LARGE-SCALE ACQUISITION OF REFINED COMMONSENSE KNOWLEDGE

By TUAN-PHONG NGUYEN

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

Saarbrücken, 2024

Date of Colloquium	24.04.2025
Dean of the Faculty	Prof. Dr. Roland Speicher

Examination Board

Chair	Prof. Dr. Vera Demberg
Reviewer	Prof. Dr. Gerhard Weikum
Reviewer	Prof. Dr. Simon Razniewski
Reviewer	Prof. Dr. Fabian M. Suchanek
Reviewer	Prof. Dr. Krishna P. Gummadi
Academic Assistant	Dr. Ha Dao

Abstract

Structured knowledge is important for many artificial intelligence (AI) applications. Commonsense knowledge (CSK) about properties of concepts and human behaviors (e.g., elephants are big and eat plants, children love visiting zoos, tipping is not a common practice in Japan) is crucial for robust human-centric AI. However, this kind of knowledge is covered by a small number of structured knowledge projects. These prior CSK resources have significant shortcomings:

1. *Expressiveness:* They are restricted in their expressiveness to subject-predicate-object (SPO) triples with simple concepts for S and monolithic strings for P and O.
2. *Lacking cultural contextualization:* They lack knowledge about human traits and behaviors conditioned on cultural contexts, which is crucial for situative AI.
3. *Assertion quality:* They suffer from either low precision or limited coverage due to imperfect sources of knowledge extraction (noisy web crawls), or approaches with limited scalability (crowdsourcing). In addition, very few have paid attention to the *saliency* of assertions.

In this dissertation, we develop methods for the automatic acquisition of *semantically refined* CSK at large scale and overcome these limitations. We tackle the CSK acquisition problem by collecting and organizing knowledge via the following entry points: (1) *concepts* (in the Ascent++ project), (2) *cultures* (in the Candle project), and (3) both *concepts and cultures* (in the Mango project).

1. *Concept-centric extraction and organization:* We introduce an expressive CSK model for everyday concepts, with: (i) *refined subjects*, including *subgroups* and *aspects* of primary subjects, (ii) *semantic facets* for assertions, and (iii) scores for *typicality* and *saliency*. Given a set of everyday concepts (e.g., `elephant`, `bicycle`), we propose Ascent++, an automated method for extracting high-quality CSK assertions from large-scale web contents. Ascent++ consists of various new techniques for aggregation and cleaning. The resulting CSK resource consists of 2M assertions for 10K important concepts, surpassing prior resources on both coverage and precision.
2. *Culture-centric extraction and organization:* Given a set of cultural groups (e.g., `Japanese`, `Buddhist`), we propose the Candle method for extracting *culture-aware commonsense knowledge* (CCSK) from a large web corpus. This method includes judicious techniques for classification-based filtering and scoring of *interestingness*, which results in a large-scale CCSK resource of 60K assertions covering 386 cultural groups, which has a significantly better quality compared to other resources of similar kind.

3. *Combining concepts and cultures:* We propose Mango, a methodology for efficiently distilling CCSK assertions from large language models (LLMs). Our method includes (i) prompt construction for large sets of concepts and cultures, and (ii) clustering assertions into topically and culturally coherent groups. Running the Mango method with GPT-3.5 as underlying LLM yields a CCSK resource of unprecedented coverage (167K assertions covering 30K concepts and 11K cultures) with even higher quality than Candle. In an extrinsic evaluation for *intercultural dialogues*, we show that the injection of Mango assertions significantly improves the specificity and cultural sensitivity of LLM responses.

Each of the constructed CSK collections is released for further research, with a web-based knowledge base browser, along with downloadable code and data.

Zusammenfassung

Strukturiertes Wissen ist wichtig für viele Anwendungen der Künstlichen Intelligenz (KI). Allgemeinwissen (*„commonsense knowledge“* – CSK) über Eigenschaften von Konzepten und menschlichem Verhalten (z. B. Elefanten sind groß und fressen Pflanzen; Kinder lieben es, Zoos zu besuchen; Trinkgeldgeben ist in Japan unüblich) ist entscheidend für robuste, menschenzentrierte KI. Diese Art von Wissen wird jedoch nur von einer kleinen Anzahl von strukturierten Wissensprojekten abgedeckt. Diese bestehenden CSK-Ressourcen weisen erhebliche Mängel auf:

1. *Ausdrucksstärke:* Ihre Ausdrucksmöglichkeiten sind auf Subjekt-Prädikat-Objekt (SPO) Tripel mit einfachen Konzepten für S und monolithischen Strings für P und O beschränkt.
2. *Fehlende kulturelle Kontextualisierung:* Sie verfügen nicht über Wissen über menschliche Eigenschaften und Verhaltensweisen, die durch kulturelle Kontexte bedingt sind, was für situationale KI entscheidend ist.
3. *Qualität der Aussagen:* Sie leiden entweder unter niedriger Präzision oder begrenzter Abdeckung aufgrund imperfekter Quellen zur Wissensextraktion (fehlerbehafte Web-Crawls) oder Ansätzen mit begrenzter Skalierbarkeit (Crowdsourcing). Darüber hinaus haben nur wenige die Prägnanz von Aussagen berücksichtigt.

In dieser Dissertation entwickeln wir Methoden zur automatischen Gewinnung von semantisch verfeinertem CSK in großem Maßstab und überwinden diese Einschränkungen. Wir gehen das Problem der CSK-Gewinnung an, indem wir Wissen über die folgenden Einstiegspunkte sammeln und organisieren: (1) Konzepte (im Projekt Ascent++), (2) Kulturen (im Projekt Candle) und (3) sowohl Konzepte als auch Kulturen (im Projekt Mango).

1. *Konzept-zentrierte Extraktion und Organisation:* Wir führen ein ausdrucksstarkes CSK-Modell für Alltagskonzepte ein, mit: (i) verfeinerten Subjekten, einschließlich Untergruppen und Aspekten von Hauptsubjekten, (ii) semantischen Facetten für Aussagen und (iii) numerischen Indikatoren für Typikalität und Prägnanz. Für eine Reihe von Alltagskonzepten (z. B. `elephant`, `bicycle`) schlagen wir Ascent++ vor, eine automatisierte Methode zur Extraktion hochwertiger CSK-Aussagen aus groß angelegten Webinhalten. Ascent++ besteht aus verschiedenen neuen Techniken zur Aggregation und Bereinigung. Die resultierende CSK-Ressource umfasst 2 Millionen Aussagen für 10.000 wichtige Konzepte und übertrifft frühere Ressourcen sowohl in Bezug auf Abdeckung als auch Präzision.
2. *Kultur-zentrierte Extraktion und Organisation:* Für eine Reihe kultureller Gruppen (z. B. `Japanese`, `Buddhist`) schlagen wir die Candle-Methode zur Extraktion von kul-

turbezogenem Allgemeinwissen („*culture-aware commonsense knowledge*" – CCSK) aus einem großen Web-Korpus vor. Diese Methode umfasst durchdachte Techniken zur klassifikationsbasierten Filterung und Bewertung der Interessantheit, was zu einer groß angelegten CCSK-Ressource mit 60.000 Aussagen für 386 kulturelle Gruppen führt, die eine signifikant bessere Qualität aufweist als andere ähnliche Ressourcen.

3. *Kombination von Konzepten und Kulturen:* Wir schlagen Mango vor, eine Methodik zur effizienten Destillation von CCSK-Aussagen aus großen Sprachmodellen („*large language models*" – LLMs). Unsere Methode umfasst (i) die Konstruktion von Prompts für große Mengen an Konzepten und Kulturen und (ii) die Clusterung von Aussagen in thematisch und kulturell kohärente Gruppen. Die Ausführung der Mango-Methode mit GPT-3.5 als zugrundeliegendem LLM ergibt eine CCSK-Ressource von beispielloser Abdeckung (167.000 Aussagen, die 30.000 Konzepte und 11.000 Kulturen abdecken) mit noch höherer Qualität als Candle. In einer extrinsischen Studie für interkulturelle Dialoge zeigen wir, dass die Einbeziehung von Mango-Aussagen die Spezifität und kulturelle Sensibilität der LLM-Antworten signifikant verbessert.

Jede der erstellten CSK-Sammlungen wird für weitere Forschungsarbeiten veröffentlicht, mit einem webbasierten Wissensdatenbank-Browser sowie herunterladbarem Code und Daten.

Acknowledgments

I would like to thank my supervisors, Simon and Gerhard, for giving me the opportunity to pursue this doctoral degree, and for their guidance, knowledge, and insights that made this dissertation possible.

I would like to thank the reviewers, Fabian and Krishna, for their insightful feedbacks.

I would like to thank my collaborators for their valuable contributions to our joint works.

I would like to thank friends at D5 and in Saarbrücken for being friends with me.

And last but not least, I would like to thank my family for their unconditional love and support.

Contents

1

INTRODUCTION

1.1 Motivation

Structured knowledge, often stored in knowledge graphs (KGs), a.k.a. knowledge bases (KBs) (Hogan et al. 2021, Weikum et al. 2021), is a key asset for many artificial intelligence (AI) applications, including search, question answering, and conversational bots. KBs cover *encyclopedic knowledge* about named entities such as singers, songs, cities, sports teams, etc. However, even large-scale KBs deployed in practice hardly touch on the dimension of *commonsense knowledge* (CSK): properties of everyday concepts (e.g., elephants are big and eat plants, buses carry passengers and drive on roads), behaviors and emotions of humans (e.g., children love visiting zoos, children enter buses to go to school), and more. Such knowledge can benefit a wide range of AI applications as it enables systems to reason about the world in a more human-like way.

An important property of CSK is that, unlike encyclopedic knowledge which should hold universally (e.g., "The Lion King" was either produced by Disney, or it was not), the plausibility of CSK can vary depending on the context. For instance, elephants drinking milk only holds for baby elephants; tipping is customary in the USA, but it can be considered rude in Japan. This crucial aspect of CSK is often overlooked in prior CSK acquisition efforts.

Large language models (LLMs), such as the GPT models (Radford et al. 2019, Brown et al. 2020, Ouyang et al. 2022, OpenAI 2023), are machine learning models that learn to predict the next token given a sequence of previous tokens. LLMs are trained on corpora consisting of trillions of tokens and have shown impressive performances on various natural language processing (NLP) tasks that they were not directly trained for. Through the training process, these models have encoded a vast amount of knowledge in their parameters (Petroni et al. 2019, West et al. 2022), however, they still struggle with tasks that require rich world knowledge, where external knowledge sources can fill gaps. A recent study (Mallen et al. 2023) has measured the effectiveness of parametric and non-parametric memories when using LLMs for knowledge-intensive tasks (mostly on encyclopedic knowledge). They found that LLMs can be benefited from explicit knowledge retrieved from external sources by using retrieval-augmented generation (RAG) techniques (Guu et al. 2020, Lewis et al. 2020), especially for long-tail entities. While large-scale and systematic studies for CSK-augmented LLMs are still lacking, we anticipate similar effects as in Mallen et al.'s study on encyclopedic knowledge.

Let us consider two short exchanges between a human and the GPT-3.5 chatbot[1] in Figure 1.1.

- In *Exchange 1*, the bot totally missed the fact that unlike adult elephants, baby elephants do drink milk. Instead, it concluded that the user might have been mistaken with a different type of animal, or they have had a misleading perception.
- In *Exchange 2*, the bot complimented the user as they were giving a generous tip at a restaurant in Japan, where tipping is actually considered rude, unlike in many other parts of the world. The bot was not aware of local etiquette in this example.

These issues can be mitigated by providing the bot with explicit knowledge from external CSK resources (see Figure 1.2). Hence, having a KB of such information along with those powerful LLMs can be particularly beneficial in situations requiring commonsense reasoning. As opposed to learning-only approaches, an essential advantage of a materialized CSK resource is that it is scrutable, meaning that humans can manually inspect its assertions in order to add or remove contents when necessary, which is important for debugging complex downstream use cases and is highly relevant in applications where errors are costly.

This dissertation aims to advance the automatic acquisition of semantically-refined and culturally-contextualized CSK assertions via: (1) extraction from large-scale web contents, and (2) knowledge distillation from LLMs. In our extrinsic evaluations, the resulting CSK resources will be shown to benefit AI systems in various downstream applications.

1.2 Prior Work and Its Limitations

Large KBs like DBpedia (Auer et al. 2007), Wikidata (Vrandečić and Krötzsch 2014), or YAGO (Suchanek et al. 2007) focus on encyclopedic knowledge of named entities like people, organizations or places and are sparse on commonsense relations of general concepts. For

[1]Executed at https://platform.openai.com/playground/chat?models=gpt-3.5-turbo in May 2024

Exchange 1

User

I just saw an elephant drinking milk.

GPT-3.5

It's highly unlikely for an elephant to drink milk as they are lactose intolerant and their digestive system is not able to process milk in the same way as humans do. Elephants primarily drink water to stay hydrated and sustain their large bodies. If you saw something similar to an elephant drinking milk, it might have been a different type of animal or a misleading perception.

Exchange 2

User

I just had a great dinner in Tokyo with my friend. The bill is 8,000 yen. I'll leave 10,000 yen then!

GPT-3.5

That's generous of you! Leaving a 2,000 yen tip is a nice gesture to show your appreciation for the great dinner and service. Your friend and the restaurant staff will surely appreciate it!

Figure 1.1: Human-bot exchanges *without* explicit CSK.

example, from the Wikidata knowledge graph, Ilievski et al. (2020) identified 106K edges between 71K nodes that express commonsense relations. Whereas, 68% of these edges are `InstanceOf` and `SubclassOf` relations, which concern taxonomical information. This set constitutes a tiny fraction of Wikidata, which consists of more than 100M nodes and 17B edges.

Notable projects that compiled large sets of CSK include ConceptNet (Speer et al. 2017), WebChild (Tandon et al. 2014a), TupleKB (Dalvi Mishra et al. 2017), Quasimodo (Romero et al. 2019), ATOMIC (Sap et al. 2019a), TransOMCS (Zhang et al. 2020a), and ATOMIC-10x (West et al. 2022). In terms of knowledge representation, they all use subject-predicate-object triples, which have *limited expressiveness* (Section 1.2.1). Furthermore, these resources also *lack cultural contextualization* for assertions (Section 1.2.2), and suffer from either *low precision* or *limited coverage* (Section 1.2.3).

Exchange 1++

User

I just saw an elephant drinking milk.

GPT-3.5 + CSK*

It's highly unlikely that you saw an adult elephant drinking milk, as they are lactose intolerant and lack the enzyme needed to digest lactose after weaning. Baby elephants, on the other hand, do drink their mother's milk until they are weaned.

** Baby elephants drink milk.* [Inserted in system message]

Exchange 2++

User

I just had a great dinner in Tokyo with my friend. The bill is 8,000 yen. I'll leave 10,000 yen then!

GPT-3.5 + CSK*

In Japan, tipping is not expected or required, and leaving extra money can sometimes be seen as rude. When paying the bill, it is generally best to simply pay the amount on the bill without leaving additional money as a tip. If you had a great dining experience, it would be more appreciated to show your appreciation by thanking the restaurant staff or chef directly.

** In Japan, tipping can be considered rude.* [Inserted in system message]

Figure 1.2: Human-bot exchanges *with* explicit CSK.

1.2.1 Limited Expressiveness

Most prior commonsense knowledge bases (CSKBs) are restricted in their expressiveness to subject-predicate-object (SPO) triples with simple concepts for S and monolithic strings for P and O.

Expressiveness for Subjects. Prior CSKBs are typically centered on simple concepts represented by single nouns (e.g., `bus`, `car`, `elephant`, `trunk`). This is problematic because it misses semantic refinements (e.g., `diesel bus` vs. `electric bus`), which can lead to different properties (e.g., `polluting` vs. `ecofriendly`).

In addition, it leads to challenges in word-sense disambiguation, for instance, `elephant trunk` vs. `car trunk`. Even when multi-word phrases are considered, semantic relations among concepts are often not captured. Lexical resources like WordNet (Miller 1995) or Wiktionary (https://www.wiktionary.org/) also have limited coverage for multi-word concepts. With these limitations, word-sense disambiguation does not work robustly, as prior attempts like WebChild and TupleKB showed mixed results.

Expressiveness for Predicates and Objects. Most prior CSKBs treat predicates and objects as monolithic strings. For example, let us consider the following triples:

- A_1: `<bus; is used for; transporting people>`
- A_2: `<bus; is used for; bringing children to school>`
- A_3: `<bus; carries; passengers>`
- A_4: `<bus; drops; visitors at the zoo on the weekend>`.

This approach misses the equivalence between assertions A_1 and A_3 and does not capture the semantic refinement from A_1 to A_2. Moreover, the spatial facets of A_2 and A_4 (i.e., "to school", and "at the zoo"), and the temporal facet in A_4 (i.e., "on the weekend") are cluttered into unrelated strings.

Alternatively, some projects such as ConceptNet and WebChild restrict predicates to a small set of pre-defined relations (e.g., `AtLocation`, `HasTaste`, `HasPart`), and objects to concise phrases. However, it comes at the cost of much lower coverage.

1.2.2 Lacking Cultural Contextualization

Mainstream KGs do not cover *culture-aware commonsense knowledge* (CCSK) at all, and major CSK collections like ConceptNet contain only very few culturally contextualized assertions. As these resources focus solely on "universal CSK" which is agreed upon by almost all people, their assertions are viewed as "globally true".

