groups have been campaigning on Donna's behalf since the verdict two years ago. • After a three-day trial and only two and a half hours of deliberation, a jury of five men and one woman convicted Wade of second-degree murder. • Circuit Judge Lynn Tepper sentenced Smith, 30, to life in prison for killing Robert Crawford in 1999, breaking with a split jurys recommendation that Smith be sentenced to death.

Both methods select few relevant excerpts into the digest. This is due to low coverage of the U.K.-based event in the U.S. news dataset. However, the text-only method selects excerpts based on term matching that leads to better Rouge scores. *Div-txtEST* method suffers due to the sparsity of annotations in the event dimensions.

4.4 Design of Test Collection for Coherence Evaluation

Automatic text summarization has been traditionally considered an effective tool to help users cope with large textual data [181]. The extractive text summarization task is often cast into a task of selecting sentences from a given set of documents and presenting them in a meaningful order [23, 83, 116, 221]. In this realm, though a lot of focus has been given to the problem of selecting informative sentences to improve the content-quality of a summary [169], relatively less attention [22, 28, 89, 109] has been put into improving its structure in terms of sentence ordering for better readability.

Traditionally, evaluation of automatically generated extractive text summaries has been considered a difficult task. This is primarily due to the absence of an "ideal" summary that can be used for comparison. On the one hand, there exist measures to estimate the content quality of text summaries, like Rouge [117] and Pyramid [151], that are computed by comparing against multiple human-written *reference* summaries. On the other hand, text-quality measures [169], like readability, are often estimated by obtaining human preference judgments, which proves to be non-scalable.

What is missing is a corpus containing various orderings of fixed-length human-written focused text summaries on a large number of independent topics where the variants are annotated with human preference judgments based only on their readability. With such a corpus made available, further insights can be obtained on regularities, like conventional proximity between sentences that make a summary more readable. Moreover, focused studies are needed to be conducted to analyze the impact of sentence ordering on coherence and readability of text summaries.

We perform an empirical study through a crowdsourcing platform to get deeper insights into what makes a text summary of an event more readable and coherent. Crowdsourcing platforms, like Crowdflower, have made it possible to efficiently gather human judgments for various tasks. Typically, on these platforms any customer can design a job which is then performed by so-called contributors. We use this platform to find answers to the following questions:

- **Q1** What is the impact of summary structure in terms of sentence order on the readability of summaries for past events?
- **Q2** Does changing the proximity between sentences in a coherent human-written summary affect its readability?

Q3 How feasible is it to evaluate the structure of a fixed-length summary for a past event through crowdsourcing?

To answer **Q1**, we design our first experiment where we present the contributors with two summaries that are differently ordered. Their task is to decide the one that is more readable and coherent. From this experiment, we intend to isolate the effect of sentence ordering on the overall quality of a summary. Insights gathered from this experiment can be used to design scalable methods to explicitly evaluate the structure of a summary. To answer **Q2**, we first generate a predetermined ordering for each summary that maximizes the *gap* (distance in terms of sentence positions) between original sentences. Then, in our second experiment, we present this variant alongside a randomly ordered summary and ask the contributors to choose the more coherent variant. Insights from this experiment can be used to infer inter-sentence relations that result in a better summary structure. To answer **Q3**, we design our third experiment to gather additional statistics to evaluate the feasibility of conducting such studies. For this, we analyze the difficulty, interestingness, rate of progress, quality of contributors, and overall contributor satisfaction level of the study. Insights gathered from this experiment can be used to design better studies on crowdsourcing platforms for other text summarization tasks.

Challenges

Addressing our problem includes the following key challenges:

- Preparing the test data containing suitable variants of fixed-length summaries for a set of past events.
- Designing suitable user interfaces with appropriate quality control measures to ensure good judgments and filter out the contributors that try to cheat.
- Finally, cleaning and performing suitable analysis on human assessments to answer **Q1**, **Q2**, and **Q3** that are described above.

4.4.1 Crowdflower Task Design

In general, it is important for any crowdsourcing-based evaluation to carefully design the experiments. For this work, we choose to use the Crowdflower platform for conducting experiments. Thus from here on, we use the terminology that is consistent with this platform. In this work, we focus on summarization of news events, since in the past a lot of importance has been given to summarizing either blogs or news articles [119].

Event Selection

To generate a test set of fixed-length text summaries for our experiments, we begin by first selecting a set of seminal events in the past which have received considerable media

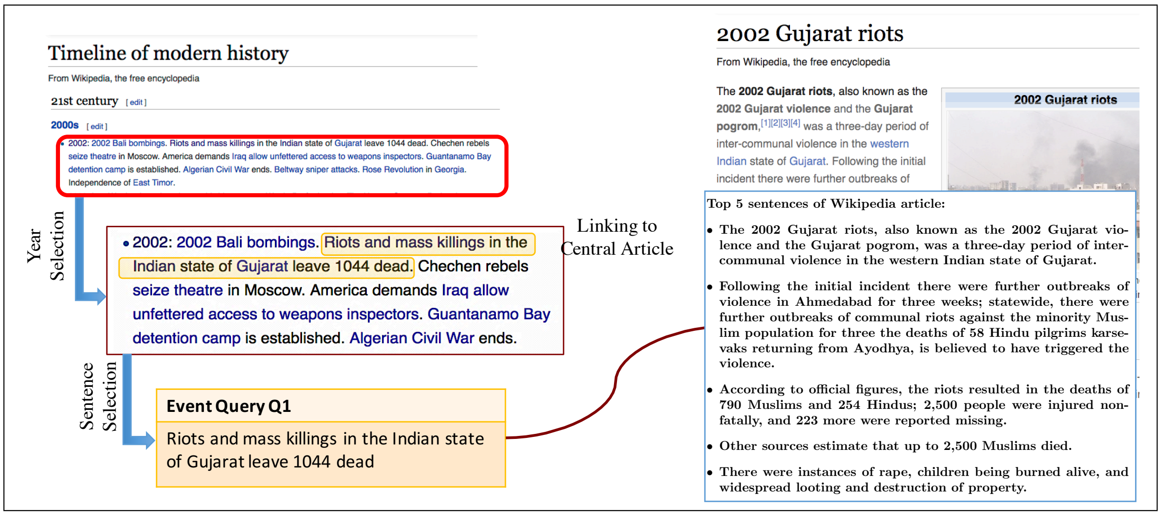


Figure 4.3 Illustration of different test data preparation stages.

coverage. This focuses the study on news events, and also enables to simulate query-focused summarization task by assuming the selected events as user queries. Moreover, it can be assumed that a Crowdflower contributor can better judge the structure of a summary if she has some prior knowledge on the event. Thus, to generate our test-event set, we leverage the Wikipedia page titled, *Timeline of modern history*[1] that lists a selected number of seminal events that occurred since the year 1901. We first split the textual description associated with each year into sentences, each of which describes a single independent event. We then randomly sample a set of events and treat them as our test queries. An illustration of a concrete example is given in Figure 4.3.

Summary Generation

We generate a fixed-length (in terms of sentence counts) summary for each of the events that are sampled from Wikipedia. For an unbiased study, it is crucial that each generated summary is: **1)** human-written, i.e., it should not have any biases from any automatic method; **2)** neutral, i.e., it should not reflect any point of view on the subject; **3)** linguistically simple, i.e., it should be understandable by non-expert contributors. Thus, we leverage Wikipedia articles as a source to generate the test summaries. For each event selected for our study, we first manually identify the Wikipedia page that centrally describes the event in the query. We then select the first 10 sentences from the lead section[2] and treat them as a fixed-length summary of the event. This process is illustrated in Figure 4.3.

[1] https://en.wikipedia.org/wiki/Timeline_of_modern_history
[2] https://en.wikipedia.org/wiki/Wikipedia:Policies_and_guidelines

Interface Design

The interface design is regarded as the most important aspect of a crowdsourcing experiment and may not lead to any results if designed in an ad-hoc manner [12]. There are two main components of a user interface in Crowdflower. First, a questionnaire, and second, a set of instructions given to the contributor for completing the job.

There are several design decisions that have significant impact on the study: **1)** number of summaries to compare in a single unit; **2)** ease of access to the summary text; **3)** the scale of graded judgments; **4)** reasons behind a contributor's decision; and **5)** finally, the difficulty of a judgment. To address these, our interface exhibits the layout illustrated in Figure 4.4. *Event Query* section displays the query as to convey the general topic. The *Summary Container* compares two summaries as *Summary A* and *Summary B*, to gather preference judgments. The text is directly embedded into the interface to minimize the access time to the summaries. Graded judgments are gathered based on the option selected in the *Choice Section* where: *A is Equally Readable to B* denotes *0*; *A is More Readable than B* denotes *1*; and *B is More Readable than A* denotes *2*. We provide a text area where a contributor can specify the reason behind each decision and refer to it as *Reason Box*. Finally, we provide a set of radio buttons that are labeled with different difficulty levels to the contributors for each unit they judge, as illustrated in Figure 4.4.

Instructions given to the contributors to guide them in accomplishing the task, should be clear, concise, and written in simple language without any jargon [12]. With this idea, our instruction set is divided into the following three sections: **1)** *Help* defines various abstract concepts such as coherence, readability, and the different difficulty levels; **2)** *Process* lists step-by-step algorithm; and **3)** *Pro Tips* conveys the general dos and don'ts while making a judgment. The full instruction set is illustrated in Table 4.4.

Quality Control Mechanisms

For this study, we take several measures to ensure good result quality. First, we request contributors that are regarded as *highly performing* and account for 60% of monthly judgments across a variety of Crowdflower jobs. Second, we design a set of test questions. Crowdflower offers a *Quiz Mode* where a small set of units are presented to the contributors before they go into the *Work Mode*. As test questions, we ourselves judge for about 10% of the total units by providing specific reasons for each, which is displayed to the contributors in case they commit a mistake. In both modes, we set the minimum accuracy to be achieved as 70%. Though it is possible to set this higher [123], excusing few mistakes also gives the contributors an opportunity to educate themselves through the reasons specified by us. Third, we set the payment per unit relatively low as to attract contributors that are interested in the task as described by [123]. Finally, in order to detect contributors that passed the quiz mode and later tried to cheat, we introduce *traps* [199] with units containing exactly the same summaries. f

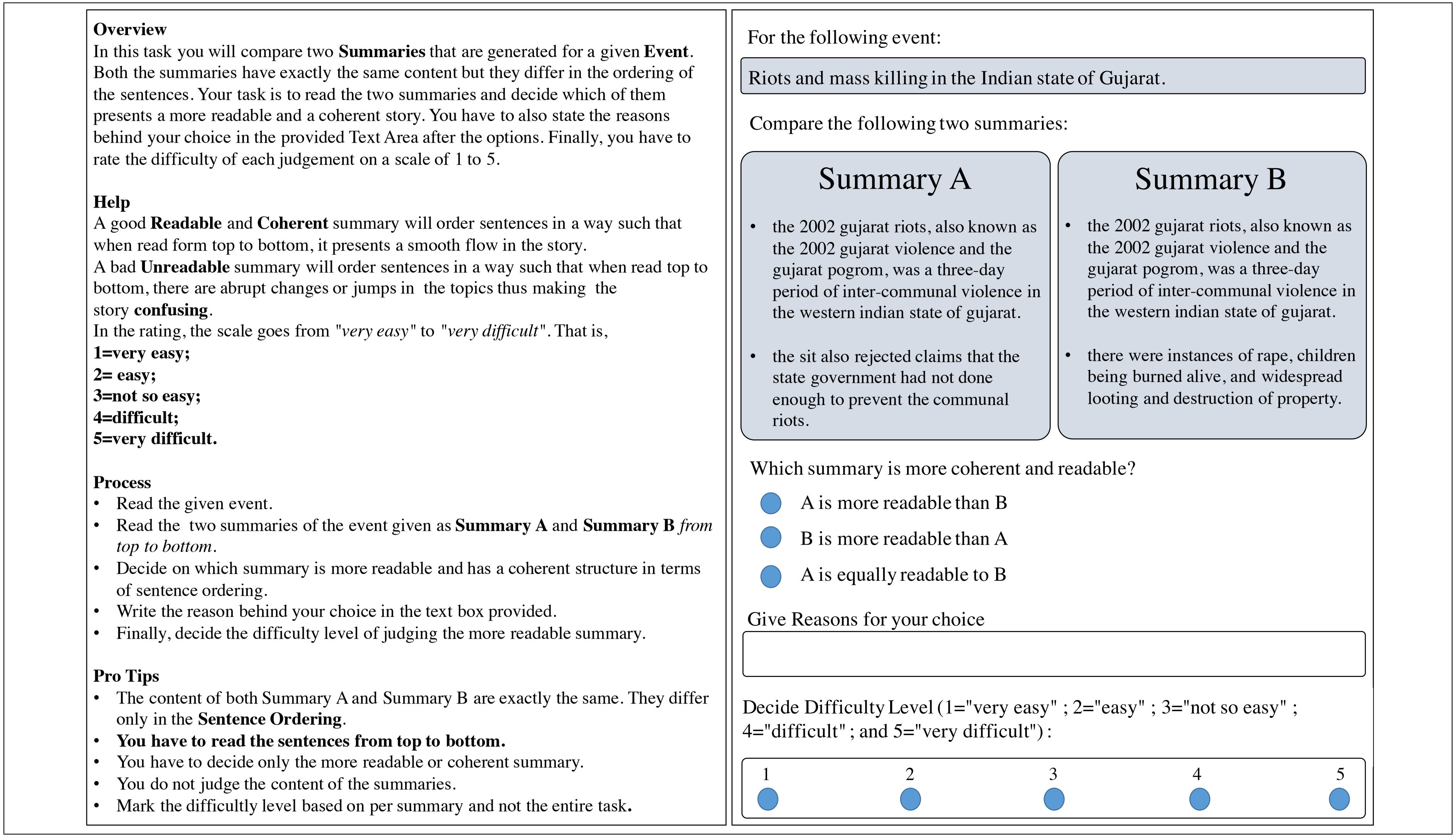


Figure 4.4 Instructions given to the contributors and interface layout.

4.4.2 Experimental Setup

We design the following three experiments to answer **Q1**, **Q2**, and **Q3** in Section 4.4:

- *Experiment 1* analyzes the impact of sentence order on coherence.
- *Experiment 2* analyzes the impact of sentence proximity on coherence.
- *Experiment 3* analyzes the viability of using crowdsourcing for our study.

We make all our experiment data and results publicly available[1].

Test-Event Set

Following the method described in Section 4.4.1, we generate a test set containing 100 randomly sampled events that happened between 1987 and 2007. This time range was chosen considering the coverage of other public document corpora (like the New York Times annotated corpus).

Test-Summary Sets

To generate the test sets, we preprocess the raw summaries extracted from Wikipedia as described in Section 4.4.1. It is important to process the raw text summaries to make the sentences independent [21]. This is to prevent the contributors from basing their decisions on straightforward syntactic cues that originate from across-sentences dependencies. We perform two preprocessing steps: firstly, we resolve all co-references. Secondly, we transform all the sentences to lowercase. Finally, leveraging the preprocessed summaries, we generate the following four test sets:

1. **Original set** $\mathscr{O}$ summary, illustrated in Table 4.9, retains the original ordering.
2. **Reverse set** $\mathscr{R}$ summary, illustrated in Table 4.10, exhibits a reversed ordering.
3. **Shuffled set** $\mathscr{S}$ summary, illustrated in Table 4.11, contains randomly shuffled ordering.
4. **Proximity maximizing set** $\mathscr{P}$ summary, illustrated in Table 4.12, exhibits an ordering that is generated by placing originally consecutive sentences as far as possible.

As an additional step, we keep the position of the first sentence across all the summaries unaltered. This is because we find that spotting the first sentence became a very easy cue for the contributors who desire to cheat by simply spotting the position of this sentence.

[1] http://resources.mpi-inf.mpg.de/d5/txtCoherence

Table 4.9 Original Set $\mathscr{O}$ Summary

[1] the 2002 gujarat riots, also known as the 2002 gujarat violence and the gujarat pogrom, was a three-day period of inter-communal violence in the western indian state of gujarat. **[2]** following the initial incident there were further outbreaks of violence in ahmedabad for three weeks ; statewide, there were further outbreaks of communal riots against the minority muslim population for three months. **[3]** the burning of a train in godhra on 27 february 2002, which caused the deaths of 58 hindu pilgrims karsevaks returning from ayodhya, is believed to have triggered the violence. **[4]** according to official figures, the **communal riots** resulted in the deaths of 790 muslims and 254 hindus ; 2,500 people were injured non-fatally, and 223 more were reported missing. **[5]** other sources estimate that up to 2,500 muslims died. **[6]** there were instances of rape, children being burned alive, and widespread looting and destruction of property. **[7]** the chief minister at that time, **narendra modi**, has been accused of initiating and condoning the violence, as have police and government officials who allegedly directed the rioters and gave lists of muslim-owned properties to them. **[8]** in 2012, **narendra modi** was cleared of complicity in the violence by a special investigation team (sit) appointed by the supreme court of india. **[9]** the sit also rejected claims that the state government had not done enough to prevent the **communal riots**. **[10]** while officially classified as a communalist riot, the events of 2002 have been described as a pogrom by many scholars, with some commentators alleging that the attacks had been planned, were well orchestrated, and that the attack on the **train in godhra on 27 february 2002** was a " staged trigger " for what was actually premeditated violence.

Table 4.10 Reverse Set $\mathscr{R}$ Summary

[1] the 2002 gujarat riots, also known as the 2002 gujarat violence and the gujarat pogrom, was a three-day period of inter-communal violence in the western indian state of gujarat. **[2]** while officially classified as a communalist riot, the events of 2002 have been described as a pogrom by many scholars, with some commentators alleging that the attacks had been planned, were well orchestrated, and that the attack on the train in godhra on 27 february 2002 was a " staged trigger " for what was actually premeditated violence. **[3]** the sit also rejected claims that the state government had not done enough to prevent the communal riots. **[4]** in 2012, narendra modi was cleared of complicity in the violence by a special investigation team (sit) appointed by the supreme court of india. **[5]** the chief minister at that time, narendra modi, has been accused of initiating and condoning the violence, as have police and government officials who allegedly directed the rioters and gave lists of muslim-owned properties to them. **[6]** there were instances of rape, children being burned alive, and widespread looting and destruction of property. **[7]** other sources estimate that up to 2,500 muslims died. **[8]** according to official figures, the communal riots resulted in the deaths of 790 muslims and 254 hindus ; 2,500 people were injured non-fatally, and 223 more were reported missing. **[9]** the burning of a train in godhra on 27 february 2002, which caused the deaths of 58 hindu pilgrims karsevaks returning from ayodhya, is believed to have triggered the violence. **[10]** following the initial incident there were further outbreaks of violence in ahmedabad for three weeks ; statewide, there were further outbreaks of communal riots against the minority muslim population for three months.

Table 4.11 Shuffled Set $\mathscr{S}$ Summary

[**1**] the 2002 gujarat riots, also known as the 2002 gujarat violence and the gujarat pogrom, was a three-day period of inter-communal violence in the western indian state of gujarat. [**2**] there were instances of rape, children being burned alive, and widespread looting and destruction of property. [**3**] following the initial incident there were further outbreaks of violence in ahmedabad for three weeks ; statewide, there were further outbreaks of communal riots against the minority muslim population for three months. [**4**] while officially classified as a communalist riot, the events of 2002 have been described as a pogrom by many scholars, with some commentators alleging that the attacks had been planned, were well orchestrated, and that the attack on the train in godhra on 27 february 2002 was a " staged trigger " for what was actually premeditated violence. [**5**] the sit also rejected claims that the state government had not done enough to prevent the communal riots. [**6**] in 2012, narendra modi was cleared of complicity in the violence by a special investigation team (sit) appointed by the supreme court of india. [**7**] the burning of a train in godhra on 27 february 2002, which caused the deaths of 58 hindu pilgrims karsevaks returning from ayodhya, is believed to have triggered the violence. [**8**] the chief minister at that time, narendra modi, has been accused of initiating and condoning the violence, as have police and government officials who allegedly directed the rioters and gave lists of muslim-owned properties to them. [**9**] according to official figures, the communal riots resulted in the deaths of 790 muslims and 254 hindus ; 2,500 people were injured non-fatally, and 223 more were reported missing. [**10**] other sources estimate that up to 2,500 muslims died.

Table 4.12 Proximity $\mathscr{P}$ Summary

[**1**] the 2002 gujarat riots, also known as the 2002 gujarat violence and the gujarat pogrom, was a three-day period of inter-communal violence in the western indian state of gujarat. [**2**] there were instances of rape, children being burned alive, and widespread looting and destruction of property. [**3**] other sources estimate that up to 2,500 muslims died. [**4**] in 2012, narendra modi was cleared of complicity in the violence by a special investigation team (sit) appointed by the supreme court of india. [**5**] the burning of a train in godhra on 27 february 2002, which caused the deaths of 58 hindu pilgrims karsevaks returning from ayodhya, is believed to have triggered the violence. [**6**] while officially classified as a communalist riot, the events of 2002 have been described as a pogrom by many scholars, with some commentators alleging that the attacks had been planned, were well orchestrated, and that the attack on the train in godhra on 27 february 2002 was a " staged trigger " for what was actually premeditated violence. [**7**] following the initial incident there were further outbreaks of violence in ahmedabad for three weeks ; statewide, there were further outbreaks of communal riots against the minority muslim population for three months. [**8**] the sit also rejected claims that the state government had not done enough to prevent the communal riots. [**9**] according to official figures, the communal riots resulted in the deaths of 790 muslims and 254 hindus ; 2,500 people were injured non-fatally, and 223 more were reported missing. [**10**] the chief minister at that time, narendra modi, has been accused of initiating and condoning the violence, as have police and government officials who allegedly directed the rioters and gave lists of muslim-owned properties to them.

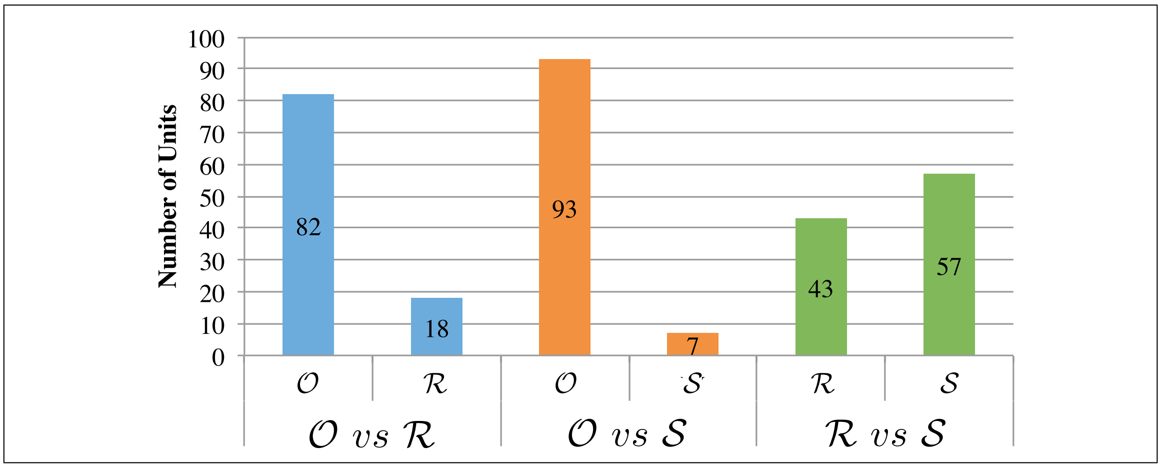


Figure 4.5 Experiment 1 Results

Crowdflower Settings

We created a single job with 700 units out of which 600 compared unique pairs of summaries for 100 queries from each set. The remaining 100 were introduced as *traps* that compared two identical summaries. In addition, our job had randomly selected 53 test units that were judged by us. For each unit we collected three judgments which summed up to a total of 2100. In a single *task* (page), we showed five units (rows) to the contributors, and paid $0.024 per unit per judgment, thus amounting to $0.012 per page. This amount was decided by first running the job on 10% of the data and then monitoring the contributor response. Setting the payment too low would result in less contributor participation and slow down progress. The total cost of the job was $83.47 and it took 77 hours to complete. The language requirement was set to *English* to ensure appropriate contributors. The performance setting was *high speed*, and the minimum accuracy in the test questions was set to 70%.

4.4.3 Experiment 1: Impact of Sentence Order

The main objective of this experiment is to analyze the impact of sentence order on the readability and coherence of fixed-length text summaries on past events. Summaries from the set $\mathscr{O}$, written by Wikipedians, should be most coherent. Reversing the order of the sentences as in set $\mathscr{R}$, should drastically affect the coherence of the text. We generate *300* units that pair-wise compare summaries from the sets $\mathscr{O}$, $\mathscr{R}$, and $\mathscr{S}$. A comparison with the set $\mathscr{S}$ summaries acts as a random test. We refer to the corresponding subsets, each containing 100 units as $\mathscr{O}\ vs\ \mathscr{R}$, $\mathscr{O}\ vs\ \mathscr{S}$, and $\mathscr{R}\ vs\ \mathscr{S}$.

Results

Results of our experiment are illustrated in Figure 4.5. The final preference label of a unit is selected based on majority voting with three judgments. Across all the subsets under comparison, the contributors judge summaries from $\mathscr{O}$ to be the most coherent.

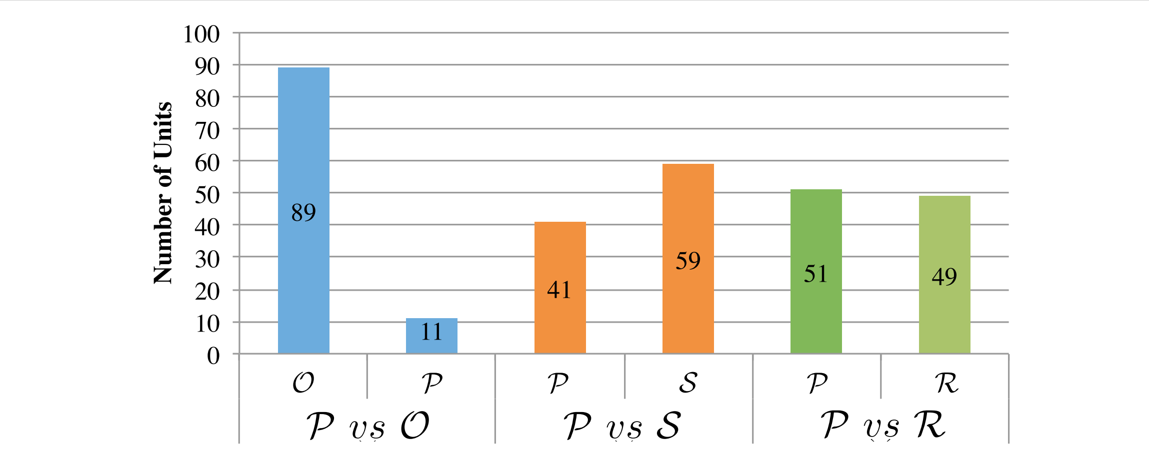


Figure 4.6 Experiment 2 Results

Among the 100 units in $\mathscr{O}$ *vs* $\mathscr{R}$, 82 units set $\mathscr{O}$ summary are more coherent. For the $\mathscr{O}$ *vs* $\mathscr{S}$ subset, in 93 units the original summary was found to be better. We obtain an interesting result for the subset $\mathscr{R}$ *vs* $\mathscr{S}$ where in 57 units, the randomly shuffled set $\mathscr{S}$ summary was found to be more coherent. We found moderate agreement for the subsets $\mathscr{O}$ *vs* $\mathscr{R}$ and $\mathscr{O}$ *vs* $\mathscr{S}$, and fair agreement for $\mathscr{R}$ *vs* $\mathscr{S}$ with Fleiss' kappa scores as 0.42, 0.47, and 0.19 respectively. The longest text obtained from *Reason Box* across the 300 units under consideration consists of 54 words. However, the shortest description for a judgment is found to be just one word. The average length is 5.6 words.

Qualitative Analysis

It is concluded that sentence ordering has significant impact on the coherence quality of fixed-length summaries generated for past events. Reversing the original sentence ordering in fact proves to be the worst among the orders under comparison. Upon closer examination, we find that many of the summaries in the set $\mathscr{S}$ partially preserve the original ordering. This seems to make the summaries more coherent as compared to those in the $\mathscr{R}$ set where the ordering is completely reversed. This is also revealed by analyzing the comments of contributors from the *Reason Box*. It can be assumed that the original summaries follow an inherent structure that best conveys information on a specific event at hand. Reversing the original structure makes the summary more confusing. In Ex. 1 from Table 4.13, a contributor finds that topical jumps make a summary less coherent. However, sometimes a reverse chronological ordering seems to be better as in Ex. 3.

4.4.4 Experiment 2: Impact of Sentence Proximity

The main objective of this experiment is to analyze the impact of changing the proximity of the sentences that originally occur next to each other on the readability of fixed-length text summaries on past events. The set $\mathscr{O}$ summaries are written by Wikipedians and can be assumed to exhibit a sentence grouping such that they present a coherent flow. For example, sentences on a single event aspect are placed next to each other. Thus, altering the ordering where such sentences are separated and the gap between them is maximized, should deteriorate the coherence of the summary. We generate 200 units that pair-wise compare summaries from the sets $\mathscr{P}$, $\mathscr{O}$, and $\mathscr{S}$. We refer to the corresponding subsets with 100 units as $\mathscr{P}$ *vs* $\mathscr{O}$, and $\mathscr{P}$ *vs* $\mathscr{S}$. We generate an additional subset of 100 units as $\mathscr{P}$ *vs* $\mathscr{R}$ subset to isolate the impact of proximity by comparing to the worst ordering in set $\mathscr{R}$ summaries.

Results

Results from our experiments are illustrated in Figure 4.6. We select the final preference label for each unit based on a *majority vote* with three judgments. We find that the contributors judge the set $\mathscr{O}$ summaries to be the most coherent. Amongst the 100 $\mathscr{P}$ *vs* $\mathscr{O}$ subset units, in 89 the $\mathscr{O}$ summary is found to be more coherent. Only in 11 units, the set $\mathscr{P}$ summary is better. Across the $\mathscr{P}$ *vs* $\mathscr{S}$ subset, the randomly shuffled summary from $\mathscr{S}$ is found as more coherent in 59 units. An interesting result is obtained from the $\mathscr{P}$ *vs* $\mathscr{R}$ subset, where the summaries from the $\mathscr{P}$ and $\mathscr{R}$ sets are judged to be coherent in an almost equal number of units, i.e., 51 and 49 respectively. We found moderate agreement across the $\mathscr{P}$ *vs* $\mathscr{R}$, $\mathscr{P}$ *vs* $\mathscr{O}$, and $\mathscr{P}$ *vs* $\mathscr{S}$ subsets with Fleiss' kappa score of 0.27, 0.50, and 0.21 respectively. The final judgment for a single unit is obtained from Crowdflower using their aggregated report generation tool which uses a contributor confidence score thus giving more weight to a judgment given by a trusted contributor. In general, this results in resolution of ties, though in our case we did not find any. Here the confidence score is computed by Crowdflower based on the job history of a contributor. Analyzing the text obtained from the *Reason Box* across all the units, we find the longest to be of 50 words and the shortest is a single word. The average length is found to be 5.6 words.

Qualitative Analysis

We conclude that altering the proximity of the sentences in a summary reduced the coherence of the text. Results obtained from the $\mathscr{P}$ *vs* $\mathscr{R}$ show that contributors were divided between deciding the more coherent summary between these sets. Thus, we can conclude that altering the proximity deteriorates the coherence as much as reversing the order. The reason given by a contributor in Example 5 of Table 4.13 clearly indicates that both the orderings are equally bad. Another instance that specifically highlights the effect of proximity is given in Example 6 in Table 4.13. Consistent with the first experiment, contributors find the more number of randomly shuffled summaries to be more coherent than the proximity altered ones. Upon closer examination, we find that

Ex.	Unit Id	A	B	Reason
1	1052032164	$\mathscr{O}$	$\mathscr{S}$	**Preference**: $\mathscr{O}$. **Reason**: The story makes sense. First, talk about the date, then about the consequences of the attacks and finish about the attack itself. B jumps from one issue to other, the link does not make sense sometimes.
2	1052262816	$\mathscr{O}$	$\mathscr{R}$	**Preference**: $\mathscr{O}$. **Reason**: The summary A describes correctly the order in which the ministers of economy were named and replaced, while summary B is talking about what the third minister of economy made without referring to his predecessors.
3	1051055147	$\mathscr{S}$	$\mathscr{R}$	**Preference**: $\mathscr{R}$. **Reason**: Again, difficult to choose but B starts with the cut of the power and finish with the restored, explaining the story in between.
4	1052081752	$\mathscr{P}$	$\mathscr{O}$	**Preference**: $\mathscr{O}$. **Reason**: the order makes sense. It starts with the flight, the number of passengers and then talk about the flight. In text A, the author jump from the flight to the pilot to come back to the airplane to come back to the pilot.
5	1051055286	$\mathscr{R}$	$\mathscr{P}$	**Preference**: $\mathscr{P}$. **Reason**: The two summaries have all paragraphs in wrong order. Both describes the causes of the accident at the end of the text when it should be at the beginning and both of them speaks about the doubts on the number of casualties before saying the official report of such amount.
6	1051054921	$\mathscr{S}$	$\mathscr{P}$	**Preference**: $\mathscr{S}$. **Reason**: The “tower commission” is the main element around which everything revolves around in these summaries. In the Summary A. any sentence about "tower commission" is near the other about it so that’s why I chose it..

Table 4.13 Hand-picked examples of Reason Box text. Unit Id in the second column links to the released data for further reference of the readers. Summary A and B presented to contributor from the set indicated. The last column specifies the summary set that the contributor finds more coherent with the reason for the judgment.

in some randomly shuffled summaries few sentences retain their proximity by chance. These are marked as more coherent.

4.4.5 Experiment 3: Feasibility of using Crowdflower

The main objective of this experiment is to evaluate the viability of conducting a Crowdflower study for evaluating a summarization task. To analyze the difficulty of the job, we ask the contributors to specify the difficulty level for each unit they judge. In the user interface this is given as a set of five radio buttons representing different difficulty levels. In addition, we analyze the interestingness of the job, rate of progress, and contributor satisfaction based on a survey provided by Crowdflower.

Table 4.14 Experiment 3 Crowdflower job statistics.

(a) Contributor funnel statistics

Quiz Mode (passed)	55
Quiz Mode (failed)	79
Work Mode (passed)	32
Work Mode (failed)	23
Trusted Judgments	2099
Untrusted Judgments	414

(b) Inter Quantile Mean (IQM) temporal statistics

Description	Time
Job Run	77 hours
IQM Trusted Judgment	1m 41s
IQM Untrusted Judgment	1m 11s
IQM Task by Trusted Contributors	8m 27s
IQM Task by Untrusted Contributors	5m 58s

(c) Contributor satisfaction

Job Aspects	out of 5
Overall	2.9
Instruction Clear	2.7
Ease of Job	2.6
Pay	2.6
Test Question Fair	2.8

(d) Reason text distribution

Length in Words	Count
≤ 5	1529
6 to 10	569
11 to 15	112
16 to 20	38
> 20	29

Results

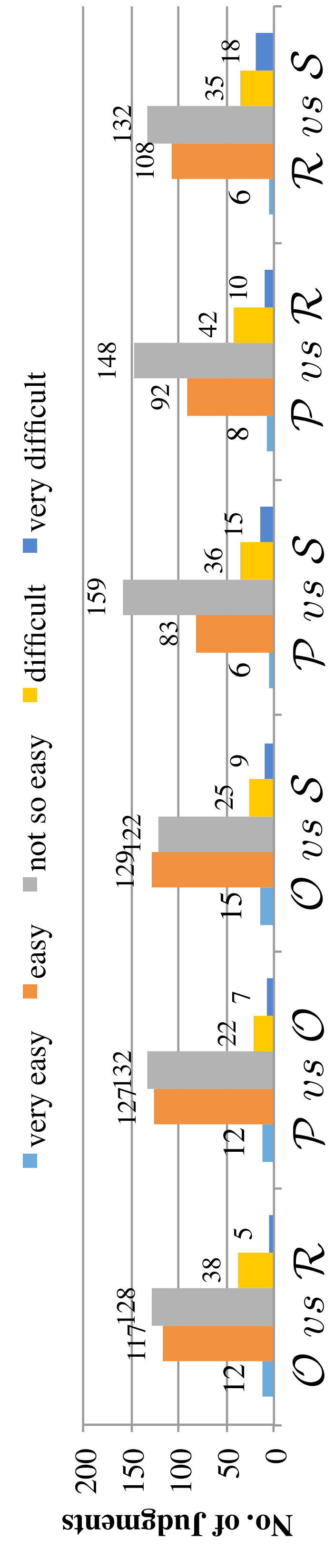
Firstly, we look into the distribution of the difficulty for each unit specified by the contributors. Over all the judgments, we find that 13% of the units are marked as *very easy*, 33% are *easy*, 39% are *not so easy*, 10% are *difficult*, and only 3.9% are marked as *very difficult*. The distribution for each of the unit subsets are illustrated in Figure 4.7a.

The entire job took approximately **77** hours to complete. Figure 4.7b illustrates the rate of the judgments acquired over this time interval as obtained from Crowdflower. As shown in Table 4.14b, a single trusted judgment took about 1 minute and 41 seconds on an average. An untrusted judgment takes comparable time of about 1 minutes and 11 seconds. The average time spent by a trusted contributor is 8 minutes and 27 seconds to judge 5 units, whereas an untrusted contributor spends about 6 minutes.

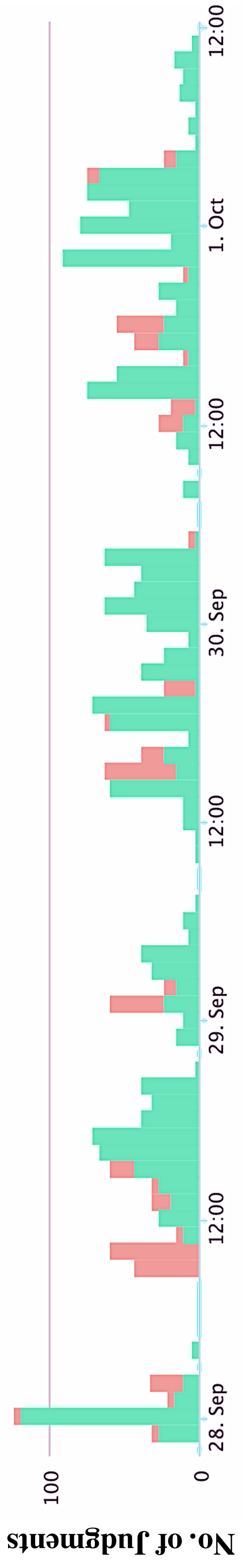
We consider the length of the textual description provided by the contributors as an indicator of the effort put in by the contributors. The more interesting a contributor finds the job, the more effort she will put for the task. This effort includes giving specific and descriptive reasons for her decision in the *Reason Box*. Table 4.13 shows the distributions of the textual description for all the units in terms of their word counts.

To assess the quality control mechanisms that we design for this task, we look into the number of contributors that were evicted from the job during the *Quiz* and *Work* modes of the job. These statistics are presented in Table 4.14a. We find no contributor falls into the *traps* set to judge identical summaries indicating the good quality of judgments.

Finally, we find above average contributor satisfaction for the job setting which was determined through a survey given to the contributors at the end of job.



(a) Distribution of the Difficulty level submited by the contributors across the unit subsets.



(b) Illustration of Judgments per hour (Time zone: GMT+2)

Figure 4.7 Experiment 3 judgments distribution and run time statistics.

Qualitative Analysis

We conclude that a crowdsourcing service like Crowdflower, with correctly designed jobs, can successfully be leveraged to gather preference judgments for comparing summary structures for a text summarization task. Though the task was not found to be very easy by the contributors, they found the difficulty level within an acceptable range. The contributors spend reasonable time for judging per unit which indicates that they show good interest in the job. This is also reflected from the quality of the textual descriptions where they give reason for each decision. However, we find a lot of they textual descriptions were very short (≤ 5 words) indicating that the contributors were reluctant to spend additional time for writing well formatted and grammatically correct text. Instead, they preferred to use as less words as possible. This suggests additional measures that need to be taken for future studies to improve its quality, like setting a minimum text-length requirements. However, we find that the current quality control mechanisms for getting the preference judgments work reasonably well. Out of 133 contributors that entered the quiz mode, 40% were successful. Among those, 60% successfully completed the job and the rest were evicted owing to their drop in accuracy on the test questions.

4.5 Summary & Future Directions

In this chapter we have addressed the problem of connecting Wikipedia events to excerpts from news articles. For this, we cast the problem into an event-focused multi-document summarization task and refer to the output summaries as event digests. In our approach, we proposed a novel divergence-based framework for selecting excerpts from pseudo-relevant news articles such that the global divergence between the digest and the query is minimized. The problem was formulated as an ILP for global inference to maximize the overall relevance of the digest to the input query, while reducing inter-excerpt redundancies all the event dimensions. In our experiments, we compared several methods and found that our divergence-based method that considers all dimensions of an event proves to be the most appropriate for our task.

To be able to understand textual patterns that make a summary more readable, in this chapter we leveraged a crowdsourcing platform and generated a corpus containing human preference-judged variants of fixed-length summaries on past events. We specifically analyzed the impact of altering the ordering and proximity between sentences in a summary that is collaboratively authored and popularly accepted as coherent. For this, we conducted two experiments on the Crowdflower platform and concluded that sentence ordering and proximity infact have significant impact on coherence. As a third experiment, we analyzed the feasibility of the study, and found that it can be successfully conducted with sufficient quality control mechanisms.

Future Directions

This work makes the first attempt to explicitly combine text, time, geolocations, and named entities for event digest generation. We speculate on how plausible result improvements can be achieved and plan to design such methods as future work:

- In this chapter, we leverage probabilistic models that are estimated for the event-query and the excerpts taken from news articles to capture their semantics and thus information content. We speculate on how further improvements over the current approach can be achieved in the previous Chapter in Section 3.6.
- As contrast to our approaches in Chapter 3, for the event digest generation, probabilistic models have to be estimated at excerpt (sentence) level. However, in comparison to whole documents, excerpts come with fewer annotations leading to sparsity issues while estimating the models. For example, while estimating time models, all excerpts that do not come with temporal expressions result in identical time models. Thus, addressing the sparsity issue can lead to estimation of better models which can in-turn improve the digest quality. We address the sparsity issue in the next chapter. As future direction, our sparsity reduction methods can be integrated with event digest generation.
- We find that the test query set extracted from Wikipedia consist of a diverse events with varying spatiotemporal scopes, and number of entity mentions. For further explanation, consider the example events in Table 4.1. The first event is about a riot that spanned across a two month time period and had even longer ramifications. This event also involved a lot of named entities (person, locations, and organizations). However, the second event refers to a specific crime committed by one person on a single day. In our methods, we treat both methods equally. Though our generic methods on average outperform the baselines, further improvements can be achieved by designing methods that leverage individual event characteristics and appropriately combine the discriminative event dimensions. As a future work, plausible adaptive learning approach can be designed based on the work presented by Jacobs et al. [86] that first predicts which event dimensions are discriminative and then combines them to estimate event models.
- So far, we have addressed the event digest problem in an unsupervised setting. By addressing a supervised event-focused summarization with access to fixed-length example summaries, better models may be learned to improve the quality. However, designing learning frameworks and evaluation methods for afore the mentioned diverse Wikipedia events becomes a challenge and opens new directions for future work.

Chapter 5

Estimating Time Models for Short News Excerpts

5.1 Motivation & Problem Statement

An *event* as defined by the Topic Detection and Tracking [9] literature is something that happens in the real world at a specific point in time and place. Time especially is an important dimension as it aids to disambiguate, understand, and retrospect newsworthy events that happened in the past. In applications such as temporal information retrieval [13] and event-focused multi-document summarization [11], time in combination with text is known to improve the result quality. Time also helps in chronological ordering of events to understand its causality, evolution, and ramifications [196]. Considering the importance of time in such applications, methods to automatically resolve accurate temporal scopes of events become essential.

News events are often described using short text with only a few sentences. These descriptions may occur on their own (e.g., in social media) or as part of a larger document such as a news article or a Wikipedia page. In such cases, the event descriptions may have different focus times than that of their host documents. Therefore, works on document dating which focus on estimating the document creation time [96, 220] are orthogonal to determine event focus time. Even the works which estimate the focus time of the document content [87], suffer due to sparsity when considering short event descriptions (e.g., passages or single sentences) as they rely on full content.

In temporal information retrieval, often the goal is to estimate the temporal scope of a given query and the documents in a target corpus to determine their relevance along the time dimension. In this context, most approaches make use of the metadata like creation time or publication dates associated with documents to model their temporal scope. For example, to estimate the relevance of a document for a given query, approaches [115, 161] model the scope of documents by applying an exponential decay

Table 5.1 Example of redundant news article excerpts describing the same event, along with their temporal annotations.

Excerpts (Events)	Time
ε_1: The electronic producers Skrillex and Diplo, who under the name Jack Ü had one of the biggest hits *last year* with "Where Are Ü Now", featuring Justin Bieber, won both best dance recording for that song and best electronic/dance album.	T_1: "last year" = [01-01-2015, 31-12-2015]
ε_2: On Monday, Skrillex and Diplo won the best dance/electronic album for "Skrillex and Diplo Present Jack and best dance recording for Where Are Ü Now"	T_2: "Monday" = [02-02-2016, 03-02-2016]
ε_3: Diplo and Skrillex took home the gold twice at the Grammy Awards night winning Best Dance Recording and Best Dance/Electronic Album.	

function to smooth their publication dates. There are few approaches [15, 26, 93, 162] that leverage temporal expressions embedded in text of the documents to estimate their temporal scopes. In Chapter 3, we present an approach to estimate query-time models which are combined with language models to improve document retrieval effectiveness. However, most of these approaches that rely on temporal expressions are not easy to extend to short sentences or passage with few or missing temporal expressions.

In extractive summarization, time has been leveraged to order sentences in textual summaries. For a summary generated in the context of an event, it is intuitive to present a chronological ordering of the sentences extracted from the news articles. Some approaches [22] leverage the publication dates of the source news articles as a proxy to model the temporal scope of sentences extracted from them. However, this strategy assumes that the documents are highly coherent and focus on a single time period indicated by their publication dates. Other approaches [28, 157] combine chronological with other strategies such as *precedence relations* [156] between sentences positions while ordering multiple sentences from the same document. In our work presented in Chapter 4, we estimate time models for sentences by considering temporal expressions in a set of pseudo-relevant documents. However, in our approach presented in the previous chapter, all sentences from a single document that do not come with any temporal expression end up having similar time models (due to smoothing with a background model). Though this approach proves to be effective than a text-only approach, it can be further improved with more accurate time models estimated for sentences with missing annotations during summarization.

Temporal information extraction [204] focuses on recognizing and normalizing temporal expressions embedded in text into precise time points or intervals. The normalized time intervals are usually represented in a standard format like TIMEX3 [204] of the

TimeML markup language. Commonly, temporal expressions are categorized into four types: *explicit*, *relative*, *implicit*, and *free-text*. Most of the approaches [13, 105, 193] operate only on individual terms or phrases, and associating temporal information to larger textual units (e.g., sentences) is out of scope. It is difficult to ground such larger textual units to precise time intervals.

It can be noted that approaches addressing the above mentioned tasks severely suffer from high sparsity of temporal information when extended to passages or single sentences in contrast to whole documents. We refer to the short textual units simply as *excerpts* in contrast to documents. In this chapter, we propose to estimate an *excerpt-time model* that captures the temporal scope of a given excerpt describing an event. An excerpt-time model can be understood as a probability distribution over *time units* or *chronons* representing days, months, or years. Thus, the temporal scope of excerpts is represented as a probabilistic time model in a time domain instead of precise intervals on a time scale. To further explain this, consider the first example excerpt ε_1 from Table 5.1. The single excerpt gives information on two independent time periods. First, the song "Where Are Ü Now" becoming a hit in the year "2015" and second, *Skrillex* and *Diplo* winning the *Grammy award* which took place on "February 2, 2016". Temporally tagging ε_1 will associate the entire excerpt to the year "2015". However, the desideratum is that the excerpt is associated with a probabilistic model in a time domain where the salient time units pertaining to the year "2015" and the day "February 2, 2016" receive higher probabilities. Also, even though ε_3 does not mention any expression, it still conveys temporal information with a scope that needs to be estimated as its probabilistic time model that additionally captures the uncertainty.

One plausible approach to estimate a more accurate time model for a given excerpt is to leverage the redundancy in the data. For example, given both ε_1 and ε_2 from Table 5.1, one can note that the second part of ε_1 highly overlaps with the event described in ε_2. Since ε_2 comes with an expression "Monday" that is normalized to "February 2, 2016" with a standard temporal tagger, we can use this redundancy to estimate a more accurate time model for ε_1. Similarly, the time model of ε_3 that does not come with any temporal expression can be estimated by combining the time models of ε_1 and ε_2. This is because the information described in ε_3 overlaps with both ε_1 and ε_2 .

Motivated from the above examples, we address the following problem:

> **Problem Statement:** *From a given set of excerpts from news articles describing an event, for each excerpt automatically estimate a probabilistic excerpt-time model capturing its temporal scope.*

As input, we consider: **1)** a set of excerpts along with their source documents; **2)** a set of normalized temporal expressions extracted from the input set of excerpts. As output, our method automatically estimates a probabilistic time model for each excerpt capturing the temporal scope of the information.

Challenges

Addressing the above problem includes the following key challenges:

- We aim to estimate probability distributions that capture the temporal scope of information in excerpts. For this, it is required to design methods that can appropriately leverage temporal expressions that come with the textual descriptions of the excerpts. This becomes our first challenge.
- Secondly, as intuitively described before, sparsity of temporal expression in an excerpt can be reduced by looking into other excerpts in a large corpus that come with additional time information. Thus, designing methods to improve the quality of estimated excerpt-time models for excerpts that lack temporal expressions based on those that give redundant information and come with temporal expressions is a challenge.
- To effectively leverage information redundancy across excerpts, it becomes important to model inter-excerpt relationships. That is, strongly related excerpts indicate higher redundancy. Designing such inter-excerpt relations so as to model their redundancy becomes a challenge.
- Finally, since there are no existing benchmarks to evaluate the quality of estimated probabilistic time models for excerpts capturing their temporal scope, it becomes a challenge to design an evaluation framework to evaluate their quality. Further, the estimated time models need to be evaluated through intrinsic and extrinsic experiments.

Approach Overview

In our approach, we propose a semi-supervised *distribution propagation* framework that leverages redundancy in the data to improve the quality of estimated time models. Our method generates an *event graph* with excerpts as nodes and models various inter-excerpt relations as edges. It then propagates *empirical* excerpt-time models estimated for temporally annotated excerpts, to those that are strongly related but lack annotations. In our experiments, we first generate a test set by randomly sampling 100 Wikipedia events as queries. For each query, making use of a standard text retrieval model, we then obtain top-10 documents with an average of 150 excerpts. From these, each temporally annotated excerpt is considered as a gold standard. The evaluation measures are first computed across the gold standard excerpts for a single query by comparing the estimated time model with our method to the empirical time model from the original expressions. Final scores are reported by averaging over all the test queries. Experiments on the English Gigaword corpus [5] show that our method estimates significantly better time models than several state-of-the-art baselines taken from the literature.

Potential Applications

We motivate that excerpt-time models can be leveraged in various applications. In temporal information extraction, these models can lead to improvements in accuracy of free text temporal expression normalization. For example, ε_3 contains the expression "Grammy Awards Night" whose normalization can be improved by combining the other two excerpts in Table 5.1. For information retrieval, accurate excerpt-time models can be used to estimate better query-time models as described in Chapter 3. A direct application in extractive summarization is to leverage the time models to improve the summary quality as described in Chapter 4, and chronologically ordering excerpts. Time has also been considered as an important information associated with entities and events that exist in knowledge bases like YAGO [82]. In this direction, temporal entity surface form as longer textual phrases or "temponymns" [105] are annotated with normalized time interval indicating their scope. In this context, our time models can be leveraged for better estimation of temporal scopes, thus improving the knowledge base quality. Several other tasks in timeline generation, question answering, and event detection also benefit from the time models (discussed in Section 5.2).

Contributions

We make the following key contributions in this chapter:

- We propose the problem of temporally scoping excerpts by estimating probabilistic time models by leveraging textual redundancy in a large document collection.
- We propose a semi-supervised distribution propagation framework that leverages several inter-excerpt relations, and then estimates excerpt-time models by propagating information from related excerpts that come with temporal expressions.
- Finally, we propose two new measures, namely Model Quality and Generative Power, to evaluate our method on a real-world data set.
- We also address the problem of estimating focus time for a given textual event description and compare with state-of-the-art method [87] on our test collection.

Organization

In Section 5.2, we review related work from the literature. In Section 5.4 we describe the different excerpt models, inter-excerpt relations, and give details of the distribution propagation framework that leverages the models and relations. Intrinsic experiments and their results are described in Section 5.5. In Section 5.6, we address the problem of estimating focus time of a given textual event description. Finally, we conclude the chapter in Section 5.7 and motivate future directions.

5.2 Related Work

We classify the related work mainly into the following four categories: **1)** classified as temporal information extraction, we first look into methods that focus on recognizing and normalizing temporal expressions that typically occur as words or phrases in text. We also look into methods that attempt to estimate temporal scopes of excerpts as a precise time interval. **2)** Second, we look into methods designed to address temporal information retrieval by estimating query and document temporal scopes. **3)** As third category, we look into extractive multi-document summarization where methods use time as a dimension to organize sentences to improve readability of an automatically generated summary. **4)** Finally, since in our approach we use a semi-supervised framework based on label propagation, we look into various methods from the literature in this direction.

Temporal Information Extraction

A temporal expression [204] is a term or a phrase that refers to a precise time point or an interval. In the information extraction community, recently a lot of attention has been given to recognizing and normalizing temporal expressions. TempEval [204] competitions are held with a specific goal to evaluate temporal taggers on different document collections. Some existing systems are HeidelTime [193], SUTime [35], and the Tarsqi toolkit [205]. Though all these systems mainly target explicit and relative temporal expressions, a recent work by Kuzey et al. [105] propose to automatically normalize free-text temporal expressions or *temponyms*.

Some attention has also been given to annotating text segments that do not contain any expressions. Chasin et al. [36] temporally anchor events to a time period between the prior time stamp and subsequent time stamp occurring in text. However, they themselves consider this as a simple algorithm, which is also acknowledged by Gung et al. [72] as a good baseline. Gung et al. use a clustering-based method to annotate textually similar sentences with the same time-stamps. Though this seems to be a reasonable assumption, hard clustering may lead to overly-specific or generic temporal scopes for specific incidents of larger events. More complex approaches [29, 108, 220] make use of additional linguistic features to anchor time. Bramsen et al. [29] break complex sentences into *clauses* and look for temporal markers like *after*, *in next year*, etc., to detect breaks in temporal continuity. Jatowt et al. [87] and Kotsakos et al. [101] attempt to estimate the precise focus time of text documents in a news corpus at the year granularity by leveraging temporal language models. More recently, Laparra et al. [108] use event detection algorithms to propagate time stamps.

In the context of our problem, recognizing and normalizing temporal expressions is not sufficient to estimate excerpt-time models. This is because the majority of excerpts extracted from documents do not come with temporal expressions. Also, the temporal expressions often suffer from uncertainty [26]. Other approaches mentioned above rely on corpus statistics and temporal relations. Firstly, these methods are not comparable as they aim at annotating text segments with a single precise time point or interval as opposed to a distribution like in our problem. Secondly, there exists a difference in their notion for "events" which makes them incomparable. Most commonly in prior works, events are described as topics with few keywords while we consider a sentence from a news article. We use the SUTime toolkit to annotate the temporal expressions in the excerpts that are extracted from news articles.

Temporal Information Retrieval

Information retrieval is another research area where leveraging temporal information associated with documents has received much attention. Among the first works, Li and Croft [115] proposed a method to combine a document's language model with a time dependent prior indicating its temporal scope. This approach was extended by Peetz et al. [161] by investigating different priors based on cognitive models. Berberich et al. [26] was the first to present a two dimensional representation to model the time domain and design a query-likelihood based method to rank documents by combining text and time. Recently, Efron et al. [51] presented a non-parametric kernel density estimation method to incorporate temporal relevance feedback from users to improve retrieval quality. In Chapter 3, we present a method to link Wikipedia events to news articles by treating a given event as a query and estimating query-time models from a set of pseudo-relevant documents. Most of the approaches mentioned above rely only on the meta-data associated with the documents like their creation times. In addition, they focus on document retrieval and may not extend to excerpts due to high sparsity.

In this work, we adopt the two-dimensional representation of time as presented by Berberich et al. [26]. This allows us to estimate two-dimensional time models by capturing the uncertainty in time. In Chapter 3 we proposed a method to estimate a query time model that captures the temporal scope of the query. In this work, we adopt a methodology in the previous to estimate excerpt-time models from a set of temporal expressions. First, we aim to estimate time models for excerpts, instead of entire documents, which may not come with expressions. Second, we propose a distribution propagation framework to estimate the time models more accurately instead of relying on pseudo-relevant documents.

Extractive Summarization

Many of the studies that leveraged time, either used document publication date as a proxy for estimating sentence temporal scope [22, 28, 157] or created handcrafted rules to identify temporal annotations for relations [57, 128, 129]. Later works [72] leveraged rule-based taggers to tag and normalize temporal expressions. However, as an issue pointed out by Mani et al. [128] temporal expressions only constitute about 25% of the temporal information in a typical news corpus and are insufficient for any learning-based method. In Chapter 4, we address the problem of event digest generation from an input set of news articles. We presented a framework to diversify across text, time, geolocations, and entities for the digest generation.

In this chapter, we adopt the concept of excerpt-time models introduced in Chapter 3. In the previous chapter, we attempt to estimate a probabilistic time model for each excerpt that capture the temporal scope of the events described by them. However, the focus of our problem was different. Moreover, there we relied on the temporal expressions that are mentioned in the source documents of the excerpt to estimate the time models. Excerpts taken from the same documents that do not come with any temporal expression get associated with similar indiscriminate time models estimated from the entire corpus as a background model. This stands as a contrast to our problem in this chapter where we focus on estimating an accurate time model for each excerpt by leveraging the redundancy of information in a set of documents.

Semi-Supervised Label Propagation

Label propagation [25, 99, 222, 224, 225] aims at labeling data by propagating class labels from labeled data. For this, the geometry of the data is leveraged by first generating a data graph where the nodes represent data points, and the edge between two nodes represents similarity between them. This paradigm was introduced by Zhu and Ghahramani [224]. In their later work [225], they proposed another formulation of the propagation algorithm that models it as a *Harmonic Gaussian Field*. A similar label propagation algorithm proposed by [222] introduces self-loops so that during each iteration nodes also receive a small contribution from their initial values. This is in contrast to the previous paradigm where the initial values are re-clamped (restored) in each iteration. Recently, Karasuyama et al. [99] proposed an algorithm to efficiently identify the manifold structure of a data graph and at the same time learn the hyper-parameters by a novel feature propagation method.

In this work, among the several approaches, we choose to adopt the algorithm proposed by Zhu and Ghahramani [224] for designing our distribution propagation framework. However, there are a few fundamental differences. Firstly, their algorithm is designed for data points in Euclidean space. However, in the text space, this type of similarity (distance function) is known to not perform so well. Therefore, we incorporate JS-divergence as our distance metric. Secondly, in their original objective function, they introduce a hyper-parameter σ as an importance parameter for each dimension. This is learned independently for each dimension. However, we linearly combine all the inter-excerpt relation weights and have a single scaling parameter that is set to the average of the total weight for simplicity.

Table 5.2 List of notations

Notations	Descriptions
d, C, V	Document, Collection, and Vocabulary
d_{event}, $d_{context}$	Wikipedia Current Events portal, and source document
ε, ε_{text}, ε_{time}	Excerpt, excerpt text part, excerpt time part
$\mathscr{E}_{text}$, $\mathscr{E}_{time}$	Excerpt-text, and excerpt-time model
t, τ	Temporal expression, time unit
$[tb, te]$, tb, te	Time interval, begin time point, and end time point
tb_l, te_l	Lower bounds on begin and end time points
tb_u, te_u	Upper bound on begin and end time points
R, R_{seed}, R_{test}	Input set, seed set with annotated excerpts, and test set with a single excerpt whose time model is to be estimated
min_t	Earliest time unit in R_{seed}
max_t	Latest time unit in R_{seed}
G_{event}, w_{ij}	Event graph, and edge weight between two nodes
δ_{ij}	Total similarity between two excerpts
δ_{text}, δ_{pos}, δ_{cep},δ_{ctx}	Textual, positional, conceptual, and contextual similarity

5.3 Notation & Representations

Before going into details of our approach, we begin by describing the notations and representations used to design our methods. Table 5.2 summarizes our notation.

Excerpt

An excerpt ε is a single unit of an input document d that gives information on an event. For this work, we fix an excerpt to a single sentence extracted from the input set of documents. Each excerpt ε can be described as having two parts: text ε_{text} and time ε_{time} derived from the excerpt description. As a bag of words, ε_{text} is drawn from a fixed vocabulary V derived from the entire collection C. Similarly, ε_{time} is a bag of temporal expressions derived from the textual description of the excerpt.

We treat the input documents as a set of excerpts referred to as R. In our algorithm, we distinguish excerpts that originally come with temporal expressions for those which lack expressions. The set of excerpts extracted from the set of input documents that originally come with temporal expressions is referred to as the seed set R_{seed}.

In our methods, we leverage a background language model of a given excerpt estimated from its source document referred to as $d_{context}$. Additionally, to stress upon more discriminative terms in a given excerpt, we use a second background model generated by coalescing the Wikipedia Current Events Portal[1] into a single document which is referred to as d_{event}.

[1] https://en.wikipedia.org/wiki/Portal:Current_events

Time Domain and Temporal Expression

Time unit or *chronon* τ indicates the time passed (to pass) since (until) a reference date such as the UNIX epoch. Every temporal expression t is treated as an interval $[tb, te]$ in our time domain $T \times T$ with begin time tb and end time te. Further, an interval is described as a quadruple $[tb_l, tb_u, te_l, te_u]$ [26] where tb_l gives the lower bound and tb_u gives the upper bound of begin time tb. Analogously, te_l and te_u give the bounds for the end time te of the interval. However, compared to [26], we adopt a simpler representation of the time to make the time models applicable in our approaches described in Chapter 3 and 4. The simplification of the time model is described in Section 3.5.1.

Event Graph Generated from Excerpts

An event graph $G_{event}(\varepsilon, r)$ is similar to the widely accepted notion of a *sentence network* [29, 38, 88, 89] where each excerpt describes an event. We generate a graph $G_{event}(\varepsilon, r)$ where each excerpt $\varepsilon \in R$ becomes a node and inter-excerpt relationship r is modeled as a weighted edge. Since any excerpt ε_i can be related to any other excerpt ε_j, the event graph G_{event} becomes a complete graph. An illustration of an event graph generated from three excerpts given in Table 5.1 is shown in Figure 5.1. In our experiments described later in Section 5.5, we design a set of methods that make use of this simple graph with different inter-excerpt relations modeled as weighted edges to estimate the temporal scope of excerpts.

5.4 Temporal Scoping Framework

We design a distribution propagation algorithm that is based on label propagation proposed by Zhu et al. [224]. From the input set of pseudo-relevant excerpts R, the subset of excerpts that come with temporal expressions are treated as a seed set R_{seed} as a subset of R. For each excerpt in R_{seed}, we estimate an empirical excerpt-time model from their original expressions. The time models for excerpts with missing expressions are initialized to a uniform distribution. An event graph is generated by considering all excerpts in R as nodes and models their relationships as weighted edges. Finally as an iterative process, the empirical time models from excerpts with expressions are propagated to those that are strongly related but lack temporal expressions.

A time model for an excerpt can be understood as a probability distribution in our time domain that captures the true temporal scope of the excerpt. We leverage redundancy in data to estimate a time model for excerpts that do not come with explicit information. We make the assumption that strongly related excerpts refer to the same event with similar temporal scope.

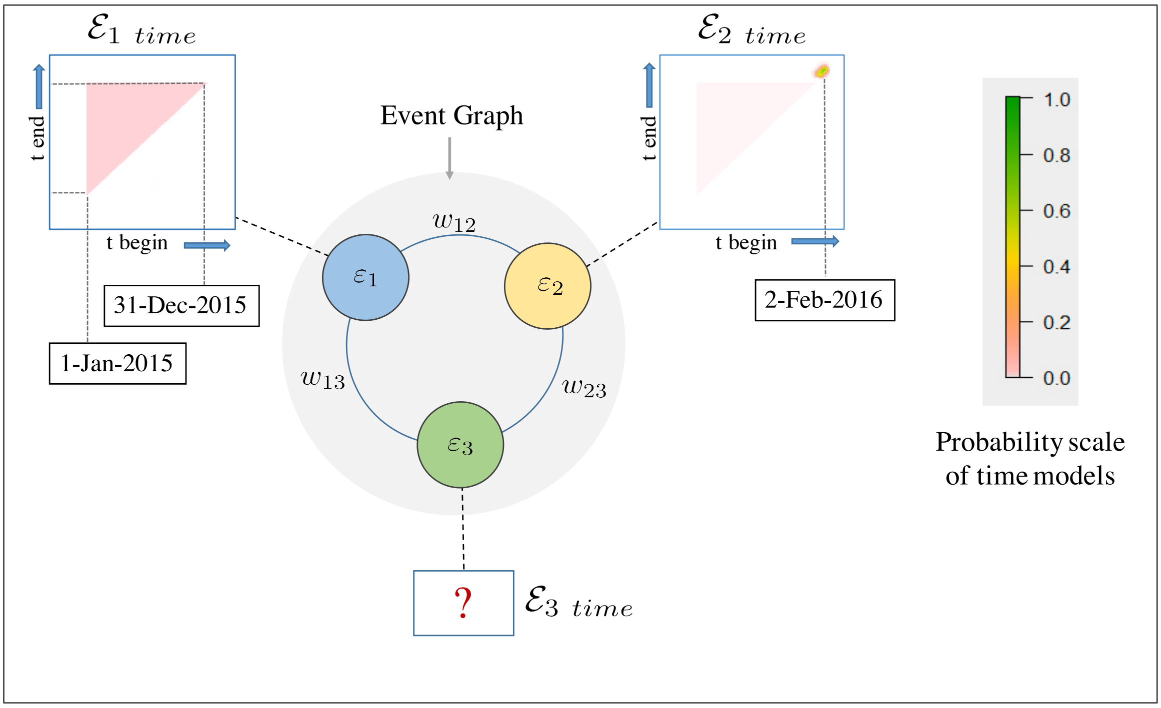


Figure 5.1 An event graph generated from the excerpts where ε_1 and ε_2 are associated with their two-dimensional empirical time models estimated from annotated temporal expressions.

5.4.1 Excerpt Models

Next we describe the different excerpt models estimated for text and time.

Excerpt-Text Model

Excerpt-text model $\mathscr{E}_{text}$ refers to a unigram language model estimated from ε_{text} that captures the event described in the excerpt. It can be observed that any ε_{text} comes with two types of terms: first, those that convey background information, and second, those that describe a specific event. To stress on the terms salient to the described event, we combine the empirical excerpt-text model with a background model estimated from: **1)** the textual content of the source news article, $d_{context}$; and **2)** the textual descriptions of events from the Wikipedia Current Events portal[1] coalesced into a single document, d_{event}. In ε_1 from Table 5.1, the $d_{context}$ background model puts emphasis on the contextual terms, like {*Skrillex, Diplo, Grammy, album*}. The d_{event} background model emphasizes on {*won, award*} that are discriminative for the specific event.

We combine the excerpt-text model with a background model by linear interpolation. The generative probability of a word w from the excerpt-text model $\mathscr{E}_{text}$ is estimated as,

$$P(w|\mathscr{E}_{text}) = (1-\lambda)\cdot P(w\,|\,\varepsilon_{text}) + \lambda\cdot\left[\beta\cdot P(w\,|\,d_{event}) + (1-\beta)\cdot P(w\,|\,d_{context})\right]. \quad (5.1)$$

[1] https://en.wikipedia.org/wiki/Portal:Current_events

A term w is generated from the background model with probability λ and from the original excerpt with probability $1-\lambda$. Since we use a subset of the available terms, we finally re-normalize the excerpt-text model. The new probability $\hat{P}(w|\mathscr{E}_{text})$ is estimated as,

$$\hat{P}(w\,|\,\mathscr{E}_{text}) = \frac{P(w\,|\,\mathscr{E}_{text})}{\sum_{w'\in V} P(w'\,|\,\mathscr{E}_{text})}\,. \tag{5.2}$$

Excerpt-Time Model

Excerpt-time model $\mathscr{E}_{time}$ can be understood as probability distribution that captures the salient periods for an event described in the excerpt. Further, we assume that a temporal expression $t \in \varepsilon_{time}$ is sampled from the excerpt-time model $\mathscr{E}_{time}$. The generative probability of any time unit τ from the time model $\mathscr{E}_{time}$ is estimated by iterating over all the temporal expressions $t = [tb_l, tb_u, te_l, te_u]$ in ε_{time} as,

$$P(\tau\,|\,\mathscr{E}_{time}) = \sum_{[tb_l,tb_u,te_l,te_u]\in\varepsilon_{time}} \frac{\mathbb{1}(\tau \in [tb_l, tb_u, te_l, te_u])}{|[tb_l, tb_u, te_l, te_u]|} \tag{5.3}$$

where the $\mathbb{1}(\cdot)$ indicator function indicates containment of a time unit t within an interval that is represented as $[tb_l, tb_u, te_l, te_u]$, i.e., does the point t lie within the interval. The denominator computes the area of the temporal expression in $T \times T$. For any given temporal expression, we can compute its area and its intersection with other expressions as described in [26]. Intuitively, the above equation assigns higher probability to time units that overlap with a larger number of specific (smaller area) intervals in ε_{time}. Finally, we re-normalize as per Equation 5.2.

5.4.2 Inter-Excerpt Relations

Edge weights in an event graph denote the relationship between two excerpts. Larger weights between two excerpts indicate closer relation with more informational overlap, and hence point to the fact that excerpts may focus on the same time periods. In our method, the edge weights are computed by enforcing an exponential function [224] over the total similarity δ_{ij} between two excerpts ε_i and ε_j, thus allowing propagation from similar excerpts more freely since. Formally the edge weights are computed as,

$$w_{ij} = exp\left(\frac{\delta_{ij}^2}{\sigma^2}\right) \tag{5.4}$$

where σ is the scaling parameter and is set to the average similarity between the excerpts. By considering the squared similarity scores, we exaggerate small differences between excerpts. We model δ_{ij} as a linear combination of four similarity scores capturing different relationships as,

$$\delta_{ij} = \hat{\delta}_{text\ ij} + \hat{\delta}_{cep\ ij} + \hat{\delta}_{pos\ ij} + \hat{\delta}_{ctx\ ij}\,. \tag{5.5}$$

In the above equation, each of the factors are normalized across the excerpts using min-max normalization. Formally this is computed as,

$$\hat{\delta}_{ij} = \frac{\delta_{ij} - min(\delta_{ij})}{max(\delta_{ij}) - min(\delta_{ij})} \,. \tag{5.6}$$

Finally, as motivated by Zhu et al. [224], we additionally smooth the weight matrix with a uniform transition probability matrix $\mathscr{U}$ where $U_{ij} = 1/|R|$ to compute $\mathscr{W}$ as,

$$\mathscr{W} = \gamma \cdot \mathscr{U} + (1-\gamma) \cdot W \,. \tag{5.7}$$

This step is motivated from the PageRank algorithm [152] that allows random jumps. In our context, the uniform weights come into play when an excerpt missing time does not have strong relations to excerpt with temporal expressions. We next propose the different inter-excerpt relations designed using the excerpt models in Section 5.4.1.

Text similarity

Text similarity $\delta_{text\ ij}$ between two excerpts can be a strong indicator of their information overlap. Leveraging this idea, we state the following hypothesis:

If two excerpts are textually similar, then they most likely discuss the same event and time period.

In order to estimate the text similarity between any two excerpts ε_i and ε_j, we compute Jensen-Shannon divergence (JSD) (described before in Section 2.1.3, Equation 2.9) between their excerpt-text models which is estimated as per Equation 5.1. Formally this is computed as,

$$\delta_{text\ ij} = -JSD(\mathscr{E}_{i\ text} || \mathscr{E}_{j\ text}) \,. \tag{5.8}$$

Positional similarity

Positional similarity $\delta_{pos\ ij}$ can be assumed that the source news articles from which the excerpts have been extracted exhibit a coherent structure. This observation to a certain degree can be generalized to the temporal dimension of the articles. With this assumption, we state the following hypothesis:

If two excerpts occur in the same document, and have higher positional proximity in the document, then they most likely discuss the same event and time period.