To the best of our knowledge, the only prior works with data that have specifically addressed the socio-cultural dimension are the projects Quasimodo (Romero et al. 2019), StereoKG (Deshpande et al. 2022), and the work of Acharya et al. (2021). The latter merely contains a few hundred assertions from crowdsourcing, StereoKG uses a specialized way of automatically extracting stereotypes from QA forums and is still small in size, and Quasimodo covers a wide mix of general CSK and a small fraction of culturally relevant assertions.

1.2.3 Assertion Quality

Most prior CSKBs have prioritized either precision (i.e., the validity of the assertions) or coverage, but not both.

Some major CSKBs such as ConceptNet and TupleKB prioritize precision but have fairly limited coverage. On the other hand, some others have broader coverage but include many noisy if not implausible assertions, for example, WebChild, Quasimodo, and TransOMCS, which rely on knowledge extraction from web contents.

Moreover, the *saliency* of assertions, i.e., the degree to which statements are common knowledge, is considered by only a few projects like ConceptNet and ATOMIC, which rely on crowdsourcing knowledge acquisition and have limited coverage.

1.3 Challenges

While popular CSK resources (e.g., ConceptNet, ATOMIC) rely on human annotation, which has limited scalability, we opt for designing methods to acquire CSK automatically at scale. For better coverage of CSK, we leverage large-scale web texts and LLMs as the sources of CSK acquisition. Each approach represents a set of particular challenges, and there are shared challenges for knowledge organization.

1.3.1 Challenges of CSK Extraction from Web Texts

Large web crawls can be a rich source of CSK. However, the extraction of CSK from web texts poses several challenges:

1. *Sparsity of CSK in texts.* CSK is implicitly agreed upon by humans when they communicate. Such knowledge is often not explicitly written in texts either, as people assume that the knowledge is possessed by all (human) readers.
2. *Noisy and inaccurate information.* Besides the valuable knowledge presented in web contents, we also have to deal with noisy and inaccurate information written by web users.
3. *Biases and stereotypes.* Along with incorrect information, web texts also contain biased or stereotypical perspectives. Such contents should be eliminated in the extraction process so as not to reinforce stereotypes and offensive materials in downstream applications.

1.3.2 Challenges of CSK Distillation from LLMs

As LLMs are trained on large corpora of trillions of tokens, they implicitly encode a vast amount of knowledge in their parameters, and distilling knowledge from LLMs has shown promising results (Petroni et al. 2019, West et al. 2022). Although it is quite straightforward to generate knowledge using LLMs, this approach comes with the following challenges:

1. *Hallucinations.* Making up information, or hallucinating, is a well-known issue of LLMs (Bang et al. 2023, Zhang et al. 2023). Detecting and eliminating such information remains a particularly challenging task.
2. *Biases and stereotypes.* As LLMs are trained on web contents, its output can also reflect the biases and stereotypes presented in the training data.
3. *Political correctness.* To overcome the issue of biases and stereotypes in training data, LLMs are usually finetuned on smaller datasets to align with human preference. We have observed that such models tend to generate overly generic statements if the prompt contains socio-cultural groups. This can be problematic given our goal of acquiring salient culture-aware knowledge.

1.3.3 Challenges of CSK Organization

Organizing the collected CSK assertions (either from text extraction or LLM distillation) poses challenges of knowledge consolidation and ranking.

1. *CSK consolidation.* A single CSK assertion can be expressed in different ways. For instance, “tipping is not customary in Japan” and “tipping is not a common practice in Japan” convey the same meaning. Consolidating such assertions is crucial for downstream applications, as well as for eliminating redundancy in the output.
2. *Ranking of CSK assertions.* Unlike encyclopedic knowledge, CSK can be subjective and context-dependent. Ranking the significance of CSK assertions poses a challenging task.

1.4 Contributions

This dissertation tackle the problem of acquiring CSK with refined semantics at scale. In order to overcome the limitations of prior work and to address the challenges of automated CSK acquisition, we design expressive knowledge models with contextualized assertions, and propose acquisition methods to populate these models with high-precision and high-coverage CSK.

First, we address the expressiveness issues by introducing an advanced knowledge model for CSK about everyday concepts in the Ascent++ project (Chapter 3). Then, culture-aware CSK is acquired in the two projects Candle (Chapter 4) and Mango (Chapter 5), addressing the issue of lacking cultural contextualization for CSK assertions. Our methods consist of both simple techniques based on hand-crafted rules and dictionaries, traditional machine learning models (such as logistic regression, linear regression) based on simple features, as well as high-performance models based on fine-tuning LLMs. All of the methods strive for reconciling both high precision and wide coverage with salient assertions.

Ascent++, Candle, and Mango acquire refined CSK assertions from different entry points (concepts, cultures, or both). In each project:

1. We propose a *methodology* for large-scale CSK acquisition from its respective entry points.
2. We implement the method and produce a *CSK resource* that outperforms prior resources of similar kind in both intrinsic and extrinsic evaluations.

We summarize the main contributions of these projects in Table 1.1.

Lessons Learned. There are lessons learned across three main projects:

1. *Achieving high precision and wide coverage with automated CSK acquisition is possible.* We show that it is feasible to achieve high precision and wide coverage with automated CSKB construction if given sufficient thoughts about knowledge organization, source selection, and method design.
2. *CSKBs are beneficial for downstream applications.* In each project, we show that the resulting CSKB can benefit AI systems in various downstream applications, and that the quality of the CSKB is crucial for the performance of these applications.
3. *Standard data models are limitedly useful for CSK.* Standard data models like SPO triples have been used in most prior CSKBs, but they are limited in their expressiveness for CSK. We show that CSKBs using more expressive models with advanced semantics and contextualized assertions can outperform those using standard data models in both intrinsic and extrinsic evaluations.
4. *Clustering is important for frequency signals, and dealing with the heterogeneity of natural language.* As we collect CSK assertions from large web crawls and LLMs, duplicates and near-duplicates are common. While untreated redundancy is undesirable, we show that clustering assertions into coherent groups can help to identify frequency signals, which are useful for ranking and filtering.

Table 1.1: A summary of this dissertation's contributions.

Project	Publication	Entry points	Main contributions	
Ascent++ (Chapter 3)	In TKDE (Nguyen et al. 2023a)	Concepts	*Methodology*	We propose Ascent++, a methodology to extract refined CSK assertions (with expressive subjects and semantics facets for triples) from the large web crawl C4 (Raffel et al. 2020). Our method consists of various techniques for aggregation and cleaning.
			Resource	We construct and release a high-quality CSK collection of 2M assertions for 10K important concepts.
Candle (Chapter 4)	In WWW (Nguyen et al. 2023b)	Cultures	*Methodology*	We propose Candle, an end-to-end methodology for extracting high-quality culture-aware CSK (CCSK) from the C4 crawl, with new techniques for judiciously classifying and filtering culture-relevant text snippets and scoring assertions by their interestingness.
			Resource	We construct and release a large collection of 60K CCSK assertions covering 386 cultural groups of three domains (geography, religion, occupation), and several cultural facets (food, drinks, clothing, traditions, rituals, behaviors).
Mango (Chapter 5)	In CIKM (Nguyen et al. 2024)	Concepts *and* Cultures	*Methodology*	We propose Mango, a methodology for efficiently distilling culture-aware CSK from LLMs at high precision and high recall. Our method includes prompt construction for large sets of concepts and cultures, and clustering assertions into topically and culturally coherent groups.
			Resource	We construct and release a CCSK collection of 167K assertions for 30K concepts and 11K cultural groups, which substantially surpasses prior CCSK resources in size and quality, using GPT-3.5 as the underlying LLM for Mango.

1.5 Publications

The contributions of this dissertation are reflected in the following publications:

1. (Nguyen et al. 2023a) Tuan-Phong Nguyen, Simon Razniewski, Julien Romero, and Gerhard Weikum. "Refined Commonsense Knowledge from Large-Scale Web Contents." In: *IEEE Transactions on Knowledge and Data Engineering*, TKDE 2023.
2. (Nguyen et al. 2023b) Tuan-Phong Nguyen, Simon Razniewski, Aparna Varde, and Gerhard Weikum. "Extracting Cultural Commonsense Knowledge at Scale." In: *Proceedings of the ACM Web Conference*, WWW 2023.
3. (Nguyen et al. 2024) Tuan-Phong Nguyen, Simon Razniewski, and Gerhard Weikum. "Cultural Commonsense Knowledge for Intercultural Dialogues." In: *Proceedings of the ACM International Conference on Information and Knowledge Management*, CIKM 2024.

Other Publications. Besides the publications contributing to this dissertation, the author contributed to the following related papers:

- (Nguyen et al. 2021a) Tuan-Phong Nguyen, Simon Razniewski, and Gerhard Weikum. "Advanced Semantics for Commonsense Knowledge Extraction." In: *Proceedings of the Web Conference*, WWW 2021.
- (Nguyen et al. 2021b) Tuan-Phong Nguyen, Simon Razniewski, and Gerhard Weikum. "Inside ASCENT: Exploring a Deep Commonsense Knowledge Base and its Usage in Question Answering." In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, ACL Demos 2022.
- (Nguyen and Razniewski 2022) Tuan-Phong Nguyen and Simon Razniewski. "Materialized Knowledge Bases from Commonsense Transformers." In: *Proceedings of the First Workshop on Commonsense Representation and Reasoning*, CSRR 2022.
- (Singhania et al. 2022) Sneha Singhania, Tuan-Phong Nguyen, and Simon Razniewski. "LM-KBC: Knowledge base construction from pre-trained language models." In: *The Semantic Web Challenge on Knowledge Base Construction from Pre-trained Language Models*, LM-KBC 2022.
- (Arnaout et al. 2023) Hiba Arnaout, Tuan-Phong Nguyen, Simon Razniewski, and Gerhard Weikum. "UnCommonSense in Action! Informative Negations for Commonsense Knowledge Bases." In: *Proceedings of the 16th ACM International Conference on Web Search and Data Mining*, WSDM Demos 2023.

All resources and demonstrations along with related commonsense projects are available at https://www.mpi-inf.mpg.de/commonsense.

1.6 Outline

Chapter 2 gives background on commonsense knowledge, and reviews prior work on commonsense knowledge acquisition and applications of commonsense knowledge in AI. Chapter 3, Chapter 4, and Chapter 5 present our main projects Ascent++, Candle, and Mango, respectively. We conclude the dissertation in Chapter 6 with discussion on what we achieved from the projects, limitations of the proposed methods, and future research opportunities.

2

BACKGROUND AND RELATED WORK

In this chapter, we discuss the notion of commonsense knowledge and its characteristics (Section 2.1), prior work on commonsense knowledge acquisition (Section 2.2), and applications of commonsense knowledge in various areas of AI (Section 2.3).

2.1 Commonsense Knowledge

In our daily life, we constantly use common sense to make judgments and decisions about the world around us. This kind of knowledge is often implicit when people communicate with each other, because it is assumed to be known by most humans. For example, when someone says "I am going to the store to buy some milk", their partner would not ask if it is their neighborhood grocery store or the hardware store next door. In another example, when going to a Chinese restaurant, one would not be surprised that food is eaten using chopsticks and spoons, not forks and knives.

In artificial intelligence (AI), commonsense knowledge (CSK) usually refers to knowledge about general concepts (e.g., `elephants`, `bicycles`, `laptops`) and activities (e.g., `feeding dogs`, `fixing laptops`, `going to the zoo`), rather than details about named entities (e.g., `Germany`, `Kylian Mbappé`, `Max Planck Society`), which is often referred to as encyclopedic knowledge. Some example assertions of commonsense and encyclopedic knowledge are:

Commonsense knowledge	Encyclopedic knowledge
Elephants eat plants.	Berlin is the capital of Germany.
Bicycles have two wheels.	Kylian Mbappé plays for Real Madrid.
Feeding dogs requires dog foods.	Max Planck Society was founded in 1911.

Both commonsense and encyclopedic knowledge are foundational cornerstones for AI applications (Davis and Marcus 2015, Chen et al. 2020), but they have significantly different characteristics. While encyclopedic knowledge is well established, and it can be structured into a well-defined ontology with canonical relations (e.g., `CapitalOf`, `PlaysForTeam`, `FoundedInYear`), there is no universally accepted definition for CSK. "Common sense" is at best a vague term for humans and for AI.

Nevertheless, there are characteristics of CSK in AI that have been commonly pointed out in prior studies:

- *CSK is common.* CSK is fundamentally shared by large groups of people. A piece of knowledge only known to a few people is not considered common sense. For example, the fact that elephants have trunks is CSK, while the fact that elephants have 40,000 muscles in their trunks is expert knowledge.
- *CSK is context-dependent.* The correctness of a CSK assertion can vary between different cultures, times, locations, or social situations (Anacleto et al. 2006, Shwartz 2022). For instance, the statement "one should tip the waiter" is true in the United States, but not in Japan.
- *CSK does not concern individual entities.* CSK does not describe facts about individual entities (e.g., Berlin is the capital of Germany), but rather general concepts and activities (e.g., a country typically has a capital city).
- *CSK covers a wide range of topics.* CSK has a broad scope, covering a wide range of topics and facets, from physical properties of objects (Speer et al. 2017) to social norms (Forbes et al. 2020) and cultural practices (Acharya et al. 2021).

In this dissertation, we view CSK as *basic knowledge of broad scope about general concepts and activities that is possessed by large groups of people*, which is often implicit in human communication and varies within contexts. This understanding of CSK aligns with theories of common sense in sociology, anthropology, and psychology (Schütz 1944, Geertz 1983, McRae et al. 2005, Devereux et al. 2014, Tomasello 2014).

2.2 Prior Work on Commonsense Knowledge Acquisition

Commonsense knowledge is broad and diverse, covering a wide range of topics and facets. The acquisition of CSK has a long tradition in AI (Lenat 1995, Singh et al. 2002, Liu and Singh 2004, Gordon et al. 2010) with different focuses and methodologies. We provide a summary of notable CSK acquisition projects in Table 2.1.

In Section 2.2.1, we analyze these prominent projects by two dimensions, acquisition approach and knowledge representation, before detailing each project further in Section 2.2.2.

In Section 2.2.3, we discuss other related projects that have also contributed to the theme of CSK acquisition, including projects that focus on specialized CSK domains or expressive knowledge representations. Finally, in Section 2.2.4, we summarize the key points of the discussed projects.

2.2.1 Overview

From Table 2.1, it can be seen that prominent acquisition approaches are manual knowledge engineering, information extraction from texts, and generative methods using large language models. Most projects represent knowledge in the form of subject-predicate-object triples, and lack contextualization for assertions.

Acquisition Approach. Early projects on CSK acquisition, such as Cyc (Lenat 1995) and Open Mind Common Sense (OMCS) (Singh et al. 2002), were based on manual knowledge engineering, where human experts or volunteers entered assertions into the systems. This approach was later revived in the ATOMIC project (Sap et al. 2019a). While this approach ensures assertions with high precision, it does not scale well, hence the coverage of these resources is limited.

With the advances in information extraction (IE), more recent projects have focused on extracting CSK from large text corpora, such as WebChild (Tandon et al. 2014a), TupleKB (Dalvi Mishra et al. 2017), Quasimodo (Romero et al. 2019), and ASER (Zhang et al. 2020b). The advantage of this approach over manual KB construction is that it can capture a much larger amount of knowledge at a lower cost, however, the quality of the extracted assertions is often lower than that of hand-crafted ones.

Large language models (LLMs), which are trained on a vast amount of data and latently encode world knowledge into their parameters, have been used to generate CSK assertions, e.g., in the projects COMET (Bosselut et al. 2019) and ATOMIC-10x (West et al. 2022). These methods can create a large number of assertions without abundant engineering efforts, but need to face the issue of hallucination, where the models generate incorrect or nonsensical assertions. Another important downside of these methods is that the generated assertions are not traceable to specific sources, making it difficult to verify the correctness of the assertions.

Knowledge Representation. Most prominent CSK acquisition methods (10 out of 13 methods listed in Table 2.1) represent knowledge in the logical form of subject-predicate-object (SPO) triples, following the established knowledge representation in encyclopedic knowledge graphs. Predicates in these triples can either be pre-specified, such as in ConceptNet, WebChild, and ATOMIC, or in free-form, such as in Quasimodo. Subjects are usually simple

Table 2.1: Notable CSK acquisition projects sorted by first publication date.

Project	**Approach**	**Source**	**Domain**	**Representation**	**Context**
Cyc (Lenat 1995)	Manual	Experts (ontologists)	Broad	Formal language	Micro-theories (assumptions)
OMCS (Singh et al. 2002)	Manual	14K volunteers	Broad	Short sentences	N/A
ConceptNet (Liu and Singh 2004)	Semi-automated	OMCS, WordNet, etc.	Broad	Triples (fixed predicates)	N/A
WebChild (Tandon et al. 2014a)	Extractive	Texts, scripts, image tags	Narrow refined relations	Triples (fixed predicates)	N/A
TupleKB (Dalvi Mishra et al. 2017)	Extractive	Web texts (via search APIs)	Elementary science	Triples (fixed predicates)	N/A
Quasimodo (Romero et al. 2019)	Extractive	Query logs & QA forums	Broad	Triples (free predicates)	N/A
ATOMIC (Sap et al. 2019a)	Manual	MTurk workers	Event-centric	Triples (fixed predicates)	N/A
COMET (Bosselut et al. 2019)	Generative	Finetuned LMs	Broad	Triples (fixed predicates)	N/A
GenericsKB (Bhakthavatsalam et al. 2020)	Extractive	Cleaned sets of web texts	Broad	Generic sentences	Surrounding sentences
ASER (Zhang et al. 2020b)	Extractive	Web texts	Event-centric	Triples (fixed predicates)	N/A
TransOMCS (Zhang et al. 2020a)	Extractive	ASER	Broad	Triples (fixed predicates)	N/A
CSKG (Ilievski et al. 2021)	Data integrating	7 CSK resources	Broad	Triples (fixed predicates)	N/A
ATOMIC-10x (West et al. 2022)	Generative	General LLMs	Broad	Triples (fixed predicates)	N/A

concepts, and objects are often monolithic strings. These relational triples are stored in a graph, which can be used by graph-based reasoning methods, but they are often limited in expressiveness. In contrast, OMCS (in its native form, i.e., inputs from volunteers) and GenericsKB store knowledge in the form of short natural-language sentences, which can capture more complex knowledge and can be directly used by LLMs.

Most methods are limited to capturing knowledge that is assumed to hold universally (i.e., assertions are not contextualized). The only methods that consider collecting context of CSK are Cyc and GenericsKB. However, the latter only captures surrounding sentences of the extracted statements without any further semantic annotations, while the former does not provide a publicly accessible dataset or system.

2.2.2 Prominent CSK Acquisition Projects

In this subsection, we review 13 prominent CSK acquisition projects listed in Table 2.1.

Cyc. The Cyc project, starting in the 1980s up until the 2000s, was the first to construct a large knowledge base (KB) that captures commonsense knowledge, based on hand-crafting assertions by a team of ontology experts and knowledge engineers (Lenat 1995, Panton et al. 2006).

The Cyc KB is codified in a syntax language called CycL, a higher-order logic language which can theoretically capture any natural-language statement. The KB is organized into micro-theories, which are sets of specific assumptions shared by a group of assertions. While many succeeding CSK acquisition projects followed the triple-based knowledge model, Cyc has been clearly ahead of its time in terms of the expressiveness of its knowledge representation. However, there have been reports criticizing its usability due to the complexity and understandability of its syntax (see, e.g., (Conesa et al. 2008)).

According to their white paper (Cycorp 2021), the Cyc KB includes more than 40K predicates, more than 1.5M concepts, and more than 25M assertions. The average number of assertions entered by an expert is 25 assertions per hour. However, it is unclear how much CSK is covered by this KB, as it also includes domain-specific knowledge and knowledge about individual instances.

A non-commercial version of Cyc, called OpenCyc (Cycorp n.d.), had a much smaller size. The last version of OpenCyc (v4.0) was released in 2012. However, this project has been discontinued without any research licenses available. For these reasons, we could not compare Cyc or OpenCyc with our methods.

OMCS. Open Mind Common Sense (OMCS) (Singh et al. 2002) is a project that aimed to collect commonsense knowledge in the form of English statements from the general public, in contrast to Cyc, which was based on expert knowledge and restricted by formal languages. This crowdsourcing approach was made possible by the increasing number of Internet users and collaborative projects on the web in the early 2000s. Compared to Cyc, which costed several tens of millions of dollars and took more than 15 years to build, OMCS's approach was less expensive and acquired knowledge faster, although it comes with the cost of losing expressiveness in the knowledge representation and the quality of the assertions. On the other hand, by crowdsourcing to a large number of contributors (more than 14,000 volunteers), it created frequency signals for consolidating assertions (more than a million English statements were entered by volunteers, resulting in more than 100,000 consolidated assertions). There were two carefully designed user interfaces used by OMCS, namely OMCS-1 and OMCS-2.

The first version, OMCS-1, asked users to enter free-form English sentences that describe CSK given a short simple story. For instance, when prompted with *"Bob had a cold. Bob went to the doctor"*, users would enter *"Bob was sick"* or *"The doctor helped Bob feel better"*. The collected statements were also used to extract CSK triples based on simple syntactic hand-crafted patterns.

The second version, OMCS-2, was designed to collect more structured knowledge by asking users to enter the missing subject or predicate or object of a CSK triple given a predefined template. For example, given the template *"A <u>?</u> can have a <u>?</u>"*, users would enter *"A <u>mother</u> can have a <u>baby</u>"* or *"A <u>dog</u> can have a <u>tail</u>"*. OMCS-2 also included a peer-review mechanism between contributors to ensure the quality of the collected assertions.

The CSK assertions in OMCS were then integrated into the ConceptNet knowledge graph (Liu and Singh 2004), which has been one of the most widely used CSK resources.

ConceptNet. ConceptNet (Liu and Singh 2004, Havasi et al. 2007, Speer and Havasi 2013, Speer et al. 2017) combined CSK collected by human crowdsourcing and knowledge from existing resources such as WordNet (Miller 1995), OMCS (Singh et al. 2002), DBpedia (Auer et al. 2007), OpenCyc (Cycorp n.d.), and Wiktionary. This KB uses a triple-based data model, and it contains highly salient information for a few pre-specified predicates (e.g., `IsA`, `PartOf`, `UsedFor`, `CapableOf`, `LocationOf`, plus lexical relations such as `HasSynonymy`, `HasEtymology`, `HasDerivedTerm`, etc.).

ConceptNet has gone through several versions, each expanded the KB with more data from other resources.

- The original release of ConceptNet (Liu and Singh 2004) was built as a graph by parsing the OMCS assertions based on hand-crafted lexico-syntactic patterns, rule-based normalization, and relaxation rules to increase coverage (e.g., `<apple; IsA; red fruit>` and `<apple; IsA; red round object>` imply `<apple; HasProperty; red>`).

- ConceptNet 3 (Havasi et al. 2007) introduced a significant reorganization of ConceptNet with a more principled data model, including defining concepts, predicates and relations (e.g., for the `CapableOf` relation, the head concept should be a noun phrase, and the tail concept should be a verb phrase). Furthermore, the authors collected more data using Open Mind Commons (Speer 2007), an update of the original OMCS interfaces which aimed to discover new connections between existing nodes in the knowledge graph. This version also introduced assertion scores based on crowdsourcing inputs, as well as assertion polarity (representing negation) based on extraction patterns.
- ConceptNet 5.2 (Speer and Havasi 2013) extended the knowledge graph with more data from other resources, including OMCS's sister projects in other languages, WordNet 3.0, DBpedia (Auer et al. 2007), "games with a purpose" (Ahn et al. 2006) and relational statements mined from Wikipedia using ReVerb (Etzioni et al. 2008) - an open information extraction (OpenIE) system. It also introduced multilingual data by translating assertions into different languages.
- The latest release, ConceptNet 5.5 (Speer et al. 2017), further expanded the knowledge graph with more data from Wiktionary, a collaborative project to produce free-content multilingual dictionaries, and OpenCyc, the open version of the Cyc KB. Excluding the lexical relations, the English part of ConceptNet 5.5 contains more than 1.6 million CSK assertions.

A simple yet well-defined knowledge model and high-quality assertions have contributed to the popularity of ConceptNet in many applications. However, it has limited coverage on many concepts, and its ranking of assertions, which is based on the number of crowdsourcing inputs, is very sparse and unable to discriminate salient properties against atypical or exotic ones (e.g., listing `tree`, `garden`, and `the bible` as locations of `snake`, with similar scores). ConceptNet does not properly disambiguate concepts, leading to incorrect assertion chains like `<elephant; HasPart; trunk>` and `<trunk; LocationOf; spare tire>`. Furthermore, the distinction between CSK and encyclopedic knowledge in ConceptNet 5.5 has become unclear, as it has included relations extracted from Wikipedia and DBpedia, for example, geo-knowledge relations like `<Berlin; PartOf; Germany>`.

WebChild. WebChild (Tandon et al. 2014b, 2014a, 2015a, 2015b, 2016, 2017) was one of the first attempts to automatically extract CSK at a large scale, based on information extraction from texts. WebChild's knowledge representation is based on SPO triples, similar to ConceptNet but with more refined relations and disambiguated concepts mapped to WordNet senses. WebChild's knowledge extraction pipelines used pattern-based OpenIE and semantic parsing techniques on web texts (e.g., Wikipedia dump), image tags, and movie scripts. Concept sense disambiguation was based on judiciously designed integer linear programming.

The final WebChild 2.0 KB (Tandon et al. 2017), which contains more than 18 million CSK assertions covering over 2 million concepts and activities, is composed of results from several sub-projects:

- Tandon et al. (2014a) expanded the `HasProperty` relation in ConceptNet with more specific properties, namely `HasShape`, `HasTaste`, and `EvokesEmotion`. Assertions were extracted from the Google Web 1T N-Gram Dataset (Brants and Franz 2006).
- Tandon et al. (2014b) extracted comparative commonsense knowledge (e.g., `<steel; sharper than; wood>`, `<car; faster than; bicycle>`) using pattern-based OpenIE on web texts.
- Tandon et al. (2015a) and Tandon et al. (2015b) extracted activity-centric CSK from movie scripts, i.e., annotating activities with their (typical) participants, location, time, etc., and linking them to previous or next activities. For example, the activity `climbing up a mountain` is annotated with the participating agents `climber` and `rope`, the location `mountain`, the times `daylight` and `holiday`, and is linked to the previous activity `packing a backpack` and the next activity `drinking water`.
- Tandon et al. (2016) mined refined part-whole commonsense relations, namely `PhysicalPartOf`, `MemberOf`, and `SubstanceOf`, from web texts and image tags, refining the general `PartOf` predicate in ConceptNet.

Although WebChild's relations are more refined than ConceptNet's and it comes with disambiguated concepts, the KB suffers from quality issues due to the noisy nature of web texts (see (Romero et al. 2019) and (Nguyen et al. 2021a) for comparisons of intrinsic quality between CSK resources). In addition, it has a limited scope due to the specific focuses.

TupleKB. TupleKB (Dalvi Mishra et al. 2017) aimed to extract triples of elementary science knowledge from web texts. The TupleKB construction pipeline used a fixed domain vocabulary (i.e., elementary science concepts), their corresponding types (e.g., animal, plant, body part, etc.), and commercial search APIs with handcrafted queries to retrieve relevant sentences. It then used OpenIE to extract triples from the retrieved sentences. Next, a subset of the extracted triples was verified by Amazon MTurk workers for their plausibility. Finally, these annotated samples were used to train a regression model to score the remaining triples.

TupleKB construction also consisted of a step for canonical schema induction, which grouped semantically similar relations into a canonical, generalized relation, for example, it maps: `<(type:animal); munch on; (type:animal)>`, `<(type:animal); chew; (type:animal)>`, and `<(type:animal); consume; (type:animal)>` to a canonical form: `<(type:animal); eat; (type:animal)>`. The canonical schema induction step was based on integer linear programming with carefully designed constraints.

TupleKB contains more than 340K triples restricted to elementary science knowledge and the triple-based knowledge model.

Quasimodo. Quasimodo (Romero et al. 2019) extracted CSK from query logs (via Google and Bing search engines) and post titles from question-answering forums (e.g., Quora, Reddit). The Quasimodo KB is based on SPO triples for knowledge representation, similar to ConceptNet and others, but it drops the pre-specified predicates and uses free-form predicates, which allows it to capture CSK at a broader scope.

The motivation behind the use of Internet users' questions as the knowledge source in Quasimodo is twofold: (1) these questions implicitly convey knowledge (e.g., when someone asks Google *"why do dogs bark?"*, the question implies that dogs bark), and (2) these questions are a good source of salient knowledge, as they are often about things that many people care about.

Based on those observations, the authors utilized a set of handcrafted patterns, to collect candidate queries using search engines' auto-complete feature. These patterns were constructed by concatenating a question word (*"why"* or *"how"*) with one of the verbs *"is"*, *"do"*, *"are"*, *"does"*, *"can"*, *"can't"* followed by a subject of interest (e.g., *"cats"*). For example, given the initial query *"why do cats"*, the method collected the auto-complete suggestions *"why do cats purr"*, *"why do cats like boxes"*, etc. A similar pattern-based approach was used to collect relevant questions from question-answering forums.

Given the collected human-written questions, Quasimodo used OpenIE to extract triple candidates, which would go through a corroboration step for verifying their plausibility. The corroboration step was based on collecting occurrences of the triple candidates from various sources, such as Wikipedia, Google Books, and image tags from Flickr. These evidences were then used to train a regression model for ranking the triple candidates.

The Quasimodo KB v4.3 contains 6.2 million CSK assertions covering 148K concepts. It offers a more diverse set of CSK compared to prior resources such as ConceptNet and TupleKB, however, due to noise, biases and emotion-driven queries in the search engine logs and QA forums, many of the extracted triples are incorrect or nonsensical, e.g., `<elephant; be; the best>`, or `<musician; be; aged>`.

ATOMIC. In the ATOMIC project (Sap et al. 2019a), the authors focused on collecting CSK about if-then relations, for example, if PersonX pays PersonY a compliment, then PersonY will likely feel happy. Although some activity-centric relations are present in previous resources such as ConceptNet (e.g., `Causes`, `HasSubEvent`, `MotivatedByGoal` relations), ATOMIC was the first to systematically collect such relations at a large scale. ATOMIC also uses a triple-based knowledge representation with a fixed set of nine predicates (e.g., `xNeed`, `xIntent`, `xWant`).

ATOMIC knowledge was collected through a crowdsourcing framework with Amazon MTurk workers. First, the authors collected a set of 24K events from various sources such as books, Google n-grams, and Wiktionary, followed by post-processing steps for disambiguation and normalization. Then, given an event (e.g., “PersonX pays PersonY a compliment”), MTurk workers were asked to write possible causes or effects of the event (e.g., “PersonY will likely feel happy”).

The ATOMIC knowledge base contains 877K if-then assertions. This KB was used to train a generative model, namely COMET (Bosselut et al. 2019), to generate new CSK triples. A later version of the KB, called ATOMIC$^{20}_{20}$ (Hwang et al. 2021), was created by extending the original ATOMIC KB with more data from ConceptNet as well as by collecting more crowdsourcing inputs. ATOMIC$^{20}_{20}$ contains 1.3M assertions spanning three genres of CSK, namely social-interaction, event-centered, and physical-entity knowledge. It was subsequently extended to a much larger KB called ATOMIC-10x (West et al. 2022) by judiciously prompting large language models.

COMET. COMET (Bosselut et al. 2019) was the first notable work to finetune a language model (LM) in order to generate new CSK triples. The COMET model was based on the GPT-2 architecture (Radford et al. 2019), and it was finetuned on ConceptNet and ATOMIC data. GPT-2 is a transformer-based causal language model that predicts the next token in a sequence given the previous tokens. The finetuning process was done by training the model to predict the object tokens given the subject and predicate tokens in a CSK triple. For example, given the subject `going to mall` and the predicate `xIntent`, the model should predict object `buying clothes` among others.

Although the model achieved promising results, its generated triples are often of lower quality compared to the training resources (Nguyen and Razniewski 2022). In addition, the generated triples are not traceable to their sources, in contrast to the extractive methods.

GenericsKB. GenericsKB (Bhakthavatsalam et al. 2020) dropped attempts at structuring assertions and instead focused on collecting generic sentences on a per-subject basis, for example, “Dogs bark”, or “Trees remove carbon dioxide from the atmosphere”. This KB was created by extracting generic statements from three corpora, namely the Waterloo corpus of 280GB of plain English text crawled from educational domains in 2001, filtered SimpleWikipedia (https://simple.wikipedia.org) pages, and the ARC corpus of 14M general sentences (Clark et al. 2018). GenericsKB’s extraction method was based on handcrafted patterns for cleaning, lexico-syntactic rules for selecting generic sentences, and a BERT-based classifier for filtering out irrelevant sentences.

The GenericsKB contains more than 3.4M generic sentences. A smaller version of GenericsKB that contains more than 1M statements, called GenericsKB-Best, was created by selecting the top-scoring generic sentences combined with sentences synthesized from WordNet and ConceptNet assertions.

ASER. ASER (Zhang et al. 2020b) is another knowledge graph (KG) extracted automatically from web texts, however, it focuses on events and relations between them. For example, in the ASER graph, the event “I am hungry” is connected to the event “I have lunch” by the relation `Result`. The authors defined 15 relation types for the KG, such as `Result`, `Reason`, `Condition`, etc.

This KG was constructed by extracting pairs of events using dependency patterns on several text sources, such as Yelp, New York Times, Wikipedia, and Reddit. The relations between the events were then automatically predicted by using a bootstrap-based method that leveraged seed training instances inferred from a manually labeled corpus (the Penn Discourse TreeBank 2.0 (Prasad et al. 2008)).

The resulting KB contains 194M events and 64M unique edges among them. Although it provides a rich source of event-centric CSK, ASER suffers from high noise and redundancy.

TransOMCS. The TransOMCS project (Zhang et al. 2020a) was an attempt to distill commonsense knowledge from ASER, mapping to canonical relations in ConceptNet. First, the authors matched ConceptNet triples with equivalent assertions in the ASER graph to learn extraction patterns, and then used these patterns to extract more triples from ASER. The extracted triples were ranked by a classifier based on BERT embeddings, graph-attention embeddings and frequency features, which was trained on 1,000 examples annotated by MTurk workers.

The TransOMCS knowledge base contains more than 18M triples. Although being one of the largest CSK resources in terms of the number of assertions, studies have shown that TransOMCS has a high level of noise and redundancy (see e.g., (Hwang et al. 2021)).

CSKG. The CSKG project (Ilievski et al. 2021) aimed to integrate seven existing CSK resources into a single knowledge graph, whereas ConceptNet and ATOMIC were the main sources that made up the majority of the integrated commonsense knowledge graph (called CSKG) - approximately 87% of the nodes and 70% of the edges came from these two KBs. CSKG has a canonical set of relations, and provides a higher coverage of CSK compared to individual resources (6M edges and 2.2M nodes). However, as the sources are all based on manual labor, the coverage of CSKG is still limited compared to automated methods.