Unlike the two prior similarities for calculating the positional similarity, we compute the absolute distance between the sentence-positions of two excerpts and apply an exponential decay function over it as shown by Tao et al. [201]. Formally this is computed as,

$$\delta_{pos\ ij} = \begin{cases} log(a + exp(-Dist(\varepsilon_i, \varepsilon_j))) & if \varepsilon_i, \varepsilon_j \in d \\ 0 & otherwise \end{cases} \tag{5.9}$$

where a is set to 0.3 [201], and $Dist()$ function returns the difference in positions as

$$max(pos(\varepsilon_i), pos(\varepsilon_j) - min(max(pos(\varepsilon_i), pos(\varepsilon_j)) .$$

Conceptual similarity

Conceptual similarity $\delta_{cep\ ij}$ between two excerpts can be computed as the noun-phrase overlap between them. Intuitively, noun phrase extraction can be considered as a soft form of concept or entity recognition. Higher similarity between excerpt-noun-phrase models indicates a stronger relationship. We state the following hypothesis:

If two excerpts contain similar noun phrases, then they most likely discuss the same event and time period.

To estimate $\delta_{cep\ ij}$, we first run it through an open source part-of-speech based noun phrase extractor. Then for an excerpt ε_i, we estimate its excerpt-noun-phrase model $\mathcal{E}_{i\ np}$ according to Equation 5.1 by simply treating each noun phrase as a distinct term. Finally, conceptual similarity between any two excerpts ε_i and ε_j is computed as Jensen-Shannon divergence (JSD) between their excerpt-noun-phrase models. Formally this is computed as,

$$\delta_{cep\ ij} = -JSD(\mathcal{E}_{i\ np} || \mathcal{E}_{j\ np}) . \tag{5.10}$$

Contextual similarity

Contextual similarity $\delta_{ctx\ ij}$ becomes an important indicator of the relationship between two excerpts if their textual description is sparse. It may also indicate if two events are part of a common larger event and should have happened in a similar time period. Considering this idea, we state the following hypothesis:

If two excerpts have higher contextual similarity, then most likely they are a part of the same event and time period.

In order to estimate $\delta_{ctx\ ij}$, we leverage the lead (first) paragraphs of the source news articles of ε_i and ε_j. For each excerpt, we first estimate an excerpt-context model $\mathcal{E}_{ctx}$ from the lead paragraph of their source news articles. This is done similar to

Equation 5.1. With this the contextual similarity can simply by defined as the Jensen-Shannon divergence between the excerpt-context models as,

$$\delta_{ctx\ ij} = -JSD(\mathscr{E}_{i\ ctx} || \mathscr{E}_{j\ ctx}) \,. \tag{5.11}$$

5.4.3 Distribution Propagation

Our distribution propagation algorithm is based on label propagation as proposed by Zhu et al. [224]. Intuitively, we leverage the idea that if two excerpts have a strong inter-excerpt relationships between them, then they may refer to the same event, and hence have a similar temporal scope.

Pseudo-code of our method is illustrated in Algorithm 7. As the first step, we extract a set of excerpts R from a given set of input documents. We then extract the earliest and latest time unit at a fixed time granularity from R. This fixes the total scope of our time domain $T \times T$. In the next step, we estimate the excerpt-time model for each excerpt that comes with temporal expression in the seed set R_{seed}. We add these excerpts to the labeled class Y_l. For the rest of the excerpts, we assume that the generative probability of any time unit τ is uniform, and we add these to the unlabeled class Y_u. Given all excerpts in R, we then construct an event graph G with excerpts as nodes. The weight matrix $\mathscr{W}$ models the relationship between every excerpt- pair and treats them as weighted edges. The algorithm then performs the following two steps until convergence: first, all excerpts propagate their time models. Second, instead of letting the generative probabilities in excerpt-time models in Y_l readjusted, we reinitialize their original distributions. Finally, we retrieve the excerpt-time models of the unlabeled excerpts from $Y^{(\infty)}$. The algorithm is guaranteed to converge as proven by Zhu et al. [224].

5.5 Experimental Evaluation

In this section, we describe details of the conducted experiments. We make all our experimental data publicly available[1].

We note that there is no ready-to-use ground truth for our task. In order to generate a test set of excerpts, we adopt a query-driven methodology where we randomly select a set of Wikipedia events and treat them as a user query. We then retrieve top-K documents relevant to an event and treat them as a set of pseudo-relevant excerpts. This methodology has two effects. First, the search space for our distribution propagation algorithm is reduced. Second, this method can be considered as filtering out noisy temporal expressions. However, it is worthwhile to highlight that our method is not dependent on this filtering step and can be generalized to any set of excerpts. This

[1] http://resources.mpi-inf.mpg.de/d5/excerptTime/

Algorithm 7 Temporal Distribution Propagation

Construct set R from input documents
Initialize the min_t and max_t as earliest and latest day in R, respectively
for excerpt $\varepsilon_i \in R$ **do**
  if $|\varepsilon_{i\ time}| > 0$ **then**
    Estimate $\mathscr{E}_{i\ time}$ (in Section 5.4.1)
    $Y_l \leftarrow Y_l \cup \mathscr{E}_{i\ time}$
  else
    Initialize $P(\tau|\mathscr{E}_{i\ time}) \leftarrow \frac{1}{|max_t - min_t|}$
    $Y_u \leftarrow Y_u \cup \mathscr{E}_{i\ time}$
  end if
end for
Generate an event graph G (in Section 5.3)
Compute the affinity matrix $\mathscr{W}$ (in Section 5.4.2)
Compute diagonal degree matrix $D_{ii} = \sum_j w_{ij}$
Initialize $Y^{(0)} \leftarrow (Y_l, Y_u)$
while not convergence to $Y^{(\infty)}$ **do**
  $Y^{(t+1)} \leftarrow D^{-1} \cdot \mathscr{W} \cdot Y^t$
  $Y_l^{(t+1)} \leftarrow Y_l^{(t)}$
end while
Final $\mathscr{E}_{i\ time}$ for ε_i is then obtained from $Y^{(\infty)}$

method, however enables us to systematically evaluate the effectiveness of our method as compared to other baselines.

5.5.1 Setup

Next, we describe our document collection, queries, ground truth, and implementation details.

Document Collection

To test our methods, an appropriate collection is one which contains documents describing events. Additionally, it is desired that the collection should contain sufficient redundancy which can be then leveraged by our method. Thus, we perform experiments on the English Gigaword corpus with about 9 million news articles published between 1994 and 2010. We process the queries in our test set with a standard query-likelihood document retrieval model. The top-10 retrieved documents with an average of 150 excerpts are considered pseudo-relevant and input into our methods.

Test Queries

Test queries are generated from the *Timeline of Modern History*[1] in Wikipedia that enlists the most prominent news events in the 20th and the 21st centuries. We randomly sample 100 events that took place between 1987 and 2007, and treat them as test queries. Each query comes with a short textual description and a time interval indicating the occurrence of the event. For the experiments, we ignore the time interval and only leverage the textual expression as a keyword query to retrieve the pseudo-relevant documents. We include a full list of the test queries for experiments in Appendix A.4.

Ground Truth

We rely on the empirical temporal expressions that originally occur in the excerpts to evaluate our methods. In our evaluation methodology, we perform leave-one-out cross validation by randomly selecting excerpts that come with temporal expressions into ground truth. The evaluation measure is computed based on how well the estimated time model for an excerpt describes the empirical model that is estimated from the original expressions in textual description of the excerpt. For a use-case experiment, we make use of the time interval associated with each query as ground truth. We describe the use-case experiment in Section 5.6.

Implementation

All our methods are implemented in Java. For the temporal annotation and noun-phrase extraction, we use Stanford SUTime toolkit [35]. We additionally use the Weka toolkit[2] to implement the cross-validation framework for the evaluation.

5.5.2 Goals, Measures, and Methodology

As motivated before, the excerpt-time models estimated by our distribution propagation method can simply be understood as a two-dimensional probability distributions that capture the temporal scope of corresponding excerpts. In this distribution, the time units with high probability represent the temporal focus of the event in the excerpt. To evaluate quality of the time models estimated for the excerpts, we propose the following steps: **1)** From the set of excerpts retrieved for each query, we first distinguish those that originally come with an expression. As described in Section 5.4, these are then considered as an initial seed set for the distribution propagation. **2)** We then adopt leave-one-out cross-validation style methodology by randomly selecting one of these excerpts and ignoring temporal expressions occurring in them. This is considered as set R_{test}. For each excerpt, we run our algorithm as described in Section 5.4 by considering

[1] https://en.wikipedia.org/wiki/Timeline_of_modern_history
[2] http://www.cs.waikato.ac.nz/ml/weka/

rest of the excerpts as seed set R_{seed}. **3)** We truncate and re-normalize the estimated model based on the scope of empirical model to make them comparable and then compute the evaluation measures. Final scores are reported by first averaging across the folds for a single query and then taking the mean across all the test queries.

In addition to evaluating the excerpt-time model quality, we also investigate the effects of varying the time granularity for the modeling. Thus, we perform experiments at three fixed time granularities: *day*, *month*, and *year*. For this, we fix the granularity of time and represent the original expression at that granularity. The finest granularity in our experiments is the day level. For experiments at the month granularity, we additionally perform a preprocessing step where we relax an expression originally at day granularity to its month. For example, *"February 2, 2016"* is relaxed to *"February 2016"*. Similarly, for the experiments at year granularity, we relax the original expression occurring at the day and the month granularity into its year. For example, *"February 2, 2016"* is relaxed to *"2016"*.

We define the following evaluation goals for our experiments:

1. We aim to evaluate the quality of estimated excerpt-time models for the excerpts as nodes in the event graph.

2. We aim to compare the inter-excerpt relations so as to evaluate their effectiveness while estimating excerpts from news articles from our target document collection.

3. We aim to evaluate the quality of excerpt-time models at different time-granularities.

We next define the measures to compare our methods and achieve our evaluation goals.

Model Quality

We compute how close an excerpt-time model $\mathscr{E}_{time}$ estimated after propagation is to the empirical model ε_{time}. We define model quality MQ as the KL-divergence between the two. Formally,

$$MQ = \frac{1}{|R_{test}|} \sum_{\varepsilon \in R_{test}} -KLD(\varepsilon_{time} || \mathscr{E}_{time}) \; . \tag{5.12}$$

Intuitively, a method estimating $\mathscr{E}_{time}$ more accurately should have a smaller divergence to ε_{time} and hence higher MQ. In our experiments where we consider a single excerpt at a time, $|R_{test}| = 1$. As a relative measure, the methods can be ranked according to the model quality.

Generative Power

With the assumption that the empirical time part of an excerpt is generated from the excerpt-time model, we define the generative power GP as the likelihood of generating the original time intervals in the time part ε_{time} from the estimated time model $\mathscr{E}_{time}$

after propagation. To compute GP, we adopt the approach to estimate the generative probability of time intervals proposed by Berberich et al. [26] and compute the generative probability as,

$$GP = \sum_{\varepsilon \in R_{test}} \left(\frac{1}{|\varepsilon_{time}|} \sum_{t \in \varepsilon_{time}} P(t|\mathscr{E}_{time}) \right). \tag{5.13}$$

Intuitively, the better the estimation of $\mathscr{E}_{time}$, the higher the likelihood of generating the empirical time part ε_{time}.

Precision

We test how well our estimated model can predict the empirical time-part of an excerpt. For this, we note that temporal expressions associated to excerpts in the R_{test} come at a certain granularity, i.e., year, month, or day. For each excerpt, we generate a ranked list $\mathscr{R}_t$ of temporal intervals at a fixed granularity, where the ranking is based on their generative probabilities from the estimated excerpt-time models. Finally, we define a notion of relevance for the time intervals and compute Precision indicating the quality of the estimated model.

A generated interval is considered relevant if it overlaps with the empirical time-part of the excerpt. Formally, we define the binary relevance $Rel()$ between an empirical time interval t and an estimated time interval $\hat{t}$ as,

$$Rel(t, \hat{t}) = \begin{cases} 1 & if \quad min(t_e, \hat{t_e}) - max(t_b, \hat{t_b}) > 0, \text{ i.e., if they overlap} \\ 0 & otherwise \end{cases} \tag{5.14}$$

where t_b and t_e are begin and end time points of t. This formulation is also used in [87]. Using this notion for relevance we formally define precision M for each excerpt as,

$$P = \frac{1}{|\mathscr{R}_t|} \sum_{t \in \varepsilon_{time}} \sum_{\hat{t} \in \mathscr{R}_t} Rel(t, \hat{t}) \tag{5.15}$$

where $|\mathscr{R}_t|$ is the total number of time units in $\mathscr{R}_t$. Using the same notion for relevance, we additionally compute measures like Recall, Mean Average Precision *MAP*, and Normalized Discounted Cumulative Gain *NDCG* scores which are standard in IR.

Methods under Comparison

We categorize our methods into three major types: Local Time-based (*LT*), Nearest-Neighbor-based (*NN*), and Distribution Propagation-based (*DP*) methods. The *LT* methods take into account only the information in the source documents to estimate excerpt-time models. We consider two methods that use the publication date (*pd*) and the surrounding temporal expressions *(S)* as described later. On the other hand, both NN and DP methods estimate excerpt-time models by leveraging the event graph generated as described in Section 5.4.1. We define several variants of the methods that consider different combinations of the inter-excerpt relations, i.e., text (T), position (P), conceptual (N), and contextual (X) similarities as indicated by their suffixes. Finally, we compare a state-of-the-art method proposed by Jatowt et al. [87] that uses term-time associations to predict focus time period of an excerpt. Next, we describe the different methods that we compare in detail.

- ***LT-pd*** method assumes that each excerpt taken from a news article describes an event from the time period indicated by its publication date as motivated by several works [22, 28, 115, 157, 161].

- ***LT-pdS*** method in addition to the publication date takes into consideration the surrounding temporal expressions of a given excerpt in the source document. Intuitively, temporal expressions denote temporal context change. Thus, the closest mentioned temporal expression prior to a given excerpt, and the closest subsequent mentioned expression can indicate its temporal scope. This method however assumes that all excerpts in a document are temporally coherent. This method is motivated from [36, 72].

- ***NN-T*** method estimates excerpt-time model from time models of textually similar excerpts. This can be understood as a two-step method where in the first step, for a given excerpt, we compute its textual similarity (as described in Section 5.4.2) to the other excerpts. Then in the second step, we estimate the excerpt-time model by interpolating empirical models of excerpts weighted by their textual similarity. Similar methods have been used in the past for estimating query-time models. The two-step method is simply implemented by iterating once in our distribution propagation algorithm (Algorithm 7) on an event graph that models on the textual relationship between the excerpts. This is motivated from the pseudo-relevance based method presented in [143, 147].

- ***NN-TPNX*** method is analogous to the simpler *NN-T* however, as an extended method, it leverages all the relationship described in Section 5.4.2 between the excerpts.

- ***DP-TPNX*** method implements the distribution algorithm described in Section 7 by leveraging all the relations described in Section 5.4.2 to generate the event graph. We additionally compare the ***DP-T***, ***DP-P***, ***DP-N***, ***DP-X***, ***DP-TPN***, ***DP-PNX***, and ***DP-TNX*** variants of this method that leverage different combinations of inter-excerpt relations as indicated by their suffixes.

- ***ADJ*** refers to the method proposed by Jatowt et al. [87] for estimating the temporal focus of documents. As a baseline, we extend their approach to estimate focus time for excerpts from news articles. Briefly, this method first generates an undirected weighted graph $G(V, E)$ where V denotes the unique words, and E denotes word relationships. This graph is then used to estimate word-time association scores. Using these scores, a temporal weight is estimated for each word which is then used to estimate the focus time of each excerpt. From several variants of the method proposed by Jatowt et al., we select their best performing method on Web data that uses the $A_{dir}(w, t)$, ω_w^{temK}, and $S_{TF}(\varepsilon, t)$. We refer to their prior work [87] for a full description of the method. We first use their method to predict the focus time interval of an excerpt in the year, month, and day time granularities. We then estimate an excerpt-time model with the predicted time interval as described in Section 5.4.1.

- ***Rand*** method randomly sets the edge weights in an event graph. This method highlights the quality of the temporal expressions in the seed set.

Parameters

We set the following parameters for our methods. To combine a background model with our query model in Equation 5.1, we choose standard settings from the literature [216] and set $\lambda = 0.85$. Further in this equation, we set $\beta = 0.5$ thus giving equal importance to the background models. In Equation 5.4, σ is set to the average inter-excerpt relation weight estimated for each query. For smoothing the estimated weight matrix in Equation 5.7, we set $\gamma = 0.0005$ as motivated by [224]. In all the *DP* methods, we set the number of iterations to 15.

5.5.3 Results & Analysis

We compare the different methods and report the results in terms of the various measures introduced earlier.

Overall Results

Results from all our methods are shown in Table 5.3. We find the distribution propagation method *DP-TPNX* proves to be the most effective method for estimating time models for excerpts across all measures. We test the significance of the results by comparing all pairs of methods with two-tail paired t test. We find the MQ result difference between all pairs of methods to be significant at $\alpha < 0.001$ except the difference between *NN-T* and *NN-TPNX* is found to be insignificant.

Firstly, we find that the *ADJ* proves to be the weakest method to estimate temporal models for excerpts across all granularities. The simplest *LT-pd* method is the second weakest at the day and year granularities across all metrics. However, it performs significantly better than the *ADJ* method at all three granularities. Leveraging the

Table 5.3 Comparison of results from all methods over 100 queries.

	Method	MQ	GP	P@5	MAP	Recall	NDCG@5
DAY	**DP-TPNX**	**-2.227**	**0.754**	**0.828**	**0.836**	**0.902**	**0.847**
	NN-TPNX	-14.082	0.395	0.408	0.182	0.404	0.406
	NN-T	-14.120	0.389	0.407	0.182	0.404	0.404
	Rand	-14.212	0.388	0.407	0.182	0.404	0.405
	LT-pdS	-14.614	0.471	0.472	0.286	0.332	0.467
	LT-pd	-17.807	0.368	0.368	0.185	0.192	0.362
	ADJ	-23.120	0.033	0.033	0.016	0.017	0.032
MONTH	**DP-TPNX**	**-2.278**	**0.477**	**0.499**	**0.503**	**0.518**	**0.504**
	NN-TPNX	-10.326	0.328	0.337	0.330	0.309	0.333
	NN-T	-10.327	0.327	0.336	0.330	0.309	0.333
	Rand	-10.355	0.323	0.334	0.327	0.309	0.331
	LT-pdS	-10.438	0.332	0.334	0.333	0.290	0.330
	LT-pd	-13.104	0.280	0.280	0.281	0.219	0.277
	ADJ	-27.889	0.018	0.018	0.018	0.017	0.017
YEAR	**DP-TPNX**	**-1.497**	**0.751**	**0.722**	**0.741**	**0.747**	**0.731**
	NN-TPNX	-2.384	0.714	0.691	0.691	0.696	0.695
	NN-T	-2.415	0.707	0.686	0.690	0.696	0.691
	Rand	-2.449	0.698	0.675	0.689	0.696	0.684
	LT-pdS	-3.064	0.717	0.708	0.673	0.675	0.702
	LT-pd	-4.266	0.655	0.654	0.602	0.602	0.642
	ADJ	-24.462	0.159	0.160	0.119	0.122	0.138

closest prior and subsequent temporal expressions around an excerpt along with the publication date for estimation of its time model as the *LT-pdS* method shows significant improvement over the simpler *LT-pd* method. We observe a gain of 10% over *LT-pdS* in GP at day granularity. We also observe a similar improvement across other measures. The nearest-neighbor methods *NN-T* and *NN-TPNX* perform significantly better than the *LT* methods at day and year granularities. The more complex *NN-TPNX* shows marginal improvement over the *NN-T* method at all granularities in terms of GP. Finally, the distribution propagation method *DP-TPNX* outperforms the other methods across all granularities in terms of all the measures. We find more than 35% improvement from the *NN* methods in the day granularity in terms of GP. The method also shows similar significant improvements at the month, and year granularities over the other methods. The *Rand* method gives the worst results at the month granularity as compared to the year or day. This is indicative of sparsity of annotations in the input set at this granularity and may not be generalizable.

Ablation Test

To get an insight into the importance of different inter-excerpt relations, we compare several variants of our distribution propagation method as shown in Table 5.4. We find that the most effective inter-excerpt relation proves to be text similarity as shown by the *DP-T* method. The next best relation is found to be conceptual similarity as indicated by the *DP-N* method. As a difference to the simple full-text, the conceptual similarity is estimated by regarding only the noun phrases in the excerpt descriptions. Intuitively, the *DP-N* method suffers due to increase in sparsity as compared to considering all terms in case of the *DP-T* method. Position similarity seems to negatively affect the quality. Intuitively, this relation makes strong coherence assumption on the structure of news articles which may not hold for all news articles. Moreover, this relation can only be estimated between excerpts from a single document. *DP-X* proves to be the worst of the inter-excerpt relations in terms of MQ at all granularities. However, a combination of contextual, conceptual, and text similarity to model the inter-excerpt relations as leveraged by *DP-TNX* is observed to be more effective for estimating time models. Finally, we observe that the *DP-TPNX* is fifth best method.

Varying Graph Size

Intuitively, a larger number of strongly related excerpts with temporal expressions for a given excerpt in the event graph should lead to better estimation of its time model. We test this by increasing the event graph size by considering more input pseudo-relevant documents. Figure 5.2 illustrates the effects of increasing the event graph size on the average quality of excerpt-time model estimated by the *DP-TPNX* and *NN-T* methods in terms of Generative Power, Mean Average Precision, and Precision at 5 and 10 cut-off

Table 5.4 Ablation test in Model Quality and Generative Power.

Methods	Day		Month		Year	
	MQ	GP	MQ	GP	MQ	GP
DP-T	**-1.829**	**0.794**	**-2.082**	**0.488**	-1.463	0.749
DP-TNX	-2.019	0.775	-2.128	0.483	**-1.450**	**0.751**
DP-N	-2.058	0.780	-2.142	0.485	-1.463	0.749
DP-TPN	-2.147	0.759	-2.194	0.478	-1.452	0.751
DP-TPNX	-2.227	0.754	-2.278	0.477	-1.452	0.751
DP-P	-2.413	0.742	-2.292	0.473	-1.458	0.749
DP-PNX	-2.428	0.744	-2.267	0.474	-1.459	0.749
DP-X	-2.578	0.766	-2.393	0.483	-1.511	0.744

levels. We find that both methods show an improvement in the quality as more data is considered for the excerpt-time model estimation. However, our method consistently outperforms the *NN-T* method even as the event graph size grows. Moreover, we find that both the methods show large improvements by increasing the graph size up to 375 nodes. Further increase in the graph size shows smaller improvements in the average result quality. This is because considering lower ranked documents, we add more irrelevant excerpts to the event graph for a given query. They often do not pose strong relation to the other excerpts, and hence have a negative impact on the overall result quality.

Gain/Loss Analysis

To get insights into the individual queries for which our *DP-TPNX* shows the highest gain and worst loss in terms of P@5 against the best baseline at all three temporal granularities, i.e., day, month, and year.

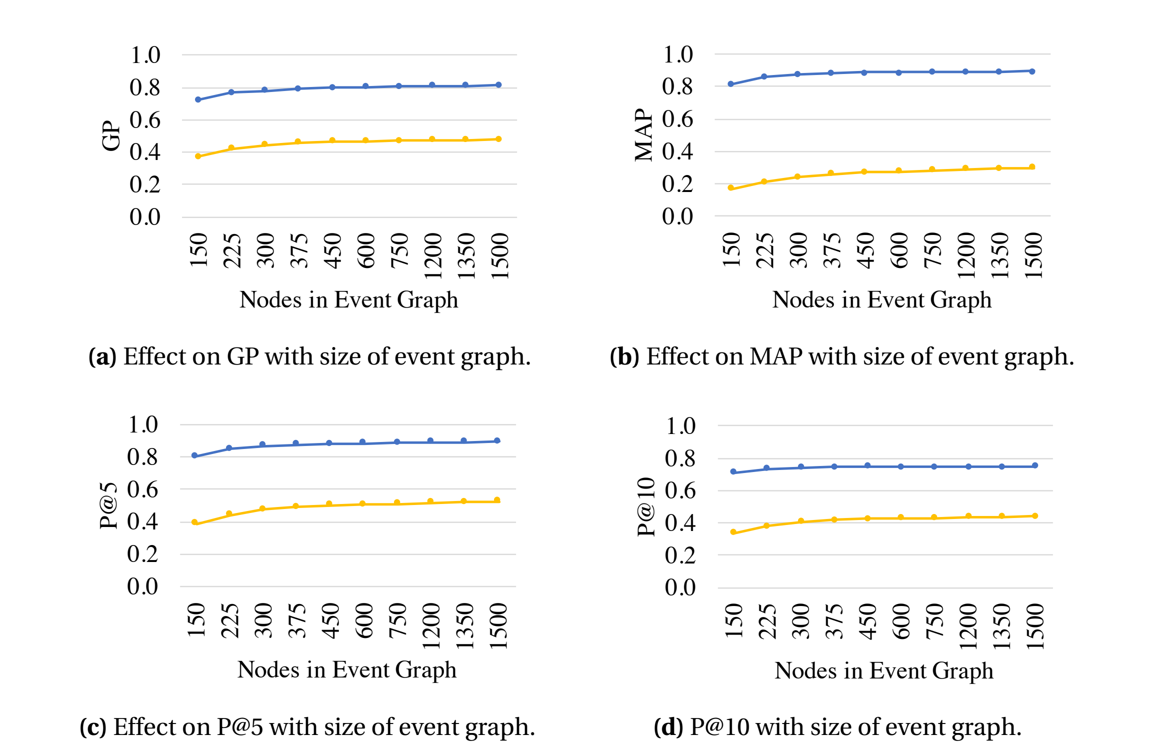
At the day granularity, the *DP-TPNX* method gets the highest gain of +0.63 with a score of 0.88 over the *LT-pd* method which proves to be the best performing baseline with 0.25 for the following query: It achieves the highest gain in P@5 from the best text-only method for the query:

Example 5.1: Rwandan President Juvénal Habyarimana and Burundi President Cyprien Ntaryamira die when a missile shoots down their jet near Kigali, Rwanda. This is taken as a pretext to begin the Rwandan Genocide.

It suffers the worst loss of −0.05 with a score equal to 0.88 while the best performing method for this query with a 0.93 score is *LT-pd* method for the following query:

Example 5.2: Ten-Day War: Fighting breaks out when the Yugoslav People's Army attacks secessionists in Slovenia.

The *NN-T* method is the next best baseline with a score of 0.42.



(a) Effect on GP with size of event graph. **(b)** Effect on MAP with size of event graph.

(c) Effect on P@5 with size of event graph. **(d)** P@10 with size of event graph.

Figure 5.2 The effect of increasing the size of the event graph on *DP-TPNX* (blue) and *NN-T* (yellow) methods.

At month granularity, the query for which the *DP-TPNX* shows the largest gain of +0.43 with a score of 0.70 is as follows:

Example 5.3: Troops of Laurent Kabila march into Kinshasa. Zaire is officially renamed Democratic Republic of the Congo.

The best baseline for this query is the *LT-pdS* with 0.26 and the next best method is *NN-T* with 0.21 P@5 scores. *DP-TPNX* suffers the worst loss of −0.23 with a score of 0.11 for the following query:

Example 5.4: Cold War: The leaders of the Yemen Arab Republic and the People's Democratic Republic of Yemen announce the unification of their countries as the Republic of Yemen.

We find that for this query, the *LT-pdS* method that gets a score of 0.33 and become the best method.

At year granularity, the *DP-TPNX* gets the most gain +0.17 with a score of 0.83 for the following query:

Example 5.5: Several explosions at a military dump in Lagos, Nigeria kill more than 1,000.

We find that all the other baselines receive an equal score of 0.67 for this query. Our method suffers the worst loss of −0.89 with score 0 for the following query:

Example 5.6: Dissolution of Czechoslovakia: The Czech Republic and Slovakia separate in the so-called Velvet Divorce.

For this query, we find that the *LT-pdS* is the best method with a score of 0.91 and *LT-pdS* is the second best with a score of 0.89.

The event, i.e., assassination of Juvénal Habyarimana and Burundi President Cyprien Ntaryamira described in Example 5.1 is an event that occurred on April 6, 1994. For this query, all the baselines get a low score. The *DP-TPNX* method considers the whole event graph and is able to generate more accurate time models for the excerpts considered for this query in comparison to the baselines including the nearest neighbor methods. In contrast, Example 5.2 is an aspect, i.e., the first attack, of a longer spanning event, Ten-Day War, with huge ramifications. Thus, we find that the *DP-TPNX* method ends up propagating time models from excerpts describing different event aspects, thereby reducing the quality of time models estimated from the excerpts. However, simply considering the most syntactically similar excerpts in case of the *NN-T* method results in the best time models. We find this effect is repeated when we look into the queries for which we get the highest gain and loss at the month and year granularities. Example 5.3 and Example 5.5 are short spanning events, while Example 5.4 and Example 5.6 describe aspects of long spanning events. Thus, we find that our method critically depends on the redundancy in the input set.

Discussion

First, we discuss the performance of different methods. The *ADJ* method turns out to be the least successful method for the time model estimation. The single temporal expression predicted by this method often does not overlap with the expressions in the ground truth. Since our evaluation metrics rely on the overlap, this method receives very low scores at all granularities. The publication date-based *LT-pd* method is observed to be less effective for estimating excerpt-time models. Similar to the *ADJ*, this method also relies on a single expression to estimate time models and suffers from sparsity. Further, the assumption that the news articles present information only on events occurring around its publication date is repudiated. The *LT-pdS* method which is a simple extension to the *LT-pd* estimates much better excerpt-time model. The *Rand* method benefits over the *LT* methods from a larger number of temporal expressions randomly selected. Among the methods that leverage the event graph, the *NN-T* is motivated from the popular pseudo-relevance feedback models [143] for estimating excerpt models. As expected, this method estimates more accurate time models as compared to the simpler publication date-based methods in terms of MQ. Finally, the distribution propagation improves the estimates of the excerpt-time model through multiple iterations (ideally until convergence).

Table 5.5 Results of estimating the focus time of 100 queries.

	Method	P@1	P@5	P@10	MAP	NDCG@5	NDCG@10
Year	DP-TNX	**0.46**	**0.13**	**0.07**	**0.52**	**0.54**	**0.55**
	NN-T	0.36	0.13	0.08	0.46	0.49	0.51
	ADJ	0.18	0.09	0.06	0.30	0.32	0.36
Month	DP-TNX	**0.28**	**0.09**	**0.05**	**0.31**	**0.31**	**0.32**
	NN-T	0.22	0.08	0.05	0.27	0.28	0.29
	ADJ	0.09	0.05	0.04	0.16	0.16	0.19
Day	DP-TNX	**0.19**	**0.16**	**0.12**	**0.19**	**0.17**	**0.21**
	NN-T	0.10	0.10	0.10	0.15	0.10	0.15
	ADJ	0.02	0.01	0.01	0.01	0.01	0.01

Next, we discuss the different granularities. At the day granularity, larger interval are represented as a set of days with uniform probability (as described in Section 5.5.2). Similarly, at year granularity all temporal expressions are relaxed. However, at the month granularity, year-level expression are expanded while the days level expressions are relaxed. Due to this mixed effect, we find that on average, quality of the time models across all the methods is reduced. We find that the results are more pronounced at the day as compared to the year granularity. At the year, due to the relaxation, we find that it becomes easier to estimate time models as indicated by all the baselines getting much better scores.

5.6 Estimating Event Focus Time

As described in Section 5.5.1, we randomly sample 100 Wikipedia events that come with a textual description and a time interval indicating their occurrence period. So far, we ignore the time part of the query and leverage only the text part as a keyword query to retrieve a set of documents which are then regarded as pseudo-relevant and input to the system. In this section, as an application-oriented use-case, we slightly shift the focus to the following problem: *for a given event as a user query that comes with a textual description, estimate its occurrence time period.* As input, our method takes an event query; as output a time interval in the day, month, and year granularity is returned indicating its occurrence period.

An event query q is described with two parts: text q_{text} as a bag of words derived from the textual description of the query; and time q_{time} as a single temporal expression that is explicitly given. The single temporal expression in the query is then represented in our time domain as described in Section 5.5.1.

We design a two-stage approach. In Stage 1, leveraging q_{text}, we make use of a standard KL-divergence based retrieval model to retrieve a set of top-100 documents. We then generate a set of excerpts ε from these documents by fixing an excerpt to a single sentence. In Stage 2, we then generate an event graph $G_{event}(\varepsilon, r)$ as described in Section 5.4.1. However, as a slight variation to the previous method, we inject the q_{text} as a special node. Next, we run our distribution propagation algorithm as described in Section 5.4 to estimate a query-time model Q_{time}. Finally, we compare the Q_{time} estimated with the event graph to q_{time} containing the temporal expression originally given as input thereby treating as the ground truth.

We compare our distribution propagation method *DP-TNX* that takes into consideration the textual (T), conceptual (N), and contextual (X) similarity. Positional similarity becomes inapplicable in this setting. Further, we compare the query text against the lead paragraph of the source document for an excerpt to estimate their contextual similarity. As baselines, we consider the nearest neighbor *NN-T* method, and the generic method proposed by Jatowt et al. [87] to estimate document focus time.

Analogous to an excerpt-time model, a query-time model Q_{time} can be understood as a probability distribution over the time units that captures the temporal scope of the event in the query. Intuitively, the units that occur with high probability indicate the salient time period associated with the event. From an estimated query-time model, we generate a ranked list of time units with decreasing probabilities in Q_{time}. However, unlike previous experiments, we do not truncate but consider the full estimated query-time model where the temporal scope of the query is set to $[min_t, max_t]$. As ground truth, we leverage the time part q_{time} in the original query. We compare the methods at three time granularity, year, month, and day using the measures define in Section 5.5.2.

Table 5.5 shows the results of the experiment. The results are found to be statistically significant with paired two-tail t test at $\alpha = 0.05$. The *DP-TNX* method is able to best estimate the occurrence time period at all time granularities across all measures. The result quality is lower, as compared to the main experiments because of mainly two reasons. Firstly, we note that most queries come with a single specific temporal expression usually at the day granularity which is considered as ground truth. This is in contrast to the previous setting where pseudo-relevant excerpts may come with multiple expressions at different granularities. Secondly, we find that the accuracy of query focus time is strongly dependent on the quality of the input pseudo-relevant excerpts. Topic drifts in the input excerpts will result in a query focus time that does not match with the ground truth. This makes the use-case problem even harder where our method is able to beat the state-of-the-art *ADJ* and *NN-T* methods.

5.7 Summary & Future Directions

In this chapter, we proposed a novel problem to estimate time models for excerpts that are extracted from news articles. To address this problem, we presented a distribution framework that extends a popular semi-supervised label propagation algorithm [224] to propagate time models from excerpts that come with temporal expressions to those that

Table 5.6 Example of incidents extracted from excerpts describing events.

Excerpts (Events)	Incidents	Time
ε_1: The electronic producers Skrillex and Diplo, who under the name Jack Ü had one of the biggest hits *last year* with "Where Are Ü Now", featuring Justin Bieber, won both best dance recording for that song and best electronic/dance album.	I_1: "had" ("Jack", "The electronic producers Skrillex and Diplo one of the biggest hits *last year* with Where Are Ü Now featuring Justin Bieber)	T_1: "last year" = [01-01-2015, 31-12-2015]
	I_2: "be featuring" ("one of the biggest hits *last year* with Where Are Ü Now","Justin Bieber")	T_1: "last year" = [01-01-2015, 31-12-2015]
	I_3: "won" ("The electronic producers Skrillex and Diplo", "both best dance recording for that song and best electronic/dance album")	

are strongly related but missing annotations. To capture the inter-excerpt relations, we designed several measures that define the relation in terms of syntactic and semantic redundancies across excerpts. We conducted elaborate experiments to evaluate the estimated time models for the excerpts. In our experiments, we found that our method estimates the most accurate time models as compared to several baselines. We compare the methods in terms of existing precision, recall, NDCG, and two new Model Quality and Generative Power measures.

As a use-case experiment intended to perform an extrinsic evaluation of the estimated time models, we also tested the effectiveness of the time models in predicting the focus time of the textual event descriptions. For this, the focus time was estimated at different time granularities namely year, month and days. In our experiment, we compared our method against the state-of-the-art methods from the literature and found that our method proves to be the most effective for the task in terms of Precision and Mean Reciprocal Rank.