ATOMIC-10x. ATOMIC-10x (West et al. 2022) consists of 6.4M CSK triples collected by prompting GPT-3 (Brown et al. 2020) with judiciously designed prompts with in-context examples selected from ATOMIC$^{20}_{20}$ (Hwang et al. 2021).

The generation process consisted of two steps: (1) event generation, where GPT-3 was asked to generate a new event given a set of example events picked from ATOMIC$^{20}_{20}$, and (2) relation generation, where the LLM was used to generate tail events given a head event and a relation, with examples provided in the prompt. The generated triples were then filtered by a RoBERTa classifier (Liu et al. 2019) finetuned on a small set of labeled data.

ATOMIC-10x was shown to have a higher quality compared to ATOMIC$^{20}_{20}$, and is ten times larger in size (6.4M vs. 0.64M triples for the six predicates: `HinderedBy`, `xNeed`, `xWant`, `xIntent`, `xReact`, `xAttr`, and `xEffect`), hence the suffix "10x".

Generating knowledge from LLMs has the advantage of being able to capture a wider range of CSK, but it also faces the issue of hallucination, and provides no traceability to sources of the generated knowledge, which is crucial for debugging and verification, especially in safety-critical applications.

2.2.3 Other Related Work

There are other projects that have contributed to the acquisition of CSK but were not covered in the subsection above. These are projects that focus on special CSK domains, such as cultural and social knowledge, and projects of related fields, such as expressive knowledge representation and extraction.

Cultural CSK Acquisition. There are only few works on culture-aware knowledge: (Anacleto et al. 2006), (Acharya et al. 2021), (Shwartz 2022), StereoKG (Deshpande et al. 2022), NormsKB (Fung et al. 2023), (CH-Wang et al. 2023), and GD-COMET (Bhatia and Shwartz 2023). These methods are based on extraction from text (Deshpande et al. 2022, Shwartz 2022), norm discovery from multi-lingual conversations (Fung et al. 2023), or fine-tuning LLMs on cultural knowledge bases (Bhatia and Shwartz 2023). On the other hand, CH-Wang et al. (2023) aligned social situations from an English knowledge base (Forbes et al. 2020) and Chinese QA forums to mine 3,069 social norms for these two cultures. All of these prior works are very limited in scope and scale; some even suffer from high noise. In addition, many of them do not release their resources publicly.

Our projects Candle (Chapter 4) and Mango (Chapter 5) aim to acquire culture-aware CSK with high precision and wide coverage with publicly accessible resources.

Social CSK Acquisition. The Social-Chem-101 (Forbes et al. 2020) dataset, built by crowdsourcing annotations, consists of 292K rule-of-thumb assertions around social situations (e.g., “It is rude to interrupt someone”) via crowdsourcing.

Another social CSK resource is NormBank (Ziems et al. 2023), which contains 155K situational social norms. Situations were generated by an LLM, including a setting (e.g., `cafe`, `classroom`), a person's behavior (e.g., `drinking hot coffee`) and the person's role (e.g., `barista`, `customer`) and attributes (e.g., `child`, `adult`). These situations were then annotated by humans to indicate if the behavior followed social norms (labeled as *expected*, *okay*, or *unexpected*).

Taxonomy and Meronymy Induction. There has been great attention in NLP and web mining to the organization of concepts in terms of subclass and part-whole relationships, termed hypernymy and meronymy, which are also considered commonsense relations. Notable works include (Etzioni et al. 2004, Girju et al. 2006, Pantel and Pennacchiotti 2006, Snow et al. 2006, Pasca and Durme 2008, Ponzetto and Strube 2011, Wu et al. 2012, Hertling and Paulheim 2017). A widely used resource, the manually curated WordNet lexicon (Miller 1995), organizes over 100K synonym sets to capture these relationships, though it contains relatively few entries for meronymy.

Recent approaches for large-scale taxonomy induction from web data include WebIsADB (Seitner et al. 2016, Hertling and Paulheim 2017), which builds upon Hearst patterns and other techniques, and the industrial GIANT ontology (Liu et al. 2020), which utilizes neural learning from user-action logs and additional sources.

Large-scale meronymy induction has been tackled by Tandon et al. (2016) and Bhakthavatsalam et al. (2020) using pre-defined and automatically learned patterns, targeting specific relations like `PhysicalPartOf`, `MemberOf`, and `SubstanceOf`.

Our Ascent++ methodology (Chapter 3) will incorporate both kinds of relations by extracting knowledge about salient subgroups and aspects of subjects. Unlike conventional taxonomies and part-whole datasets, our subgroups include many multi-word phrases: composite noun phrases (e.g., `forest elephant`, `elephant keeper`) and adjectival and verbal phrases (e.g., `male elephant`, `working elephant`). The aspects in our approach cover further refinements of subjects that do not fit into taxonomy or meronymy (e.g., `elephant's diet`, `elephant habitat`).

Expressive Knowledge Representation and Extraction. Most prior works on CSK acquisition rely on the traditional triple-based data model as knowledge representation, which comes with expressiveness issues. Although expressive knowledge representations that capture semantic frames (Hogan et al. 2021) or modal logics such as `always`, `often`, `rarely`, and `never`

(Gabbay 2003) have been around for decades, research on populating these refined models has been limited. Notable exceptions include small-scale projects like Knext (Schubert 2002) and OntoSenticNet (Dragoni et al. 2018), which focuses on affective valence annotations.

Few studies on contextualizing CSK assertions include (Zhang et al. 2017), which scored natural language sentences on an ordinal scale ranging from `very likely` to `impossible`; (Chen et al. 2020), which gave assertions probabilistic scores; and the Dice project (Chalier et al. 2020) which ranked assertions by four dimensions: plausibility, typicality, remarkability, and saliency.

In the task of semantic role labeling (SRL), sentences are mapped onto frames (often corresponding to specific types of events), with respective slots (e.g., agent, participant, instrument) filled with values extracted from the input text (Palmer et al. 2010, Clarke et al. 2012, Stanovsky et al. 2018). Facet-based open information extraction (Cetto et al. 2018, Prasojo et al. 2018), built on this paradigm, extract tuples with qualifying semantic facets such as time, location, and mode. Our Ascent++ methodology (Chapter 3) will extent this approach in various ways geared for the case of CSK. Specifically, we focus on facets specifically relevant to CSK, refine subjects by subgroups and aspects, and strive to reconcile both high precision and wide coverage for CSK extraction.

Commonsense Benchmarks. A different kind of commonsense dataset is commonsense reasoning benchmarks, which often have smaller sizes and are used to evaluate the commonsense reasoning abilities of models in specific tasks. Some text-based benchmarks are CommonGen (Lin et al. 2020), CommonsenseQA (Talmor et al. 2019), CommonsenseQA 2.0 (Talmor et al. 2021), HellaSwag (Zellers et al. 2019), NumerSense (Lin et al. 2020), PIQA (Bisk et al. 2020), RiddleSense (Lin et al. 2021), Social IQa (Sap et al. 2019b), SWAG (Zellers et al. 2018), Winograd Schema Challenge (Levesque et al. 2012), and Winogrande (Sakaguchi et al. 2021). Notable image-based benchmarks include Adversarial VQA (Li et al. 2021), COFAR (Gatti et al. 2022), KB-VQA (Wang et al. 2017), VCR (Zellers et al. 2019), Visual Genome (Krishna et al. 2017), and WinoGAVil (Bitton et al. 2022). We refer readers to the survey of more than 100 commonsense reasoning benchmarks by Davis (2023).

In terms of culture-aware NLP, there have been a few datasets used for benchmarking commonsense reasoning abilities of models in cultures beyond the Western world. These include MaRVL (Liu et al. 2021), GD-VCR (Yin et al. 2021), GeoMLAMA (Yin et al. 2022), CALI (Huang and Yang 2023), FORK (Palta and Rudinger 2023), CultureAtlas (Fung et al. 2024), and CulturalTeaming (Chiu et al. 2024).

2.2.4 Summary

Notable CSK acquisition efforts have been made over the past few decades, with different focuses and methodologies. The early projects, such as Cyc and OMCS, relied on manual knowledge engineering, while recent projects, such as WebChild and Quasimodo, have focused on extracting CSK from large text corpora. Most recently, large language models have been used to generate CSK assertions, as in COMET and ATOMIC-10x. These projects mostly use a triple-based knowledge representation, with significant limitations in expressiveness. In addition, these prominent resources do not capture culture-aware commonsense knowledge. The few projects focusing on cultural CSK, on the other hand, suffer from low coverage or high noise.

The quality of the extracted assertions varies across projects, with manual methods often producing higher-quality assertions but at a lower scale, and automated methods producing a larger number of assertions but often suffering from lower precision. Among the automated methods, those based on large language models often lead to CSK of a wider coverage and sometimes higher quality compared to extractive methods, but they also face the issue of hallucination and lack of traceability to knowledge sources.

Our work in this dissertation strives for reconciling both high precision and high recall of CSK acquisition, based on new techniques for information extraction from large web crawls, LLM prompting, assertion clustering and consolidation. We also introduce expressive knowledge models that capture advanced semantic facets and cultural contextualization for CSK assertions.

2.3 Applications of Commonsense Knowledge

Although AI has beaten humans in expert-level games like Chess and Go (e.g., IMB's Deep Blue and Google's AlphaGo systems), tasks that require common sense, which humans find trivial, remain challenging for such advanced systems. State-of-the-art AI bots like ChatGPT, despite implicitly possessing commonsense knowledge, continue to be brittle in commonsense reasoning tasks. Building a general AI with common sense remains a complex and long-term challenge, which requires not only compiling large sets of commonsense knowledge, but also developing reasoning mechanisms that can effectively utilize this knowledge (Brachman and Levesque 2022).

Nevertheless, materialized CSK resources have been used in many smaller applications across major areas of AI, including natural language processing, computer vision, and robotics. In the following subsections, we briefly discuss some of the recent efforts on CSK-enhanced AI applications, which usually involve deep learning models integrated with CSK resources to improve their robustness and generalizability. For more dated applications in the eras before deep learning, we refer readers to surveys by Lieberman et al. (2004) and Tandon (2016).

2.3.1 Natural Language Processing

Deep neural networks (DNNs), including large language models (LLMs), have marked a significant advancement in natural language processing (NLP) in the last decade. DNNs rely on learning latent representations of texts from large corpora, showing superior performance in various NLP tasks. However, the downside of these models is that they lack interpretability compared to traditional NLP approaches based on explicit lexical and syntactic features.

In particular, LLMs, whose goal is to predict the next token in a sequence given the previous tokens, are trained on corpora of trillion-token scale. LLMs have been achieving the highest scores on multiple NLP benchmarks including ones aimed at commonsense reasoning. For example, GPT-4 (OpenAI 2023) performs on par with humans on the HellaSwag benchmark (Zellers et al. 2019), which tests commonsense inference via the task of sentence completion. However, doing well on such benchmarks does not mean these models "have common sense". As pointed out by Davis (2023), existing commonsense reasoning benchmarks are often flawed and unreliable as they consist of erroneous test samples, and still many aspects of commonsense reasoning have not been addressed. Indeed, we have seen that state-of-the-art LLMs can be easily attacked by simple adversarial examples despite having nearly perfect scores on several commonsense benchmarks.

In spite of relying on learning-only approaches, there have been works that utilized commonsense resources, such as ConceptNet (Speer et al. 2017) and ATOMIC (Sap et al. 2019a), to improve commonsense reasoning ability of NLP models in specific tasks. Using such resources provides better interpretability and scrutability to the systems.

The applications of CSK in NLP include:

- *question answering* (Lin et al. 2019, Talmor et al. 2019, Chen et al. 2020, Lv et al. 2020, Wang et al. 2020, Bosselut et al. 2021),
- *sentiment analysis* (Ofek et al. 2016, Dragoni et al. 2018, Ma et al. 2018),
- *emotion detection* (Ghosal et al. 2020, Li et al. 2021, 2022, Yang et al. 2023),
- *dialogue and response generation* (Zhou et al. 2018, Zhou et al. 2022, Wu et al. 2020, 2020, Zhang et al. 2020c, Varshney et al. 2022, Cai et al. 2023, Kim et al. 2023, Li et al. 2023),

- and others such as *stance detection* (Liu et al. 2021), *title-to-essay generation* (Yang et al. 2019), *sarcasm generation* (Chakrabarty et al. 2020), *sarcasm detection* (Li et al. 2021), *intent detection* (Siddique et al. 2021), *sentence ordering* (Ghosal et al. 2021), and *fake news detection* (Gao et al. 2023), to name a few.

Let us consider an example for the task of dialogue response generation:

- Context: John, an American, is visiting his friend Kenji, who lives in Tokyo. They are paying their bill for dinner at a restaurant.
- Ongoing dialogue:
  - John: That's a great meal, Kenji. I really liked the sushi.
  - Kenji: My pleasure, John. I'm glad you enjoyed it.
  - John: Let me see the bill. It is 8,000 yen. I'm gonna leave 10,000 yen.

In this scenario, a dialogue system that is aware of common sense in Japan, i.e., tipping is not customary and can be considered rude, should generate a response for Kenji that indicates John's intention to leave a tip is unnecessary, instead of a generic response like "Thank you, John.", or "That's a very generous tip. Thanks, John!".

We summarize the four popular approaches to incorporate CSK in these NLP applications:

- *Retrieval-augmented generation (RAG):* This approach is used with LLMs, where the model is explicitly prompted with the task and relevant CSK assertions retrieved from external resources to support its commonsense reasoning, e.g., (Chen et al. 2020, Kim et al. 2023).
- *Embeddings integration:* CSK assertions or subgraphs are encoded as embeddings, which are integrated with other input embeddings of the model to improve its performance, e.g., (Liu et al. 2021).
- *Fine-tuning with CSK:* DNNs are trained on CSK resources to improve their commonsense reasoning abilities, e.g., (Ghosal et al. 2020).
- *Reasoning over CSKGs:* This approach is often used by question-answering systems, where the model reasons over a CSKG to answer questions, e.g., (Lin et al. 2019, Lv et al. 2020). It can also be used to identify relevant evidence for other classification tasks, e.g.,. (Siddique et al. 2021).

2.3.2 Computer Vision

With the advancement of NLP models and their promising performance on natural language understanding and generation, there have been efforts to integrate these models to solve computer vision tasks that require semantic understanding of objects and their relations and interactions in images or videos.

One of the most important tasks in visual understanding and reasoning is *scene graph generation* (SGG), where the goal is to construct a graph that represents the objects in the input image and their relationships. The scene graph, which provides rich semantic information of an image, can be used to answer questions about the image (Zhang et al. 2022), generate image captions (Yang et al. 2019), support image retrieval (Gatti et al. 2022), image generation (Gu et al. 2019, Fu et al. 2024), visual story telling (Chen et al. 2021), and more. SGG methods often consist of two main components: object detection and localization, and relationship prediction. The latter is where CSK resources can be beneficial, as they provide prior knowledge about the relationships between objects, instead of relying solely on limited training data. For example, the fact that dog has head and tail is CSK that most humans possess and would constantly use if they were to manually sketch a scene graph of a given photo of a dog catching a frisbee. However, such knowledge is sparse in SGG training data like Visual Genome (Krishna et al. 2017), making it challenging for models to learn such relationships from data alone. Integrating an external CSK resource into such situations can boost the generalizability and robustness of these models.

A popular approach is to integrate CSK graph embeddings into SGG models to enhance their commonsense reasoning (Gu et al. 2019, Guo et al. 2020, Kan et al. 2021, Khan et al. 2022, Zhang et al. 2022). On the other hand, Zareian et al. (2020) uses graph-based neural networks (GNNs) to bridge a scene graph proposal generated by a simple model and a CSK graph of detected objects gathered from external sources.

Another approach to visual question answering (VQA) that does not rely on scene graph generation is (Ravi et al. 2023), which proposes to search for relevant CSK assertions (from external sources) given names of objects in the input image and the input question. These assertions are then encoded and integrated into a transformers-based model to improve the performance of VQA.

2.3.3 Robotics

In robotics, CSK has been used to help robots perform tasks in dynamic or unknown environments, and to optimize their efficiency. That is inspired by the fact that humans usually adapt to new situations using their commonsense knowledge. For example, when a robot is asked to find a cup, it should head to the kitchen and search in the cupboard or cabinet, like how humans would typically do, instead of immediately looking for it in the bathroom. This example is a simple case of the *object localization* task, where the robot needs to find the location of an object as per request. Recent works (e.g., (Daruna et al. 2019, Zhang et al. 2019, Chernova et al. 2020)) have integrated CSK resources such as ConceptNet (Speer et al. 2017) and LabelMe (Russell et al. 2008), via graph embeddings or as prior knowledge in a probabilistic model, leading to improved performance in this task.

Similarly, CSK resources are used in a wide range of other robotics applications, such as *object recognition* (Pratama et al. 2014, Kümpel et al. 2020, Chiatti et al. 2022), *object delivery* (Al-Moadhen et al. 2015, Zhang and Stone 2015, Wang et al. 2019), *pick and place* (Al-Moadhen et al. 2013, Javia and Cimiano 2016, Mitrevski et al. 2021), *warehousing* (Ayari et al. 2015, Pradeepani et al. 2022), *cooking* (Nyga and Beetz 2012, Agostini et al. 2015), to name a few. Interested readers are referred to the survey by Töberg et al. (2024) for a comprehensive review of recent CSK applications in robotics.

In this dissertation, we develop methods to produce CSK resources of high quality and high coverage. Importantly, we will show that our CSK resources can be used to boost the performance of AI models in various extrinsic use cases, such as question answering and dialogue generation, and outperform other resources as evaluated by human annotators.

3

CONCEPT-CENTRIC EXTRACTION AND ORGANIZATION

In this chapter, we acquire commonsense knowledge (CSK) for the first type of entry points: everyday concepts (e.g., `elephant`, `bicycle`, `beer`).

To address the expressiveness limitation of the triple-based data model in prior commonsense knowledge bases (CSKBs), we introduce an expressive knowledge model that captures composite concepts with subgroups and aspects, as well as refines assertions with semantic facets expressing temporal and spatial validity and further qualifiers. Our proposed method, called ASCENT++, combines open information extraction (OpenIE) with judicious cleaning and ranking by typicality and saliency scores to extract high-precision CSK assertions from general web contents. As knowledge source, we tap into the large-scale crawl C4 for high coverage of CSK. The evaluation with human judgments shows the superior quality of the KB, and an extrinsic evaluation for QA-support tasks underlines the benefits of ASCENT++.

The project website is hosted at https://ascentpp.mpi-inf.mpg.de, including downloadable code and data.

3.1 Introduction

Motivation. Mainstream KBs such as DBpedia (Auer et al. 2007), Wikidata (Vrandečić and Krötzsch 2014), and YAGO (Suchanek et al. 2007) have their main focus on encyclopedic knowledge of named entities (e.g., Berlin is the capital of Germany) and are sparse on knowledge of general concepts (e.g., a country typically has a capital city). On the other hand, there have been projects concentrating on CSK, notably ConceptNet (Speer et al. 2017), TupleKB (Dalvi Mishra et al. 2017), WebChild (Tandon et al. 2014a), Quasimodo (Romero et al. 2019), ATOMIC (Sap et al. 2019a), TransOMCS (Zhang et al. 2020a) and ATOMIC-10x (West et al. 2022). However, they all use subject-predicate-object (SPO) triples to represent CSK, which have significant shortcomings in expressiveness.

- *Expressiveness for S:* Prior works on CSK acquisition have a strong focus on single-noun concepts such as car, elephant, trunk. A major disadvantage of this approach is that it misses refined concepts which may have different properties, e.g., diesel cars are polluting, but electric cars are ecofriendly. Furthermore, ineffective disambiguation of concepts (e.g., elephant trunk vs. car trunk) can lead to incorrect assertion chains like <elephant; HasPart; trunk> and <trunk; LocationOf; spare tire>.
- *Expressiveness for P and O:* Since predicates and objects are treated as monolithic strings, it creates redundancies and cannot capture semantic relations between assertions. For example, these two assertions are equivalent but both present in ConceptNet: <bus; CapableOf; carry passengers>, <bus; UsedFor; transportation>. Furthermore, useful facets concerning spatial and temporal information are often cluttered into unrelated strings, e.g., <bus; carries; visitors to the zoo on the weekend>.

Assertion quality is another issue of prior CSKBs, as some of them prioritize precision but have limited coverage (e.g., ConceptNet (Speer et al. 2017), TupleKB (Dalvi Mishra et al. 2017)), while others such as WebChild (Tandon et al. 2014a), Quasimodo (Romero et al. 2019) and TransOMCS (Zhang et al. 2020a) have better coverage but contain many noisy assertions, as they are based on automated knowledge extraction from web contents and lack appropriate consolidation.

The *saliency* of assertions (i.e., the degree to which statements are common knowledge) has been often overlooked by prior works. For example, in ConceptNet, tree, garden, and the bible are all listed as locations of snake, with similar scores. This makes it difficult for downstream applications to pull out relevant assertions from the KB.

Our goal is to overcome these limitations of prior works while retaining their positive characteristics. In particular, we aim to reconcile high precision with wide coverage and saliency. Like TupleKB (Dalvi Mishra et al. 2017) and Quasimodo (Romero et al. 2019), we desire to acquire open assertions (as opposed to pre-specified predicates only in ConceptNet (Speer et al. 2017)) but strive for more expressive representations by refining subjects and capturing semantic facets of assertions.

Besides, we also provide a canonicalized version of the resulting CSKB in the ConceptNet schema, arguably the most widely used CSK resource with canonical predicates such as `CapableOf`, `HasProperty`, `UsedFor`, `HasPart`, `AtLocation`, etc., thus enabling direct use in applications relying on this fixed schema (e.g., (Lin et al. 2019)).

Approach. We present a methodology, called Ascent++ (**a**dvanced **s**emantics for **c**ommonsens**e** k**n**owledge ex**t**raction), for acquiring CSK assertions about everyday concepts with refined semantics from large-scale web contents. Our method operates in two phases: (i) scalable extraction from a large web corpus, and (ii) aggregation and consolidation.

In the first phase, Ascent++ processes the C4 crawl (Raffel et al. 2020), a collection of 365 million English web pages. We extract OpenIE-style tuples by using carefully designed dependency-parse-based rules, taking into account assertions for subgroups and aspects of target subjects. The extractor uses cues from prepositional phrases and adverbs to detect semantic facets and uses supervised classification for eight facet types.

In the second phase, on a per-subject basis, Ascent++ identifies relevant web pages based on embedding similarity to reference Wikipedia articles, this way being able to distinguish homonyms like `bus (public transport)` versus `bus (network topology)`. Assertions are iteratively grouped and organized using embedding-based similarity. OpenIE-style assertions are canonicalized into the established ConceptNet schema. Finally, a supervised machine learning model ranks the resulting statements by saliency and typicality scores.

We ran Ascent++ on the C4 crawl for 10,000 salient concepts from ConceptNet as target subjects. To evaluate the intrinsic quality of the resulting CSKB, we obtained human judgments for a large sample. Our CSKB significantly improves over automatically-built state-of-the-art CSK collections in terms of precision and relative recall.

In addition, we performed an extrinsic evaluation in which commonsense knowledge was used to support language models in question answering tasks. Using three different settings and six different CSKBs, Ascent++ significantly outperformed language models without this commonsense background knowledge in two of the three settings, and was best or second best among all six CSKBs in all three cases.

Contributions. This project's key contributions are:

1. *Knowledge model* (Section 3.2): We introduce an expressive model for commonsense knowledge with advanced semantics, subgroups and aspects of subjects and faceted assertions as first-class citizens, and scores for typicality and saliency.
2. *Methodology* (Section 3.3): We propose an automated method for populating the model with high-quality CSK assertions by large-scale web content extraction and various techniques for aggregation and cleaning.
3. *Resource* (Section 3.4): We construct and publicly release a high-quality CSKB with 2 million assertions for 10,000 important concepts.

The evaluation with human judgments shows that the Ascent++ assertions are of significantly higher quality than those from prior works (Section 3.5). An extrinsic evaluation for QA-support tasks underlines the benefits of Ascent++ (Section 3.6). Code and data can be accessed at https://ascentpp.mpi-inf.mpg.de.

3.2 Knowledge Representation

In the traditional triple-based data model, subjects and objects are linked via predicates. Most prior CSKBs follow this model, typically with single nouns as subjects, free-form or pre-specified phrases as predicates, and words or phrases as objects. Typical examples from ConceptNet (Speer et al. 2017) are `<bus; AtLocation; road>` and `<bus; UsedFor; get to work>`. Few projects, such as WebChild (Tandon et al. 2014a) and TupleKB (Dalvi Mishra et al. 2017), have attempted to refine these assertions by employing word sense disambiguation (Navigli 2009), in order to distinguish terms such as buses on the road from computer buses. Similarly, a few other projects (Gordon and Schubert 2010, Zhang et al. 2017, Romero et al. 2019, Chalier et al. 2020) have tried to identify salient assertions against correct ones that are unspecific (e.g., buses used for getting to a place), atypical (e.g., buses used for showcasing local artists' work), or even misleading (e.g., buses used for getting time to read).

We extend this prevalent paradigm in three significant ways: adding refined subjects (Section 3.2.1), refining triples with semantic facets (Section 3.2.2), and quantitative scoring for typicality and saliency (Section 3.2.3). Finally, in Section 3.2.4, we formally define advanced commonsense assertions that we aim to acquire in this work.

3.2.1 Expressive Subjects

The acquisition of commonsense knowledge usually starts by collecting assertions for single-noun subjects. This approach has two main limitations: (1) it does not distinguish different meanings of the same word, and (2) it misses out on refined concepts and variants of word senses. Although word sense disambiguation (WSD) has been used to address the first issue in a few projects such as WebChild (Tandon et al. 2014a) and TupleKB (Dalvi Mishra et al. 2017), they were inherently limited because their underlying word-sense lexicons (WordNet and Wiktionary) mainly focus on single nouns. For example, phrases like `tourist bus` or `newborn elephant` are not present.

Our approach to rectify this problem is twofold:

1. When discovering source documents for a target subject, we compare the documents with its reference Wikipedia article, and we only retain documents with high similarity. This way, we can disentangle different senses, for example, of `bus` as in public transport and network topology themes.
2. During the knowledge extraction phase, we also consider multi-word phrases as candidates for refined subjects. This allows us to acquire `IsA`-like refinements, creating *subgroups* of broader subjects such as `school bus`, `city bus`, `circus elephant`, or `elephant cow`; and other kinds of relevant *aspects* such as `bus' route`, `bus capacity`, `elephant tusk`, or `elephant habitat`. In the following, we will elaborate on the notions of *subgroups* and *aspects*.

Subgroups. Our notion of *subgroups*, which can be thought of as an inverse `IsA` relation, goes beyond traditional taxonomies by better coverage of multi-word composites (e.g., `circus elephant`, `school bus`), enabling us to better represent specialized assertions such as `<circus elephants; catch; balls>` and `<school bus; transports; students>`.

Aspects. Our notion of *aspects* includes part-whole relations (such as `PartOf`, `MemberOf`, `SubstanceOf`) (Girju et al. 2006, Tandon et al. 2016, Shwartz and Waterson 2018, Bhakthavatsalam et al. 2020), as well as additional aspects that go beyond hypernymy and meronymy (e.g., `bus accident`, `elephant habitat`). Unlike single nouns, these compound phrases are rarely ambiguous, providing crisp concepts without requiring explicit WSD.

3.2.2 Semantic Facets

The validity of CSK assertions often depends on specific temporal and spatial contexts. For instance, elephants scare away lions only in Africa, or elephants bathe in rivers only during the daytime. In addition, assertions often become crisper when framed with causes, effects, or instruments, for example, children ride the bus … to go to school, or circus elephants catch balls … using their trunks.

We integrate such information into our expressive model by contextualizing SPO triples with *semantic facets* based on ideas from research on semantic role labeling (SRL) (Palmer et al. 2010, Clarke et al. 2012, Stanovsky et al. 2018). Initially, SRL was developed to annotate hand-crafted frames (e.g., purchase) with values for frame-specific roles (e.g., buyer, goods, price, etc.).

We start with 35 labels proposed in prior study (Prasojo et al. 2018), which combines 13 labels from the Illinois Curator SRL (Clarke et al. 2012), and 22 additional labels crafted by analyzing semantic roles of prepositions in Wiktionary. Because many of these labels are very special, we consolidate them into eight widely useful roles that are CSK-relevant.

- Labels that qualify the validity of assertions: `DEGREE`, `LOCATION`, `TEMPORAL`, `OTHER-QUALITY`.
- Labels that capture other dimensions: `CAUSE`, `MANNER`, `PURPOSE`, `TRANSITIVE-OBJECT`.

3.2.3 Quantitative Scoring

Prior works typically quantify the quality of assertions by a single numeric score. ConceptNet (Speer et al. 2017), for instance, scores its assertions essentially by the number of annotators that stated them (in most cases, 1). TupleKB (Dalvi Mishra et al. 2017) employs a supervised model that predicts a $[0, 1]$ score capturing statement plausibility. With Ascent++, we want to empower downstream users to remain flexible in how to rank and use the data. We thus propose a scoring mechanism along two dimensions:

Saliency. This score captures how spontaneous an assertion comes to the human mind. For example, elephants being used for tourist rides is quite salient, while elephants sleeping at night is less so. Saliency is important to understand which statements matter to humans (e.g., in conversational agents).

Typicality. This dimension captures the degree to which an assertion applies to individual instances of a concept, on a per-subject basis. For example, most elephants sleep in most nights, whereas only few elephants give tourists a ride. Typicality is important to understand which utterances make sense (e.g., in question answering).

Typicality and saliency are thus orthogonal dimensions, allowing to capture finer properties of commonsense assertions than just frequency or plausibility. Further dimensions could be considered (Chalier et al. 2020), though we found *plausibility* not to be a dimension of high discriminative utility (implausible statements should rather not even enter CSKBs).

3.2.4 Advanced Modeling for Commonsense Knowledge

Given those design considerations, we propose the following knowledge model for commonsense knowledge of everyday concepts.

Let C_0 be a set of primary concepts of interest, which could be manually defined or taken from a dictionary. Subjects for assertions include all $c_0 \in C_0$ and judiciously selected multi-word phrases containing some c_0. Subjects are interrelated by *subgroup* and *aspect* relations: each c_0 can be refined by a set of subgroup subjects denoted $sg(c_0)$ and by a set of aspect subjects denoted $asp(c_0)$. The overall set of subjects is $C \coloneqq C_0 \cup sg_{C_0} \cup asp_{C_0}$.

A commonsense assertion for subject $s \in C$ is a sextuple $\langle s, p, o, F, \pi, \theta \rangle$ with single-noun or noun-phrase subject s, short phrases for predicate p and object o, a set F of semantic facets, and two $[0, 1]$ scores: π for saliency and θ for typicality. Each facet $(k, v) \in F$ is a key-value pair with one of eight possible keys k and a short phrase as v. Note that a single assertion can have multiple key-value pairs with the same key (e.g., different spatial phrases).

For example, given the primary concept $c_0 =$ `elephant`, its subgroups include `adult elephant`, `newborn elephant`, `elephant cow`, etc.; its aspects include `elephant trunk`, `elephant diet`, `elephant ear`, etc. Some example assertions (without scores) are:

- `<elephant; sucks; water; PURPOSE:{drink; spray on its body}>`,
- `<newborn elephant; is; blind>`,
- `<elephant trunk; is; sensitive; DEGREE:extremely>`.

3.3 Methodology

We propose the Ascent++ methodology to populate the advanced knowledge model by extracting CSK assertions from large-scale web contents.

3.3.1 Architecture Overview

3.3.1.1 Design Considerations

Three major design points of CSK acquisition are: (1) the choice of knowledge sources, (2) the choice of the extraction techniques, and (3) the choice of cleaning or consolidating the extracted candidate assertions.

Sources. Popular sources in prior works on automated CSK acquisition include:

- *Carefully selected high-quality texts:* These sources include book n-grams (Tandon et al. 2014a), concept definitions in encyclopedic sources, and school text corpora about science (Clark et al. 2018). However, these sources cover only a limited scope of CSK, and surprisingly high noise and bias are also found in seemingly clean texts like book n-grams (Gordon and Durme 2013).

- *Relevant web pages retrieved by search engines:* Retrieving relevant web pages were used by Dalvi Mishra et al. (2017) and Nguyen et al. (2021a). However, the manual query formulations in (Dalvi Mishra et al. 2017) required non-negligible effort. On the other hand, the simple query templates that incorporate hypernyms in (Nguyen et al. 2021a) led to limited coverage for many subjects where named instances are generally more prominent than the general concepts (e.g., `laptop`, `university`).
- *Questions asked by web users:* Romero et al. (2019) tapped into query auto-completion from search engines and questions posted on question-answering forums. Although these sources gave access to highly salient assertions, they also present heavily biased and sensational contents (e.g., search-engine auto-completion for *"elephants eat"* suggesting *"… plastic"* and *"… poop"*).

In Ascent++, we opt for directly using a huge corpus as an extraction source for wide coverage, and devise techniques for quality assurance. ASER (Zhang et al. 2020b) and TransOMCS (Zhang et al. 2020a) also used large-scale web contents as sources, however, these approaches are recall-oriented and lack appropriate consolidation.

Extraction Techniques. Prior works have employed various extraction techniques such as co-occurrence- and pattern-based methods (e.g., (Elazar et al. 2019)), OpenIE (e.g., (Dalvi Mishra et al. 2017), (Romero et al. 2019), (Nguyen et al. 2021a)), and supervised learning for classification and sequence tagging. While co-occurrence is fairly effective for a limited set of well-defined predicates using distant seeds, supervised extractors are constrained by the need for training data specific to each predicate. Therefore, more recent approaches prefer OpenIE techniques, which our extractors also adopt.

Knowledge Consolidation. Early approaches retained all assertions from the ingest process, for example, crowdsourcing in ConceptNet (Speer and Havasi 2013). However, recent projects have employed supervised classifiers or rankers for cleaning (Dalvi Mishra et al. 2017, Zhang et al. 2017, Romero et al. 2019, Chalier et al. 2020), and also limited forms of clustering (Dalvi Mishra et al. 2017, Romero et al. 2019) to reduce semantic redundancy.

In Ascent++, we leverage language models (Reimers and Gurevych 2019) to cluster assertions of identical meanings and reinforce the frequency signals of those assertions. Furthermore, after clustering, we do a mapping from our open-schema CSKB to the well-established ConceptNet schema as it is favorable to many researchers (e.g., (Lin et al. 2019, Feng et al. 2020, Hwang et al. 2021)). Finally, a heuristic-based cleaning approach is applied to eliminate other remaining noise in the resulting KB.