Future Directions

In this chapter, we motivate estimation of time models for textual excerpts extracted from news articles to better estimate their temporal scope. In our approach, we design a simple semi-supervised distribution propagation framework based on the popular label propagation that leverages redundancy in text based on several inter-excerpt relations. We next speculate on how plausible improvements in the quality of the estimated time models can be achieved to design future work.

- So far in our work, we considered a simple graphical representation of the excerpts extracted from new to constitute the event graph. However, we note that often a single excerpt describes multiple incidents that may be part of a larger event. With this perspective, the excerpts extracted from news articles obtained by leveraging the document structure can be made more fine-grained by extracting independent clauses in the excerpt descriptions. We can thus define the fine-grained segments as incidents that refer to a single unit or aspect within a larger event. Such fine-grained incidents extracted can be used to generate more elaborate incident graphs [57] for estimating incident-focus time. The final excerpt time model can be then estimated based on the associated incidents (e.g., with interpolation). A concrete example of incidents extracted as clauses from an excerpt is illustrated in Table 5.6. One approach to extracting incidents can be designed with the following steps: first, pass the excerpt describing an event through an open clause extractor [41] that extracts relations and their arguments (as subject and predicate) which we refer to as clauses. We obtain the three clauses I_1, I_2, and I_3 with *governing verb* as a function with two arguments. Next, pass the incidents through a tagger to recognize and normalize the temporal expressions that occur in the description. In case of our running example, we recognize the relative temporal expressions *"last year"* in the incidents I_1 and I_2, and normalize it with the publication date of the source input document as the reference date. The extracted incidents I_1 and I_2 can be then used to construct the incident graph in distribution propagation algorithm. We plan to explore the idea of incident graphs with novel inter-incident relations as a future work.

- In our approach, we leverage short-context unigram language models [214] for modeling text similarity as an inter-excerpt relation. In our experiments, though we find estimating inter-excerpt similarity based on unigram language models, recently Guo et al. [73] argue their short context forces a lack of semantic knowledge. On the other hand, long-span LSTM-based and n-gram-based across-sentence-boundary model [149] additionally capture semantic relations in text [73] and have shown impressive gains in semantically-motivated NLP tasks, like Question Answering [170]. Thus, leveraging long-span language models like the neural network based long short-term memory (LSTM) language models [195] may improve inter-excerpt redundancy estimation thereby improving the time models. Thus, as a future work it remains to incorporate and test the long-span language models to better estimate semantic inter-excerpt redundancies for our task.

- We introduced the probabilistic time models estimated for documents in Chapter 3. Our experiments described in Section 3.5.5 showed that the time models when used in combination with document language models improved retrieval quality. In Chapter 4, we estimate similar time models for generation of event digests. We speculate that both applications can be further improved by leveraging more accurate time models as shown in this chapter. However, so far we have leveraged a simpler time model which can be further improved by considering uncertainty to time [26], and temporal relations such as "after" and "before". We speculate how we can further improve the time models by considering uncertainty of temporal expressions in Section 3.6.

- In our approach, time models are represented as probability distributions over the time domain that need to be estimated on-the-fly so as to be applied in context of different applications. However, from a practicality and scalability stand point, it is required to design efficient data structures that can operate on the distributions. Thus, as a future work, we plan to look into the efficiency side of the problem by designing methods to incorporate the distributions into indexing systems. One plausible approach can be to extend indexing systems such as the APLA [122] proposed by Ljosa and Singh that facilitate efficient processing of k-NN queries.

Chapter 6

Conclusion

In this work, we have addressed three problems towards the direction of organizing ever growing online information available on news worthy events. In detail:

- We proposed a novel linking task of connecting Wikipedia past event descriptions to online news articles. We cast the linking problem into an information retrieval task thereby considering the excerpt as a user query. As two instances of such event descriptions, we consider those that are listed in special Wikipedia year pages, and those that occur as arbitrary passages within general articles. In our approach, we proposed several time-aware language models estimated from the temporal expressions in meta data like document publication dates to connect Wikipedia year page events to news articles. Experimental evaluations performed with 50 randomly sampled events on the New York Times corpus showed that our two-stage cascade approach proves to be the most effective as compared to several baselines.

- To illustrate a practical application of the time-aware retrieval models capturing different temporal intents, we presented a demonstration of EXPOSÉ, a time-aware exploratory search system for past events.

- To connect Wikipedia excerpts extracted from general pages, analogous to before, we presented a retrieval framework that leverages additional semantics, namely time, geolocations, and named entities that come with a given Wikipedia excerpts. Under this framework our method estimates independent query-text, -time, -space, and -entity models by considering them as event dimensions. Finally, documents are ranked by comparing them to the query across all the dimensions. Experimental evaluations with 150 excerpts randomly sampled from Wikipedia articles on the New York Times and the ClueWeb 12-B13 TREC corpus shows that our method considering a combination of all the event dimensions proves to be the most effective.

- We have addressed the problem of generating event digests with the goal of providing holistic view on past events. As a special case of the multi-document extractive text summarization problem, the event digest problem additionally aimed at diversifying across the event dimensions namely, text, time, geolocations, and named entities. In our approach, we proposed a novel divergence-based framework for global inference where the overall relevance of the digest to input event was maximized while redundancy was avoided across all the four event dimensions. To implement the idea, we used integer linear programming with necessary constraint. In our experimental evaluation, we compared several state-of-the-art methods from the literature and found that our divergence-based method that considers all event dimensions is most effective.

- With the goal of making an event digest coherent and readable, we aimed to study textual patterns within short summaries that makes it coherent. For this, we leveraged a crowdsourcing platform and generated a corpus containing human preference-judged variants of fixed-length summaries on past events. More specifically, we analyzed the impact of altering the ordering and proximity between sentences and studied its effect on human perception for coherence of the text. Moreover, we analyzed the feasibility of the study and found that it can be successfully conducted with sufficient quality control.

- We addressed the problem of estimating time models for excerpts that are extracted from news articles. We extended the label propagation algorithm [224] to design a distribution propagation algorithm. Our approach propagated time models from excerpts that come with temporal expressions to those that are strongly related but missing annotations. We designed several inter-excerpt relations in terms of syntactic and semantic redundancies across excerpts. Our elaborate experiments to evaluate the estimated time models for the excerpts were performed on the English Gigaword corpus and we found that our method estimates the most accurate time models as compared to several baselines. We compare the methods in terms of existing Precision, Recall, NDCG, and two new Model Quality and Generative Power measures. As an extrinsic experiment, we also tested effectiveness of the time models in predicting the focus time [87] of the textual event descriptions. We compared our method against the state-of-the-art methods from the literature and found that our method proves to be the most effective.

6.1 Outlook

The problems addressed in this work are defined with the goal of better organizing information on past events so as to aid retrospection. However, towards achieving the larger goals, there are several open problems that need to be addressed, thus presenting plenty research opportunities. Next, we outline some directions that we consider promising and important.

Linking Information on Past Events

In Chapter 3 we motivate connecting different information sources on past events to facilitate an effective and efficient retrospection of past events. As two instances, we began by considering Wikipedia events and news articles. The linking problem proposed in this chapter open several research directions. As the first direction, the linking problem can be defined to connect social media to Wikipedia and news articles. With this, a user retrospecting on past events will be able to acquire three views: **1)** the constantly evolving and collectively authored Wikipedia as the *collective memory*, **2)** punctually published news articles as *objective and unbiased*, and **3)** the social media posts relevant to the event as *opinionated democratic view*.

So far we considered representing the event dimensions by estimating probabilistic models leveraging the semantic annotations in text. as the second research direction, we propose to design better representations for the event dimensions, including text, time, geolocation and entity, can be considered. However, as an additional challenge more complex representations also affect the efficiency of systems. A third research direction is to estimate more dense representations for the event models. In our proposed approaches, the text, time, space, and entity models are probability distribution over the entire scope of the corpus. If viewed as vectors, they can be sparse (specially for events with smaller scope). Appropriate techniques to represent the models as dense vectors may lead to quality and efficiency improvements. As a final direction, indexing systems that can deal with efficient storage of elaborate multiple query and document models can be designed to improve query time efficiency and aid interactive analytical tasks. Addressing the several future directions delineated to further improve the current techniques developed by us Section 3.6 in Chapter 3 presents new challenges.

Event Digest Generation

In Chapter 4, we motivate generation of an event digest to connect Wikipedia events to excerpts taken from news articles. We find several directions to proceed from here. Firstly, the event digest generation with explicitly considering the text, time, geolocations, and entities can be defined on a stream of documents, thus addressing the problem of update event summarization. Secondly, additional signal can be leveraged while generating an event digest to capture more information. For example, affective norms have been used for viewpoint summarization. Incorporating such methods can make the task more event focused.

Additionally, in Chapter 4, we design a dataset that can be used to develop automatic methods for summary-text coherence evaluation. In this context, we plan to extend the corpus so as to analyze more syntactic and semantic regularities that make a summary more coherent. We find that the current measures for evaluating text summarization task do not consider structural quality. In the future, we intend to look into the problem of designing measures that explicitly evaluate summary structures for the text summarization task.

Temporally Scoping News Excerpts

In Chapter 5, we motivated estimating time models that capture temporal information associated with excerpts taken from news articles. Our time models were designed as probability distributions over a time domain where salient time point receive higher probability thus capturing the temporal scope. We find several research directions to proceed from here. We have so far considered only estimating time models for short textual news excerpts. However, the generic method can be also extended to other event dimensions such as entities, and geographic locations. The more accurate models in combination with text can be then used to improve our approaches in Chapter 3 and 4. Thus extending our approach to the event dimensions becomes the first research direction that we plan to investigate in future.

One interesting perspective on the time models can come from the direction of generating embeddings for excerpts. In recent works [113, 140], leveraging neural word embeddings have shown significant improvements in various text-based applications. The time model estimation presented may be considered as the first step of embedding excerpts into a space that models the time dimension of information content. As a second future direction, we plan to design advanced methods to identify better sources of temporal information than relying on blind feedback, and output dense representation of excerpts to better capture its temporal semantics.

List of Figures

List of Tables

Bibliography

[1] Wikipediaś current events portal, 2016. URL http://en.wikipedia.org/wiki/Portal:Currentevents.

[2] The clueweb09 collection, 2016. URL http://www.lemurproject.org/clueweb09.php/.

[3] The clueweb12 collection, 2016. URL http://www.lemurproject.org/clueweb12.php/.

[4] The new york times annotated corpus, 2016. URL http://corpus.nytimes.com.

[5] English gigaword corpus. https://catalog.ldc.upenn.edu/ldc2003t05, 2017. Accessed: 2017-01-24.

[6] Gurobi optimizer. http://www.gurobi.com, 2017. Accessed: 2017-01-24.

[7] Maristella Agosti and James Allan. Introduction to the special issue on methods and tools for the automatic construction of hypertext. *Information Processing & Management*, 33(2):129–131, 1997. ISSN 03064573. doi: 10.1016/S0306-4573(96)00057-X.

[8] J. Allan. Overview of the trec 2004 robust retrieval track. In *Proceedings of TREC*, volume 13, 2004.

[9] James Allan. *Topic detection and tracking: event-based information organization*, volume 12. Springer Science & Business Media, 2012. ISBN 1461509335.

[10] James Allan, Ron Papka, and Victor Lavrenko. On-line new event detection and tracking. In *SIGIR 1998: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 24-28 1998, Melbourne, Australia*, pages 37–45, 1998. doi: 10.1145/290941.290954.

[11] James Allan, Rahul Gupta, and Vikas Khandelwal. Temporal summaries of news topics. In *SIGIR 2001: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, September 9-13, 2001, New Orleans, Louisiana, USA*, pages 10–18, 2001. doi: 10.1145/383952.383954.

[12] Omar Alonso and Ricardo A. Baeza-Yates. Design and implementation of relevance assessments using crowdsourcing. In *Advances in Information Retrieval - 33rd European Conference on IR Research, ECIR 2011, Dublin, Ireland, April 18-21, 2011. Proceedings*, pages 153–164, 2011. doi: 10.1007/978-3-642-20161-5_16.

[13] Omar Alonso, Jannik Strötgen, Ricardo A. Baeza-Yates, and Michael Gertz. Temporal information retrieval: Challenges and opportunities. pages 1–8, 2011.

[14] Ioannis Arapakis, Mounia Lalmas, Hakan Ceylan, and Pinar Donmez. Automatically embedding newsworthy links to articles: From implementation to evaluation. *JASIST*, 65(1):129–145, 2014. doi: 10.1002/asi.22959.

[15] Irem Arikan, Srikanta J. Bedathur, and Klaus Berberich. Time will tell: Leveraging temporal expressions in IR. In *Proceedings of the Second International Conference on Web Search and Web Data Mining, WSDM 2009, Barcelona, Spain, February 9-11, 2009*, 2009.

[16] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary G. Ives. Dbpedia: A nucleus for a web of open data. pages 722–735, 2007. doi: 10.1007/978-3-540-76298-0_52.

[17] Jing Bai, Jian-Yun Nie, Guihong Cao, and Hugues Bouchard. Using query contexts in information retrieval. In *SIGIR 2007: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, The Netherlands, July 23-27, 2007*, pages 15–22, 2007. doi: 10.1145/1277741.1277747.

[18] Krisztian Balog, Pavel Serdyukov, and Arjen P. de Vries. Overview of the TREC 2010 entity track. Technical report, 2010.

[19] Siddhartha Banerjee and Prasenjit Mitra. Wikikreator: Improving wikipedia stubs automatically. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*, pages 867–877, 2015.

[20] Regina Barzilay and Michael Elhadad. Using lexical chains for text summarization. pages 111–121, 1999.

[21] Regina Barzilay, Noemie Elhadad, and Kathleen R. McKeown. Sentence ordering in multidocument summarization. In *Proceedings of the first international conference on Human language technology research*, pages 1–7, 2001. ISBN 1076-9757. doi: 10.3115/1072133.1072217.

[22] Regina Barzilay, Noemie Elhadad, and Kathleen McKeown. Inferring strategies for sentence ordering in multidocument news summarization. *J. Artif. Intell. Res.*, 17: 35–55, 2002. doi: 10.1613/jair.991.

[23] Tal Baumel, Raphael Cohen, and Michael Elhadad. Query-chain focused summarization. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 1: Long Papers*, pages 913–922, 2014.

[24] Patrice Bellot, Antoine Doucet, Shlomo Geva, Sairam Gurajada, Jaap Kamps, Gabriella Kazai, Marijn Koolen, Arunav Mishra, Véronique Moriceau, Josiane Mothe, Michael Preminger, Eric SanJuan, Ralf Schenkel, Xavier Tannier, Martin

Theobald, Matthew Trappett, Andrew Trotman, Mark Sanderson, Falk Scholer, and Qiuyue Wang. Report on INEX 2013. volume 47, pages 21–32, 2013. doi: 10.1145/2568388.2568393.

[25] Yoshua Bengio, Olivier Delalleau, and Nicolas Le Roux. Label Propagation and Quadratic Criterion. *Compute*, pages 185–207, 2006. doi: 10.7551/mitpress/9780262033589.003.0011.

[26] Klaus Berberich, Srikanta J. Bedathur, Omar Alonso, and Gerhard Weikum. A language modeling approach for temporal information needs. In *Advances in Information Retrieval, 32nd European Conference on IR Research, ECIR 2010, Milton Keynes, UK, March 28-31, 2010. Proceedings*, pages 13–25, 2010. doi: 10.1007/978-3-642-12275-0_5.

[27] Kurt D. Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2008, Vancouver, BC, Canada, June 10-12, 2008*, pages 1247–1250, 2008. doi: 10.1145/1376616.1376746.

[28] Danushka Bollegala, Naoaki Okazaki, and Mitsuru Ishizuka. A preference learning approach to sentence ordering for multi-document summarization. *Inf. Sci.*, 217: 78–95, 2012. doi: 10.1016/j.ins.2012.06.015.

[29] Philip Bramsen, Pawan Deshpande, Yoong Keok Lee, and Regina Barzilay. Inducing temporal graphs. In *EMNLP 2007, Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, 22-23 July 2006, Sydney, Australia*, pages 189–198, 2006.

[30] Marc Bron, Bouke Huurnink, and Maarten de Rijke. Linking archives using document enrichment and term selection. In *Research and Advanced Technology for Digital Libraries - International Conference on Theory and Practice of Digital Libraries, TPDL 2011, Berlin, Germany, September 26-28, 2011. Proceedings*, pages 360–371, 2011. doi: 10.1007/978-3-642-24469-8_37.

[31] James P. Callan. Passage-level evidence in document retrieval. In *Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval. Dublin, Ireland, 3-6 July 1994 (Special Issue of the SIGIR Forum)*, pages 302–310, 1994.

[32] Ricardo Campos, Gaël Dias, Alípio Jorge, and Celia Nunes. GTE: a distributional second-order co-occurrence approach to improve the identification of top relevant dates in web snippets. In *21st ACM International Conference on Information and Knowledge Management, CIKM'12, Maui, HI, USA, October 29 - November 02, 2012*, pages 2035–2039, 2012. doi: 10.1145/2396761.2398567.

[33] Ricardo Campos, Gaël Dias, Alípio M. Jorge, and Adam Jatowt. Survey of temporal information retrieval and related applications. *ACM Computing Surveys (CSUR)*, 47(2):15, 2015. ISSN 0360-0300.

[34] Jaime G. Carbonell and Jade Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *SIGIR 1998: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 24-28 1998, Melbourne, Australia*, pages 335–336, 1998. doi: 10.1145/290941.291025.

[35] Angel X. Chang and Christopher D. Manning. Sutime: A library for recognizing and normalizing time expressions. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey, May 23-25, 2012*, pages 3735–3740, 2012.

[36] Rachel Chasin, Daryl Woodward, Jeremy Witmer, and Jugal Kalita. Extracting and displaying temporal and geospatial entities from articles on historical events. *The Computer Journal*, 57(3):403–426, 2013. ISSN 1460-2067.

[37] Hai Leong Chieu and Yoong Keok Lee. Query based event extraction along a timeline. In *SIGIR 2004: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Sheffield, UK, July 25-29, 2004*, pages 425–432, 2004. doi: 10.1145/1008992.1009065.

[38] C. Ravindranath Chowdary and P. Sreenivasa Kumar. Sentence ordering for coherent multi-document summary generation. In *Sharing Data, Information and Knowledge, 25th British National Conference on Databases, BNCOD 25, Cardiff, UK, July 7-10, 2008. Proceedings*, pages 40–50, 2008. doi: 10.1007/978-3-540-70504-8_5.

[39] Gobinda G. Chowdhury. *Introduction to modern information retrieval*. Facet publishing, 2010. ISBN 185604694X.

[40] James H. Coombs. Hypertext, full text, and automatic linking. In *SIGIR 1990: Proceedings of the 13th International Conference on Research and Development in Information Retrieval, Brussels, Belgium, 5-7 September 1990, Proceedings*, pages 83–98, 1990. doi: 10.1145/96749.98010.

[41] Luciano Del Corro and Rainer Gemulla. Clausie: clause-based open information extraction. In *22nd International World Wide Web Conference, WWW '13, Rio de Janeiro, Brazil, May 13-17, 2013*, pages 355–366, 2013.

[42] Vittoria Cozza, Antonio Messina, Danilo Montesi, Luca Arietta, and Matteo Magnani. Spatio-temporal keyword queries in social networks. In *Advances in Databases and Information Systems - 17th East European Conference, ADBIS 2013, Genoa, Italy, September 1-4, 2013. Proceedings*, pages 70–83, 2013. doi: 10.1007/978-3-642-40683-6_6.

[43] W. Bruce Croft, Donald Metzler, and Trevor Strohman. *Search Engines: Information Retrieval in Practice*, volume 283. Addison-Wesley Reading, 2009. ISBN 9780136072249. doi: 10.1093/comjnl/bxq039.

[44] Jeffrey Dalton, Laura Dietz, and James Allan. Entity query feature expansion using knowledge base links. In *The 37th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '14, Gold Coast , QLD, Australia - July 06 - 11, 2014*, pages 365–374, 2014. doi: 10.1145/2600428.2609628.

[45] Supratim Das, Arunav Mishra, Vinay Setty, and Klaus Berberich. Estimating event focus time using neural word embeddings. In *Proceedings of the 26th ACM International Conference on Information and Knowledge Management, CIKM 2017, Singapore, November 6-10, 2017*, 2017.

[46] Franciska De Jong, Henning Rode, and Djoerd Hiemstra. Temporal language models for the disclosure of historical text. In *Cognitive Science*, pages 1–9. Royal Netherlands Academy of Arts and Sciences, 2005. ISBN 9069844567.

[47] Gianluca Demartini, Tereza Iofciu, and Arjen P. de Vries. Overview of the INEX 2009 entity ranking track. In *Focused Retrieval and Evaluation, 8th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2009, Brisbane, Australia, December 7-9, 2009, Revised and Selected Papers*, pages 254–264, 2009. doi: 10.1007/978-3-642-14556-8_26.

[48] Laura Dietz, Michael Schuhmacher, and Simone Paolo Ponzetto. Queripidia : Query-specific Wikipedia Construction. *NIPS Workshop for Automated Knowledge Base Construction (AKBC)*, pages 1–5, 2014.

[49] Zhicheng Dou, Sha Hu, Kun Chen, Ruihua Song, and Ji-Rong Wen. Multi-dimensional search result diversification. In *Proceedings of the Forth International Conference on Web Search and Web Data Mining, WSDM 2011, Hong Kong, China, February 9-12, 2011*, pages 475–484, 2011. doi: 10.1145/1935826.1935897.

[50] Robert Durst, Kevin Hunter, and David Stein. System and method for linking online resources to print media and authoring tool for same, 2000.

[51] Miles Efron, Jimmy J. Lin, Jiyin He, and Arjen P. de Vries. Temporal feedback for tweet search with non-parametric density estimation. In *The 37th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '14, Gold Coast , QLD, Australia - July 06 - 11, 2014*, pages 33–42, 2014. doi: 10.1145/2600428.2609575.

[52] Patrick Ernst, Arunav Mishra, Avishek Anand, and Vinay Setty. Bionex: A system for biomedical news event exploration. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7-11, 2017*, pages 1277–1280, 2017. doi: 10.1145/3077136.3084150.

[53] Hui Fang and ChengXiang Zhai. Probabilistic models for expert finding. In *Advances in Information Retrieval, 29th European Conference on IR Research, ECIR 2007, Rome, Italy, April 2-5, 2007, Proceedings*, pages 418–430, 2007. doi: 10.1007/978-3-540-71496-5_38.

[54] Hui Fang, Tao Tao, and ChengXiang Zhai. An exploration of formalized retrieval heuristics. *Urbana*, 51:61801, 2003.

[55] Mohamed Abdel Fattah and Fuji Ren. Automatic text summarization. *World Academy of Science, Engineering and Technology*, 37:2008, 2008.

[56] Mohamed Abdel Fattah and Fuji Ren. Ga, mr, ffnn, PNN and GMM based models for automatic text summarization. *Computer Speech & Language*, 23(1):126–144, 2009. doi: 10.1016/j.csl.2008.04.002.

[57] Ao Feng and James Allan. Finding and linking incidents in news. In *Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management, CIKM 2007, Lisbon, Portugal, November 6-10, 2007*, pages 821–830, 2007. doi: 10.1145/1321440.1321554.

[58] Paolo Ferragina and Ugo Scaiella. TAGME: on-the-fly annotation of short text fragments (by wikipedia entities). In *Proceedings of the 19th ACM Conference on Information and Knowledge Management, CIKM 2010, Toronto, Ontario, Canada, October 26-30, 2010*, pages 1625–1628, 2010. doi: 10.1145/1871437.1871689.

[59] Besnik Fetahu, Katja Markert, and Avishek Anand. Automated news suggestions for populating wikipedia entity pages. pages 323–332, 2015. doi: 10.1145/2806416.2806531.

[60] Elena Filatova. Event-based extractive summarization. In *ACL Workshop on Summarization*, 2004.

[61] Evgeniy Gabrilovich, Michael Ringgaard, and Amarnag Subramanya. FACC1: Freebase annotation of ClueWeb corpora, Version 1 (Release date 2013-06-26, Format version 1, Correction level 0). *Note: http://lemurproject.org/clueweb09/FACC1/Cited by*, 5, 2013.

[62] Antony Galton and Juan Carlos Augusto. Two approaches to event definition. In *Database and Expert Systems Applications, 13th International Conference, DEXA 2002, Aix-en-Provence, France, September 2-6, 2002, Proceedings*, pages 547–556, 2002. doi: 10.1007/3-540-46146-9_54.

[63] Mahak Gambhir and Vishal Gupta. Recent automatic text summarization techniques: a survey. *Artificial Intelligence Review*, 47(1):1–66, 2017. ISSN 15737462. doi: 10.1007/s10462-016-9475-9.

[64] Tao Ge, Baobao Chang, Sujian Li, and Zhifang Sui. Event-based time label propagation for automatic dating of news articles. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1–11, 2013.

[65] Fredric C. Gey, Ray R. Larson, Mark Sanderson, Kerstin Bischoff, Thomas Mandl, Christa Womser-Hacker, Diana Santos, Paulo Rocha, Giorgio Maria Di Nunzio, and Nicola Ferro. Geoclef 2006: The CLEF 2006 cross-language geographic information retrieval track overview. In *Evaluation of Multilingual and Multi-modal Information Retrieval, 7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006, Alicante, Spain, September 20-22, 2006, Revised Selected Papers*, pages 852–876. 2006. doi: 10.1007/978-3-540-74999-8_109.

[66] Fredric C. Gey, Ray R. Larson, Mark Sanderson, Kerstin Bischoff, Thomas Mandl, Christa Womser-Hacker, Diana Santos, Paulo Rocha, Giorgio Maria Di Nunzio, and Nicola Ferro. Geoclef 2006: The CLEF 2006 cross-language geographic information retrieval track overview. In *Evaluation of Multilingual and Multi-modal Information Retrieval, 7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006, Alicante, Spain, September 20-22, 2006, Revised Selected Papers*, pages 852–876. 2006. doi: 10.1007/978-3-540-74999-8_109.

[67] Fredric C. Gey, Ray R. Larson, Noriko Kando, Jorge Machado, and Tetsuya Sakai. Ntcir-geotime overview: Evaluating geographic and temporal search. In *Proceedings of the 8th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access, NTCIR-8, National Center of Sciences, Tokyo, Japan, June 15-18, 2010*, pages 147–153, 2010.

[68] Fredric C. Gey, Ray R. Larson, Jorge Machado, and Masaharu Yoshioka. Ntcir9-geotime overview - evaluating geographic and temporal search: Round 2. In *Proceedings of the 9th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access, NTCIR-9, National Center of Sciences, Tokyo, Japan, December 6-9, 2011*, 2011.

[69] Dan Gillick and Benoit Favre. A Scalable Global Model for Summarization. In *Proceedings of the Workshop on Integer Linear Programming for Natural Langauge Processing*, number June, pages 10–18, 2009. ISBN 9781932432350. doi: 10.3115/1611638.1611640.

[70] José Miguel Goñi-Menoyo, José Carlos González-Cristóbal, and Jorge Fombella-Mourelle. An optimised trie index for natural language processing lexicons. Technical report, MIRACLE Technical Report. Universidad Polit{é}cnica de Madrid, 2004.

[71] José Miguel Goñi-Menoyo, José Carlos González Cristóbal, Sara Lana-Serrano, and Ángel Martínez-González. Miracle's ad-hoc and geographical IR approaches for CLEF 2006. In *Evaluation of Multilingual and Multi-modal Information Retrieval, 7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006, Alicante, Spain, September 20-22, 2006, Revised Selected Papers*, pages 877–880, 2006. doi: 10.1007/978-3-540-74999-8_110.

[72] James Gung and Jugal Kalita. Summarization of historical articles using temporal event clustering. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 3-8, 2012, Montréal, Canada*, pages 631–635, 2012.

[73] Jiafeng Guo, Yixing Fan, Qingyao Ai, and W. Bruce Croft. A deep relevance matching model for ad-hoc retrieval. In *Proceedings of the 25th ACM International Conference on Information and Knowledge Management, CIKM 2016, Indianapolis, IN, USA, October 24-28, 2016*, pages 55–64, 2016. doi: 10.1145/2983323.2983769.

[74] Dhruv Gupta and Klaus Berberich. Identifying time intervals of interest to queries. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, CIKM 2014, Shanghai, China, November 3-7, 2014*, pages 1835–1838, 2014. doi: 10.1145/2661829.2661927.

[75] Ramaswamy Hariharan, Bijit Hore, Chen Li, and Sharad Mehrotra. Processing spatial-keyword (SK) queries in geographic information retrieval (GIR) systems. In *19th International Conference on Scientific and Statistical Database Management, SSDBM 2007, 9-11 July 2007, Banff, Canada, Proceedings*, page 16, 2007. doi: 10.1109/SSDBM.2007.22.

[76] Jiyin He, Maarten de Rijke, Merlijn Sevenster, Rob C. van Ommering, and Yuechen Qian. Generating links to background knowledge: a case study using narrative radiology reports. In *Proceedings of the 20th ACM Conference on Information and Knowledge Management, CIKM 2011, Glasgow, United Kingdom, October 24-28, 2011*, pages 1867–1876, 2011. doi: 10.1145/2063576.2063845.

[77] Marti A. Hearst and Christian Plaunt. Subtopic structuring for full-length document access. In *Proceedings of the 16th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval. Pittsburgh, PA, USA, June 27 - July 1, 1993*, pages 59–68, 1993. doi: 10.1145/160688.160695.

[78] Monika Rauch Henzinger, Bay-Wei Chang, Brian Milch, and Sergey Brin. Query-free news search. volume 8, pages 101–126, 2005. doi: 10.1007/s11280-004-4870-6.

[79] Monika Rauch Henzinger, Bay-Wei Chang, Brian Milch, and Sergey Brin. Query-free news search. *World Wide Web*, 8(2):101–126, 2005. doi: 10.1007/s11280-004-4870-6.

[80] Jerry R. Hobbs. *Literature and Cognition*, volume 18. Center for the Study of Language (CSLI), 1990. ISBN 0937073539.

[81] Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. Robust disambiguation of named entities in text. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, 27-31 July 2011, John McIntyre Conference Centre, Edinburgh, UK, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 782–792, 2011.

[82] Johannes Hoffart, Fabian M. Suchanek, Klaus Berberich, and Gerhard Weikum. YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia. In *IJCAI International Joint Conference on Artificial Intelligence*, volume 194, pages 3161–3165, 2013. ISBN 9781577356332. doi: 10.1016/j.artint.2012.06.001.

[83] Po Hu, Dong-Hong Ji, Hai Wang, and Chong Teng. Query-focused multi-document summarization using co-training based semi-supervised learning. In *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation, PACLIC 23, Hong Kong, China, December 3-5, 2009*, pages 190–199, 2009.

[84] Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry P. Heck. Learning deep structured semantic models for web search using clickthrough data. In *22nd ACM International Conference on Information and Knowledge Management, CIKM'13, San Francisco, CA, USA, October 27 - November 1, 2013*, pages 2333–2338, 2013. doi: 10.1145/2505515.2505665.

[85] David Hume. *Philosophical Essays Concerning Human Understanding.* Georg Olms Verlag, 1748.

[86] Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 1991. doi: 10.1162/neco.1991.3.1.79.

[87] Adam Jatowt, Ching-man Au Yeung, and Katsumi Tanaka. Estimating document focus time. In *22nd ACM International Conference on Information and Knowledge Management, CIKM'13, San Francisco, CA, USA, October 27 - November 1, 2013*, pages 2273–2278, 2013. doi: 10.1145/2505515.2505655.

[88] Donghong Ji and Yu Nie. Sentence ordering based on cluster adjacency in multi-document summarization. In *Third International Joint Conference on Natural Language Processing, IJCNLP 2008, Hyderabad, India, January 7-12, 2008*, pages 745–750, 2008.

[89] Paul D. Ji and Stephen G. Pulman. Sentence ordering with manifold-based classification in multi-document summarization. In *EMNLP 2007, Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, 22-23 July 2006, Sydney, Australia*, pages 526–533, 2006.

[90] Hubert Jin, Richard M. Schwartz, Sreenivasa Sista, and Frederick Walls. Topic tracking for radio, TV broadcast, and newswire. In *Sixth European Conference on Speech Communication and Technology, EUROSPEECH 1999, Budapest, Hungary, September 5-9, 1999*, 1999.

[91] H. Y. Jing. Summary generation through intelligent cutting and pasting of the input document. Technical report, Technical report, Columbia University, 1998.

[92] K. Sparck Jones, S. Walker, and S. E. Robertson. A probabilistic model of information retrieval: development and comparative experiments: Part 1. *Information Processing & Management*, 36(6):779–808, 2000. ISSN 0306-4573.

[93] Rosie Jones and Fernando Diaz. Temporal profiles of queries. *ACM Transactions on Information Systems*, 25(3):14–es, July 2007. ISSN 10468188. doi: 10.1145/1247715.1247720.

[94] K. Sparck Jonesa, S. Walkerb, and S. E. Robertsonb. A probabilistic model of information retrieval: development and comparative experiments Part 2. *Information Processing and Management*, 36(809):840, 2000.

[95] Michael Kaisser, Marti A. Hearst, and John B. Lowe. Improving search results quality by customizing summary lengths. In *ACL 2008, Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics, June 15-20, 2008, Columbus, Ohio, USA*, pages 701–709, 2008.

[96] Nattiya Kanhabua and Kjetil Nørvåg. Using temporal language models for document dating. In *Machine Learning and Knowledge Discovery in Databases, European Conference, ECML PKDD 2009, Bled, Slovenia, September 7-11, 2009, Proceedings, Part II*, pages 738–741, 2009. doi: 10.1007/978-3-642-04174-7_53.

[97] Nattiya Kanhabua and Kjetil Nørvåg. A comparison of time-aware ranking methods. In *Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2011, Beijing, China, July 25-29, 2011*, pages 1257–1258, 2011. doi: 10.1145/2009916.2010147.

[98] Nattiya Kanhabua, Roi Blanco, and Kjetil Nørvåg. Temporal information retrieval. *Foundations and Trends in Information Retrieval*, 9(2):91–208, 2015. doi: 10.1561/1500000043.

[99] Masayuki Karasuyama and Hiroshi Mamitsuka. Manifold-based similarity adaptation for label propagation. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.*, pages 1547–1555, 2013.

[100] Marcin Kaszkiel and Justin Zobel. Effective ranking with arbitrary passages. *JASIST*, 52(4):344–364, 2001. doi: 10.1002/1532-2890(2000)9999:9999<::AID-ASI1075>3.0.CO;2-%23.

[101] Dimitrios Kotsakos, Theodoros Lappas, Dimitrios Kotzias, Dimitrios Gunopulos, Nattiya Kanhabua, and Kjetil Nørvåg. A burstiness-aware approach for document dating. In *The 37th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '14, Gold Coast , QLD, Australia - July 06 - 11, 2014*, pages 1003–1006, 2014. doi: 10.1145/2600428.2609495.

[102] Wessel Kraaij, Thijs Westerveld, and Djoerd Hiemstra. The importance of prior probabilities for entry page search. In *SIGIR 2002: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 11-15, 2002, Tampere, Finland*, pages 27–34, 2002. doi: 10.1145/564376.564383.

[103] Anagha Kulkarni, Jaime Teevan, Krysta Marie Svore, and Susan T. Dumais. Understanding temporal query dynamics. In *Proceedings of the Forth International Conference on Web Search and Web Data Mining, WSDM 2011, Hong Kong, China, February 9-12, 2011*, pages 167–176, 2011. doi: 10.1145/1935826.1935862.

[104] Erdal Kuzey, Jilles Vreeken, and Gerhard Weikum. A fresh look on knowledge bases: Distilling named events from news. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, CIKM 2014, Shanghai, China, November 3-7, 2014*, pages 1689–1698, 2014. doi: 10.1145/2661829.2661984.

[105] Erdal Kuzey, Vinay Setty, Jannik Strötgen, and Gerhard Weikum. As time goes by: Comprehensive tagging of textual phrases with temporal scopes. In *Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada, April 11 - 15, 2016*, pages 915–925, 2016. doi: 10.1145/2872427.2883055.

[106] John Lafferty and Chengxiang Zhai. Probabilistic relevance models based on document and query generation. In *Language modeling for information retrieval*, pages 1–10. Springer, 2003.

[107] Sara Lana-Serrano, José Miguel Goñi-Menoyo, and José Carlos González Cristóbal. Miracle's 2005 approach to geographical information retrieval. 2005.

[108] Egoitz Laparra, Itziar Aldabe, and German Rigau. Document level time-anchoring for timeline extraction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 2: Short Papers*, pages 358–364, 2015.

[109] Mirella Lapata. Probabilistic text structuring: Experiments with sentence ordering. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, 7-12 July 2003, Sapporo Convention Center, Sapporo, Japan.*, pages 545–552, 2003.

[110] Victor Lavrenko, James Allan, Edward DeGuzman, Daniel LaFlamme, Veera Pollard, and Stephen Thomas. Relevance models for topic detection and tracking. In *Proceedings of the second international conference on Human Language Technology Research*, pages 115–121. Morgan Kaufmann Publishers Inc., 2002.

[111] Victor Lavrenko, Martin Choquette, and W. Bruce Croft. Cross-lingual relevance models. In *SIGIR 2002: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 11-15, 2002, Tampere, Finland*, pages 175–182, 2002. doi: 10.1145/564376.564408.

[112] E. L. Lawler and D. E. Wood. Branch-and-bound methods: A survey. *Operations Research*, 14(4):699–719, 1966. doi: 10.1287/opre.14.4.699.

[113] Quoc V. Le and Tomas Mikolov. Distributed representations of sentences and documents. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, pages 1188–1196, 2014.

[114] T. Leelanupab, M. Halvey, and J. M. Jose. Application and evaluation of multi-dimensional diversity. *Search*, (October):52–59, 2009.

[115] Xiaoyan Li and W. Bruce Croft. Time-based language models. In *Proceedings of the 2003 ACM CIKM International Conference on Information and Knowledge Management, New Orleans, Louisiana, USA, November 2-8, 2003*, pages 469–475, 2003. doi: 10.1145/956863.956951.

[116] Yanran Li and Sujian Li. Query-focused multi-document summarization: Combining a topic model with graph-based semi-supervised learning. In *COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, August 23-29, 2014, Dublin, Ireland*, pages 1197–1207, 2014.

[117] C. Y. Lin. Rouge: A package for automatic evaluation of summaries. In Stan Szpakowicz Marie-Francine Moens, editor, *Proceedings of the workshop on text summarization branches out (WAS 2004)*, number 1, pages 25–26, 2004.

[118] Jianhua Lin. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151, 1991. ISSN 0018-9448.

[119] Kenneth C. Litkowski. Summarization Experiments in DUC 2004. In *DUC*, 2004.

[120] Marina Litvak and Mark Last. Graph-based keyword extraction for single-document summarization. In *Multi-source Multilingual Information Extraction and Summarization*, 2008.

[121] Xiaoyong Liu and W. Bruce Croft. Passage retrieval based on language models. In *Proceedings of the 2002 ACM CIKM International Conference on Information and Knowledge Management, McLean, VA, USA, November 4-9, 2002*, pages 375–382, 2002. doi: 10.1145/584792.584854.

[122] Vebjorn Ljosa and Ambuj K. Singh. APLA: indexing arbitrary probability distributions. In *Proceedings of the 23rd International Conference on Data Engineering, ICDE 2007, The Marmara Hotel, Istanbul, Turkey, April 15-20, 2007*, pages 946–955, 2007. doi: 10.1109/ICDE.2007.367940.

[123] Elena Lloret, Laura Plaza, and Ahmet Aker. Analyzing the capabilities of crowdsourcing services for text summarization. *Language Resources and Evaluation*, 47 (2):337–369, 2013. doi: 10.1007/s10579-012-9198-8.

[124] Hans Peter Luhn. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(2):159–165, 1958. doi: 10.1147/rd.22.0159.

[125] Thomas Mandl, Fredric C. Gey, Giorgio Maria Di Nunzio, Nicola Ferro, Ray R. Larson, Mark Sanderson, Diana Santos, Christa Womser-Hacker, and Xing Xie. Geoclef 2007: The CLEF 2007 cross-language geographic information retrieval track overview. In *Advances in Multilingual and Multimodal Information Retrieval, 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Budapest, Hungary, September 19-21, 2007, Revised Selected Papers*, pages 745–772. 2007. doi: 10.1007/978-3-540-85760-0_96.

[126] I. Mani and M. Maybury. *Advances in automatic text summarization*, volume 8. MIT Press, 1999.

[127] Inderjeet Mani, David House, Gary Klein, Lynette Hirschman, Therese Firmin, and Beth Sundheim. The tipster summac text summarization evaluation. In *EACL 1999, 9th Conference of the European Chapter of the Association for Computational Linguistics, June 8-12, 1999, University of Bergen, Bergen, Norway*, pages 77–85, 1999. doi: 10.3115/977035.977047.

[128] Inderjeet Mani, Barry Schiffman, and Jianping Zhang. Inferring temporal ordering of events in news. In *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, HLT-NAACL 2003, Edmonton, Canada, May 27 - June 1, 2003*, 2003.

[129] Inderjeet Mani, Marc Verhagen, Ben Wellner, Chong Min Lee, and James Pustejovsky. Machine learning of temporal relations. In *ACL 2006, 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, Sydney, Australia, 17-21 July 2006*, 2006.

[130] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to information retrieval.* Cambridge University Press, 2008. ISBN 978-0-521-86571-5.

[131] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. The stanford corenlp natural language processing toolkit. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, System Demonstrations*, pages 55–60, 2014.

[132] Gary Marchionini. Exploratory search: from finding to understanding. *Communications of the ACM*, 49(4):41–46, 2006. ISSN 0001-0782.

[133] Alexander Markowetz, Yen-Yu Chen, Torsten Suel, Xiaohui Long, and Bernhard Seeger. Design and implementation of a geographic search engine. In *Proceedings of the Eight International Workshop on the Web & Databases (WebDB 2005), Baltimore, Maryland, USA, Collocated mith ACM SIGMOD/PODS 2005, June 16-17, 2005*, pages 19–24, 2005.

[134] Andrew McCallum, Kamal Nigam, et al. A comparison of event models for naive bayes text classification. In *AAAI-98 workshop on learning for text categorization*, volume 752, pages 41–48. Madison, WI, 1998.

[135] Michael McCandless, Erik Hatcher, and Otis Gospodnetic. *Lucene in Action, Second Edition: Covers Apache Lucene 3.0*. Manning Publications Co., 2010. ISBN 1933988177.

[136] Richard McCreadie, Craig Macdonald, and Iadh Ounis. Incremental update summarization: Adaptive sentence selection based on prevalence and novelty. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, CIKM 2014, Shanghai, China, November 3-7, 2014*, pages 301–310, 2014. doi: 10.1145/2661829.2661951.

[137] Ryan T. McDonald. A study of global inference algorithms in multi-document summarization. In *Advances in Information Retrieval, 29th European Conference on IR Research, ECIR 2007, Rome, Italy, April 2-5, 2007, Proceedings*, pages 557–564, 2007. doi: 10.1007/978-3-540-71496-5_51.

[138] Kathleen McKeown, Judith Klavans, Vasileios Hatzivassiloglou, Regina Barzilay, and Eleazar Eskin. Towards multidocument summarization by reformulation: Progress and prospects. pages 453–460, 1999.

[139] Qiaozhu Mei, Hui Fang, and ChengXiang Zhai. A study of poisson query generation model for information retrieval. In *SIGIR 2007: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, The Netherlands, July 23-27, 2007*, pages 319–326, 2007. doi: 10.1145/1277741.1277797.

[140] Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.*, pages 3111–3119, 2013.

[141] Arunav Mishra. Linking today's wikipedia and news from the past. In *Proceedings of the 7th Workshop on Ph.D Students, PIKM@CIKM 2014, Shanghai, China, November 3, 2014*, pages 1–8, 2014. doi: 10.1145/2663714.2668048.

[142] Arunav Mishra and Klaus Berberich. Exposé: Exploring past news for seminal events. In *Proceedings of the 24th International Conference on World Wide Web Companion, WWW 2015, Florence, Italy, May 18-22, 2015 - Companion Volume*, pages 223–226, 2015. doi: 10.1145/2740908.2742844.

[143] Arunav Mishra and Klaus Berberich. Event digest: A holistic view on past events. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, SIGIR 2016, Pisa, Italy, July 17-21, 2016*, pages 493–502, 2016. doi: 10.1145/2911451.2911526.

[144] Arunav Mishra and Klaus Berberich. Estimating time models for news article excerpts. In *Proceedings of the 25th ACM International Conference on Information and Knowledge Management, CIKM 2016, Indianapolis, IN, USA, October 24-28, 2016*, pages 781–790, 2016. doi: 10.1145/2983323.2983802.

[145] Arunav Mishra and Klaus Berberich. Leveraging semantic annotations to link wikipedia and news archives. In *Advances in Information Retrieval - 38th European Conference on IR Research, ECIR 2016, Padua, Italy, March 20-23, 2016. Proceedings*, pages 30–42, 2016. doi: 10.1007/978-3-319-30671-1_3.

[146] Arunav Mishra and Klaus Berberich. How do order and proximity impact the readability of event summaries? In *Advances in Information Retrieval - 39th European Conference on IR Research, ECIR 2017, Aberdeen, UK, April 8-13, 2017, Proceedings*, pages 212–225, 2017. doi: 10.1007/978-3-319-56608-5_17.

[147] Arunav Mishra, Dragan Milchevski, and Klaus Berberich. Linking wikipedia events to past news. In *SIGIR 2014 workshop on temporal, social and spatially-aware information access (TAIA2014). ACM*, 2014.

[148] Bhaskar Mitra, Fernando Diaz, and Nick Craswell. Learning to match using local and distributed representations of text for web search. In *Proceedings of the 26th International Conference on World Wide Web, WWW 2017, Perth, Australia, April 3-7, 2017*, pages 1291–1299, 2017. doi: 10.1145/3038912.3052579.

[149] Saeedeh Momtazi, Friedrich Faubel, and Dietrich Klakow. Within and across sentence boundary language model. In *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26-30, 2010*, pages 1800–1803, 2010.

[150] Greg Myers. *The discourse of blogs and wikis.* A&C Black, 2010. ISBN 1847064140.

[151] Ani Nenkova and Rebecca J. Passonneau. Evaluating content selection in summarization: The pyramid method. In *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, HLT-NAACL 2004, Boston, Massachusetts, USA, May 2-7, 2004*, pages 145–152, 2004.

[152] Andrew Y. Ng, Alice X. Zheng, and Michael I. Jordan. Link analysis, eigenvectors and stability. In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence, IJCAI 2001, Seattle, Washington, USA, August 4-10, 2001*, pages 903–910, 2001.

[153] Daan Odijk, Ork De Rooij, Maria-hendrike Peetz, Toine Pieters, Maarten De Rijke, and Stephen Snelders. Semantic Document Selection. In *TPDL 2012: Theory and Practice of Digital Libraries*, 2012.

[154] Daan Odijk, Giuseppe Santucci, Maarten de Rijke, Marco Angelini, and Guido Lorenzo Granato. Time-aware exploratory search: Exploring word meaning through time. In *SIGIR 2012 Workshop on Time-aware Information Access, Portland, OR, USA*, 2012.

[155] Paul Ogilvie and James P. Callan. Experiments using the lemur toolkit. In *Proceedings of The Tenth Text REtrieval Conference, TREC 2001, Gaithersburg, Maryland, USA, November 13-16, 2001*, 2001.

[156] Naoaki Okazaki, Yutaka Matsuo, and Mitsuru Ishizuka. Improving chronological sentence ordering by precedence relation. In *COLING 2004, 20th International Conference on Computational Linguistics, Proceedings of the Conference, 23-27 August 2004, Geneva, Switzerland*, 2004.

[157] Naoaki Okazaki, Yutaka Matsuo, and Mitsuru Ishizuka. Improving chronological ordering of sentences extracted from multiple newspaper articles. *ACM Transactions on Asian Language Information Processing (TALIP)*, 4(3):321–339, 2005. ISSN 15300226. doi: 10.1145/1111667.1111673.

[158] Kenji Ono, Kazuo Sumita, and Seiji Miike. Abstract generation based on rhetorical structure extraction. In *15th International Conference on Computational Linguistics, COLING 1994, Kyoto, Japan, August 5-9, 1994*, pages 344–348, 1994.

[159] Liang Pang, Yanyan Lan, Jiafeng Guo, Jun Xu, and Xueqi Cheng. A study of match pyramid models on ad-hoc retrieval. *CoRR*, abs/1606.04648, 2016.

[160] Liang Pang, Yanyan Lan, Jiafeng Guo, Jun Xu, Shengxian Wan, and Xueqi Cheng. Text matching as image recognition. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA.*, pages 2793–2799, 2016.

[161] Maria-Hendrike Peetz and Maarten de Rijke. Cognitive temporal document priors. In *Advances in Information Retrieval - 35th European Conference on IR Research, ECIR 2013, Moscow, Russia, March 24-27, 2013. Proceedings*, pages 318–330, 2013. doi: 10.1007/978-3-642-36973-5_27.

[162] Maria-Hendrike Peetz, Edgar Meij, and Maarten de Rijke. Using temporal bursts for query modeling. *Inf. Retr.*, 17(1):74–108, 2014. doi: 10.1007/s10791-013-9227-2.

[163] José M. Perea-Ortega and Luis Alfonso Ureña López. Geographic expansion of queries to improve the geographic information retrieval task. In *Natural Language Processing and Information Systems - 17th International Conference on Applications of Natural Language to Information Systems, NLDB 2012, Groningen, The Netherlands, June 26-28, 2012. Proceedings*, pages 94–103, 2012. doi: 10.1007/978-3-642-31178-9_9.

[164] José M. Perea-Ortega, Fernando Martínez Santiago, Arturo Montejo Ráez, and Luis Alfonso Ureña López. Geo-ner: un reconocedor de entidades geográficas para inglés basado en geonames y wikipedia. *Procesamiento del Lenguaje Natural*, 43:33–40, 2009.

[165] Jay M. Ponte and W. Bruce Croft. A language modeling approach to information retrieval. In *SIGIR 1998: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 24-28 1998, Melbourne, Australia*, pages 275–281, 1998. doi: 10.1145/290941.291008.

[166] Jeffrey Pound, Peter Mika, and Hugo Zaragoza. Ad-hoc object retrieval in the web of data. In *Proceedings of the 19th International Conference on World Wide Web, WWW 2010, Raleigh, North Carolina, USA, April 26-30, 2010*, pages 771–780, 2010. doi: 10.1145/1772690.1772769.

[167] James Pustejovsky, José M. Castaño, Robert Ingria, Roser Saurí, Robert J. Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir R. Radev. Timeml: Robust specification of event and temporal expressions in text. pages 28–34, 2003.

[168] Dragomir R. Radev, Eduard H. Hovy, and Kathleen McKeown. Introduction to the special issue on summarization. *Computational Linguistics*, 28(4):399–408, 2002. doi: 10.1162/089120102762671927.

[169] Dragomir R. Radev, Simone Teufel, Horacio Saggion, Wai Lam, John Blitzer, Hong Qi, Arda Çelebi, Danyu Liu, and Elliott Drábek. Evaluation challenges in large-scale document summarization. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, 7-12 July 2003, Sapporo Convention Center, Sapporo, Japan.*, pages 375–382, 2003.

[170] Jinfeng Rao, Hua He, and Jimmy J. Lin. Noise-contrastive estimation for answer selection with deep neural networks. In *Proceedings of the 25th ACM International Conference on Information and Knowledge Management, CIKM 2016, Indianapolis, IN, USA, October 24-28, 2016*, pages 1913–1916, 2016. doi: 10.1145/2983323.2983872.

[171] Ridho Reinanda, Daan Odijk, and Maarten de Rijke. Exploring entity associations over time. *SIGIR 2013 Workshop on Time-aware Information Access (#TAIA2013)*, 2013.

[172] Korbinian Riedhammer, Benoît Favre, and Dilek Hakkani-Tür. Long story short - global unsupervised models for keyphrase based meeting summarization. *Speech Communication*, 52(10):801–815, 2010. doi: 10.1016/j.specom.2010.06.002.

[173] Van Rijsbergen and Cornelis Joost. A theoretical basis for the use of co-occurrence data in information retrieval. *Journal of documentation*, 33(2):106–119, 1977.

[174] Stephen E. Robertson. The probability ranking principle in ir. *Journal of documentation*, 33(4):294–304, 1977.

[175] Stephen E. Robertson and Karen Spärck Jones. Relevance weighting of search terms. *JASIS*, 27(3):129–146, 1976. doi: 10.1002/asi.4630270302.

[176] Joseph John Rocchio. Relevance feedback in information retrieval. *The Smart retrieval system-experiments in automatic document processing*, 1971.

[177] Gerard Salton. *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer.* Addison-Wesley, 1989. ISBN 0-201-12227-8.

[178] Gerard Salton and Chris Buckley. Automatic text structuring and retrieval: Experiments in automatic encyclopedia searching. In *Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Chicago, Illinois, USA, October 13-16, 1991 (Special Issue of the SIGIR Forum).*, pages 21–30, 1991. doi: 10.1145/122860.122863.

[179] Gerard Salton, Chung-Shu Yang, and Clement T. Yu. A theory of term importance in automatic text analysis. *JASIS*, 26(1):33–44, 1975. doi: 10.1002/asi.4630260106.

[180] Gerard Salton, James Allan, and Chris Buckley. Approaches to passage retrieval in full text information systems. In *Proceedings of the 16th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval. Pittsburgh, PA, USA, June 27 - July 1, 1993*, pages 49–58, 1993. doi: 10.1145/160688.160693.

[181] Gerard Salton, Amit Singhal, Mandar Mitra, and Chris Buckley. Automatic text structuring and summarization. *Information Processing & Management*, 33(2): 193–207, 1997. ISSN 03064573. doi: 10.1016/S0306-4573(96)00062-3.

[182] Christina Sauper and Regina Barzilay. Automatically generating wikipedia articles: A structure-aware approach. In *ACL 2009, Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP, 2-7 August 2009, Singapore*, pages 208–216, 2009.

[183] Vinay Setty, Abhijit Anand, Arunav Mishra, and Avishek Anand. Modeling event importance for ranking daily news events. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining, WSDM 2017, Cambridge, United Kingdom, February 6-10, 2017*, pages 231–240, 2017. doi: 10.1145/3018661.3018728.

[184] Dafna Shahaf, Jaewon Yang, Caroline Suen, Jeff Jacobs, Heidi Wang, and Jure Leskovec. Information Cartography : Creating Zoomable , Large-Scale Maps of Information. pages 1097–1105, 2013. ISSN 1450321747. doi: 10.1145/2487575.2487690.

[185] Xuehua Shen, Bin Tan, and ChengXiang Zhai. Context-sensitive information retrieval using implicit feedback. In *SIGIR 2005: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Salvador, Brazil, August 15-19, 2005*, pages 43–50, 2005. doi: 10.1145/1076034.1076045.

[186] Yelong Shen, Xiaodong He, Jianfeng Gao, Li Deng, and Grégoire Mesnil. Learning semantic representations using convolutional neural networks for web search. In *23rd International World Wide Web Conference, WWW '14, Seoul, Republic of Korea, April 7-11, 2014, Companion Volume*, pages 373–374, 2014. doi: 10.1145/2567948.2577348.

[187] Luo Si, Rong Jin, James P. Callan, and Paul Ogilvie. A language modeling framework for resource selection and results merging. In *Proceedings of the 2002 ACM CIKM International Conference on Information and Knowledge Management, McLean, VA, USA, November 4-9, 2002*, pages 391–397, 2002. doi: 10.1145/584792.584856.

[188] Amit Singhal, Chris Buckley, and Mandar Mitra. Pivoted document length normalization. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'96, August 18-22, 1996, Zurich, Switzerland (Special Issue of the SIGIR Forum)*, pages 21–29, 1996. doi: 10.1145/243199.243206.

[189] Fei Song and W. Bruce Croft. A general language model for information retrieval. In *Proceedings of the 1999 ACM CIKM International Conference on Information and Knowledge Management, Kansas City, Missouri, USA, November 2-6, 1999*, pages 316–321, 1999. doi: 10.1145/319950.320022.

[190] Alexander Sorokin and David Forsyth. Utility data annotation with Amazon Mechanical Turk. *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops*, 51(c):1–8, 2008. ISSN 2160-7508. doi: 10.1109/CVPRW.2008.4562953.

[191] Josef Steinberger and Karel Jezek. Evaluation measures for text summarization. *Computing and Informatics*, 28(2):251–275, 2009.

[192] Jannik Strötgen and Mich ael Gertz. WikiWarsDE: A German Corpus of Narratives Annotated with Temporal Expressions. In *Proceedings of the Conference of the German Society for Computational Linguistics and Language Technology (GSCL)*, volume 2, pages 129–134, 2011.

[193] Jannik Strötgen and Michael Gertz. Multilingual and cross-domain temporal tagging. *Language Resources and Evaluation*, 47(2):269–298, 2013. doi: 10.1007/s10579-012-9179-y.

[194] Jannik Strötgen, Omar Alonso, and Michael Gertz. Identification of top relevant temporal expressions in documents. In *2nd Temporal Web Analytics Workshop, TempWeb '12, Lyon, France, April 16-17, 2012*, pages 33–40, 2012. doi: 10.1145/2169095.2169102.

[195] Martin Sundermeyer, Ralf Schlüter, and Hermann Ney. LSTM neural networks for language modeling. In *INTERSPEECH 2012, 13th Annual Conference of the International Speech Communication Association, Portland, Oregon, USA, September 9-13, 2012*, pages 194–197, 2012.

[196] Russell Swan and David Jensen. TimeMines: Constructing Timelines with Statistical Models of Word Usage. In *ACM SIGKDD 2000 Workshop on Text Mining*, pages 73–80, 2000.

[197] Bin Tan, Xuehua Shen, and ChengXiang Zhai. Mining long-term search history to improve search accuracy. In *Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, PA, USA, August 20-23, 2006*, pages 718–723, 2006. doi: 10.1145/1150402.1150493.

[198] Bilyana Taneva and Gerhard Weikum. Gem-based entity-knowledge maintenance. In *22nd ACM International Conference on Information and Knowledge Management, CIKM'13, San Francisco, CA, USA, October 27 - November 1, 2013*, pages 149–158, 2013. doi: 10.1145/2505515.2505715.

[199] Jiayu Tang and Mark Sanderson. Evaluation and user preference study on spatial diversity. In *Advances in Information Retrieval, 32nd European Conference on IR Research, ECIR 2010, Milton Keynes, UK, March 28-31, 2010. Proceedings*, pages 179–190, 2010. doi: 10.1007/978-3-642-12275-0_18.

[200] Tao Tao and ChengXiang Zhai. Regularized estimation of mixture models for robust pseudo-relevance feedback. In *SIGIR 2006: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Seattle, Washington, USA, August 6-11, 2006*, pages 162–169, 2006. doi: 10.1145/1148170.1148201.

[201] Tao Tao and ChengXiang Zhai. An exploration of proximity measures in information retrieval. In *SIGIR 2007: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, The Netherlands, July 23-27, 2007*, pages 295–302, 2007. doi: 10.1145/1277741.1277794.

[202] Manos Tsagkias, Maarten de Rijke, and Wouter Weerkamp. Linking online news and social media. In *Proceedings of the Forth International Conference on Web Search and Web Data Mining, WSDM 2011, Hong Kong, China, February 9-12, 2011*, pages 565–574, 2011. doi: 10.1145/1935826.1935906.

[203] Howard Turtle and W. Bruce Croft. Evaluation of an inference network-based retrieval model. *ACM Transactions on Information Systems*, 9(3):187–222, 1991. ISSN 10468188. doi: 10.1145/125187.125188.

[204] Naushad UzZaman, Hector Llorens, James F. Allen, Leon Derczynski, Marc Verhagen, and James Pustejovsky. Tempeval-3: Evaluating events, time expressions, and temporal relations. *CoRR*, abs/1206.5333, 2012.

[205] Marc Verhagen and James Pustejovsky. The TARSQI toolkit. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey, May 23-25, 2012*, pages 2043–2048, 2012.

[206] Xuanhui Wang, Hui Fang, and ChengXiang Zhai. A study of methods for negative relevance feedback. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2008, Singapore, July 20-24, 2008*, pages 219–226, 2008. doi: 10.1145/1390334.1390374.

[207] Ross Wilkinson and Alan F. Smeaton. Automatic link generation. *ACM Computing Surveys*, 31(4es):27–es, 1999. ISSN 03600300. doi: 10.1145/345966.346024.

[208] Kristian Woodsend and Mirella Lapata. Wikisimple: Automatic simplification of wikipedia articles. In *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2011, San Francisco, California, USA, August 7-11, 2011*, 2011.

[209] Kristian Woodsend and Mirella Lapata. Multiple aspect summarization using integer linear programming. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, EMNLP-CoNLL 2012, July 12-14, 2012, Jeju Island, Korea*, pages 233–243, 2012.

[210] Yu Xu, Yusi Liu, and Li Qi. Examining collective memory building in wikipedia: A multilevel network approach. In *50th Hawaii International Conference on System Sciences, HICSS 2017, Hilton Waikoloa Village, Hawaii, USA, January 4-7, 2017*, 2017. doi: 10.24251/HICSS.2017.259.

[211] Mohamed Yahya, Klaus Berberich, Shady Elbassuoni, and Gerhard Weikum. Robust question answering over the web of linked data. In *22nd ACM International Conference on Information and Knowledge Management, CIKM'13, San Francisco, CA, USA, October 27 - November 1, 2013*, pages 1107–1116, 2013. doi: 10.1145/2505515.2505677.

[212] Rui Yan, Xiaojun Wan, Jahna Otterbacher, Liang Kong, Xiaoming Li, and Yan Zhang. Evolutionary timeline summarization: a balanced optimization framework via iterative substitution. In *Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2011, Beijing, China, July 25-29, 2011*, pages 745–754, 2011. doi: 10.1145/2009916.2010016.

[213] Conglei Yao, Xu Jia, Sicong Shou, Shicong Feng, Feng Zhou, and Hongyan Liu. Autopedia: automatic domain-independent wikipedia article generation. In *Proceedings of the 20th International Conference on World Wide Web, WWW 2011, Hyderabad, India, March 28 - April 1, 2011 (Companion Volume)*, pages 161–162, 2011. doi: 10.1145/1963192.1963274.

[214] ChengXiang Zhai. Statistical language models for information retrieval. 1, 2008.

[215] Chengxiang Zhai and John Lafferty. A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems (TOIS)*, 22(2):179–214, 2004. ISSN 1046-8188.

[216] ChengXiang Zhai and John D. Lafferty. Model-based feedback in the language modeling approach to information retrieval. In *Proceedings of the 2001 ACM CIKM International Conference on Information and Knowledge Management, Atlanta, Georgia, USA, November 5-10, 2001*, pages 403–410, 2001. doi: 10.1145/502585.502654.

[217] ChengXiang Zhai and John D. Lafferty. Two-stage language models for information retrieval. In *SIGIR 2002: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 11-15, 2002, Tampere, Finland*, pages 49–56, 2002. doi: 10.1145/564376.564387.

[218] ChengXiang Zhai, William W. Cohen, and John D. Lafferty. Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. In *SIGIR 2003: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, July 28 - August 1, 2003, Toronto, Canada*, pages 10–17, 2003. doi: 10.1145/860435.860440.

[219] Ruiqiang Zhang, Yuki Konda, Anlei Dong, Pranam Kolari, Yi Chang, and Zhaohui Zheng. Learning recurrent event queries for web search. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, EMNLP 2010, 9-11 October 2010, MIT Stata Center, Massachusetts, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1129–1139, 2010.

[220] Yue Zhao and Claudia Hauff. Sub-document timestamping of web documents. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, Santiago, Chile, August 9-13, 2015*, pages 1023–1026, 2015. doi: 10.1145/2766462.2767803.

[221] Sheng Hua Zhong, Yan Liu, Bin Li, and Jing Long. Query-oriented unsupervised multi-document summarization via deep learning model. *Expert Systems with Applications*, 42(21):8146–8155, 2015. ISSN 09574174. doi: 10.1016/j.eswa.2015.05.034.

[222] Dengyong Zhou, Olivier Bousquet, Thomas Navin Lal, Jason Weston, and Bernhard Schölkopf. Learning with local and global consistency. In *Advances in Neural Information Processing Systems 16 [Neural Information Processing Systems, NIPS 2003, December 8-13, 2003, Vancouver and Whistler, British Columbia, Canada]*, pages 321–328, 2003.

[223] Yinghua Zhou, Xing Xie, Chuang Wang, Yuchang Gong, and Wei-Ying Ma. Hybrid index structures for location-based web search. In *Proceedings of the 2005 ACM CIKM International Conference on Information and Knowledge Management, Bremen, Germany, October 31 - November 5, 2005*, pages 155–162, 2005. doi: 10.1145/1099554.1099584.

[224] Xiaojin Zhu and Zoubin Ghahramani. Learning from Labeled and Unlabeled Data with Label Propagation. *Technical Report CMU-CALD-02-107, Carnegie Mellon University*, 2002. doi: 10.1.1.13.8280.

[225] Xiaojin Zhu, Zoubin Ghahramani, and John D. Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In *Machine Learning, Proceedings of the Twentieth International Conference (ICML 2003), August 21-24, 2003, Washington, DC, USA*, pages 912–919, 2003.

Appendix A

Query Workloads

A.1 Wikipedia Year Page Events

ID	Date	Event
1	Jan 3 1987	Aretha Franklin becomes the first woman inducted into the Rock and Roll Hall of Fame.
2	Oct 11 1987	The first National Coming Out Day is held in celebration of the second National March on Washington for Lesbian and Gay Rights.
3	Jan 1 1988	The Soviet Union begins its program of economic restructuring (perestroika) with legislation initiated by Premier Mikhail Gorbachev (though Gorbachev had begun minor restructuring in 1985).
4	May 15 1988	Soviet war in Afghanistan: After more than 8 years of fighting, the Soviet Army begins withdrawing from Afghanistan.
5	Jun 11 1988	Wembley Stadium hosts a concert featuring stars from the fields of music, comedy and film, in celebration of the 70th birthday of imprisoned ANC leader Nelson Mandela.
6	Aug 18 1988	The Republican National Convention in New Orleans, Louisiana nominates George H. W. Bush for President and Dan Quayle for Vice President of the United States of America.
7	Jan 15 1989	Thirty-five European nations, meeting in Vienna, agree to strengthen human rights and strengthen East-West trade.
8	May 3 1989	Cold War - Perestroika - The first McDonald's restaurant in the USSR begins construction in Moscow. It will open on 31 January 1990.
9	Jul 28 1989	In the Iranian presidential election, electors overwhelmingly elect Akbar Hashemi Rafsanjani as President of Iran and endorse changes to the Constitution of the Islamic Republic of Iran, increasing the powers of the president.

10	Feb 13 1990	German reunification: An agreement is reached for a two-stage plan to reunite Germany.
11	Apr 6 1990	Robert Mapplethorpe's "The Perfect Moment" show of nude and homoerotic photographs opens at the Cincinnati Contemporary Arts Center, in spite of accusations of indecency by Citizens for Community Values.
12	Jun 19 1990	The Communist Party of the Russian Soviet Federative Socialist Republic is founded in Moscow.
13	Jul 6 1990	Somali president Siad Barre's bodyguards massacre antigovernment demonstrators during a soccer match; 65 people are killed, more than 300 seriously injured.
14	Sep 6 1990	In Burma, the State Law and Order Restoration Council orders the arrest of Aung San Suu Kyi and five other political dissidents.
15	Oct 8 1990	Israeli-Palestinian conflict: In Jerusalem, Israeli police kill 17 Palestinians and wound over 100 near the Dome of the Rock mosque on the Temple Mount.
16	Feb 22 1991	Gulf War: Iraq accepts a Soviet-proposed cease fire agreement. The U.S. rejects the agreement, but says that retreating Iraqi forces will not be attacked if they leave Kuwait within 24 hours.
17	May 13 1991	Winnie Mandela is convicted of kidnapping. On May 14, she is sentenced to 6 years in prison.
18	Jan 18 1992	In Nairobi, Kenya, more than 100,000 attend protests demanding an end to one-party rule by the Kenya African National Union.
19	Mar 11 1992	Manuel de Dios Unanue, former editor of El Diario La Prensa, is slain in a restaurant in Queens, New York after having received death threats from the Colombian drug cartels.
20	Mar 25 1992	The International Atomic Energy Agency orders Iraq to destroy an industrial complex at Al Atheer that is being used to manufacture nuclear weapons.
21	Apr 27 1992	Betty Boothroyd becomes the first woman elected Speaker of the British House of Commons.

22	Apr 28 1992	The two remaining constituent republics of the former Socialist Federal Republic of Yugoslavia Ð Serbia and Montenegro Ð form a new state, named the Federal Republic of Yugoslavia (after 2003, Serbia and Montenegro), bringing to an end the official union of Serbs, Croats, Slovenes, Montenegrins, Bosnian Muslims and Macedonians that existed from 1918 (with the exception of the period during World War II).
23	Jul 19 1992	A car bomb placed by the Mafia (with the collaboration of Italian intelligence) kills judge Paolo Borsellino and 5 members of his escort.
24	Aug 21 1992	Events at Ruby Ridge, Idaho, are sparked by a Federal Marshal surveillance team, resulting in the death of a Marshal, Sam Weaver and his dog and the next day the wounding of Randy Weaver, the death of his wife Vicki and the wounding of Kevin Harris.
25	Sep 12 1992	STS-47: Dr. Mae Jemison becomes the first African American woman to travel into space, aboard the Space Shuttle Endeavour.
26	Dec 24 1992	President George H. W. Bush pardons 6 national security officials implicated in the Iran-Contra affair, including Caspar Weinberger.
27	Aug 4 1993	A federal judge sentences Los Angeles Police Department officers Stacey Koon and Laurence Powell to 30 months in prison for violating motorist Rodney King's civil rights.
28	Oct 21 1993	A coup in Burundi results in the death of president Melchior Ndadaye and sparks the Burundi Civil War.
29	Feb 21 1995	Steve Fossett lands in Leader, Saskatchewan, Canada, becoming the first person to make a solo flight across the Pacific Ocean in a balloon.
30	Mar 30 1995	A police officer tries to assassinate Takaji Kunimatsu, chief of the National Police Agency of Japan.
31	Dec 20 1995	HM The Queen advises "an early divorce" to Lady Diana Spencer and Charles, Prince of Wales. The divorce is finalized on 28 August 1996.
32	Apr 3 1997	The Thalit massacre in Algeria: All but one of the 53 inhabitants of Thalit are killed by guerrillas.
33	May 22 1997	Kelly Flinn, the U.S. Air Force's first female bomber pilot certified for combat, accepts a general discharge in order to avoid a court martial.
34	Dec 19 1998	The U.S. House of Representatives forwards articles of impeachment against President Clinton to the Senate, making him the second president to be impeached in the nation.

35	Sep 12 1999	Under international pressure to allow an international peacekeeping force, Indonesian president BJ Habibie announces that he will do so.
36	Nov 9 1999	TAESA Flight 725, covering the route Tijuana-Guadalajara-Uruapan-Mexico City, crashes a few minutes after takeoff from Uruapan International Airport, killing 18 people on board. This event causes the bankruptcy of the Mexican airline a few months later.
37	Sep 16 2000	Ukrainian journalist Georgiy Gongadze is last seen alive; this day is taken as the commemoration date of his death.
38	Jul 3 2001	A Vladivostokavia Tupolev Tu-154 jetliner crashes on approach to landing at Irkutsk, Russia, killing 145.
39	May 28 2003	Prometea, the first horse cloned by Italian scientists, is born.
40	Feb 27 2004	2004 SuperFerry 14 bombing: The Abu Sayyaf guerrilla group is blamed for the deadliest terrorist attack at sea in world history, which kills 116 in the Philippines.
41	Oct 27 2005	The 2005 French riots begin after 2 young immigrants die in Clichy-sous-Bois, while hiding from the police.
42	Oct 29 2005	At least 61 people are killed and many others wounded in 3 powerful blasts in the Indian capital, Delhi.
43	Jan 4 2006	Ariel Sharon, Prime Minister of Israel, suffers a severe stroke and cerebral hemorrhage.
44	Mar 16 2006	The United Nations General Assembly votes overwhelmingly to establish the United Nations Human Rights Council.
45	Apr 26 2007	Bronze Night: Russians riot in the city of Tallinn, Estonia, about moving the Bronze Soldier war memorial, a Soviet World War II memorial. One person is killed after two of the worst nights of rioting in Estonian history.
46	Feb 27 1991	President Bush declares victory over Iraq and orders a cease-fire.
47	Jan 9 1992	Bosnian Serbs declare their own republic within Bosnia and Herzegovina, in protest of the decision by Bosniaks and Bosnian Croats to seek EC recognition.
48	Jun 15 1995	A powerful earthquake, registering a moment magnitude of 6.2, hits the city of Aigio, Greece, resulting in several deaths and significant damage to many buildings.
49	Aug 6 2005	Tuninter Flight 1153 is ditched due to engine failure; 16 are killed.
50	Mar 19 2002	US war in Afghanistan: Operation Anaconda ends after killing 500 Taliban and Al-Qaeda fighters, with 11 allied troop fatalities.