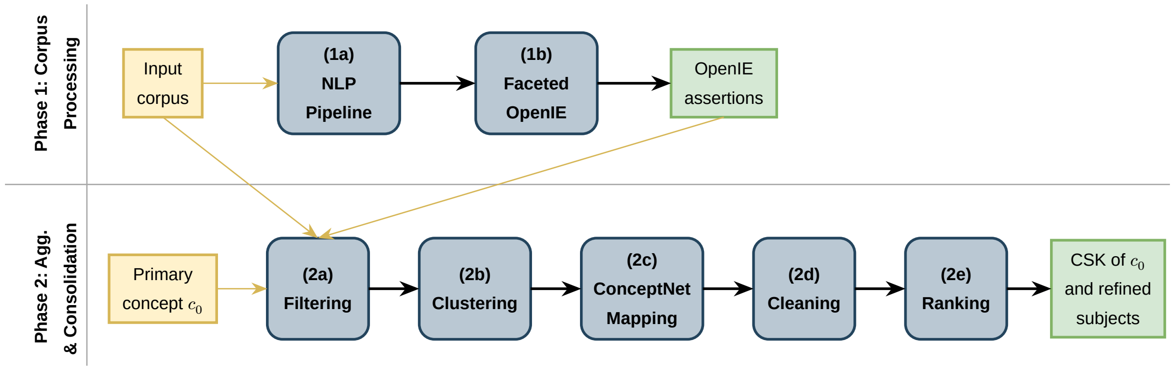


Figure 3.1: Architecture of the Ascent++ system.

3.3.1.2 Approach

The Ascent++ methodology operates in two phases (see Figure 3.1):

1. *Corpus processing:*
   - 1.a. *NLP pipeline*: Running NLP pipeline for the input corpus to get NLP features of all sentences, particularly part-of-speech tags and dependency trees.
   - 1.b. *Faceted OpenIE*: Running a faceted OpenIE system to get assertions from the processed sentences.
2. *Aggregation and consolidation:* Each subject is processed separately in this phase.
   - 2.a. *Filtering*: Performing a series of document and assertion filtering to get relevant and high-quality assertions for a given primary concept.
   - 2.b. *Clustering* of retained assertions based on sentence embeddings.
   - 2.c. *Mapping* from open assertions into ConceptNet schema.
   - 2.d. *Cleaning* based on heuristics and a dictionary of unwanted patterns in assertions.
   - 2.e. *Ranking* of assertions: Annotating assertions with complementary scores for typicality and saliency.

Both phases treat each document or subject independently and thus can be highly parallelized (see Section 3.4.3). Since commonsense knowledge evolves rather slowly compared to encyclopedic knowledge where new entities and relations emerge on a daily basis, computing a large CSKB is a one-time endeavor with long-term value. Nonetheless, whenever new or updated inputs need to be processed, Phase 1 can be run incrementally on the new inputs only. Only steps 2b and 2e require re-loading previous statements (on a per-subject basis). Re-running these steps takes about half a day for a corpus like the C4 crawl (Raffel et al. 2020). Table 3.2 gives detailed run-times (see Section 3.4.3).

In the following, we discuss steps 1a-b and 2a-2e in separate subsections.

3.3.2 Phase 1a: NLP Pipeline

The NLP pipeline consists of fundamental operations, including sentence splitting, tokenization, lemmatization, part-of-speech tagging, dependency parsing, and named entity recognition. Our extractors will use all of these basic NLP features to output faceted OpenIE tuples.

3.3.3 Phase 1b: Faceted OpenIE

Our method leverages an open information extraction (OpenIE) system developed earlier in the Ascent project (Nguyen et al. 2021a), which we refer to as AscentOpenIE.

AscentOpenIE was built upon StuffIE (Prasojo et al. 2018), a rule-based OpenIE extractor used for extracting triples and semantic facets from English sentences. The core idea of the approach is to consider each verb as a candidate predicate and then identify subjects, objects, and facets via grammatical relations, so-called dependency paths. The elaboration below uses the Universal Dependencies style format (Marneffe et al. 2021).

- *Subjects* must be connected to the candidate predicate through subject-related dependency edges (*nsubj*, *nsubjpass*, and *csubj*) or the adjectival clause edge (*acl*).
- For *objects*, the respective edges include direct object (*dobj*), indirect object (*iobj*), and nominal modifier (*nmod*).
- *Semantic facets* are identified through the following complements to the selected verb: adverbial modifiers (*advmod*), prepositional and clausal complements (*ccomp*). The facets are then labeled by a fine-tuned language model.

AscentOpenIE also extended the original set of rules in StuffIE to better deal with conjuncts and adverb facets, as well as leveraged coreference resolution to resolve pronouns. That helped to identify significantly more assertions and facets, and improve the conciseness of the output tuples. For example, given the sentence *"elephants use their trunks to pick up objects and drink water"*, the system can extract two assertions: `<elephants; use; their trunks; PURPOSE:pick up objects>` and `<elephants; use; their trunks; PURPOSE:drink water>`.

We run AscentOpenIE on all sentences in the input corpus, producing general OpenIE tuples, which will be processed in Phase 2, where we collect and consolidate CSK assertions for a target subject.

3.3.4 Phase 2a: Filtering

Using a general web corpus as an extraction source implies the presence of substantially irrelevant content. In this step, we introduce techniques to filter out irrelevant documents (a document is a web page from the input corpus in our case), and potentially noisy OpenIE assertions.

Document Filtering. Given a subject s, one might first collect all OpenIE tuples whose subject is equal s as candidate assertions, and then apply ranking or filtering techniques afterward. Similar approaches have been used in Quasimodo (Romero et al. 2019) and TupleKB (Dalvi Mishra et al. 2017). Nevertheless, such post-hoc filtering misses out on broader context from the original documents. In Ascent++, we thus employ some filters first, at the document level, to decide which documents to use for candidate extraction at all.

We only extract statements for a primary subject c_0 and its subgroups and aspects from a document d if it passes the following filters:

1. Document d is only used if it contains between 3 and 40 OpenIE tuples with subject c_0. The rationale for filters in either direction is that if c_0 occurs too rarely, d is more likely off-topic. On the other hand, if c_0 occurs too often, then d may be a noisy document such as a machine-created shopping catalog or simply a crawling error.
2. Then, we compute the cosine similarity between the embeddings of d and the Wikipedia article of the subject c_0. Document d will be retained only if the similarity is higher than 0.6 (chosen based on tuning on withheld data). This way, we can deal with ambiguous subject terms like `bus`, which can be either a vehicle or a network topology.

Assertion Filtering. After document filtering, we collect all OpenIE tuples whose subjects are either c_0 or its subgroups or aspects from the retained documents. The refined subjects, which include multi-word composites and relevant aspects of c_0, are identified using a set of lexico-syntactic heuristics adopted from Ascent (see Section 3.4.2). Finally, by counting the extracted triples, we only retain those with a frequency of at least 3, as assertions with less occurrences are either noise or unlikely commonsense knowledge.

3.3.5 Phase 2b: Clustering

Natural language is rich in paraphrases. For example, *"elephant eats plants"* can also be expressed as *"elephant feeds on plants"* or *"elephant consumes plants"*. Identifying and clustering such assertions is necessary to avoid redundancies, and to get better frequency signals for individual assertions.

Triple Clustering. For triple clustering, we use the hierarchical agglomerative clustering (HAC) algorithm along with LM-based embeddings to group semantically similar triples. Specifically, we use SentenceBert (Reimers and Gurevych 2019) to compute embeddings of CSK triples. First, given a triple, we concatenate its subject, predicate, and object. Next, we feed the whole string to SentenceBert to get its contextualized embeddings. Then, for each pair of triples, we compute the Euclidean distance between their normalized embeddings. These distances will be used as input for the HAC algorithm.

Facet Clustering. For facet clustering, we use average word2vec embeddings (Mikolov et al. 2013) and the HAC algorithm. Although more advanced language models such as Phrase-BERT (Wang et al. 2021) could be used instead of word2vec, we found that, for such a limited number of candidate facets (usually less than 10 facets per assertion), word2vec embeddings already provide good performance.

The set of chosen hyper-parameters for the clustering algorithms will be presented in Section 3.4.4.

3.3.6 Phase 2c: ConceptNet Mapping

Motivation. There are two main schools for knowledge representation in CSKBs: those relying on open predicates, and those using a fixed set of predefined predicates. Each has its strengths and challenges regarding expressiveness, redundancy, and usability. To bridge the two, we provide the Ascent++ KB in two variants: with open assertions and with canonicalized predicates. OpenIE supplies the former; the module presented in this subsection normalizes open assertions into fixed relations.

Our fixed schema of choice is the established ConceptNet schema (Speer et al. 2017), from which we use the following 19 relations: `AtLocation`, `CapableOf`, `Causes`, `CreatedBy`, `DefinedAs`, `Desires`, `HasA`, `HasPrerequisite`, `HasProperty`, `HasSubevent`, `IsA`, `MadeOf`, `MotivatedByGoal`, `PartOf`, `ReceivesAction`, `RelatedTo`, `SimilarTo`, `SymbolOf` and `UsedFor`.

Mapping open triples to a fixed schema raises several challenges. In the most straightforward case, the subject and object from the open assertion can remain unchanged, and we only need to pick one of the fixed relations. For example, `<elephant; lives in; the wild>` can be mapped to `<elephant; AtLocation; the wild>`. In some cases, part of the relation and object can be moved, e.g., `<elephant; is; a part of a herd>` can be mapped to `<elephant; PartOf; herd>`. In other cases, part or all of the predicate is in the object, like in `<circus elephant; catches; balls>` that can be mapped to `<circus elephant; CapableOf; catch balls>`.

Our normalization method consists of two steps: (i) first, we use a multi-class classifier to predict a fixed predicate for each open triple; (ii) second, we use a list of crafted rules to modify the object so that the new triple preserves the meaning of the original triple.

Supervised Classifier and Rule-Based Disambiguation. For the classification model, we fine-tune a RoBERTa model (Liu et al. 2019) to predict one of the ConceptNet predicates given an input in the following format: "[CLS] S [SEP] P O", whereas S, P, and O are the subject, predicate, and object of the open triple. The contextualized embeddings of the [CLS] token will be fed to a fully connected layer to get the prediction. The training data for this model is constructed from ConceptNet triples. Specifically, for each ConceptNet predicate, we manually compiled a set of one to six open predicates with similar meanings. For instance, `UsedFor` is aligned with "be used for", meanwhile `CapableOf` is aligned with "be capable of", "be able to", "can", "could", and an empty string. This way, we can automatically generate approximately 1.2M training examples.

Due to the nature of ConceptNet, the generated data is highly biased towards a few top predicates. The three most popular predicates are `IsA` (27.92%), `AtLocation` (20.24%) and `CapableOf` (13.76%). Meanwhile, the three least popular ones are `MadeOf` (0.20%), `CreatedBy` (0.06%) and `SymbolOf` (12 samples, less than 0.001%). This imbalance affects the predictions. The most important difficult case is distinguishing the three predicates `IsA`, `HasProperty`, and `ReceivesAction`, which can all be expressed with the open predicate word "be". In the ConceptNet-based training data, the `IsA` relation (27.92% of the data) is dominant over the other two, `HasProperty` (3.07%) and `ReceivesAction` (2.20%). Therefore, the LM occasionally classifies open triples whose predicate is "be" incorrectly as `IsA` relations. For this particular case, we have a post-processing step to adjust the predicted predicate: we only assign `HasProperty` to objects which are adjective phrases, `ReceivesAction` to objects which are verb phrases in passive form, and the rest are assigned to `IsA`. The `IsA` relation still contains considerable noise. Hence, later in the cleaning phase (see Section 3.3.7), we introduce heuristics to get high-quality assertions of this type.

An alternative could be to re-balance the training data by under-sampling the frequent classes or adopting a loss function that gives different weights to different classes. While such generic techniques could be considered, our experience is mirrored in related projects on creating and curating high-quality KBs, where injecting a modest amount of expert knowledge is often the most effective solution (Bhakthavatsalam et al. 2020, Pellissier Tanon et al. 2020). More advanced methods for relation alignment also exist (Soderland et al. 2013, Galárraga et al. 2014, Putri et al. 2019, Zhang et al. 2019). However, given that ConceptNet itself is imbalanced and has its peculiarities, our customized mapping is far superior to generic alignment methods.

Processing Objects. Once we get a predicted fixed-schema predicate, we have to produce an object for the normalized triple.

The most common case where the object needs modifications is when the predicted predicate is CapableOf. In this case, if the open predicate is neither “be capable of”, “be able to”, “can” nor “could”, the new object will be the concatenation of the original predicate and object. For example, <elephant; can; lift a tree> is mapped into <elephant; CapableOf; lift a tree>, meanwhile <elephant; carry; tree trunk> is mapped to <elephant; CapableOf; carry tree trunk>.

In some cases, a part of the original object must be cut out as it overlaps with the ConceptNet predicate. Those predicates include PartOf and SymbolOf. Open triples corresponding to those predicates usually look like <elephant; be; part of a herd> or <elephant; be; symbol of strength> which should be canonicalized into <elephant; PartOf; herd> and <elephant; SymbolOf; strength>, respectively.

Other rules deal with other predicates, including Desires, HasProperty, IsA, ReceivesAction, and UsedFor. These rules remove redundant words such as “to” and “be” from objects. We preserve the original object if a triple does not fall into one of those rules.

3.3.7 Phase 2d: Cleaning

Noise can come from various sources, including OpenIE errors, too specific or general statements, nonsensical assertions, schema normalization errors, etc. Unlike supervised classifiers employed in other projects, our cleaning module is rule-based and thus highly scrutable. Ascent++’s cleaning module consists of the following heuristics:

1. First, we verbalize the triples, and use an autoregressive model, namely GPT-2 (Radford et al. 2019), to compute their perplexity. Only triples with medium-to-low perplexity will be retained (in our experiments, we retained triples with perplexity less than 500).
2. The IsA triples produced by OpenIE are rich but rather noisy. Extracting IsA relations is a well-established research theme in entity typing and taxonomy construction and has already reached high-quality results. For each subject, we take the set of its IsA relations in ConceptNet and extract all head nouns of the objects. For example, from the ConceptNet triple <elephant; IsA; placental mammal>, we take out the object phrase “placental mammal” and extract its head noun “mammal”. We only retain the IsA assertions whose objects contain one of the trustworthy head nouns extracted from ConceptNet. This way, we not only get high-precision assertions but also better recall than ConceptNet. If a subject does not occur in ConceptNet, we remove all of its extracted IsA assertions, as in our observation, there are usually more noisy IsA assertions than

valuable ones in the original extraction set. Note that, while ConceptNet is our source of choice here, any other existing resources that provide high-quality `IsA` relations can be used, for example, WordNet (Miller 1995) or BabelNet (Navigli and Ponzetto 2012).

3. Then, we manually constructed a dictionary of unwanted objects based on heuristics. For example, we removed URLs, pronouns, numbers, only-stopwords phrases, too general/specific phrases such as "make sure" or "this case", and vague predicate-object pairs (e.g., `<SubjectX; MadeOf; part>`, or `<SubjectX; HasProperty; available>`). We also eliminated ethnicity- and religion-related assertions to avoid potentially critical biases (Mehrabi et al. 2021). Chapter 4 and Chapter 5 will focus on collecting such culture-aware knowledge.
4. Finally, we only keep the 1,000 most frequent assertions per subject. For each assertion, we only keep its three most frequent facets. Cutting the tail this way improves the precision significantly but only minorly affects recall (see Section 3.5.3).

These filters are a pragmatic technique, and alternatives are conceivable. Some parts require manual work by a knowledge engineer. This holds particularly for the domain-specific dictionary filter. However, the dictionary is small (200 entries), and a knowledge engineer can easily construct it within a day (at much lower energy consumption and carbon footprint than trying to automate everything computationally).

3.3.8 Phase 2e: Ranking

Existing resources mainly provide unidimensional rankings of their assertions by either *frequency* (e.g., ConceptNet (Speer et al. 2017), Ascent (Nguyen et al. 2021a)) or supervised models trained to predict *plausibility/typicality* (e.g., TupleKB (Dalvi Mishra et al. 2017), Quasimodo (Romero et al. 2019), TransOMCS (Zhang et al. 2020a)). However, these two dimensions are quite different, and we consider it important to differentiate their semantics.

- *Typicality* states that an assertion holds for most instances of a concept. For example, elephants using their trunks is typical, whereas elephants drinking milk holds only for baby elephants.
- *Saliency* refers to the human perspective of whether an assertion is associated with a concept by most humans more or less on first thought. For example, elephants having trunks is salient, whereas elephants passing by zebras is not.

In Ascent++, we make both dimensions first-class citizens of the CSKB and annotate each assertion with scores for both.

Saliency Ranking. For saliency, we rely on the reporting frequency of the assertions, which approximates very well how prominent an assertion is. We transform raw frequencies to a normalized log-scaled frequency as follows:

$$saliency(spo) = \frac{\log(count(spo)) - \log(min\_count(s))}{\log(max\_count(s)) - \log(min\_count(s))} \tag{3.1}$$

Whereas $count(spo)$ is the raw frequency of the triple spo, $min\_count(s)$ is the minimal frequency of a triple whose subject is s, and $max\_count(s)$ is the maximal one.

Typicality Ranking. Typicality is the most challenging dimension. Although reporting frequency correlates with typicality moderately, reporting bias in texts (Gordon and Durme 2013) means that sensational statements may be grossly overreported and commonalities underrepresented. Our qualitative facets (see Section 3.2) give us a unique handle to obtain further insights into typicality.