A.2 Wikipedia Excerpts to New York Times Articles

ID	Event
1	Beavis and Butt-head: In February 1994, watchdog group Morality in Media claimed that the death of 8-month-old Natalia Rivera, struck by a bowling ball thrown from an overpass onto a Jersey City, New Jersey highway near the Holland Tunnel by 18-year-old Calvin J. Settle, was partially inspired by Beavis and Butt-Head .
2	Enrico Fermi: While at Columbia during World War II, Fermi and his wife resided in Leonia, New Jersey.
3	Guam: The United States returned and fought the Battle of Guam on July 21, 1944, to recapture the island from Japanese military occupation. More than 18,000 Japanese were killed as only 485 surrendered. Sergeant Shoichi Yokoi, who surrendered in January 1972, appears to have been the last confirmed Japanese holdout in Guam.
4	History of Honduras: As the November 1985 election approached, the PLH could not settle on a presidential candidate and interpreted election law as permitting multiple candidates from any one party. The PLH claimed victory when its presidential candidates collectively outpolled the PNH candidate, Rafael Leonardo Callejas, who received 42 % of the total vote. JosE Azcona, the candidate receiving the most votes (27 %) among the PLH, assumed the presidency in January 1986. With strong endorsement and support from the Honduran military, the Suazo Administration ushered in the first peaceful transfer of power between civilian presidents in more than 30 years. In 1989 he oversaw the dismantling of Contras which were based in Honduras.
5	Pervez Musharraf: Pakistan and Turkey Musharraf and his family left for Pakistan on one of the last safe trains in August 1947, a few days before the partition of India took effect.
6	Revolutionary Armed Forces of Colombia: On May 5, 2003, the FARC assassinated the governor of Antioquia, Guillermo Gaviria Correa, his advisor for peace, former defense minister Gilberto Echeverri MejIa, and 8 soldiers. The FARC had kidnapped Mr. Gaviria and Mr. Echeverri a year earlier, when the 2 men were leading a march for peace from MedellIn to Caicedo in Antioquia.
7	Closings and cancellations following the September 11 attacks: After a few switching delays at 96th Street, service was changed on September 19. The train resumed local service in Manhattan, but was extended to New Lots Avenue in Brooklyn (switching onto the express tracks at Chambers Street) to replace the 3, which now terminated at 14th Street as an express. The train continued to make local stops in Manhattan and service between Chambers Street and South Ferry as well as skip-stop service remained suspended. Normal service on all four trains was restored September 15, 2002, but Cortlandt Street will remain closed while the World Trade Center site is redeveloped.
8	Santa Monica, California: In October 1998, alleged Culver City 13 gang member Omar Sevilla, 21, of Culver City was killed.

9	Shining Path: Despite these arrests, the Shining Path continues to exist in Peru . On December 22, 2005, the Shining Path ambushed a police patrol in the Hu_nuco region, killing eight.
10	Sun Myung Moon: 1990s In April 1990 Moon visited the Soviet Union and met with President Mikhail Gorbachev. Moon expressed support for the political and economic transformations under way in the Soviet Union. At the same time the Unification Church was expanding into formerly communist nations.
11	T. S. Eliot: On January 10, 1957, Eliot at the age of 68, married EsmE Valerie Fletcher, who was 32. In contrast to his first marriage, Eliot knew Fletcher well, as she had been his secretary at Faber and Faber since August, 1949. They kept their wedding secret ; the ceremony was held in a church at 6:15 A.M., with virtually no one in attendance other than his wife's parents. Since Eliot's death, Valerie has dedicated her time to preserving his legacy ; she has edited and annotated The Letters of T. S. Eliot and a facsimile of the draft of The Waste Land. Eliot never had children with either of his wives. In the early 1960s, by then in failing health, Eliot worked as an editor for the Wesleyan University Press, seeking new poets in Europe for publication.
12	Gun control: The NAACP lawsuit was dismissed in 2003.
13	Willy Brandt: Hostages in Iraq One of Brandt's last public appearances was in flying to Baghdad, Iraq, to free Western hostages held by Saddam Hussein, following the Iraqi invasion of Kuwait in 1990. Brandt secured the release of a large number of them, and on November 9, 1990, his airplane landed with 174 freed hostages on board at the Frankfurt Airport.
14	Abbie Hoffman: Back to visibility In November 1986, Hoffman was arrested along with fourteen others, including Amy Carter, the daughter of former President Jimmy Carter, for trespassing at the University of Massachusetts at Amherst .
15	Sevastopol: On July 10, 1993, the Russian parliament passed a resolution declaring Sevastopol to be a federal Russian city . At the time, many supporters of the president, Boris Yeltsin, had ceased taking part in the Parliament's work.
16	Missionary: The Muslim population of the US has increased greatly in the last one hundred years, with much of the growth driven by widespread conversion.
17	Barbara McClintock: McClintock completed her secondary education at Erasmus Hall High School in Brooklyn.
18	Charles Addams: Death Addams died September 29, 1988, at St. Clare's Hospital and Health Center in New York City, having suffered a heart attack while still in his car after parking it . An ambulance took him from his apartment to the hospital, where he died in the emergency room.

19	Mobutu Sese Seko: On 12 May 1997, as Laurent-DEsirE Kabila's Alliance of Democratic Forces for the Liberation of Congo rebels were advancing on Gbadolite, Mobutu had the remains flown by cargo plane from his mausoleum to Kinshasa where they waited on the tarmac of N'djili Airport for three days. On 16 May, the day before Mobutu fled Zaire (and the country was renamed the Democratic Republic of the Congo), Habyarimana's remains were burned under the supervision of an Indian Hindu leader.
20	John Zorn: Masada Books John Zorn recorded Kristallnacht in November 1992, his premiere work of radical Jewish culture, featuring a suite of seven compositions reflecting the infamous Night of Broken Glass in late 1938 where Jews were targets of violence and destruction in Germany and Austria.
21	Monmouth County, New Jersey: At the June 28, 1778 Battle of Monmouth, near Freehold, General George Washington's soldiers battled the British under Sir Henry Clinton, in the longest land battle of the American Revolutionary War. It was at Monmouth that the tactics and training from Friedrich Wilhelm von Steuben developed at Valley Forge during the winter encampment were first implemented on a large scale.
22	Galloway Township, New Jersey: The Garden State Parkway passes through the township. It was on this stretch of the Parkway that Governor of New Jersey Jon Corzine was involved in a serious accident on April 12, 2007.
23	Moonachie, New Jersey: The name of the borough is typically pronounced moo-NAH-kee ; however, in January 1987, then-Mayor of New York City Ed Koch pronounced it mah-NOO-chee when he made his now-famous quip that the New York Giants should hold their victory parade in the borough after the team had just won Super Bowl XXI. Koch had refused to grant the Giants permission to hold a parade within the city limits because the team plays its home games in New Jersey, not in New York City.
24	Teaneck, New Jersey: As de facto racial segregation increased, so did tensions between residents of the northeast and members of the predominantly white male Teaneck Police Department . On the evening of April 10, 1990, the Teaneck Police Department responded to a call from a resident complaining about a teenager with a gun. After an initial confrontation near Bryant School and a subsequent chase, Phillip Pannell, an African American teenager, was shot and killed by Gary Spath, a white Teaneck police officer . Spath said he thought Pannell had a gun and was turning to shoot him . Witnesses said Pannell was unarmed and had been shot in the back. Protest marches, some violent, ensued ; most African Americans believed that Pannell had been killed in cold blood, while other residents insisted that Spath had been justified in his actions. Testimony at the trial claimed that Pannell was shot in the back, and that he was carrying a gun. A fully loaded.22 caliber pistol was recovered from Pannell's jacket pocket. The gun, originally a starter's pistol, had been modified into an operable weapon that was loaded with eight cartridges.

25	Glassboro, New Jersey: The Glassboro Summit Conference between U.S. President Lyndon B. Johnson and Soviet Premier Alexei Kosygin took place in Glassboro. Johnson and Kosygin met for three days from June 23 to June 25, 1967, at Glassboro State College (later renamed Rowan University). The location was chosen as a compromise. Kosygin, having agreed to address the United Nations in New York City, wanted to meet in New York. Johnson, wary of encountering protests against the Vietnam War, preferred to meet in Washington, D.C. They agreed on Glassboro because it was equidistant between the two cities.
26	West Windsor Township, New Jersey: The West Windsor post office was found to be infected with anthrax during the anthrax terrorism scare back in 2001-2002.
27	Hillside, New Jersey: In 1991, police from both Hillside and Newark fired nearly 40 shots at a van that had rammed a Hillside police vehicle after a high-speed chase. The pursuit had started after the van had been reported stolen at gunpoint in Newark and was being followed by three Newark police cars before crossing into Hillside. Two of the people inside the vehicle were killed and four of the five other passengers were wounded, though the Union County Prosecutor indicated that there was no clear explanation for why the police had started shooting.
28	Bellaire, Texas: As of 1996 Bellaire prohibits smoking in public parks and dogs in all non-dog public parks; as of that year smoking in public parks brings a fine of $500. The ordinance was adopted around 1996 on a 4-3 vote.
29	Dean Martin: Martin was diagnosed with lung cancer at Cedars Sinai Medical Center in September 1993, and in early 1995 retired from public life. He died of acute respiratory failure resulting from emphysema at his Beverly Hills home on Christmas morning 1995, at age 78.
30	Bernard Francis Law: In May 2004, John Paul II appointed Law to a post in Rome, putting him in charge of the Basilica di Santa Maria Maggiore, with the title of Archpriest .
31	Pascal Couchepin: He was elected to the Swiss Federal Council on March 11, 1998 as a member of the Free Democratic Party (FDP/PRD) and the canton of Valais. In 1998 he took over the Federal Department of Economic Affairs, in which position he fought against the Swiss government contributing any money to the $ 1.25 billion settlement between Swiss banks and Holocaust survivors. He was quoted as saying that there is no reason for the Swiss Government to pay anything, as a government commission had shown we did what was possible in the hard times of the war.
32	Al Sharpton: In May 1990, when one of the two leaders of the mob was acquitted of the most serious charges brought against him, Sharpton led another protest through Bensonhurst. In January 1991, when other members of the gang were given light sentences, Sharpton planned another march for January 12, 1991. Before that demonstration began, neighborhood resident Michael Riccardi tried to kill Sharpton by stabbing him in the chest.

33	Media coverage of the Iraq War: The most popular cable network in the United States for news on the war was Fox News, and had begun influencing other media outlet's coverage.
34	Heather Mercer: On August 3, 2001, the Taliban arrested the two women as they preached Christianity in a private home in Kabul.
35	Gloria Foster: Williams was the one to announce her death in 2001. The cause of her death was diabetes. Though she was no longer married, her ex-husband, Clarence Williams III, was the one to announce her death. A funeral was performed at Cypress Hills Cemetery in Brooklyn on October 15, 2001.
36	Lake Nyos: A pocket of magma lies beneath the lake and leaks carbon dioxide (CO2) into the water, changing it into carbonic acid. Nyos is one of only three known exploding lakes to be saturated with carbon dioxide in this way, the others being Lake Monoun, away SSE, and Lake Kivu in Democratic Republic of Congo.
37	Edward Egan: Archbishop of New York Egan was appointed Archbishop of New York on May 11, 2000 and installed in that position on June 19, 2000.
38	Erich Mielke: Death Mielke died on 21 May 2000 aged 92 in a Berlin nursing home. An estimated 100 people reportedly attended the funeral. His remains are buried in the Zentralfriedhof Friedrichsfelde in Berlin. Mielke's unmarked grave is outside the memorial section established at the entrance in 1951 by East German leaders for communist heroes.
39	Astoria, Queens: At a period many Bangladeshi Americans settled in Astoria, Queens. Many had originated from Sylhet. By 2001 many of the Bangladeshi American people who had settled in Astoria had been moving to Metro Detroit. A survey of an Astoria-area Bengali language newspaper estimated that, in an 18 month period until March 2001, 8,000 Bengali people moved to the Detroit area.
40	MicroProse: In December 1998, Micro Prose finally managed to publish Falcon 4.0 (in development since 1992), to disappointing sales. In December 1999, Hasbro Interactive closed down former Micro Prose studios in Alameda and Chapel Hill.
41	E. Howard Hunt: According to Seymour Hersh, writing in The New Yorker, Nixon White House tapes show that after presidential candidate George Wallace was shot on May 15, 1972, Nixon and Colson agreed to send Hunt to the Milwaukee home of the gunman, Arthur Bremer, to place Mc Govern presidential campaign material there. The intention was to link Bremer with the Democrats. Hersh writes that, in a taped conversation, Nixon is energized and excited by what seems to be the ultimate political dirty trick : the FBI and the Milwaukee police will be convinced, and will tell the world, that the attempted assassination of Wallace had its roots in left-wing Democratic politics. Hunt did not make the trip, however, because the FBI had moved too quickly to seal Bremer's apartment and place it under police guard.
42	Kreuzberg: Hip hop was largely introduced to the youth of Kreuzberg by the children of American servicemen who were stationed nearby until the reunification of Germany.

43	Faberg egg: In 1989, as part of the San Diego Arts Festival, 26 Faberge eggs were loaned for display at the San Diego Museum of Art, the largest exhibition of Faberge eggs anywhere since the Russian Revolution.
44	Merton Miller: Miller was married to Eleanor Miller, who died in 1969. He was survived by his second wife, Katherine Miller, and by three children from his first marriage and two grandsons.
45	Annie (musical): The first attempt at a sequel, Annie 2: Miss Hannigan's Revenge, opened at the John F. Kennedy Center for the Performing Arts in Washington, D.C. in December 1989 to universally disastrous reviews. Extensive reworking of the script and score proved futile, and the project ended before reaching Broadway .
46	Harry Reid: Reid has supported the use of force in the Middle East but in September 2007 he called for a drastic change in strategy. In January 1991 he voted to authorize the first Gulf War.
47	Kay Bailey Hutchison: On June 10, 1993, shortly after the special election victory, Travis County authorities, led by Democratic district attorney Ronnie Earle, raided Hutchison's offices at the State Treasury. Earle failed to obtain a search warrant in conducting that raid.
48	Goodridge v. Department of Public Health: Goodridge v. Department of Public Health Goodridge v. Dept. of Public Health, 798 N.E. 2d 941 (Mass. 2003), was a landmark state appellate court case dealing with same-sex marriage in Massachusetts. The November 18, 2003, decision was the first by a U.S. state's highest court to say that same-sex couples had the right to marry.
49	Nancy Landon Kassebaum: Baker was born Nancy Landon in Topeka, Kansas, the daughter of Theo (ne Cobb) and Governor Alf Landon.
50	Abdurahman Khadr: The CIA reportedly offered him a contract in March 2003 and asked him to work as an infiltrator for American intelligence in Guantanamo, being paid $ 5,000 and a monthly stipend of $ 3000.
51	Dan Burton: In January 1997, Burton played in the AT&T Pebble Beach National Pro-Am, at the invitation of AT&T, the tournament sponsor. The day before the tournament, he played a practice round with Robert E. Allen, AT&T's chairman and chief executive, at a nearby country club. AT&T also hosted a campaign fund-raising dinner for Burton at a local restaurant. Three weeks earlier, Burton had become the chairman of the House Committee on Government Reform and Oversight, which had jurisdiction over the legislative agency scheduled to soon award at least $ 5 billion in long-distance and local telephone and telecommunications contracts with the federal government. Burton defended his participation in the tournament, saying it would not affect his objectivity when dealing with telecommunications issues. He said that he had partially paid for the trip, with his re-election campaign funds paying as well because he attended three fund-raising events while in California.

52	Fatos Nano: In 1997, the collapse of Ponzi schemes marked the beginning of an armed popular revolt against president Sali Berisha, who was forced to resign on July 1997.
53	John Lewis (U.S. politician): Protests In March 2003, Lewis spoke to a crowd of 30,000 in Oregon during an anti-war protest before the Iraq War started.
54	Roy Kinnear: On 19 September 1988, Kinnear fell from a horse during the making of The Return of the Musketeers in Toledo, Spain, and sustained a broken pelvis. He was taken to hospital in Madrid but died from a heart attack the next day. He was 54 years old.
55	LucasVarity: The Company was formed in August 1996, by the merger between Lucas Industries plc and the North American Varity Corporation.
56	Interstate 405 (California): While dangerous high-speed chases along the San Diego Freeway are not uncommon, perhaps the most famous chase in its history was also one of the slowest. On the afternoon of June 17, 1994, former football star O.J. Simpson, suspected in the murder of his ex-wife Nicole Brown Simpson and waiter Ronald Goldman, took to the freeway in a white Ford Bronco (driven by former USC teammate Al Cowlings) with police in pursuit. A bizarre, widely televised low-speed chase ensued and ended hours later when Simpson returned to his Brentwood estate via the Sunset Boulevard exit and surrendered to law enforcement.
57	Interstate 405 (California): Murder of Ennis Cosby 1997 Ennis Cosby, the only son of Bill Cosby, was murdered along I-405 in Los Angeles on January 16, 1997, while fixing a flat tire.
58	Zelimkhan Yandarbiyev: In April 1996, following the assassination of his predecessor Dzhokhar Dudayev, he became an Acting President. In late May 1996, Yandarbiyev headed a Chechen delegation that met President of Russia Boris Yeltsin and Prime Minister of Russia Viktor Chernomyrdin for peace talks at the Kremlin that resulted in the signature of a ceasefire agreement on May 27, 1996.
59	Zelimkhan Yandarbiyev: In 1997, during the signing of the Russian-Chechen Peace Treaty in Moscow, Yandarbiyev famously forced his Russian counterpart President Yeltsin to change seats at a negotiating table so he would be received like a head of sovereign state. Yandarbiyev stood in the presidential election held in Chechnya in February 1997, but was defeated by the Chechen separatist top military leader, General Aslan Maskhadov, getting 10 per cent of the votes and landing third behind Maskhadov and Shamil Basayev. Together with Maskhadov, Yandarbiyev took part of signing of the lasting peace treaty in Moscow.

60	Grant Fuhr: In 1990, Fuhr came forward about his drug use after spending two weeks in a counseling center in Florida. He admitted that he used a substance o he did not say cocaine o for some seven years, or most of the period that the Oilers rested at the top of the NHL. Details of Fuhr's drug use were supplied by the player's ex-wife, Corrine, who told the press in Edmonton that she often found cocaine hidden in his clothing and that she fielded numerous threatening telephone calls from drug dealers who had not been paid. These embarrassing details no doubt contributed to the one-year suspension handed down in September 1990 by NHL president John Ziegler, who called Fuhr's conduct dishonorable and against the welfare of the league. Once Fuhr was re-instated, fans of opposing teams taunted him at games with bags of sugar.
61	Judith Miller (journalist): Miller was the only major U.S. media reporter, and the "New York Times" the only major U.S. media organization, to be victimized by a fake anthrax letter in the fall of 2001. Miller had reported extensively on the subject of biological threats and had co-authored, with Stephen Engelberg and William Broad, a book on bio-terrorism, "" which was published on October 2, 2001. Miller co-authored an article on Pentagon plans to develop a more potent version of weaponized anthrax, "U.S. Germ Warfare Research Pushes Treaty Limits", published in the "New York Times" on September 4, 2001, weeks before the first anthrax mailings.
62	Voice of the Faithful: VOTF began when a small group of parishioners met in the basement of St. John the Evangelist Church in Wellesley, Massachusetts, to pray over allegations that a priest had abused local youngsters. Its meetings soon became well attended, as well as attracting significant media attention. A conference it held in July 2002 attracted over 4,000 lay Catholics, victims of clergy sexual abuse, theologians, priests and religious from around the United States of America and the world.
63	Therese Shaheen: After the Taiwan elections in March 2004, Shaheen resigned her position to return to her private sector businesses.
64	Mary Stuart (actress): When she died in 2002 of a stroke, it was revealed that Mary Stuart was also suffering from gastric cancer and bone cancer. She had previously undergone an endoscopy and an operation to remove a tumor in her stomach in 1999. Stuart had battled breast cancer earlier in her life.
65	Alfred Mosher Butts: In the early 1930s after working as an architect but now unemployed, Butts set out to design a board game. He studied existing games and found that games fell into three categories: number games such as dice and bingo; move games such as chess and checkers; and word games such as anagrams. A resident of Jackson Heights, it was there that the game of Scrabble was invented.
66	ConAgra Foods: Con Agra recalled 19 million pounds of ground beef in July 2002 with bacterial contamination. It was the third-largest recall up to that time . That meat was linked to the illnesses of 19 people in six Western and Midwestern states.
67	Jason Robert Brown: Since 2003 Brown has been married to fellow composer Georgia Stitt;.

68	Jaber Al-Ahmad Al-Jaber Al-Sabah: Gulf War After much discussion of a border dispute between Kuwait and Iraq, Iraq invaded its smaller neighbor on August 2, 1990 with the stated intent of annexing it . Apparently, the task of the invading Iraqi army was to capture or kill Sheikh Jaber .
69	Bernard Pivot: Pivot then created "Bouillon de culture", whose scope he tried to broaden beyond books. He eventually came back to books, however.
70	Kiwi International Air Lines: In July 1997, a Federal bankruptcy judge agreed to liquidate Kiwi in a $16.5 million deal: Joe Logan, Aviation Holdings and Dr. Charles C. Edwards, an orthopedic surgeon and entrepreneur who had led about 30 business enterprises over his 33 year career, bought Kiwi's assets, in a deal that included a Huntington Station, New York investment firm called NJS Acquisitions, which invested $3.5 million for a 20% stake.
71	Ty Inc.: Ty, Inc. has been involved in a large amount of fundraising. Some has been through the sale of certain Beanie Babies in which the proceeds have been donated to various causes. Other times, it has been through other means, such as voting for a fee.
72	Maurice Ferre: On December 20, 1995, Francisco FerrE Malaussena, Mariana Gomez de Ferre, and Felipe Antonio Ferre Gomez, the son, daughter-in-law, and grandson of Maurice Ferre, died when American Airlines Flight 965 crashed into a mountain in Colombia.
73	Supramolecular chemistry: The importance of supramolecular chemistry was established by the 1987 Nobel Prize for Chemistry which was awarded to Donald J. Cram, Jean-Marie Lehn, and Charles J. Pedersen in recognition of their work in this area.
74	Robert Indiana: Between 1989 and 1994, Indiana painted a series of 18 canvases inspired by the shapes and numbers in the "war motifs" paintings that Marsden Hartley did in Berlin between 1913-15.
75	Robert Serber: Serber died June 1, 1997, at his home in Manhattan, from complications following surgery for brain cancer.
76	Al Gore presidential campaign, 1988: On 11 April 1987, Senator Gore of Tennessee announced his candidacy. He stated that he believed he could offer, clearer goals than the other candidates.
77	Rod Lurie: His second was "The Contender" (2000), written for Joan Allen and co-starring Gary Oldman and Jeff Bridges.
78	Silviu Brucan: Romania's Front to fight elections, Guardian, January 24, 1990, Page 24 Under public allegations, Brucan resigned from the FSN in February 1990, claiming that he has accomplished his mission, to restore stability in Romania and to put the country on a course toward multi-party elections.
79	WESCO International: CD & R sold WESCO to The Cypress Group for $1.1 billion in June 1998. This group formed WESCO International, Inc., which is the current owner of WESCO Distribution.

80	Larry MacPhail: MacPhail was elected to the Baseball Hall of Fame in 1978; his son Lee MacPhail was elected to the Hall in 1998, making them the only father and son inductees.
81	Me So Horny: Then-Broward County prosecutor Jack Thompson prosecuted 2 Live Crew on obscenity charges and persuaded a Federal District judge to declare the album obscene in June 1990. 2 Live Crew performed songs from the album including Me So Horny and were prosecuted for obscenity. Record store clerks who sold copies of the album were arrested.
82	Come Blow Your Horn: The play was revived at the Jewish Repertory Theater, New York City, running in December 1987.
83	Angelo Ponte: Ponte was a target of the operation wasteland investigation. On 28 January 1997 he pled guilty to participating in a Mafia run operation to control and manipulate the cartage business in New York City.
84	Oyster Creek Nuclear Generating Station: In 1999, GPU agreed to sell the Oyster Creek Nuclear Plant to AmerGen Energy for $10 million.
85	Bain Capital: Bain, together with Thomas H. Lee Partners, acquired Experian, the consumer credit reporting business of TRW, in 1996 for more than $1 billion. Formerly known as TRW's Information Systems and Services unit, Experian is one of the leading providers of credit reports on consumers and businesses in the US.
86	Leslie Uggams: Uggams was picked to star in "Hallelujah, Baby!" after Lena Horne declined the role of Georgina. The musical premiered on Broadway in 1967 and "created a new star" in Uggams.
87	Granian: Granian was founded in 1995 in Holmdel Township, New Jersey, where Gueyikian initially used the name Grane but changed to Granian over possible copyright concerns.
88	Vitas Gerulaitis: Gerulaitis was born on July 26, 1954, in Brooklyn, NY, to Lithuanian immigrant parents, and grew up in Howard Beach, Queens.
89	Kelo v. City of New London: The decision was widely criticized.
90	Albert J. Dunlap: Dunlap was also suspected of irregularities at Scott Paper. Not long after the shareholder settlement, he agreed to pay $500,000 to settle the SEC's charges. He was also banned from serving as an officer or director of any public company.
91	Irving Trust: Merged into Bank of New York On October 7, 1988 the Irving Trust board signed an agreement to merge with Bank of New York ending a yearlong battle as Bank of New York engineered a hostile takeover. At the time of the merger the combined banks became the United States' 12th largest bank with asset of $ 42 billion.
92	Rick Benjamin (conductor): In February 1999, Benjamin and the Paragon Ragtime Orchestra premiered Oh, You Kid! at the Kennedy Center in Washington, DC, in collaboration with the Paul Taylor Dance Company.

93	Asra Nomani: Influence In November 2003, Nomani became the first woman in her mosque in West Virginia to insist on the right to pray in the male-only main hall.
94	Tom Werner: Sports San Diego Padres Werner's first attempt at owning a professional sports franchise began on June 14, 1990 when he, along with 14 other Southern California-based investors, purchased the San Diego Padres from Joan Kroc for US $ 75 million.
95	Gerald Cardinale: After Congresswoman Marge Roukema announced her retirement in 2002, she endorsed Cardinale as her successor in the Republican primary. However, Cardinale finished with 25%, a close third behind State Assemblyman Scott Garrett (the eventual winner, with 45%) and David C. Russo (who received 26% of votes cast).
96	The Golden Apple (musical): The piece continues to receive occasional productions. For example, a 1990 production by the York Theater Company in New York featured Muriel Costa-Greenspon.
97	Enterprise Oil: The Company was purchased by Royal Dutch Shell for 3.5bn in 2002.
98	Virginia Museum of Fine Arts: The Leslie Cheek Theater, the 500-seat proscenium theater constructed in 1955 within VMFA, has seen several transitions in its 60-year history. It was designed under the supervision of director Cheek, who was a Harvard/Yale-educated architect and who consulted with Yale Drama theater engineers Donald Oenslager and George Izenour to have a state-of-the-art facility.
99	Matatu: The name is a Swahili colloquialism.
100	Trafalgar House (company): In 1964 Broackes brought into the company Victor Matthews, who was fourteen years his senior. Matthews became Broackes' principal lieutenant for the next eighteen years, until parting from him to take control of Express Newspapers, which Trafalgar House had decided to spin off from the group. Nigel Broackes was knighted in 1984 for services to the development of the London docklands, and Victor Matthews was ennobled in 1980 (as Lord Matthews of Southgate) for unswerving support to the Conservative Party.
101	Min Xiaofen: Min lives in Forest Hills, Queens, New York.
102	Tommie Frazier: However, his brief professional football career came to an end with a life-threatening scare. On September 4, 1996, Frazier was admitted to Montreal General Hospital because of pneumonia.
103	Ron Carey (labor leader): Carey did take extensive measures to clean up the union, however. In September 1992, he trusteed 18,000-member Local 237 in New York City for corruption, which led to an extensive battle for control of the local .
104	Asha Puthli: In August 2006, she headlined Central Park Summerstage in New York City on an eclectic bill with DJ Spooky, Talvin Singh, Outernational and Prefuse 73 and special guests Dewey Redman and Dres (rapper) of the hip-hop group Black Sheep.

105	Bruce Murray (soccer): Murray took a break from professional soccer when he signed a contract with the U.S. Soccer Federation (USSF) to play full time with the U.S. national team. On July 30, 1993, the U.S. Soccer Federation released Murray from his national team contract in order to allow him to pursue professional opportunities in Europe.
106	Karen Akers: On September 19, 1993, she married Kevin Patrick Power, a vice president of Orion Network Systems, a satellite communications company in a Roman Catholic ceremony at St. Paul's Chapel at Columbia University in New York.
107	16th G7 summit: This was Thatcher's last opportunity G7 summit meeting.
108	Marianne Wiggins: She and Salman Rushdie wed in January 1988. On a book tour in the US, the couple learned on February 14, 1989 that Ayatollah Ruhollah Khomeini had ordered Rushdie killed for blasphemy in the book "The Satanic Verses". As a result, Wiggins went into protective hiding in Great Britain, along with Rushdie.
109	Matter of Kasinga: Matter of Kasinga The Matter of Kasinga was a legal case decided in June 1996 involving Fauziya Kassindja (surname also spelled as Kasinga), a Togolese teenager seeking asylum in the United States in order to escape a tribal practice of female genital mutilation.
110	Lanny Davis: After leaving the White House, Davis returned to Patton Boggs . There he worked as a lobbyist for the nation of Pakistan prior to the attacks of September 11, 2001.
111	United States Penitentiary, Florence High: In 2000, seven federal correctional officers who called themselves "The Cowboys" were charged with committing misconduct which occurred between January 1995 and July 1997, which included beating and choking handcuffed inmates, mixing waste into the inmates' food, and threatening other officers who objected to their actions.
112	Lawrence E. Spivak: Spivak's office was at the Sheraton-Park Hotel in Washington, D.C., which was also his home. He was widowed in 1983. Spivak died of congestive heart failure at Washington's Sibley Memorial Hospital on March 9, 1994, at the age of 93.
113	Seven Blocks of Granite: Ironically, in its final two games the 1936 team was tied by an inferior University of Georgia team and beaten by a lowly NYU team - ending their hopes of a Rose Bowl appearance.
114	Jochen Piest: On January 10, 1995, Piest was killed in a suicide attack by a Chechen rebel against a Russian mine-clearing unit in the village of Chervlyonna, about 24 kilometers northeast of the Chechen capital, Grozny . The gunman died when the locomotive collided with the military train.
115	Samuel Corsaro: On April 21, 1989, Corsaro and other New Jersey mobsters were indicted on conspiracy and racketeering charges. The most serious charge was planning to burglarize and burn down the Fairfield, New Jersey office of the Attorney General, which was the base for the Northern New Jersey Organized Crime Task Force.
116	Jama'at al-Jihad al-Islami: In November 2004 the Kazakh government disrupted a terror cell, arresting several members.

117	Benedetto Aloi: Windows case In May 1990, Aloi was indicted in the famous Windows Case along with other members of four of the New York crime families. In the Windows case, the crime families used their control over local construction unions and companies to fix the bid prices offered to the New York Housing Authority for thermal pane windows in its housing projects.
118	Franklin Huddle: While serving as the Consul General of Bombay (Mumbai), Huddle and his wife, Chanya Pom Huddle, survived the crash of Ethiopian Airlines Flight 961, which was hijacked, on November 23, 1996.
119	Boobie Clark: Clark died of a blood clot in his lung at the age of 37 on October 25, 1988 at Memorial Hospital in Jacksonville.
120	Texas Longhorns men's basketball: Hired from the University of Rhode Island on April 6, 1988 to replace Weltlich as the Texas head coach, Tom Penders rapidly revitalized the moribund Longhorn basketball program.
121	Krishnaism: Krishnaism (also Bhagavatism) is a group of Hindu denominations within Vaishnavism, centered on devotion to Radha Krishna or other forms of Krishna, identified with Vishnu.
122	Warnborough College: In 1995 Warnborough College enrolled its first group of students onsite in a four-year academic programme. It has been alleged that Warnborough misrepresented itself as being related to Oxford University. When some of the students discovered that Warnborough had no connection with Oxford University they withdrew from the college and demanded refunds. The college denied that it had represented itself as being formally associated with Oxford University.
123	Baku-Rostov highway bombing: Similar incidents A similar aerial attack on a large column of refugees fleeing Grozny fighting took place in August 1996.
124	1999 Grozny refugee convoy shooting: A similar incident involving refugees fleeing Grozny was reported in August 1996 during the First Chechen War.
125	Milton Balkany: In November 2001, auditors began a preliminary investigation of a number of Economic Development Initiative grants that had been awarded in the New York metropolitan area, including the $ 700,000 Children's Center grant. In connection with the investigation, auditors allegedly learned that the Children's Center had failed to file any of the regular progress reports required under rules and procedures of the grant . When auditors interviewed Balkany, he insisted the funds had been used to convert a Bais Yaakov building into the Children's Center but he refused to provide the auditors with access to the school's books and records, according to the Complaint.
126	Mike Huwiler: On March 25, 1997, he moved to the Milwaukee Rampage (A-League) for the 1997 and 1998 seasons.
127	Christopher Coe: Coe died of AIDS on 6 September 1994 at his home in Manhattan.

128	Sadri Gjonbalaj: Sadri Gjonbalaj is a retired American soccer forward who played professionally in the Major Indoor Soccer League, American Soccer League, American Professional Soccer League and National Professional Soccer League. He also earned five caps, scoring one goal, with the U.S. national team. Born in Albania, he grew up in Brooklyn, New York.
129	Dominick Trinchera: Trinchera's body was moved out the club front door into a Ford Econoline van and driven to a lot in Lindenwood, Queens, where Gambino crime family mobsters John Gotti and Gene Gotti arranged the burial. In December 2004, after some children discovered a body in the Lindenwood lot, FBI agents excavated the property and discovered the bodies of the three capos .
130	Philip Giaccone: In October 2004, after some children reported finding a body in the Lindenwood lot, FBI agents excavated the property and discovered the bodies of Giaccone and Trinchera. Among the personal items they unearthed was a Piaget watch that belonged to Giaccone's wife.
131	The Shops at Sunset Place: Prior to the building of Sunset Place, the property was the site of The Bakery Centre, which opened in 1986 on the site of the old Holsum Bread Bakery.
132	Delisa Walton-Floyd: In 1991 she was tested positive for amphetamine and suspended. She claimed she took a drug called "Sydnocarb" which the United States Olympic Committee's Drug Hotline had assured her to be legal after her inquiry. A law suit that she filed against the USOC had no success.
133	1989 Chicago White Sox season: Offseason Potential move to Florida In July 1988, legislators from the State of Illinois narrowly approved a proposal for a new state-financed stadium and a lease deal that would save the team $ 60 million and kept the White Sox from moving to St. Petersburg, Florida .
134	John Kao: His advisory work for Senator Hillary Rodham Clinton was described in "The New York Times" as "out of the box".
135	Louis Daidone: In March 2003, Daidone was indicted again for racketeering, loansharking, gambling and other crimes. In one of the crimes, a Brooklyn landlord was assaulted on Daidone's orders because the landlord ignored a request by Daidone to lower the volume on his home music system.
136	West Las Vegas riots: West Las Vegas riots The West Las Vegas riots were sparked on April 29, 1992, after the Rodney King verdict, where all four white Los Angeles Police Department officers were acquitted for the beating of Rodney King in Los Angeles, California. After the Los Angeles riots were sparked, Black residents of West Las Vegas had already started to loot and burned several stores. Gun battles had started with snipers on intersections and one white motorist was pulled from his vehicle and beaten.

137	Alfred Jolson: After several years of teaching in various Jesuit educational institutions in the United States (including Saint Joseph's University, Philadelphia), Italy and Iraq, H.E. Msgr. Jolson was appointed Diocese of Reykjavk by Pope John Paul II in 1987. Jolson died suddenly in 1994.
138	Charles S. Witkowski: Witkowski suffered heart failure and died on June 1, 1993, at Saint Vincent's Catholic Medical Center in Manhattan.
139	Leslie Cockburn: During the Gulf War in 1991, Cockburn reported from Israel on the Iraqi Scud attacks against Tel Aviv. Her film, shot from a high-rise building close to the impact zone, provided irrefutable evidence that contrary to official reports, the U.S.-supplied Patriot missiles were not only entirely failing to intercept the Scuds but were instead impacting on the city itself.
140	Hispanics in the United States Marine Corps: On January 22, 1991, Captain Manuel Rivera, Jr. (1959 n 1991), a Marine aviator, became the first Hispanic soldier to be killed in Operation Desert Shield . Rivera was killed during a support mission over the Persian Gulf when his TAV-8B Harrier smashed into the Omani coastline while approaching the deck of the amphibious assault ship for a landing.
141	Tiflet: Tiflet is between the cities of Rabat and Khemisset Tiflet is a town that was served by workers of the United States Peace Corps until the attack on America of September 11, 2001 .
142	Dhoruba al-Mujahid bin Wahad: Early years The shooting On May 19, 1971, Thomas Curry and Nicholas Binetti, two NYPD officers who were guarding the home of Frank S. Hogan, the Manhattan district attorney, were fired upon in a drive-by shooting, with a machine-gun.
143	Donald J. Atwood Jr.: US Defense Department On January 26, 1989, President George H. W. Bush named Atwood to the No. 2 job in the US Defense Department. Atwood's management and scientific skills were thought to be a complement to those of the Defense Secretary appointed, John G. Tower, who was expected to play a political and policy-setting role.
144	Anthony Pullard: Pullard signed with the Philadelphia 76ers of the NBA but was waived in July 1990.
145	Roman Bronze Works: After the foundry closed, an auction was staged of original plaster models of major works by American artists, Frederic Remington, Daniel Chester French, Charles Russell, Bessie Potter Vonnoh and Anna Hyatt Huntington, in New York, 17 September 1988 .
146	John DiGilio: Shortly after Di Gilio's trial ended, his wife Ellen reported him missing to police. On May 26, 1988, Di Gilio's body, with two bullet wounds to the head, was discovered floating in a bag on the Hackensack River near Carlstadt, New Jersey.
147	Guardian Royal Exchange Assurance: In February 1999 it was acquired by AXA of France for $ 5.7 bn and integrated into its Sun Life & Provincial Holdings Division.

148	Delia Brown: Life and work Best known for painted scenes that include her friends cavorting in places of specific privilege (such as art collectors' homes), Brown gained notoriety in October 2000 when the New York Times Magazine ran an 8-page spread of her watercolors posing as fashion editorial. This publication coincided with her debut exhibition at D'Amelio Terras gallery in Chelsea, titled What ! Are You Jealous ? (featuring scenes of women drinking champagne poolside in Beverly Hills-ish backyards, with the title borrowed from a Gauguin painting of Tahitian women lounging likewise), which was attacked by Times critic Michael Kimmelman who called the buzz around her work a pseudo-event .
149	John E. Sprizzo: Ultimately Sprizzo's ruling led to changes in U.S. extradition laws. Doherty was deported in February 1992.
150	Joseph C. Salema: Guilty Plea and Sentencing At the conclusion of the federal investigation, Joseph Salema was formally indicted in February 1995 for the acceptance of kickbacks in exchange for influencing New Jersey state bond deals.

A.3 Wikipedia Excerpts to ClueWeb12 Web Articles

ID	Event
1	J._B._S._Haldane: J. B. S. Haldane. Haldane died on 1 December 1964. He willed that his body be used for study at the Rangaraya Medical College, Kakinada.
2	Tengku_Razaleigh_Hamzah: Tengku Razaleigh Hamzah. But it proved a short-lived setback for Razaleigh. Within two years, a share and proxy battle orchestrated by merchant bank Rothschild which was also a part owner of Bumiputera Merchant Bankers brought Sime Darby under Malaysian control and its headquarters shifted to Kuala Lumpur. Control of London Tin Company was acquired the same year. In another coup in 1979, Malaysian money made a dawn raid on British plantation giant, Kumpulan Guthrie Bhd, which wrenched the company from British control; most of the other British owned plantations soon followed.
3	Anglo-Lutheran_Catholic_Church: Anglo-Lutheran Catholic Church. On February 21, 2011, it became public that Catholic authorities in Rome have invited the Anglo-Lutheran Catholic Church to join the Catholic Church through the provisions of "Anglicanorum Coetibus" and that the ALCC has officially and unconditionally accepted that invitation.
4	Survival_International: Survival International. Survival International was founded in 1969 after an article by Norman Lewis in the UK’s "Sunday Times" highlighted the massacres, land thefts and genocide taking place in Brazilian Amazonia.
5	Oliver!: Oliver!. The production closed on January 8, 2011, to be replaced at the theatre by the original London production of "Shrek the Musical".
6	Landmark_Education_litigation: Landmark Education litigation. After attending the Landmark Forum in Germany, Martin Lell wrote a book titled "Das Forum: Protokoll einer Gehirnwsche: Der Psycho-Konzern Landmark Education" [The Forum: Account of a Brainwashing: The Psycho-Outfit Landmark Education], Deutscher Taschenbuch Verlag, Munich, 1997, ISBN 3-423-36021-6. This book detailed Lell’s attendance at the course, and claimed that he had suffered a mental collapse directly afterwards. However, the record at the Hearing indicated that Mr. Lell did not see a doctor; was not hospitalized; did not seek or obtain medication; and was not diagnosed by a medical professional as being brainwashed or having any mental problem.
7	National_Library_of_Papua_New_Guinea: National Library of Papua New Guinea. The National Library of Papua New Guinea is, as its name suggests, the national library of Papua New Guinea. Founded in 1978, it is located in Port Moresby.
8	Sander_Kleinenberg: Sander Kleinenberg. In early 2012, Kleinenberg toured Lahore, Pakistan and performed in front of a large audience.

9	Lanzhou_Military_Region: Lanzhou Military Region. The Region includes two Group Armies (the 21st at Baoji and the 47th at Lintong) plus two Armed Police Units (the 7th and 63rd). Known smaller formations include the 12th Armoured Division ('84701 Unit') at Jiuquan, Gansu. The region also includes the Xinjiang Military District, unusual among PRC military districts in that it contains a significant number of combat troops (the 4th Infantry Division, 6th Infantry Division, 8th Infantry Division, and, apparently, the 11th Highland Motorised Infantry Division reportedly either at Urumqi or in the Karakoram Mountains (Blasko 2000).
10	Happy_Valley_Reservoir: Happy Valley Reservoir. The reservoir acts as a 'holding pond' for water directed to it from the Clarendon Weir via a five km long underground tunnel. The 1.8 m diameter tunnel was bored simultaneously from both ends and when meeting had a deviation of 25mm. Its deepest point underground is 122 m where it passes through a hill. On 7 August 1896 the tunnel's inlet valve was opened by the Governor of South Australia, Sir Thomas Fowell Buxton and the reservoir began filling.
11	Second_Battle_of_Villers-Bretonneux: Second Battle of Villers-Bretonneux. Lt Biltz and his crew reboarded "Nixe" and successfully returned to German Lines. The tank was eventually broken up for spares in June 1918. Earlier in the day, another tank in the same group as Lt Biltz , A7V No.506, "Mephisto", had fallen onto its side and was abandoned. It was captured by Australian troops when they counterattacked a few days later. Today, it is the only surviving German World War I tank and it is preserved at the Queensland Museum in Brisbane, Australia.
12	East_End_of_London: East End of London. Industries associated with the sea developed throughout the East End, including rope making and shipbuilding. The former location of roperies can still be identified from their long straight, narrow profile in the modern streets, for instance Ropery Street near Mile End. Shipbuilding was important from the time when Henry VIII caused ships to be built at Rotherhithe as a part of his expansion of the Royal Navy. On 31 January 1858, the largest ship of that time, the SS Great Eastern, designed by Isambard Kingdom Brunel, was launched from the yard of Messrs Scott Russell & Co, of Millwall. The vessel was too long to fit across the river, and so the ship had to be launched sideways. Due to the technical difficulties of the launch, this was the last big ship to be built on the River, and the industry fell into a long decline.
13	Martin_Moffat: Martin Moffat. Moffat drowned off County Sligo on 5 January 1946. His VC and other medals are on display at the Lord Ashcroft VC Gallery in the Imperial War Museum, London.
14	Demographics_of_China: Demographics of China. The 2010 Census counted 234,829 residents from Hong Kong, 21,201 residents from Macao, 170,283 residents from Taiwan, and 593,832 residents from other locations, totaling 1,020,145 residents.
15	Altadis: Altadis. In 1926, France concentrated its tobacco industry into a single state-run monopoly called Service d'Exploitation industrielle des tabacs (SEIT).

16	Jack_Bursey: Jack Bursey. During WWII he joined the U.S. Coast Guard and was captain of a F&S ship in the Philippines. Then in 1955, now a Lieutenant Commander, Bursey was assigned to Operation Deep Freeze for the International Geophysical Year.
17	Nepenthes__pyriformis: Nepenthes pyriformis. Nepenthes pyriformis (or ; from Latin: "pyrus" = pear, -"forma" = shaped) is a natural hybrid involving "N. inermis" and "N. talangensis". It is known only from Mount Talang in Sumatra, to which "N. talangensis" is endemic. "Nepenthes talangensis" was only described as a distinct species in 1994.
18	T-Mobile_(UK): T-Mobile (UK). Consumer Focus and the Communications Consumer Panel sent a joint letter to the then Competition Commissioner Neelie Kroes in December 2009 asking for the merger to be investigated by authorities in the United Kingdom, rather than Brussels.
19	Representative_peer: Representative peer. Further, the Government pointed out that, even if the election of Scottish peers were entrenched, Parliament could amend the provision under the doctrine of Parliamentary sovereignty. Though the position of the Church of Scotland was "unalterably" secured, the Universities (Scotland) Act 1853 repealed the requirement that professors declare their faith before assuming a position.
20	Picnic_(chocolate_bar): Picnic (chocolate bar). Picnic is a brand of chocolate bar consisting of milk chocolate and peanuts, covering chewy nougat, caramel, biscuit and puffed rice. Picnic bars are lumpy in shape. It is sold in Australia, New Zealand, India, Canada, Ireland, Ukraine and the United Kingdom. The UK and Indian versions differ from the Australasian version in that they also contain raisins. The Cadbury Picnic bar was first released in the UK in 1958.
21	Han_Suyin: Han Suyin. Han Suyin has funded the Chinese Writers Association to create the "National Rainbow Award for Best Literary Translation" (which is now the Lu Xun Literary Award for Best Literary Translation) to help develop literature translation in China. Han Suyin Award for Young Translators sponsored by the China International Publishing Group was also set up by Han Suyin. So far it has given out awards 21 times(in 2009).
22	A_Global_Threat: A Global Threat. A Global Threat was an American street punk band, formed in Bangor, ME in 1997.
23	Internet_in_Russia: Internet in Russia. On April 3, 2008, the RIF-2008 was opened by president-elect of Russia Dmitry Medvedev who said in the opening address to the forum that he estimates Runet to be populated by 40 million users, or 28 percent of the population. He also stated that Russian sites do $3 billion in annual transactions and have $370 million in advertising revenue.
24	Andrew_Laming: Andrew Laming. After the 2004 federal redistribution in Queensland, Sciacca nominated for the newly created seat of Bonner, which contained historically safe Labor areas formerly in Bowman, and Laming secured the now notionally Liberal seat of Bowman, centred on Redland City, with the highest swing to the Liberals of any seat in Queensland (6.06%).

25	Police_aviation_in_the_United_Kingdom: Police aviation in the United Kingdom. In 1921, the British airship R33 was able to help the police in traffic control around the Epsom and Ascot horse-racing events.
26	Primal_therapy: Primal therapy. In 1989, Arthur Janov established the Janov Primal Center in Venice (later relocated to Santa Monica) with his second wife, France.
27	Brenton's_English_Translation_of_the_Septuagint: Brenton's English Translation of the Septuagint. This version of the Old Testament was a translation of the Septuagint by Sir Lancelot Charles Lee Brenton and published by Samuel Bagster & Sons, Ltd., London, in 1844.
28	Mikoyan-Gurevich_MiG-21_operators: Mikoyan-Gurevich MiG-21 operators. Though little is known about the North Korean air force, it is known that a KPAF MiG-21PFM shot down a US Army CH-47 helicopter on July 14, 1977.
29	Gare_Montparnasse: Gare Montparnasse. The Gare Montparnasse became famous for a derailment on 22 October 1895 of the GranvilleParis Express that overran the buffer stop. The engine careened across almost of the station concourse, crashed through a thick wall, shot across a terrace and smashed out of the station, plummeting onto the Place de Rennes below, where it stood on its nose. Two of the 131 passengers sustained injuries, along with the fireman and two conductors. The only fatality was a woman on the street below who was killed by falling masonry.
30	Bi_(jade): Bi (jade). The design of the reverse side of the medals given in the 2008 Summer Olympics in Beijing, China are based on bi disks.
31	San_Guillermo_Parish_Church: San Guillermo Parish Church. The church boasts of having main retablo, side retablos and pulpit that are heavily gilded with gold leaves. The rich decorations of the church depict the Baroque style of architecture. Only half of the original facade of the church can be seen today due to the eruption of Mount Pinatubo in 1991 which half-buried the church. After the volcanic eruption, the towns people painstakingly excavated the altar and the retablo and relocated it under the dome in order for the tall wooden retablo to fit. The retablos niches are filled with centuries-old statues which were saved from destruction of the lahar. The citizens of Bacolor take pride in their rich heritage which is why they carefully excavated the ornately carved main and side altars and restored in its immaculate condition. The church is already a world famous tourist destination prior to the lahar tragedies and present has remained being so.
32	Ralph_Regenvanu: Ralph Regenvanu. In early February 2011, he was described as "instrumental", along with Minister for Trade Ham Lini, in preparing a bill to introduce a "Copyright Act" in Vanuatu, with an aim to protect the intellectual property of artists.
33	Premiership_of_Najib_Tun_Razak: Premiership of Najib Tun Razak. On January 27, 2012 Najib made a promise that the Malaysian Anti-Corruption Commission will be given more independence if the Malaysian people give the government two-thirds majority.

34	Bodmer_Papyri: Bodmer Papyri. Then, in March 2007 it was announced the Vatican had acquired the Bodmer Papyrus XIV-XV (P75), which is believed to contain the world's oldest known written fragment from the Gospel of Luke, the earliest known Lord's Prayer, and one of the oldest written fragments from the Gospel of John.
35	Robert_Caesar_Childers: Robert Caesar Childers. He then moved to Sri Lanka for an official position in the Ceylon civil service. During this period he studied Sinhalese culture, particularly the Pali language. In 1869 he published the first Pali text in Britain, and began to work on a Pali dictionary, which was published 1872-75.
36	Sharada_Srinivasan: Sharada Srinivasan. Sharada Srinivasan,Professor, National Institute of Advanced Studies, Bangalore, India, works in the field of inter-disciplinary scientific studies in art, archaeology, archaeometallurgy and culture and is also an acclaimed exponent of classical Bharata Natyam dance. She has a PhD. in Archaeometallurgy from the Institute of Archaeology, University College London (1996) on the theme of 'Archaeometallugical and art historical studies on South Indian metal icons'.
37	Zhuang_Xiaoyan: Zhuang Xiaoyan. Zhuang was one of the torchbearers in the Olympic torch relay for the 2008 Summer Olympics in Beijing.
38	Shenzhou_10: Shenzhou 10. Shenzhou 10 is a planned manned spaceflight of China's Shenzhou program that is scheduled for launch in 2012.
39	Stuart_Pearce: Stuart Pearce. His brother Dennis is a British National Party activist and was third on the BNP list for London for the European Parliament election, 2009. However, in a brief statement through the FA which has a strong anti-racism stance Pearce said: "My brother's views are his own and do not in any way reflect mine.".
40	Partition_of_Bengal_(1947): Partition of Bengal (1947). Rail and road links connecting North East India to the rest of the country passed through East Bengal territory. The lines connecting Siliguri in North Bengal to Kolkata and Assam to Chittagong were severed. The whole Assam Railway was cut off from the rest of the Indian system.
41	Royal_Scots_Greys: Royal Scots Greys. In November 1917, the Scots Greys saw a glimpse of their future when they moved to support the armoured attack at the Battle of Cambrai. Initially intended to be part of the exploitation force, as at the Battle of the Somme, the plan failed to develop the type of break through which could be exploited by the cavalry. As the fighting bogged down, Scots Greys once again found themselves fighting on foot in an infantry role. Interestingly, part of the reason that the Scots Greys were unable to advance as cavalry was due to the fact that the bridge which was crucial to the advance was accidentally destroyed when the tank crossing it proved to be too heavy.

42	HMS_Ursula_(N59): HMS Ursula (N59). On 14 December 1939 "Ursula" was on patrol off the Elbe estuary when she sighted the German light cruiser "Leipzig", escorted by six destroyers. The "Leipzig" was returning to Kiel to undergo repairs, having been torpedoed and damaged by HMS "Salmon". The waters of the Elbe estuary are shallow and to dive deep is a dangerous undertaking involving the risk of getting stuck on a sandbank. Nevertheless, "Ursula" dived beneath the destroyer screen and got within range of the cruiser, the depth of water being only just enough to allow this manoeuvre. On coming up again to periscope depth, "Ursula" was found to be within point-blank range of the "Leipzig". She fired a salvo of six torpedoes and the two resulting explosions were so close that the "Ursula" herself was badly shaken. On returning to periscope depth, there was no sign of the cruiser, but it did reveal four of her escorting destroyers closing in at high speed to attack. One of these, the destroyer escort "F9", had been hit and was sinking. Once again, risking the sandbanks, the "Ursula" went deep and managed to evade the inevitable depth charges. Of the cruiser, "Leipzig", no further trace was seen, but when the "Ursula" returned to look for evidence, two of the destroyers were still in the area and engaged, apparently, in a search for survivors. "Ursula's" commander, Lt.Cdr. G.C. Phillips, was awarded the DSO and promoted. The "Leipzig" had in fact been missed, the torpedoes had instead hit the "F 9". "Leipzig" made it to port and underwent repair.
43	Sampurnanand_Sanskrit_University: Sampurnanand Sanskrit University. In 1791, during the British Raj in India, Jonathan Duncan, resident of the East India Company proposed the establishment of a Sanskrit college for the development and preservation of Sanskrit Vangmaya to demonstrate British support for Indian education. The initiative was sanctioned by governor general Lord Cornwallis. The first teacher of the institution was Pandit Kashinath and the governor general sanctioned a budget of Rs. 20,000 per annum. The first principal of Government Sanskrit College was J. Myor, followed by J. R. Ballantyne, Ralph T. H. Griffith, G. Thevo, Arthur Venis, Sir Ganganath Jha and Gopinath Kaviraj.
44	Early_naval_vessels_of_New_Zealand: Early naval vessels of New Zealand. By 1840 several Royal Navy ships were engaged in hydrographic surveys directed by the Admiralty. Captain Owen Stanley, on HMS "Britomart", drew up an Admiralty chart of the Waitemata Harbour.
45	The_Phantom_of_the_Opera_(1986_musical): The Phantom of the Opera (1986 musical). A 25th-anniversary stage performance was held in London on 1 and 2 October 2011 at the Royal Albert Hall and was screened live in cinemas worldwide.
46	Kurt_Weill: Kurt Weill. In 1922 he joined the Novembergruppe, a group of leftist Berlin artists that included Hanns Eisler and Stefan Wolpe.
47	Primark: Primark. In December 2008, the UK charity War on Want launched a new report, Fashion Victims II, that showed terms and conditions had not improved in Bangladeshi factories supplying Primark, two years after the charity first visited them.

48	Neturei_Karta: Neturei Karta. Neturei Karta (Jewish Babylonian Aramaic: , literally "Guardians of the City") is a Litvish Jewish group formally created in Jerusalem, British Mandate of Palestine, in 1938, splitting off from Agudas Yisrael. Neturei Karta opposes Zionism and calls for a peaceful dismantling of the State of Israel, in the belief that Jews are forbidden to have their own state until the coming of the Messiah.
49	Louise_Lawler: Louise Lawler. Photographing at Art Basel and Art Basel Miami Beach fairs, the Museum of Modern Art, Christie's and various galleries, Lawler later presented a behind-the-scenes view of art: the hoisting of a Richard Serra sculpture attended by uniformed handlers; white-gloved hands carefully transporting a Richter painting; Cattelan's giant Picasso head swathed in plastic sitting on the floor behind its disconnected body; another Richter painting lying on its side propped against the wall, its public exposure at MoMA at an end; a Hirst spin-painting glimpsed through a closet door. Lawler titled her 2004 survey show at Museum fr Gegenwartskunst in Basel "Louise Lawler and Others" in acknowledgement of the artists whose artworks she photographs.
50	40th_Airlift_Squadron: 40th Airlift Squadron. As Allied troops pressed westward across New Guinea, the Squadron moved to Ward Airdrome at Port Moresby. This move, accomplished on 6 October 1943, placed the Squadron closer to combat, and eliminated many of the long flights back to Australia. The 317th Group soon became known as the "Jungle Skippers".
51	Police_van: Police van. These vehicles were usually painted black or a very dark blue. In the United Kingdom, Ireland, Australia, New Zealand and the United States, a police wagon was also sometimes called a "Black Maria" (pronounced like "Mariah"). The origin of this term is equally uncertain. The OED lists the first usage as the "Boston Evening Traveller" from 1847 which mentions them as a new type of wagon. "Brewer's Dictionary of Phrase and Fable" suggests the name came from Maria Lee, a large and fearsome black keeper of a sailors' boarding house who the police would call on for help with difficult prisoners. The term is still used today in parts of Britain for the vehicle that transports prisoners from jail to court, appearing in the songs "Guns of Brixton" by The Clash and "Adios Hermanos" by Paul Simon. Frequently, blackened-windowed buses are also used for the same purpose.
52	Mrida_Initiative: Mrida Initiative. In 2008, the ATF received 2 million USD to assist in the expansion of Spanish language eTrace software to Mexico and Central America region to assist them with firearms tracking issues, and their immediate goal is to deploy Spanish e-Trace software to all thirty-one states within Mexico.

53	Loy_W._Henderson: Loy W. Henderson. In 1948, Henderson ran afoul of domestic groups lobbying for the creation of the state of Israel. Secretary of State George C. Marshall and Henderson speaking for the Department of State, opposed the United Nations resolution dividing Palestine into Jewish and Arab states. On the other side, Presidential advisors such as David Niles and Clark Clifford, along with American Jewish groups and much of the general public, favored it. Henderson was roundly attacked in the press. In the election year of 1948, his views did not prevail and his transfer to the ambassadorship for India is said to have been a result of political pressure from the pro-Zionist groups.
54	Asian_carp: Asian carp. On September 8, 2010, the Council on Environmental Quality announced the appointment of John Goss as the Asian Carp Director. Goss' role is primarily to serve as the principal advisor to the CEQ's chair, Nancy Sutley on Asian carp issues, and oversee federal, state, and local coordination on Asian carp control efforts. Goss was previously executive director of the Indiana Wildlife Federation (a state affiliate of the National Wildlife Federation), director of the Indiana Department of Natural Resources, and vice-chairman of the Great Lakes Commission.
55	Tan_Cheng_Lock: Tan Cheng Lock. Prior to the independence of Malaya, he was also a member of the Legislative Council of the Straits Settlements. In 1952, Tan Cheng Lock and the United Malays National Organisation (UMNO) under Tunku Rahmans leadership contested the election as partners. He was best remembered for his contributions in the business and political arenas and his work for integrating between the Chinese and the Indian communities to the nascent Malayan society.
56	Cabmen's_Shelter_Fund: Cabmen's Shelter Fund. The Cabmen's Shelter Fund was established in London in 1875 to run shelters for the drivers of hansom cabs and later hackney carriages (taxicabs).
57	Malnutrition_in_India: Malnutrition in India. Subodh Varma, writing in The Times of India, states that on the Global Hunger Index India is on place 67 among the 80 nations having the worst hunger situation which is worse than nations such as North Korea or Sudan. 25% of all hungry people worldwide live in India. Since 1990 there has been some improvements for children but the proportion of hungry in the population has increased. In India 44% of children under the age of 5 are underweight. 72% of infants and 52% of married women have anemia. Research has conclusively shown that malnutrition during pregnancy causes the child to have increased risk of future diseases, physical retardation, and reduced cognitive abilities.
58	BC_Partners: BC Partners. BC Partners is a private equity firm specialising in buyouts and acquisitions financing in Europe and the United States. The firm invests across all industries. BC Partners was founded in 1986 and is based in London, United Kingdom with additional offices in Paris, France; New York City; Milan and Hamburg.
59	Hook_Norton_Brewery: Hook Norton Brewery. Hook Norton Brewery uses a Buxton & Thornley steam engine that has powered most of the machinery in the brewery since 1899. It is the last commercially working open crank stationary steam engine in the UK.

60	Sergi_Bruguera: Sergi Bruguera. Outside tennis, Bruguera is a long-time fan of the Los Angeles Lakers and would often attend their games while playing at tournaments in the United States. In Miami on 28 March 1997, during the same tournament where he defeated World No. 1 Sampras in the semifinals, Bruguera sank three shots (layup, free throw, top of key) during a time-out of a game between the Lakers and the Miami Heat to earn US$500. This money was given to ATP Charities in his name. Bruguera has also played semi-professional Football in his native Spain.
61	Tupolev_Tu-144: Tupolev Tu-144. The only Tu-144 on display outside the former Soviet Union was acquired by the Auto & Technikmuseum Sinsheim in Germany, where it was shipped not flown in 2001 and where it now stands, in its original Aeroflot livery, on display next to an Air France Concorde.
62	Abu_Dhabi_Film_Festival: Abu Dhabi Film Festival. Abu Dhabi Film Festival (ADFF) () is an international film festival created in 2007. The ceremony is held annually in October in Abu Dhabi, United Arab Emirates by the Abu Dhabi Authority for Culture and Heritage (ADACH), under the patronage of Sheikh Sultan Bin Tahnoon Al Nahyan, Chairman of the ADACH. The ADFF aims to encourage and foster the growth of filmmaking in the Arab world by showcasing movies from the region alongside standout productions from prominent international filmmakers.
63	Jeremy_Sims: Jeremy Sims. Sims graduated from in 1990 from Sydney, Australias' National Institute of Dramatic Art (NIDA) with a degree in Performing Arts (Acting).
64	Amy_Roselle: Amy Roselle. Her first role, as a juvenile, was Constance in a version of "King Arthur". After this, her father leased the Cardiff and Swansea theatres for two years. At these theatres, Roselle played in Shakespeare and other productions. She debuted in London at the Haymarket Theatre. There, at age 16, she played Lady Teazle and then, opposite Samuel Phelps, numerous leading parts. She replaced Madge Kendal in "Diplomacy" and played Esther Eccles in "Caste", by T. W. Robertson. She appeared opposite Mary Anderson as Cynisca in "Pygmalion and Galatea" by W. S. Gilbert and created the role of Darine in Gilbert's "The Wicked World" in 1873. She also performed at the Adelphi Theatre and other London theatres.
65	Argo: Argo. A replica of a Greek penteconter was completed in 2008, which was named Argo. This vessel, with a 50-oar crew made up from all 27 European Union member countries, sailed from Jason's hometown of Volos to Venice, stopping at 23 cities "en route".
66	Irish_Civil_War: Irish Civil War. In July 1922 a Protestant orphanage near Clifden, County Galway, housing 58 children was burnt by the anti-treaty side. The children were subsequently transferred to England on board a British destroyer as the Provisional government was unable to rescue them.

67	Platt_Brothers: Platt Brothers. An interesting link exists between the history of Platt Brothers and that of the Toyota company of Japan. In 1929, Platt Brothers paid 100,000 for the patent rights for an innovative automatic weaving loom designed by Sakichi Toyoda himself. The Toyoda Model G loom featured mechanical sensors that automatically shut down the loom if a warp thread snapped. The thinking behind this feature was jidoka which translates as automation with a human touch. Thus workers were freed from being monitors of automatic looms and mill owners could achieve a dramatic increase in labour productivity with one worker able to operate up to 30 machines. Ironically it was the money from the sale of rights that was the start-up capital for the Toyota automobile endeavour. The name change was done for phonetic reasons so although Toyota is now best known as an automotive company, it actually began as Toyoda the textile machinery manufacturer.
68	Charles_R._Middleton: Charles R. Middleton. Charles R. Middleton has served as the President of Roosevelt University since 2002.
69	Masakatsu_Funaki: Masakatsu Funaki. Despite Funaki's body being very broken down from injuries, he returned for a fight against the legendary Rickson Gracie at Colosseum 2000 held at the Tokyo Dome. The show was almost canceled due to Rickson trying to change the rules to make knees and strikes to the head illegal, but the problems were overcome and the show continued.
70	Awa_Marie_Coll-Seck: Awa Marie Coll-Seck. Coll-Seck has been awarded the following professional and academic honours: the Chevalier de lOrdre du Mrite de la Rpublique Francaise, Chevalier des Palmes Acadmiques Francaises, Officier de lOrdre du Mrite Sngalais and Chevalier de lOrdre du Mrite du Burkina Faso. She was elected as chairperson of Commission B of the 2002 World Health Assembly and as President of the Assembly of the Ministries of Health of the West African Health Organization (WAHO, 20022003) and is currently a member of the Academy of Sciences and Technologies of Senegal.
71	Tom_Eyen: Tom Eyen. In 1970, Eyen had his biggest commercial success to date with "The Dirtiest Show in Town", a satiric response to, but also an example of, the era's plays featuring sexual situations and nude actors, which ran for two seasons with later versions off-Broadway and in London's West End.
72	As'ad_AbuKhalil: As'ad AbuKhalil. In a televised debate which aired on Al-Jazeera TV on February 23, 2010 (as translated by MEMRI), AbuKhalil stated that US President Barack Obama "has given free rein to the Zionist lobby to do whatever it likes, both in terms of foreign policy and domestic policy." AbuKhalil also stated that "The Zionists want to muzzle us, so that we won't oppose the wars, violence, or hatred of Israel." In the same interview, Abukhalil sharply criticized MEMRI, stating that it is "a rude, propaganda-spreading organization... which was established by a former Israeli intelligence official." (alluding to MEMRI founder, Yigal Carmon).

73	Big_River_(musical): Big River (musical). Based on Mark Twain's classic 1884 novel, "Adventures of Huckleberry Finn", it features music in the bluegrass and country styles in keeping with the setting of the novel. The Broadway production ran for over 1,000 performances and it remained one of the few very successful American musicals amongst the emerging successes coming from Great Britain.
74	National_Library_of_New_Zealand: National Library of New Zealand. Established in 2004 the NDHA is a partnership between the National Library of New Zealand, Ex Libris and Sun Microsystems to develop a digital archive and preservation management system.
75	Sylhet_International_University: Sylhet International University. Sylhet International University (SIU) is a private university located at Shamimabad, Sylhet, Bangladesh. The university was established by The Private University Act 1992.
76	Popular_Resistance_Committees: Popular Resistance Committees. On March 9, 2012 an Israeli Air Force strike in Gaza killed the secretary-general of the Popular Resistance Committees, Zuhir al-Qaisi (Zuhair al-Qaissi).
77	Irish_Guards: Irish Guards. Historically, Irish Guards officers were often drawn from British public schools, particularly those with a Roman Catholic affiliation, such as Ampleforth College, Downside School and Stonyhurst College. This is less common in recent times. In November 1942 Jean, Grand Duke of Luxembourg joined the British Army as a volunteer in Irish Guards.
78	Kurt-Heinz_Stolze: Kurt-Heinz Stolze. In 1968 he appeared as harpsichordist with the Wrttemberg Chamber Orchestra conducted by Jrg Faerber in the Queen Elizabeth Hall in London.
79	Junoon_(band): Junoon (band). On April 28, 2012, Junoon travelled on a tour to India performing at Mumbai.
80	Theatre_of_Australia: Theatre of Australia. The Theatre Royal, Hobart, opened in 1837 and it remains the oldest theatre in Australia.
81	Transport_in_Hobart: Transport in Hobart. Ferry services operate twice-daily express river crossing services for commuters in morning and evening rush-hour slots from Bellerive Quay to Sullivans Cove, and return. The trip takes approximately seven minutes operated by Hobart Yellow Water Cab. The trip allows commuters from Clarence to arrive just a five minute walk from the CBD. Since the opening of the Museum of Old and New Art in January 2011, a regular daily ferry operates from Sullivan's Cove to the museum's location in the northern suburb of Berriedale.
82	Foreign_relations_of_Libya_under_Muammar_Gaddafi: Foreign relations of Libya under Muammar Gaddafi. In October 1978, Gaddafi sent Libyan troops to aid Idi Amin in the Uganda-Tanzania War when Amin tried to annex the northern Tanzanian province of Kagera, and Tanzania counterattacked. Amin lost the battle and later fled to exile in Libya, where he remained for almost a year.

83	Human_Rights_Party_Malaysia: Human Rights Party Malaysia. The Human Rights Party Malaysia is a Malaysian human rights-based political party founded on 19 July 2009, led by human rights activist P.Uthayakumar.
84	Monty_Python_&_the_Holy_Grail_in_Lego: Monty Python & the Holy Grail in Lego. Monty Python & the Holy Grail in Lego (also known as Lego Knights or "It's Only a Model") is a Lego stop motion animated version of the Camelot dance sequence from the film "Monty Python & the Holy Grail". It was created by Spite Your Face Productions on commission from the Lego Group and Python Pictures and released in 2001 as an extra feature on the special edition DVD of "Monty Python & the Holy Grail".
85	Adelaide_Hills_bushfires_(1939): Adelaide Hills bushfires (1939). Damage was assessed at 650,000, including the destruction of ninety houses. No lives were lost, but the fire highlighted the inadequacy of South Australia's fire-fighting capability. Six thousand city volunteers had helped to combat the fires, using mainly branches and wet bags. As a result, the Emergency Fire Service was set up, the precursor to the modern Country Fire Service.
86	Foreign_Service_Institute,_India: Foreign Service Institute, India. The Foreign Service Institute was established by the Government of India in 1986 primarily to cater to the professional training needs of the trainees of the Indian Foreign Service and use to run from two rooms in Akbar Bhawan. The training programme of the Indian diplomats goes on for about a year, during which they are taught various aspects of India's foreign policy, international relations, Indian history and culture, Indian and the world economic scenario, communication and interpersonal skills, and the like, before they take up posting within the Ministry of External Affairs and sent abroad later.
87	Fred_Kavli: Fred Kavli. Through The Kavli Foundation, Kavli established scientific prizes in the fields of Astrophysics, Nanoscience, and Neuroscience. The Kavli Prizes are presented in cooperation with the Norwegian Academy of Science and Letters and the Norwegian Ministry of Education and Research, and have been awarded biennially at a ceremony in Oslo since 2008.
88	Llandysul: Llandysul. The Welsh Harp Centre, Telynau Teifi is a community business, set up with help from Ceredigion County Council and the European Union in 2004. It is believed to be the only harp-making business in Wales, a country with a traditional association with the instrument. Celtic and folk harps are made here as well as a 20 string lap harp, and the whole manufacturing process takes place on the site.
89	Gauntlet_track: Gauntlet track. Close to where the borders of Belgium, Germany and the Netherlands come together, the interlaced tunnel section provides an important connection between Germany and the Belgian harbour at Antwerp. After completing the installation in 1991, trains with an oversize loading gauge were rerouted over this line, and the lightly used (but tunnel-free) secondary line between Stolberg and Welkenraedt (crossing the border at Raeren) was closed to freight traffic. Trains requiring use of the central track must be diesel hauled as electrification only currently reaches the tunnel mouth on the German side to allow for banking.