We use a linear regression model on three features:

- *Modifier score.* This feature is based on adverbs and quantifiers in facets and in subjects. Specifically, we assign each frequency-related modifier a specific numeric score (see Table 3.1), then average all scores in each assertion cluster. We consider two types of modifiers: adverbs (e.g., *"always"*, *"often"*, *"sometimes"*) that occur in semantic facets, and subject quantifiers (e.g., *"all"*, *"few"* or *"some"*). We assign a default score of 0.5 to assertions without any modifiers.
- *Neutrality.* We use a sentiment analysis model (Barbieri et al. 2020) to compute the probability of a source sentence being positive, neutral, or negative. For each assertion, we consider the average polarity over all of its source sentences as its polarity. The value of this neutrality feature for an assertion is 1.0 if it is classified as neutral. Otherwise, a value of zero reflects a polarized assertion.
- *Normalized frequency.* The value for this feature is computed as in Equation (3.1).

3.4 Implementation

In this section, we present the implementation of Ascent++, which includes the choice of input corpus, input subjects, and hyperparameters, the processing time of each module, and the size of the resulting CSKB.

3.4.1 Input Corpus

Choosing the input text corpus is essential because large text corpora, especially those scraped from the web, often come with a large portion of noise and irrelevant contents. Our

Table 3.1: Frequency modifiers and their scores.

Frequency adverbs	**Subject quantifiers**	**Score**
always	*all, every*	1.0
typically, mostly, mainly	*most*	0.9
usually, normally, regularly, frequently, commonly		0.8
	many	0.7
often		0.6
	some	0.5
sometimes		0.4
occasionally	*few*	0.3
		0.2
hardy, rarely		0.1
	no, none	0.0

input corpus of choice is the Colossal Clean Crawled Corpus (C4), created to train the T5 model (Raffel et al. 2020).

C4 was created by intensively cleaning the Common Crawl's web crawl corpus. The filtering process included deduplication, English-language text detection, removing pages containing source code, offensive language, or pages with too few contents and lines, removing lines that did not end with a terminal punctuation mark, etc. That resulted in a large corpus of 750GB of text comprising reasonably clean and natural English text.

The version we use, C4.EN, consists of 365M English articles. Each comes with its text, URL, and crawling timestamp. This amount of text enabled us to collect CSK of significantly better coverage than the manually-built ConceptNet and any other automated CSK resource when limiting to top-100 assertions per subject (see Section 3.5.2). Our judicious rule-based approaches for extraction, filtering, cleaning, and unsupervised ranking helped to produce CSK assertions of higher saliency than any other automated CSK resource while maintaining a very high precision (see Section 3.5.1).

3.4.2 Input Subjects

We executed the ASCENT++ pipeline for 10K popular concept-centric subjects taken from ConceptNet, which we treated as primary concepts. For each primary concept, we took its top-10 most frequent aspects and subgroups previously extracted by ASCENT (Nguyen et al. 2021a) as high-quality fine-grained subjects. These subjects were collected from top-ranked web pages retrieved by the Bing Search API by using the following heuristics.

Table 3.2: Processing time and output size of each step in ASCENT++.

#	Step	Processing time	Output
1a	NLP pipeline	1.5 days	365M processed documents
1b	Faceted OpenIE	20 hours	8B OpenIE tuples
2a	Filtering	2 hours	165M tuples (15M unique triples)
2b	Clustering	10 hours	7.5M clusters
2c	ConceptNet mapping	30 minutes	7.5M canonicalized assertions
2d	Cleaning	10 minutes	2M assertions
2e	Ranking	2 hours	2M ranked assertions
-	*Total*	*~ 3 days*	-

Heuristics for Extracting Refined Subjects. Given a primary concept c_0, the following heuristics were used to extract its subgroups and aspects:

- *Subgroups* could be sub-species (in the case of animals), different types of the primary concept, or refer to the primary concept in different states, e.g., `African elephant`, `newborn elephant`, `male elephant`, `working elephant`. All noun chunks ending with c_0 or any of its WordNet lemmas were collected as potential candidates. Then, these terms were clustered based on word2vec embeddings, and only the most frequent term of each cluster was selected. In addition, WordNet was leveraged to distinguish antonyms, with which the latent representations typically struggle, in order to avoid antonyms being merged into the same cluster.
- *Aspects* were extracted from noun chunks collected from two sources: (i) *possessive noun chunks*, for example, *"elephant's diet"* and *"their diet"* (with a resolution to c_0 = *"elephant"*); and (ii) objects of OpenIE tuples whose subjects are equal c_0 or any of its WordNet lemmas, and p is one of the following verb phrases: *"have"*, *"contain"*, *"be assembled of"*, *"be composed of"*.

Finally, the extracted subgroups and aspects were normalized and cleaned in order to avoid spurious extractions or overly specific terms.

3.4.3 Processing Time and Output Size

In Table 3.2, we provide details on each step's processing time and outputs in the pipeline.

The corpus processing phase (Phase 1) is executed once, independent of any choice of subjects. We used spaCy (https://spacy.io/), a popular Python-based NLP library, as our NLP pipeline. First, we ran spaCy on 365M C4 documents in parallel, on a cluster of 6,400 CPU cores (AMD EPYC 7702 64-core processors), which took 1.5 days to complete. Then,

Table 3.3: Size statistics of the Ascent++ KB.

Subject type	**#Subjects**	**#Assertions**	**#Facets**
Primary subject	8,067	1,651,455	1,975,385
Subgroup subject	10,191	80,176	62,581
Aspect subject	5,843	323,257	312,004
All	*24,101*	*2,054,888*	*2,349,970*

it took 20 hours for AscentOpenIE to digest all 356M processed documents. The result of this step was a set of 8B OpenIE assertions.

The filtering process took two hours and resulted in 165M tuples for the selected subjects, which accounted for 15M unique triples.

The clustering process, including precomputing embeddings and running the HAC algorithm, took ten hours. Embeddings were computed on a cluster of 150 NVIDIA Quadro RTX 8000 GPUs. The HAC algorithm led to 7.5M clustered assertions. The average size of a cluster is two assertions.

The ConceptNet mapping module took around 30 minutes. The rule-based cleaning step was the lightest component, taking less than ten minutes on a personal laptop. It resulted in 2M assertions in the final Ascent++ CSKB, whose size is reported in Table 3.3. There are about 10K subgroups and 5.8K aspects for 8K primary subjects in the KB. We collected 1.6M assertions for primary subjects, 80K assertions for subgroups, and 323K assertions for aspects. The total number of semantic facets is 2.3M. Hence, each assertion has more than one facet on average.

Finally, the ranking module took less than two hours to complete.

In summary, running the Ascent++ pipeline on the C4 corpus took approximately three days with our computing resources.

3.4.4 Hyperparameters

For triple clustering, we use the SentenceBert *paraphrase-mpnet-base-v2* model to compute embeddings of the verbalized triples. The embeddings are normalized before being fed to the HAC algorithm, for which we used the Ward's linkage (Ward 1963). We only merged two clusters when their Euclidean distance was less than 0.5. For facet clustering, this distance threshold was set at 1.0. These thresholds were chosen based on manual tuning on a small set of validation data points.

For ConceptNet mapping, we fine-tuned the RoBERTa-base model (Liu et al. 2019) for three epochs using the AdamW optimizer (Loshchilov and Hutter 2019) with a batch size of 64 samples, a learning rate of $5e^{-5}$, and a weight decay of 0.001.

For typicality scoring, we trained a regression model on 500 manually-annotated examples. The resulting formula is:

$$\text{typicality} = 0.324 \times \text{ms} + 0.428 \times \text{fr} + 0.088 \times \text{nt} \tag{3.2}$$

whereas ms is modifier score, fr is frequency, and nt is neutrality score (cf. Section 3.3.8).

Experiment Overview

The evaluation of Ascent++ is centered on three research questions:

- *RQ1:* Is the CSKB of higher quality than existing resources? (Section 3.5)
- *RQ2:* Does (structured) CSK help in extrinsic use cases? (Section 3.6)
- *RQ3:* What are the quality and extrinsic value of facets? (Section 3.7)

3.5 Intrinsic Evaluation

To investigate *RQ1*, we instantiate *quality* with the standard notions of precision and recall, splitting precision further up into the dimensions of *typicality* and *saliency*, measuring this way the degree of truth and the degree of relevance of assertions (see Section 3.2.3 and also (Romero et al. 2019)).

Evaluation Metrics. We employ three evaluation metrics:

1. *typicality* (Section 3.5.1),
2. *saliency* (Section 3.5.1),
3. *relative recall* (Section 3.5.2).

Knowledge base construction is typically evaluated by precision and recall. Following earlier work (Romero et al. 2019), we split *precision* into two dimensions: *typicality* (the degree of truth) and *saliency* (how readily a statement is available to a human). These two dimensions are generally independent, as salient statements need not be typical, and vice versa. Furthermore, as the absolute recall is difficult to establish (e.g., there is no way to obtain a complete set of all assertions for what elephants are capable of doing), we use a *relative recall* metric, measuring the fraction of statements from a human-built resource that are captured in the respective CSKB.

Compared Resources. We compare Ascent++ with six prominent resources:

1. ConceptNet (Speer and Havasi 2013),
2. TransOMCS (Zhang et al. 2020a),
3. TupleKB (Dalvi Mishra et al. 2017),
4. Quasimodo (Romero et al. 2019),
5. Ascent (Nguyen et al. 2021a) (the earlier version of Ascent++, see below),
6. COMET-ATOMIC$^{20}_{20}$ (Hwang et al. 2021) (with a caveat, see below).

Ascent (Nguyen et al. 2021a) was the predecessor of Ascent++. These two share the same OpenIE system (cf. Section 3.3.3). However, the key difference is that Ascent extracted CSK from a small set of top-ranked results retrieved by a search engine, while Ascent++ directly tapped into a huge web crawl. Besides the comparison of seven resources, we will present a head-to-head comparison between Ascent++ and Ascent in Section 3.5.3, where Ascent++ shows superior quality compared to its predecessor.

The ATOMIC (Sap et al. 2019a), ATOMIC-10x (West et al. 2022), and ASER (Zhang et al. 2020b) projects do not qualify for comparison as they do not contain concept-centric assertions.

ATOMIC$^{20}_{20}$ (Hwang et al. 2021) has a portion for physical commonsense relations, but most of those assertions come directly from ConceptNet (except for the `ObjectUse` relation for which more human-annotated data was collected). Therefore, we do not include ATOMIC$^{20}_{20}$ directly in this evaluation. Instead, we compare other CSKBs with a generative model trained on that resource, the COMET-ATOMIC$^{20}_{20}$ model (Hwang et al. 2021).

WebChild (Tandon et al. 2017) targeted special commonsense relations including comparative, part-whole relations, and fine-grained object properties (`HasShape`, `HasSize`, `HasColor` relations), which is generally of narrower scope compared to our competing resources. Comparisons of CSK resources in other studies (Romero et al. 2019, Nguyen et al. 2021a) reported that WebChild was inferior to Ascent and Quasimodo in assertion quality and in the extrinsic use case of question answering. We did not include this resource in our evaluation to limit our cost for crowdsourced assessments.

3.5.1 Precision: Typicality and Saliency

Evaluation Scheme. Unlike the precision of encyclopedic knowledge ("The Lion King" was either produced by Disney or not), the precision of CSK is generally not a binary concept, calling for more refined evaluation metrics. Hence, we assessed the *typicality* and *saliency* of triples, following the Quasimodo project (Romero et al. 2019).

Table 3.4: Crowdsourcing questions for assertion quality evaluation.

Dimension	**Question and answer options**
Typicality	*Is this a correct assertion about <subject>?*
	1. *Always/Often* - the knowledge assertion presented is always or often true 2. *Sometimes/Likely* - it is sometimes or likely true 3. *Farfetched/Never* - it is false or farfetched at best 4. *Invalid* - it is invalid or makes no sense
Saliency	*Imagine you have 2 minutes time to explain a kid about <subject>, would you mention the following information?*
	1. *Absolutely* - the information is very interesting/important for that concept 2. *Probably* - it is quite good to know 3. *Maybe not* - it is not interesting or too obvious/uninteresting/boring 4. *Definitely not* - it is completely irrelevant/unimportant/wrong or makes no sense

Given a CSK triple, we asked annotators on Amazon MTurk to rate the triple along a 4-point Likert scale on each of the two dimensions. We present the crowdsourcing questions and answer options in Table 3.4. The crowdsourcing templates are inspired by those introduced in COMET-ATOMIC$^{20}_{20}$ (Hwang et al. 2021). We performed two separate MTurk tasks on the two dimensions. For each task, we randomly sampled 500 triples among 200 prominent subjects in four common domains: animal, occupation, engineering, and geography. For the saliency task, the sampling pool was restricted to the top-10 assertions per subject.

Each MTurk task contained five CSK triples and was assigned to three different workers. Following Hwang et al. (2021), we also used human-readable language forms for triples in the fixed-schema CSKBs (i.e., ConceptNet, TransOMCS, and COMET). For Ascent++, we used the open triples as the prompt display. Triples that received the first two labels for a dimension (see Table 3.4) are marked as *typical/salient*. The final judgment for a triple is based on a majority vote over the choices provided by its three annotators. Annotation quality was ensured by requiring the MTurk annotators to be *Master workers* with an *all-time approval rate* of over 90% (this rate is provided by the platform). The inter-rater agreement on the three labels measured by Fleiss' κ (Fleiss and Cohen 1973) is 0.33 (i.e., fair agreement).

Table 3.5: Intrinsic evaluation results of ASCENT++ versus prior CSK resources.

CSK resource	Precision (%)		Relative recall (%)						Size
	Saliency@10	Typicality	Top-100 assertions/subject			All assertions			#spo
			$\tau = 0.96$	$\tau = 0.98$	$\tau = 1.0$	$\tau = 0.96$	$\tau = 0.98$	$\tau = 1.0$	
Crowdsourcing									
ConceptNet	79.2	96.0	5.29	3.39	1.10	5.47	3.53	1.13	0.5M
Generative									
COMET-ATOMIC$^{20}_{20}$	55.2	78.9	-	-	-	-	-	-	-
Extractive									
TransOMCS	40.4	51.4	4.04	3.24	1.85	19.70	16.47	9.06	18.5M
TupleKB	36.0	**92.0**	3.57	1.99	0.41	4.32	2.49	0.58	0.3M
Quasimodo	38.8	67.8	11.05	9.47	5.17	21.87	**19.36**	**10.88**	6.3M
ASCENT	60.0	79.2	9.60	7.02	3.17	17.09	12.62	5.82	8.6M
ASCENT++	**68.8**	88.4	**12.70**	**10.70**	**5.90**	17.54	14.64	7.95	2.0M
ASCENT++^{large}	-	82.2	-	-	-	**22.13**	18.02	9.37	7.5M

Note: ASCENT++^{large} *is constructed by dropping* ***Phase 2d: Cleaning***.

Results. We report the precision evaluation results in the left part of Table 3.5.

- *Comparison with extractive methods:* Among CSKBs constructed using extractive methods (TransOMCS (Zhang et al. 2020a), TupleKB (Dalvi Mishra et al. 2017), Quasimodo (Romero et al. 2019), Ascent (Nguyen et al. 2021a), and Ascent++), Ascent++ yields the most salient statements by a large margin (9 percentage points over its predecessor, and more than 13 percentage points over all others).

  For typicality, the new Ascent++ resource outperforms all but the domain-specific TupleKB (10 percentage points over the others). TupleKB still wins by 4 percentage points, yet produces unsalient statements (-32 percentage points) and for the science domain only, based on high-quality textbooks, with no obvious way to scale beyond (see also recall evaluation, where Ascent++ has 4x more recall than TupleKB). For example, while the top assertions for `elephant` in Ascent++ include `<elephant; IsA; social animal>` and `<elephant; HasProperty; intelligent>`, those in TupleKB include `<elephant; HasPart; skin cell>` and `<elephant; HasPart; cell membrane>`.

- *Comparison with a generative method:* We compared Ascent++ KB with a generative CSKB construction method, COMET (Bosselut et al. 2019), specifically the latest COMET-ATOMIC$^{20}_{20}$ model (Hwang et al. 2021). COMET is an autoregressive language model fine-tuned on existing CSK resources and can be used to predict possible objects of given subject-predicate pairs. Since COMET does not come with a materialized resource, we had to generate assertions ourselves. As there is no obvious stop criterion, we only evaluated the precision of top assertions but could not evaluate COMET's recall. We used the provided BART model, which was trained on the ATOMIC$^{20}_{20}$ dataset (Hwang et al. 2021), which includes ConceptNet assertions and crowdsourced assertions about events and social interactions. For each pair of subject and predicate, we asked COMET to predict top-5 objects. We used the same sampling processes and MTurk templates described above for typicality and saliency evaluation.

  The evaluation results of COMET are included in Table 3.5. Ascent++ clearly performs better than this generative model, even though both have seen comparable amounts of texts. Some examples of the top assertions in Ascent++ and the ones generated by COMET-ATOMIC$^{20}_{20}$ are shown in Table 3.6. Although COMET has the flexibility of generating objects to any given subject-predicate pair, it makes many incorrect predictions and produces notable redundancies. On the other hand, Ascent++, which collected OpenIE assertions from several sources and aggregated them through various steps such as filtering, clustering, and cleaning, produced more correct and complementary assertions.

Table 3.6: Top-5 assertions of selected subject-predicate pairs.