90	John_Gifford_Bellett: John Gifford Bellett. It was in Dublin that, as a layman, he first became acquainted with John Nelson Darby, then a minister in the established Church of Ireland, and in 1829 the pair began meeting with others such as Edward Cronin and Francis Hutchinson for communion and prayer.
91	LOFAR: LOFAR. On April 26, 2005, an IBM Blue Gene/L supercomputer was installed at the University of Groningen's math center, for LOFAR's data processing. At that time it was the second most powerful supercomputer in Europe, after the MareNostrum in Barcelona.
92	All_Tripura_Tiger_Force: All Tripura Tiger Force. The All Tripura Tiger Force (ATTF) is an isolationist group from the Tripura region of India. It was founded on July 11, 1990, by a group of former Tripura National Volunteer members under the leadership of Ranjit Debbarma. The ATTF is currently considered a terrorist organization by India.
93	Mad_Sin: Mad Sin. Founded 1987 by Koefte De Ville, who had just dropped out of school, punk and rockabilly guitarist Stein and four-week-bass-playing Holly, they struggled around with the help of some friends, who organized gigs in several shady bars of Berlin. They played as street musicians in shopping malls, where they played rockabilly, country, and blues songs to get the tourist's money.
94	Air_Southwest: Air Southwest. An alliance with UK regional carrier Eastern Airways was announced on 25 February 2010. As a result of the alliance Air Southwest will join a Global Distribution Systems (GDS) which will enable them to sell tickets through a number of external sources like travel agents and increase their market presence. It will also pave the way for the introduction of codeshare agreements between the two airlines.
95	London_Buses_route_422: London Buses route 422. Route 422 was formed on 16 January 1988 from the split of route 122, which was withdrawn between Bexleyheath and Plumstead. The 422 was initially run by Bexleybus between Bexleyheath and Woolwich, providing an overlap with route 122 between Woolwich and Plumstead.
96	Racism_in_Asia: Racism in Asia. The Khmer Rouge regime in Cambodia disproportionately targeted ethnic minority groups. These included ethnic Chinese, Vietnamese and Thai. In the late 1960s, an estimated 425,000 ethnic Chinese lived in Cambodia, but by 1984, as a result of Khmer Rouge genocide and emigration, only about 61,400 Chinese remained in the country. The Cham, a Muslim minority who are the descendants of migrants from the old state of Champa, were forced to adopt the Khmer language and customs. A Khmer Rouge order stated that henceforth The Cham nation no longer exists on Kampuchean soil belonging to the Khmers (U.N. Doc. A.34/569 at 9). Only about half of the Cham survived.

97	Turtle_Islands,_Tawi-Tawi: Turtle Islands, Tawi-Tawi. Together with three islands of neighbor country Malaysia and the surrounding coral waters, Turtle Islands are the only living areas for the Green Sea Turtles in Asia and in the whole world. In 1996, the islands were declared as Turtle Islands Heritage Protected area by the governments of the Philippines and Malaysia as the only way to guarantee the continued existence of the green sea turtles and their nesting sites.
98	Socialist_Alternative_(Australia): Socialist Alternative (Australia). In 2004, Liz Ross wrote [http://catalogue.nla.gov.au/Record/3081017 "Dare to struggle, dare to win! Builders Labourers fight deregistration, 1981-94"] published by Vulgar Press, which outlined the history of the militant Builders Labourers Federation. David Renton of "Labour History" said the book "takes seriously the challenge of understanding the past".
99	Imperial_Bank_of_Persia: Imperial Bank of Persia. The Imperial Bank of Persia (Persian: "Bank Shahanshah") was an Iranian bank. It was established in 1885 with a concession from the government of Persia to Baron Julius De Reuter, under a Royal charter from Queen Victoria.
100	Chickpea: Chickpea. In 1793, ground-roast chickpeas were noted by a German writer as a coffee substitute in Europe. In the First World War, they were grown for this use in some areas of Germany. They are still sometimes brewed instead of coffee.
101	Japanese_occupation_of_Cambodia: Japanese occupation of Cambodia. The Japanese occupation of Cambodia ended with the official surrender of Japan in August 1945. After Allied military units entered Cambodia, the Japanese military forces present in the country were disarmed and repatriated. The French were able to reimpose the colonial administration in Phnom Penh in October the same year. After arresting Son Ngoc Thanh for collaboration with the Japanese, the French colonial authorities exiled him to France, where he lived in house arrest. Some of his supporters went underground and escaped to Thai-controlled northwestern Cambodia, where they were eventually to join forces in a pro-independence group, the Khmer Issarak. This anti-French, politically heterogeneous nationalist movement was organized with Thai backing, but would later split into factions.
102	Helicopter: Helicopter. Meanwhile, Juan de la Cierva was developing the first practical rotorcraft in Spain. In 1923, the aircraft that would become the basis for the modern helicopter rotor began to take shape in the form of an autogyro, Cierva's C.4.
103	Northampton_Meadowlarks: Northampton Meadowlarks. A new Northampton Meadowlarks team solicited monies, including $2000 from President Calvin Coolidge, who was former mayor of Northampton, presumably at the behest of his wife Grace, "The First Lady of Baseball".

104	Federalism: Federalism. In Europe, "federalist" is sometimes used to describe those who favor a common federal government, with distributed power at regional, national and supranational levels. Most European federalists want this development to continue within the European Union. European federalism originated in post-war Europe; one of the more important initiatives was Winston Churchill's speech in Zurich in 1946.
105	Setanta_Sports_USA: Setanta Sports USA. As of July 27, 2009 the ownership structure of Setanta Sports North America Limited has changed. Setanta Sports North America Limited was owned by Sabloss. Both Setanta Sports Channel Ireland Limited and Setanta Sports North America Limited were licensed by the Broadcasting Commission of Ireland.
106	Guy_Gavriel_Kay: Guy Gavriel Kay. When Christopher Tolkien needed an assistant to edit his father J. R. R. Tolkien's unpublished work, he chose Kay, then a student at the University of Manitoba, whose parents were friends of Baillie Tolkien's parents. Kay moved to Oxford in 1974 to assist Tolkien in the editing of "The Silmarillion". There he learnt a lot about writing and editing, and later admitted of Tolkien's influence, "to be successful in fantasy, you have to take the measure of Tolkien work with his strengths and away from his weaknesses".
107	Macroom: Macroom. Macroom () is a market town in Ireland located in a valley on the River Sullane, a tributary of the River Lee, between Cork and Killarney. It is one of the key gateways to the tourist region of West Cork. The town recorded a population on 3,553 in the 2006 national census. The name in Irish Gaelic may mean 'meeting place of followers of the god Crom or 'crooked plain', the latter derived from a large oak tree at one time growing in the town-square.
108	History_of_Israel: History of Israel. In the Fall of 2000, talks were held at Camp David to reach a final agreement on the Israel/Palestine conflict. Ehud Barak offered to meet most of the Palestinian teams requests for territory and political concessions, including Arab parts of east Jerusalem; however, Arafat abandoned the talks without making a counterproposal.
109	Anna_Journey: Anna Journey. Her critical essay on Sylvia Plath ("'Dragon Goes to Bed With Princess': F. Scott Fitzgerald's Influence on Sylvia Plath") appears in "Notes on Contemporary Literature". Her essay, "Lost Vocabularies: On Contemporary Elegy" appears in Parnassus: Poetry In Review." In 2006, Journey discovered the unpublished status of Plath's early sonnet "Ennui" that was published in "Blackbird".
110	Timeline_of_nuclear_program_of_Iran: Timeline of nuclear program of Iran. Mar 23, 2005: Iran offers a proposal to the EU including: Irans adoption of the IAEA Additional Protocol and continuous on-site inspections at key facilities; as well as limiting the expansion of Irans enrichment program, and a policy declaration of no reprocessing.
111	Calverley: Calverley. The village was part of the Municipal Borough of Pudsey alongside Farsley until 1974, though for centuries previously both Pudsey and Farsley were part of the Calverley parish.

112	James_Morgan_(engineer): James Morgan (engineer). Although not a civil engineer of anything like the stature of contemporaries such as Thomas Telford, Morgan rose to the challenge of designing and supervising bridges, tunnels, basins and docks. Perhaps his most notable achievements were the Maida Hill and Islington tunnels the latter was the longest on the canal, and when Telford himself was invited to inspect it in 1818, he opined: "materials and workmanship excellent, and its direction perfectly straight".
113	Catholic_Church_in_India: Catholic Church in India. All the bishops in India, both Western and Eastern, form the Catholic Bishops' Conference of India, which was founded in 1944.
114	Faneuil_Hall: Faneuil Hall. Faneuil Hall (or ; previously), located near the waterfront and today's Government Center, in Boston, Massachusetts, has been a marketplace and a meeting hall since 1742. It was the site of several speeches by Samuel Adams, James Otis, and others encouraging independence from Great Britain, and is now part of Boston National Historical Park and a well-known stop on the Freedom Trail. It is sometimes referred to as "the Cradle of Liberty".
115	Sino-British_Joint_Declaration: Sino-British Joint Declaration. In 1999 the government of the HKSAR asked Chinas State Council to seek an interpretation of a provision in the Basic Law by the National People's Congress Standing Committee. The Chinese government said that a decision by Hong Kongs Court of Final Appeal would allow 1.6 million mainland immigrants to enter Hong Kong. As a result the Chinese authorities obliged and the Hong Kong judgment was overturned.
116	Michael_Madhusudan_Dutt: Michael Madhusudan Dutt. Dutt is widely considered to be one of the greatest poets in Bengali literature and the father of the Bengali sonnet. He pioneered what came to be called "amitrakshar chhanda" (blank verse). Dutt died in Kolkata, India on 29 June 1873.
117	Frances_Cairncross: Frances Cairncross. Cairncross became Rector of Exeter College, Oxford in October 2004.
118	Manoa_Dobui: Manoa Dobui. In the previous election of 1999 he unsuccessfully contested the Lautoka City Open Constituency for the Christian Democratic Alliance (VLV).
119	Mary_Leakey: Mary Leakey. In the spring of 1926, in Mary's 13th year, her father died of cancer. The services were read by Lemozi. Erskine's brother, Percy, came to take them back to London. Cecilia sold Erskine's paintings and moved to a boardinghouse in Kensington. She placed Mary in a local Catholic convent to be educated, following the example of her own life. Later, Mary boasted of never passing an examination there.
120	Tina_Cardinale: Tina Cardinale. Tina Cardinale-Beauchemin is a member of the Northeastern University athletics Hall of Fame. Cardinale was elected in 2002 for her excellence in women's ice hockey. Cardinale was also a pioneer as a member of the first ever United States women's national hockey team where she was captain.

121	Black_Dyke_Band: Black Dyke Band. Black Dyke Band, formerly Black Dyke Mills Band, is one of the oldest and best-known brass bands in the world. The band has won many prizes and competitions over the years. In 2009, the band won the National Brass Band Championships of Great Britain for a record 22nd time.
122	Jacques_Henri_Lartigue: Jacques Henri Lartigue. By then as he received stints for fashion magazines, he was famous in other countries other than his native France, when until 1974 he was commissioned by the newly elected President of France Valry Giscard d'Estaing to shoot an official portrait photograph. The result was a simple photo of him without the use of lighting utilising the national flag as a background.
123	Mahmood_Hussein_Mattan: Mahmood Hussein Mattan. The trial took place at the Glamorgan Assizes in Swansea in July 1952. The main witness for the prosecution was Harold Cover, a Jamaican with a history of violence who later received a share of a reward of 200 offered by the Volpert family. Cover claimed to have seen Mattan leaving Volpert's shop, though it later emerged that he had previously identified another Somali living in the area at the time, Taher Gass, as the man he had seen.
124	Jalen_Rose: Jalen Rose. Rose's tenure with the Knicks was uneventful and prior to the start of the 200607 NBA season on October 30, 2006, the Knicks cut ties with Rose by waiving him. He was courted by several teams including the Phoenix Suns, Detroit Pistons and Miami Heat. On November 3, 2006, Rose announced he would sign with the Suns on his blog at jalenrose.com.
125	Georgia_v._Brailsford_(1794): Georgia v. Brailsford (1794). Explicitly acknowledging jury nullification, the first Chief Justice, John Jay, wrote: "It is presumed, that juries are the best judges of facts; it is, on the other hand, presumed that courts are the best judges of law. But still both objects are within your power of decision you [juries] have a right to take it upon yourselves to judge both, and "to determine the law as well as the fact in controversy"".
126	Inklings: Inklings. Until late 1949, Inklings readings and discussions usually occurred during Thursday evenings in C. S. Lewis's college rooms at Magdalen College. The Inklings and friends were also known to gather informally on Tuesdays at midday at a local public house, The Eagle and Child, familiarly and alliteratively known in the Oxford community as The Bird and Baby, or simply The Bird.
127	Norway: Norway. In its 2007 Worldwide Press Freedom Index, Reporters Without Borders ranked Norway at a shared 1st place (with Iceland) out of 169 countries.
128	Staples_Inc.: Staples Inc.. During the 2008 holiday season, Staples advertising for the first time engaged Facebook, Twitter, YouTube and other social media platforms. The company created a ,character named "Coach Tom" to promote its Gift it for Free sweepstakes, in which 10,000 Staples customers won up to $5,000 in merchandise.

129	Dormant_Commerce_Clause: Dormant Commerce Clause. The idea that regulation of interstate commerce may to some extent be an exclusive Federal power was discussed even before adoption of the Constitution, though the framers did not use the word "dormant." On September 15, 1787, the Framers of the Constitution debated in Philadelphia whether to guarantee states the ability to lay duties of tonnage without Congressional interference, in order for states to finance the clearing of harbors and the building of lighthouses.
130	Richard_Codey: Richard Codey. In 2011, former Gov. Codey also published his autobiography entitled Me, Governor? a humorous autobiography in which he discusses various chapters of his life as governor of New Jersey. In a section of his book, he cites a FDU PublicMind poll study that was conducted about three years after he left office. According to the results from the poll on March 4, 2009, he was still the most popular politician in the state of New Jersey with 41% favorable over 15% unfavorable views.
131	Name_change: Name change. The Supreme Court of the Philippines Justice Leonardo Quisumbing on September 12, 2008, allowed Jeff Cagandahan, 27, who has congenital adrenal hyperplasia, to change his birth certificate name from Jennifer to Jeff, and his gender from female to male.
132	Liberal_Parliamentarians_for_Israel: Liberal Parliamentarians for Israel. Liberal Parliamentarians for Israel is an organization of pro-Israel parliamentarians in the Liberal Party of Canada. The group was founded in 2002, and includes MPs such as Irwin Cotler, Joe Volpe and Carolyn Bennett. Former parliamentarians Stephen Owen and Jim Peterson were also members.
133	Michael_Dickinson_(artist): Michael Dickinson (artist). In July 2007, Dickinson's collages were displayed at the A Gallery, London, in the Stuckist show "I Won't Have Sex with You as long as We're Married".
134	Mark-8: Mark-8. Although not very commercially successful, the Mark-8 prompted the editors of "Popular Electronics" magazine to consider publishing a similar but more easily accessible microcomputer project, and just six months later, in January 1975, they went through with their plans announcing the Altair 8800.
135	Patricia_Racette: Patricia Racette. In 2000 Racette made her debut at La Scala as Ellen Orford and her debut at the Lyric Opera of Chicago as Jenfa. She has since returned to Chicago numerous time to portray such roles as Mim, Marguerite in Charles Gounod's "Faust", Micala, Li, and Madame Lidoine.
136	Kristofer_Janson: Kristofer Janson. Kristofer Janson was born in Bergen, Norway. His father, a prominent merchant, was also the American consul. Janson graduated with a degree in theology from the University of Christiania during 1865. Although he had been trained in theology, he was not ordained into the Church of Norway. He traveled extensively in Europe and upon his return to Norway became popular as a teacher and author. Janson married the former Drude Krog and became the father of seven children.

137	Muslim_Brotherhood: Muslim Brotherhood. After the 1967 Six Day War, as Israel's occupation started, Israel may have looked to cultivate political Islam as a counterweight to Fatah, the main secular Palestinian nationalist political organization.
138	Mustek_Systems: Mustek Systems. Mustek Systems Inc. is a company based in Hsinchu, Taiwan, established in October 1988.
139	Michigan_State_University: Michigan State University. MSU is also ranked number four in several other fields, including international/intercultural communication, mass communication, and interpersonal communication based on the November 2004 NCA report.
140	Southern_Tagalog: Southern Tagalog. Consequently, Administrative Order No. 129 was issued on August 19, 2005 to address this backlash directing the abeyance of Executive Order 429, pending the approval of an implementation plan for the orderly transfer of Palawan from MIMAROPA to Region VI. Presently, Palawan is still considered part of MIMAROPA.
141	Flag_of_South_Australia: Flag of South Australia. The flag is based on the defaced British Blue Ensign with the state badge located in the fly. The badge is a gold disc featuring a Piping Shrike with its wings outstretched. The badge is believed to have been originally designed by Robert Craig, a teacher at the School of Arts in Adelaide, and officially gazetted on 14 January 1904.
142	Counterfeit_money: Counterfeit money. On 3 May 1999 the New Zealand Reserve Bank started circulating polymer banknotes printed by Note Printing Australia Limited.
143	Bill_T._Jones: Bill T. Jones. Jones is the co-creator, director and choreographer of the musical "Fela!", which ran Off-Broadway in 2008 and opened on Broadway in previews in October 2009. Jones won the Lucille Lortel Award as Outstanding Choreographer for his work as well as the Tony Award for Best Choreography.
144	Sir_Edward_Troubridge,_2nd_Baronet: Sir Edward Troubridge, 2nd Baronet. Troubridge was the only son of Rear-Admiral Sir Thomas Troubridge, 1st Baronet. He entered the Royal Navy in 1797 and was present at the Battle of Copenhagen as a Midshipman in HMS "Edgar".
145	Lewis_Evans_(collector): Lewis Evans (collector). Through the efforts of his friend Robert Gunther, Evans donation helped in the founding of the Museum of the History of Science, Oxford in 1930 by providing what became known as the Lewis Evans Collection of Historic Scientific Instruments, the core of the museum's initial collection. His library is also owned by the museum.
146	Wayne_Kasserman: Wayne Kasserman. Wayne Kasserman is a New York city based stage and screen actor, most notable for his role in the 2008 remake of Knight Rider as mechanic Dylan Fass.

147	Ellen_DeGeneres: Ellen DeGeneres. On December 3, 2011, DeGeneres headlined the third annual Change Begins Within gala for the David Lynch Foundation held at the Los Angeles County Museum of Art.
148	Cessna: Cessna. In 1960, Cessna affiliated itself with Reims Aviation of Reims, France.
149	Westville,_Florida: Westville, Florida. American author Laura Ingalls Wilder along with husband Almanzo and her daughter Rose lived here for a short time in 1891. Laura's cousin Peter lived here as well. Peter married and stayed in the region, but the Wilders quickly determined the move had been a mistake for them, and left the area after about a year, eventually settling in Mansfield, Missouri.
150	History_of_plug-in_hybrids: History of plug-in hybrids. In 1994, the Esoro H301 two-door, four passenger plug-in hybrid sedan was built in Switzerland by the vehicle prototyping company Esoro AG.

A.4 Wikipedia Seminal Events of Past

ID	Event	Gold Standard
1	Riots and mass killings in the Indian state of Gujarat	2002 Gujarat riots
2	The International Astronomical Union defines ’planet’ at its 26th General Assembly, demoting Pluto to the status of ’dwarf planet’ more than 70 years after its discovery	IAU definition of planet
3	Four terror attacks (3 on the London Underground and 1 on a bus) rock the transport network in London, killing 52 (not including the 4 bombers) and injuring over 700	7 July 2005 London bombings
4	One of the worst natural disasters in recorded history hits Southeast Asia, when the strongest earthquake in 40 years, measuring 9.3 on the Richter scale, hits the entire Indian Ocean region, which generates an enormous tsunami that crashes into the coastal areas of a number of nations includingThailand, India, Sri Lanka, the Maldives, Malaysia, Myanmar, Bangladesh, and Indonesia. The official death toll in the affected countries stands at 186,983 while more than 40,000 people are still missing.	2004 Indian Ocean earthquake and tsunami
5	The Orange Revolution begins when the government of the Ukraine is accused of electoral fraud against presidential candidate Viktor Yushchenko.	Orange Revolution
6	Chechen terrorists take 1,128 people hostage, mostly children, in a school in the Beslan school hostage crisis. The hostage-takers demand the release of Chechen rebels imprisoned in neighbouring Ingushetia and the independence of Chechnya from Russia.	Beslan school siege
7	America demands Iraq allow unfettered access to weapons inspectors.	Iraq disarmament crisis
8	The Beltway sniper attacks begin with 5 shootings in Montgomery County, Maryland.	Beltway sniper attacks
9	Four terror attacks (3 on the London Underground and 1 on a bus) rock the transport network in London, killing 52 (not including the 4 bombers) and injuring over 700	September 11 attacks
10	An earthquake hits Gujarat, India, killing almost 20,000.	2001 Gujarat earthquake

11	Sierra Leone Civil War: The British Armed Forces launch Operation Palliser to support the Sierra Leone government to counter the Revolutionary United Front.	British military intervention in the Sierra Leone Civil War Operation Palliser
12	East Timor votes for independence from Indonesia in a referendum.	1999 East Timorese crisis
13	Columbine High School massacre: Two Littleton, Colorado teenagers, Eric Harris and Dylan Klebold, open fire on their teachers and classmates, killing 12 students and 1 teacher, and then themselves.	Columbine High School massacre
14	Good Friday: 1 hour after the end of the talks deadline, the Belfast Agreement is signed between the Irish and British governments and most Northern Ireland political parties, with the notable exception of the Democratic Unionist Party.	Good Friday Agreement
15	Diana, Princess of Wales is killed in a car accident in Paris	Death of Diana, Princess of Wales
16	The United Kingdom hands sovereignty of Hong Kong to the People's Republic of China.	Transfer of sovereignty over Hong Kong
17	The Australian government introduces a nationwide ban on the private possession of both automatic and semi-automatic rifles, in response to the Port Arthur massacre.	Port Arthur massacre (Australia)
18	Srebrenica massacre: Units of the Army of Republika Srpska, under the command of General Ratko Mladic, enter Srebrenica with little resistance from Dutch peacekeepers of the United Nations Protection Force, going on to kill thousands of Bosniak men and boys and rape many women.	Srebrenica massacre
19	South Africa holds its first fully multiracial elections, marking the final end of apartheid. Nelson Mandela wins the elections and is sworn in as the first democratic president.	Negotiations to end apartheid in South Africa

20	Dissolution of Czechoslovakia: The Czech Republic and Slovakia separate in the so-called Velvet Divorce.	Dissolution of Czechoslovakia
23	Cold War: The leaders of the Yemen Arab Republic and the People's Democratic Republic of Yemen announce the unification of their countries as the Republic of Yemen.	Yemeni unification
24	German reunification: An agreement is reached for a two-stage plan to reunite Germany.	German reunification
25	Cold War and Fall of the Berlin Wall: G¸nter Schabowski accidentally states in a live broadcast press conference that new rules for traveling from East Germany to West Germany will be put in effect "immediately". East Germany opens checkpoints in the Berlin Wall, allowing its citizens to travel freely to West Germany for the first time in decades (November 17 celebrates Germans tearing the wall down).	Berlin Wall
26	The Soviet Union begins its program of economic restructuring (perestroika) with legislation initiated by Premier Mikhail Gorbachev (though Gorbachev had begun minor restructuring in 1985).	Perestroika
27	Black Monday: Stock market levels fall sharply on Wall Street and around the world.	Black Monday (1987)
28	A peaceful student demonstration in Prague, Czechoslovakia, is severely beaten back by riot police. This sparks a revolution aimed at overthrowing the Communist government (it succeeds on December 29).	Velvet Revolution
29	The Woodstock Festival is held in upstate New York, featuring some of the top rock musicians of the era.	Woodstock
30	Two trains carrying explosives and fuel collide in Ryongchon, North Korea, killing 161 people, injuring 1,300 and destroying thousands of homes.	Ryongchon disaster
31	Seung-Hui Cho, a South Korean expatriate student, shoots and kills 32 people at the Virginia Polytechnic Institute and State University, before committing suicide, resulting in the deadliest shooting incident by a single gunman in United States history.	Virginia Tech shooting
32	Homeland Security police detain workers at 6 meatpacking plants in the midwestern U.S.	Swift raids
33	An explosion takes place at one of BP's largest oil refineries in Texas City, killing 15 and injuring more than 170.	Texas City Refinery explosion

34	a $1.73 billion research program to study both the North Pole and South Pole, is launched in Paris.	International Polar Year
35	In Casablanca, Morocco, 33 civilians are killed and more than 100 injured in the Casablanca terrorist attacks.	2003 Casablanca bombings
36	Terrorists detonate bombs in 2 nightclubs in Kuta, Bali, killing 202 and injuring over 300.	2002 Bali bombings
37	Terrorists†detonate bombs†in 2 nightclubs in Kuta,†Bali, killing 202 and injuring over 300.	2002 Bali bombings
38	Kenya Airways Flight 431 crashes off the coast of CÙte d'Ivoire into the Atlantic Ocean, killing 169.	Kenya Airways Flight 431
39	in the Federal Republic of Yugoslavia, 3 Chinese embassy workers are killed and 20 others wounded when a NATO B-2 aircraft mistakenly bombs the Chinese Embassy in Belgrade.	Kosovo War
40	The bombings of the United States embassies in†Dar es Salaam, Tanzania, and†Nairobi, Kenya, kill 224 people and injure over 4,500; they are linked to terrorist†Osama bin Laden, an exile of†Saudi Arabia.	1998 United States embassy bombings
41	Rosemary West is sentenced to life for killing 10 women and girls, including her daughter and stepdaughter, after the jury returns a guilty verdict at Winchester Crown Court. The trial judge recommends that she should never be released from prison, making her only the second woman in British legal history to be subjected to a whole life tariff (the other is Myra Hindley).	Rosemary West
42	A massive anti-poll tax demonstration in Trafalgar Square, London, turns into a riot; 471 people are injured, and 341 are arrested.	Poll Tax Riots
43	F. W. de Klerk announces the unbanning of the African National Congress and promises to release Nelson Mandela.	Apartheid
44	Patrick Edward Purdy kills 5 children, wounds 30 and then shoots himself in Stockton, California.	Stockton schoolyard shooting
45	A British Midland Boeing 737 crashes on approach to East Midlands Airport, leaving 47 dead.	Kegworth air disaster
46	Lieutenant Colonel Oliver North and Vice Admiral John Poindexter are indicted on charges of conspiracy to defraud the United States.	IranñContra affair
47	Republic of China Army execute 19 unarmed Vietnamese refugees on Donggang beach, Lieyu, Kinmen off Mainland China	1987 Lieyu massacre

48	The Tower Commission rebukes U.S. President Ronald Reagan for not controlling his National Security Council staff.	Tower Commission
49	An Amtrak train en route from Washington, D.C. to Boston collides with Conrail engines at Chase, Maryland, killing 16	1987 Maryland train collision
50	Vietnam War: In Washington, D.C., 250,000ñ500,000 protesters stage a peaceful demonstration against the war, including a symbolic "March Against Death".	Vietnam War
52	STS-31: The Hubble Space Telescope is launched aboard Space Shuttle Discovery.	Hubble Space Telescope
53	The Intergovernmental Panel on Climate Change releases its first assessment report, linking increases in carbon dioxide in the Earth's atmosphere, and resultant rise in global temperature, to human activities.	IPCC First Assessment Report
54	Rwandan President JuvÈnal Habyarimana and Burundi President Cyprien Ntaryamira die when a missile shoots down their jet near Kigali, Rwanda. This is taken as a pretext to begin the Rwandan Genocide.	Assassination of JuvÈnal Habyarimana and Cyprien Ntaryamira
55	The bombings of the United States embassies in Dar es Salaam, Tanzania, and Nairobi, Kenya, kill 224 people and injure over 4,500; they are linked to terrorist Osama bin Laden, an exile of Saudi Arabia.	1998 United States embassy bombings
57	8888 Uprising: Thousands of protesters in Burma, now known as Myanmar, are killed during anti-government demonstrations	8888 Uprising
58	The death of Hu Yaobang sparks the beginning of the Tiananmen Square protests of 1989.	Tiananmen Square protests of 1989
59	AndrÈs RodrÌguez, who had seized power and declared himself President of Paraguay during a military coup in February, wins a landslide election in a general election marked by charges of fraud.	AndrÈs RodrÌguez (politician)
60	Exxon Valdez oil spill: In Alaska's Prince William Sound, the Exxon Valdez spills 240,000 barrels (38,000 m3) of oil after running aground.	Exxon Valdez oil spill
61	Ten-Day War: Fighting breaks out when the Yugoslav People's Army attacks secessionists in Slovenia.	Ten-Day War

62	In Paris, the Vietnam-backed government of the state of Cambodia signs an agreement with the Khmer Rouge to end the civil war and bring the Khmer Rouge into government in spite of its role in the Cambodian genocide which ends the CambodianñVietnamese War. The deal results in the creation of the United Nations Transitional Authority in Cambodia.	United Nations Transitional Authority in Cambodia
63	Bosnian Serbs declare their own republic within Bosnia and Herzegovina, in protest of the decision by Bosniaks and Bosnian Croats to seek EC recognition.	Bosnian War
64	Ethnic Russians riot in Tallinn and other towns in Estonia, about moving the Bronze Soldier, a Soviet World War II memorial	Bronze Night
65	Operation Orchard Israeli airplanes strike a suspected nuclear site in Syria	Operation Orchard
66	World Trade Center bombing: In New York City, a van bomb parked below the North Tower of the World Trade Center explodes, killing 6 and injuring over 1,000.	1993 World Trade Center bombing
67	Bureau of Alcohol, Tobacco, Firearms and Explosives agents raid the Branch Davidian compound in Waco, Texas, with a warrant to arrest leader David Koresh on federal firearms violations. Four agents and five Davidians die in the raid and a 51-day standoff begins.	Waco siege
68	At least 50 people are killed and more than 120 injured in a series of coordinated suicide bombings in Amman, Jordan.	2005 Amman bombings
69	Hurricane Katrina makes landfall along the U.S. Gulf Coast, causing severe damage. At least 1,836 die in the aftermath.	Hurricane Katrina
70	Eritrea: Eritrean independence is declared as a result of a referendum held with United Nations verification.	Eritrean War of Independence
71	Israeli Kahanist Baruch Goldstein opens fire inside the Cave of the Patriarchs in the West Bank; he kills 29 Muslims before worshippers beat him to death.	Cave of the Patriarchs massacre
72	The Bombay Riots take place in Mumbai, in December 1992 and January 1993, in which around 900 people died	Bombay Riots
73	A bomb explodes outside the First National Bank in Oshakati, Namibia, killing 27 and injuring 70.	1988 Oshakati bomb blast
74	Oklahoma City bombing: 168 people, including 8 Federal Marshals and 19 children, are killed at the Alfred P. Murrah Federal Building and 680 wounded by a bomb set off by Timothy McVeigh and one of his accomplices, Terry Nichols.	Oklahoma City bombing

75	The Dayton Agreement to end the Bosnian War is reached at Wright-Patterson Air Force Base near Dayton, Ohio (signed December 14).	Dayton Agreement
76	Third Taiwan Strait Crisis: The Chinese People's Liberation Army fires missiles into the waters north of Taiwan.	Third Taiwan Strait Crisis
77	Several explosions at a military dump in Lagos, Nigeria kill more than 1,000.	2002 Lagos armoury explosion
78	The Republic of China (Taiwan) holds its first direct elections for president; Lee Teng-hui is re-elected.	Republic of China presidential election, 1996
79	Dolly the sheep, the first mammal to be successfully cloned from an adult cell, is born at the Roslin Institute in Midlothian, Scotland.	Dolly (sheep)
80	Tropical Storm Allison produces 36 inches (900 mm) of rain in Houston, killing 22, damaging the Texas Medical Center, and causing more than 5 billion American dollars of damage overall.	Tropical Storm Allison
82	The Parliament of India is attacked; 12 are killed.	2001 Indian Parliament attack
83	Troops of Laurent Kabila march into Kinshasa. Zaire is officially renamed Democratic Republic of the Congo.	First Congo War
84	The Versailles wedding hall disaster kills 23 in Jerusalem, Israel.	Versailles wedding hall disaster
85	The Euro currency is introduced in 11 countries - members of the European Union (with the exception of the United Kingdom, Denmark, Greece and Sweden).	Euro Introduction
86	The Northeast blackout of 2003 cuts power to an estimated 10 million people in Ontario, Canada, and 45 million people in eight U.S. states.	Northeast blackout of 2003
87	In Phnom Penh, Cambodia, the Thai embassy is burned and commercial properties of Thai businesses are vandalized.	2003 Phnom Penh riots
88	armed conflict between India and Pakistan that took place between May and July 1999 in the Kargil district of Kashmir and elsewhere along the Line of Control (LOC).	Kargil War

89	This is the final date during which there is no human presence in space; on October 31, Soyuz TM-31 launches, carrying the first resident crew to the International Space Station. The ISS has been continuously crewed since.	Expedition 1
90	In Aden, Yemen, USS Cole is badly damaged by two Al-Qaeda suicide bombers, who place a small boat laden with explosives alongside the United States Navy destroyer, killing 17 crew members and wounding at least 39.	USS Cole bombing
91	2,977 victims are killed in the September 11 attacks at the World Trade Center in New York City, New York, The Pentagon in Arlington County, Virginia, and in rural Shanksville, Pennsylvania after American Airlines Flight 11 and United Airlines Flight 175 are hijacked and crash into the World Trade Center's Twin Towers, American Airlines Flight 77 is hijacked and crashes into the Pentagon, and United Airlines Flight 93 is hijacked and crashes into grassland in Shanksville, due to the passengers fighting to regain control of the airplane. The World Trade Center towers collapse as a result of the crashes.	September 11 attacks
92	A passenger train derails in Amagasaki, Hyogo Prefecture, Japan, killing 107 people and injuring another 562	Amagasaki rail crash
93	Uzbek Interior Ministry and National Security Service troops kill up to 700 during protests in eastern Uzbekistan over the trials of 23 accused Islamic extremists. President Islam Karimovdefends the act.	Andijan massacre
94	A sudden storm engulfs Mount Everest with several climbing teams high on the mountain, leaving 8 dead. By the end of the month, at least 4 other climbers die in the worst season of fatalities on the mountain to date.	1996 Mount Everest disaster
95	During economic crisis in Argentina government effectively froze all bank accounts for twelve months which led to riots.	December 2001 riots in Argentina
96	The Space Shuttle Columbia takes off for mission STS-107 which would be its final one. Columbia disintegrated 16 days later on re-entry.	Space Shuttle Columbia disaster
97	Ten-Day War: Fighting breaks out when the Yugoslav People's Army attacks secessionists in Slovenia.	Ten-Day War

98	One of the worst natural disasters in recorded history hits Southeast Asia, when the strongest earthquake in 40 years, measuring 9.3 on the Richter scale, hits the entire Indian Ocean region, which generates an enormous tsunami that crashes into the coastal areas of a number of nations including Thailand, India, Sri Lanka, the Maldives, Malaysia, Myanmar, Bangladesh, and Indonesia. The official death toll in the affected countries stands at 186,983 while more than 40,000 people are still missing.	2004 Indian Ocean earthquake and tsunami
99	Terrorists execute simultaneous attacks, with bombs in 4 rush-hour trains in Madrid, killing 191 people.	2004 Madrid train bombings
100	The State of Arizona executes Karl LaGrand, a German national involved in an armed robbery in 1982 that led to a death. Karl's brother Walter is executed a week later, in spite of Germany's legal action in the International Court of Justice to attempt to save him.	LaGrand case
101	he 2007 European heatwave was a heat wave that affected most of Southern Europe and the Balkans.	2007 European heat wave
102	World stock markets plummet after China and Europe release less-than-expected growth reports.	Chinese stock bubble of 2007
103	Sixty-five miners die after becoming trapped underground, following an explosion in Nueva Rosita, Mexico	Pasta de Conchos mine disaster
104	Israeli troops invade Lebanon in response to Hezbollah kidnapping 2 Israeli soldiers and killing 3. Hezbollah declares open war against Israel 2 days later	2006 Lebanon War
105	The Abu Sayyaf guerrilla group is blamed for the deadliest terrorist attack at sea in world history, which kills 116 in the Philippines.	2004 SuperFerry 14 bombing
107	Al Ayyat railway accident at Reqa Al-Gharbiya in Egypt: a fire on a train running from Cairo to Luxorkills at least 383 and injures over 65.	2002 Al Ayyat railway accident
108	A lone bomber explodes a home-made bomb in the Myyrmanni shopping mall north of Helsinki, Finland; the casualties include himself.	Myyrmanni bombing

109	A Chinese fighter jet bumps into a U.S. EP-3E surveillance aircraft, which is forced to make an emergency landing in Hainan, China. The U.S. crew is detained for 10 days and the F-8 Chinese pilot, Wang Wei, goes missing and is presumed dead.	Hainan Island incident
110	A sideswipe collision of 2 Tokyo Metro trains kills 5 people.	Naka-Meguro train disaster