Subject - Predicate	ASCENT++	COMET-ATOMIC$^{20}_{20}$
elephant - CapableOf	● perform trick ● eat grass ● eat fruit ● become agitated ● give ride	● climb tree ● walk on land ● climb tree trunk ● walk on tree ● eat elephant
beer - MadeOf/MadeUpOf	● hop ● water ● barley ● yeast ● grain	● beer ● alcohol ● drunk ● drinking ● drink
laptop - UsedFor/ObjectUse	● work ● gaming ● office work ● email ● social media	● browse the internet ● use as a coaster ● play games on ● use as a weapon ● browse the web

3.5.2 Relative Recall

Ground Truth. Evaluating recall requires a notion of ground truth. We use a *relative recall* notion w.r.t. the statements contained in the CSLB property norm dataset (Devereux et al. 2014), which consists of short human-written sentences about salient properties of general concepts. There are 22.6K sentences expressing properties of 638 concepts in the dataset. The CSLB dataset could also be considered a CSK resource. However, due to its limited size, we did not include it in the comparisons with other CSKBs. Instead, we used it as ground truth for evaluating relative recall and the mask prediction task (see Section 3.6).

Assertion Matching. Since there are always different expressions of a CSK assertion in natural language, we allow soft matching between assertions in CSK resources and sentences in the ground-truth CSLB dataset. Specifically, for each CSLB sentence, we find the most similar assertion in the target CSK resource based on the cosine similarity between their embeddings (computed by SentenceBert (Reimers and Gurevych 2019)). That assertion will be considered a true positive w.r.t the given CSLB fact only if their cosine similarity is greater than or equal to a predefined threshold τ. When $\tau = 1.0$, only exact-match assertions are considered true positives. When lowering τ, we can match, e.g., *"elephant eats grass"* and *"elephant feeds on grasses"*.

Results. The relative recall evaluation results are shown in the right part of Table 3.5, where we once show relative recall using the top-100 assertions per subject and once all. We report results at three cosine similarity thresholds: $\tau = 0.96$, $\tau = 0.98$ and $\tau = 1.0$.

We find that Ascent++ yields the highest relative recall among all automated resources of the same size (columns under "Top-100 assertions/subject" in Table 3.5), outperforming the next best KB, Quasimodo, by 14-15% in relative recall. The gap to the manually built ConceptNet is even larger, where Ascent++ achieves two to five times higher relative recall, depending on the similarity threshold.

Considering all statements from each resource, Quasimodo and TransOMCS appear to have a slight edge; yet this is only due to the precision-oriented thresholding of Ascent++ (2M assertions vs. 18.5M for TransOMCS and 6.3M for Quasimodo). Without the cleaning phase (see Section 3.3.7), the unfiltered Ascent++ variant (which we denote as Ascent++^{large} in Table 3.5) would be the size-wise better point of comparison: with a size of 7.5M assertions. This resource achieves better relative recall scores than Quasimodo at $\tau =$ 0.96, and TransOMCS at all three selected similarity thresholds. Moreover, Ascent++^{large} still outperforms both TransOMCS and Quasimodo in terms of typicality.

In Section 3.5.3, we will show that when increasing the size of Ascent++ to reach that of Ascent, we still achieve competitive typicality scores. This gives us the flexibility to tune for either precision or recall, by adjusting the cleaning phase.

These results confirm that large-scale extraction from web crawls can significantly outperform the recall of resources built from smaller, specifically selected document collections (Ascent, TupleKB) without sacrificing precision. Furthermore, the significant gains in precision over TransOMCS and Quasimodo come at a much lower decrease in recall.

3.5.3 Ascent++ versus Ascent

The Ascent++ method presented in this chapter substantially extended a prior work, Ascent (Nguyen et al. 2021a). These two pipelines leverage the same OpenIE system. However, Ascent operated solely on top-ranked query results from a search engine, whereas Ascent++ processes a massive web crawl, involving new techniques for scalability and quality control in the presence of very noisy web contents. Besides, our data model has extended the one in Ascent to include a new dimension of typicality scores. We also improved our techniques for canonicalizing assertions with enhanced clustering techniques.

In the following, we make a head-to-head quality comparison between the two resources.

Quality Comparison. To compare Ascent++ and Ascent, in addition to *saliency@10* and *typicality (all assertions)* for the 200 common subjects, as well as the relative recall previously

Table 3.7: Quality comparison between ASCENT and ASCENT++.

Resource	**#s**	**#spo**	**Typicality @10**	**Typicality (Random)**	**Saliency (Random)**
ASCENT^{filtered}	500	86K	87.6%	65.6%	40.6%
ASCENT++^{filtered}	500	86K	**93.6%**	**82.0%**	**44.6%**

presented in Table 3.5, we perform three more evaluations: *typicality@10*, *typicality (random assertions)* and *saliency (random assertions)*, whose results are presented in Table 3.7.

For *typicality@10*, we perform the same sampling process as for *saliency@10*; the only difference is that ASCENT++ assertions are now sorted by typicality score instead of saliency score.

For *typicality (random)* and *saliency (random)*, a fair comparison must consider the different sizes of the two resources. Thus, we randomly sampled 500 subjects from both resources. For these 500 subjects, ASCENT has 86,054 CSK assertions. For ASCENT++ to have a similar number of assertions, we increased the limit for the maximal number of assertions per subject in ASCENT++ (i.e., the last filter of the pipeline, see Section 3.3.7) from 1,000 to 7,250. That resulted in 86,065 CSK assertions in ASCENT++ for the 500 random subjects. Now that the two resources have comparable sizes, we randomly picked up 500 triples from each resource and used them for the two crowd-sourcing evaluations, measuring *typicality (random)* and *saliency (random)*.

The results in Table 3.7 show that ASCENT++ clearly outperforms the prior ASCENT KB by a large margin regarding the typicality of both top-ranked statements and randomly sampled ones. Although the saliency scores of random samples drop significantly compared to those of the top-10 statements (see Table 3.5), ASCENT++ still outperforms ASCENT by 4 percentage points. Combined with the results in Table 3.5, the new ASCENT++ KB consistently shows better quality than its predecessor ASCENT KB, on both precision and relative recall.

3.5.4 Precision of ConceptNet Mapping

To evaluate the ConceptNet mapping module, we manually annotated 100 random samples from the final ASCENT++ KB. For each sample, we marked it as a correct mapping if the fixed-schema triple preserved the meaning of the original triple. We obtained a precision of 96%.

3.6 Extrinsic Evaluation

3.6.1 Setup

To answer *RQ2*, we conduct a comprehensive evaluation of the contribution of commonsense knowledge to question answering (QA) via three different setups, all based on the idea of priming pre-trained language models (LMs) with context (Guu et al. 2020, Lewis et al. 2020, Petroni et al. 2020).

1. In *masked prediction*, we ask language models to predict single tokens in generic sentences (Petroni et al. 2019).
2. In *generative QA*, we provide questions and let autoregressive LMs generate arbitrary answer sentences (Lewis et al. 2020).
3. In *span prediction*, LMs select the best answers from provided CSKB content (Lan et al. 2020).

We illustrate all settings in Table 3.8. In every setting, LMs are provided with a context in the form of assertions taken from competitor CSKBs. These setups are motivated by the observation that priming language models with context can significantly influence their predictions. Previous works on language model priming mainly focused on evaluating retrieval strategies. In contrast, our comprehensive test suite focuses on the impact of utilizing different CSK resources while leaving the retrieval component constant.

Masked prediction comes with the advantage of allowing automated evaluation, although automated evaluation may unfairly discount sensible alternative answers. Also, masked prediction is limited to single tokens. Autoregressive LM-based generations circumvent this restriction, although they necessitate human annotations and can be prone to evasive answers. They are thus well complemented by extractive answering schemes, limiting the language models' abstraction abilities but providing the cleanest way to evaluate the context alone.

Base Models. Following standard usage, we use RoBERTa-large (Liu et al. 2019) for masked prediction, GPT-3 (Brown et al. 2020) for the generative setup, and ALBERT-xxlarge (Lan et al. 2020), fine-tuned on SQuAD 2.0 (Rajpurkar et al. 2018), for span prediction.

Context Retrieval. Using the SentenceBert model *msmarco-distilbert-base-v3*, we compute embeddings of the given query and all verbalized triples in the CSKBs. Then we use cosine similarity to select the top-5 most similar triples to the query as context.

Table 3.8: Examples of three QA settings in the extrinsic evaluation.

Setting	Input	Sample output
Masked prediction	Elephants eat [MASK]. [SEP] Elephants eat roots, grasses, fruit, and bark, and they eat a lot of these things.	everything (15.52%), trees (15.32%), plants (11.26%)
Generative QA	Context: Elephants eat roots, grasses, fruit, and bark, and they eat a lot of these things. Question: What do elephants eat? Answer:	Elephants eat a variety of different foods.
Span prediction	question=*"What do elephants eat?"* context=*"Elephants eat roots, grasses, fruit, and bark, and they eat a lot of these things."*	start=14, end=46, answer=*"roots, grasses, fruit, and bark"*

Models used (top to bottom): RoBERTa, GPT-3, and ALBERT.

Task Construction. Previous work has generated masked sentences based on templates from ConceptNet triples (Petroni et al. 2019). However, the resulting sentences are often unnatural, following the idiosyncrasies of the ConceptNet data model.

We leveraged a new dataset of natural commonsense sentences for *masked prediction* that is based on an independent human-annotated dataset, the aforementioned CSLB dataset (Devereux et al. 2014), and consists of 19.6K masked sentences processed by Nguyen et al. (2021a).

For the *generative* and *extractive* settings, we used the Google Search auto-completion functionality to collect commonsense questions about popular subjects by feeding the API with six prefixes: *"what/when/where are/do <subject> ..."*. That process returned 8,098 auto-completed queries for 200 popular subjects.

Next, we drew samples from the query set, then manually removed jokes and other noise (e.g., *"where do cows go for entertainment"*), obtaining 100 questions for evaluation.

Evaluation Scheme. For commonsense topics, questions often have multiple valid answers. Additionally, given that answers in our generative and extractive QA settings are very open, creating an automated evaluation is difficult. Therefore, we use human judgments for evaluating all settings except masked prediction. Specifically, given a question and set of answers, we ask human annotators to assess each answer based on two dimensions, *correctness*, and *informativeness*, each on a 4-point Likert scale from 0 (lowest) to 3 (highest) (see Table 3.9 for the question templates given to the annotators).

Table 3.9: Questions for the human evaluation of context-augmented QA.

Dimension	Question and answer options
Correctness	*How often does this answer hold true?*
	3 - Always/Often | The answer is always or often true.
	2 - Sometimes/Likely | It is sometimes or likely true.
	1 - Farfetched/Never | It is false or farfetched at best.
	0 - Invalid | It is invalid or makes no sense.
Informativeness	*How informative is the answer?*
	3 - Highly | It provides very useful knowledge w.r.t the question.
	2 - Moderately | It is moderately useful.
	1 - Slightly | It is slightly useful.
	0 - Not at all | It is too general or makes no sense.

These are used for the human evaluation of Generative QA and Span Prediction tasks.

- *Correctness* indicates whether an answer holds true.
- *Informativeness* reflects that the information conveyed by an answer is helpful.

An answer could be correct and uninformative at the same time. For example, given the question *"What do elephants use their trunk for?"*, the answers *"For breathing"* and *"To suck up water"* are both correct and informative. However, *"To do things"* is a correct answer but not informative at all. On the other hand, if an answer is incorrect, it should automatically be uninformative. Three annotators evaluate each question in Amazon MTurk.

We use the mean precision at k ($P@k$) metric for evaluating *masked prediction*, following Petroni et al. (2019).

3.6.2 Results

The extrinsic evaluation results are shown in Table 3.10.

- For *masked prediction*, all CSKBs contribute useful contexts that substantially improve the quality of LM responses. Ascent++, along with the only manually-constructed KB, ConceptNet, statistically significantly outperforms all other KBs at every threshold k (all p-values of paired Student's t-test below 0.05).
- For *generative QA*, markedly, we find that GPT-3 performs on average better without any context. However, in combination with Ascent++, it still performs better than with any other CSKB. More research is needed to design methods to pull relevant context from CSKBs and decide when to use it in LM-based QA and when to rely on the LM's knowledge alone.

Table 3.10: Results of context-augmented QA evaluation.

Context	**Masked Prediction**			**Generative QA**		**Span Prediction**	
	P@1	**P@5**	**P@10**	**Corr**	**Info**	**Corr**	**Info**
No context	8.10	17.16	21.37	**2.47**	**2.01**	-	-
ConceptNet	**14.41**	**27.08**	32.16	2.22	1.70	1.74	1.52
TransOMCS	7.08	15.42	19.99	1.32	0.86	0.99	0.85
TupleKB	11.61	24.76	30.36	2.22	1.51	1.70	1.38
Quasimodo	12.11	22.75	27.71	2.03	1.51	1.75	1.44
Ascent	11.95	24.70	29.70	2.25	1.76	1.88	1.60
Ascent++	13.30	27.03	**32.90**	2.32	1.71	**1.94**	**1.63**

Metrics: ***P@k*** *- mean precision at k (%),* ***Corr*** *- correctness ([0..3]),* ***Info*** *- informativeness ([0..3]).*

- For *span prediction*, where answers come directly from retrieved contexts, Ascent++ also outperforms all other competitors. Ascent++ obtained statistically significant gains over Quasimodo, TupleKB, and TransOMCS on both metrics, and over ConceptNet on correctness. This indicates that our Ascent++ assertions have high quality compared to others.

3.7 Evaluation of Semantic Facets

To answer *RQ3*, we evaluate semantic facets both intrinsically and extrinsically.

3.7.1 Intrinsic Evaluation of Semantic Facets

As there are no existing CSKBs coming with semantic facets, we provide comparisons with a strong LM baseline, GPT-2 (Radford et al. 2019). First, we randomly drew 300 assertions along with their top-1 facets from our KB. Next, we translated each statement into a sentence prefix and ask GPT-2 to fill in the remaining words to complete the sentence. For example, given the quadruple `<elephant; uses; their trunks; PURPOSE:to suck up water>`, the sentence prefix would be *"Elephants use their trunk to"*. For that, GPT-2's continuation is *"to move around."* Then, each sentence prefix along with the two answers (from Ascent++ and GPT-2) were shown to a human annotator (without knowing the source of the answers) who annotated if each answer was *correct/incorrect* and *informative/uninformative*, following similar metrics used for the QA evaluation in Section 3.6. The results are reported in Table 3.11. Ascent++ achieves 70.1% correctness and 54.15% informativeness, both significantly better than the values for the GPT-2 model.

Table 3.11: Assessment of Ascent++ facets and LM-generated facets.

Source	Correctness (%)	Informativeness (%)
GPT-2	61.79	48.84
Ascent++	**70.10**	**54.15**

Table 3.12: Extrinsic evaluation of semantic facets.

Ascent++ assertions	Masked Prediction	Generative QA		Span Prediction	
	P@1	Corr	Info	Corr	Info
Without facets	13.30	**2.32**	1.71	1.94	1.63
With facets	**13.38**	2.26	**1.79**	**2.12**	**1.77**

*Metrics: **P@1** - mean precision at one (%), **Corr** - correctness ([0..3]), **Info** - informativeness ([0..3]).*

3.7.2 Extrinsic Evaluation of Semantic Facets

We reused the three question answering tasks from Section 3.6. The results of this evaluation are shown in Table 3.12. Expanding the triples with semantic facets gives a consistent improvement in four out of five evaluation metrics (precision@1 in masked prediction, informativeness in generative QA and span prediction, and correctness in span predication), with the most prominent effect being observed for span prediction (8.6% and 9.3% relative improvements over the no-facet context in informativeness and correctness, respectively).

3.8 Summary

In this chapter, we acquired refined commonsense knowledge for the first type of entry points: everyday concepts. We presented Ascent++, a methodology to extract and semantically organize refined commonsense knowledge from large-scale web contents. Our expressive knowledge representation allowed us to identify considerably more informative assertions, overcoming the limitations of prior works. The techniques for filtering, aggregating, and consolidating extracted tuples showed that CSK extraction from broad web content is feasible at scale, with both high precision and high recall. Intrinsic and extrinsic evaluations confirmed that the resulting CSKB is a significant advance over existing CSK collections and provides an edge over recent LM-based approaches. Code, data, and a web interface are accessible at https://ascentpp.mpi-inf.mpg.de.

4

CULTURE-CENTRIC EXTRACTION AND ORGANIZATION

In this chapter, we acquire commonsense knowledge for the second type of entry points: cultural groups (e.g., `Japanese`, `Hispanic`, `Buddhist`).

Our proposed method, called Candle, discovers cultural commonsense knowledge (CCSK), including human traits and behaviors conditioned on cultural contexts, from large-scale web contents. Considering three domains of cultural groups (geography, religion, occupation), Candle extracts high-quality CCSK assertions for a number of general cultural facets (food, drinks, clothing, traditions, rituals, behaviors), and organizes them into coherent clusters. Candle includes judicious techniques for classification-based filtering and scoring of interestingness. The evaluation with human judgements shows the superiority of the resulting CCSK collection over prior works, and an extrinsic use case demonstrates the benefits of CCSK for the GPT-3 language model.

The project website is hosted at https://candle.mpi-inf.mpg.de, including downloadable code and data.

4.1 Introduction

Motivation. Major CSK acquisition projects, such as ConceptNet (Speer et al. 2017) and ATOMIC (Sap et al. 2019a), have compiled large sets of CSK assertions, but are solely focused on "universal CSK": assertions that are agreed upon by almost all people and are thus viewed as "globally true". In the Ascent++ project (Chapter 3), our attempt to refine CSK assertions was restricted to general semantic facets such as spatial or temporal conditions, and other dimensions like cause and purpose. What is missing, though, is that CSK must often be viewed in the *context of specific social or cultural groups*: the world view of a European teenager does not necessarily agree with those of an American business person or a Far-East-Asian middle-aged factory worker.

The work presented in this chapter addresses this gap by automatically compiling CSK that is conditioned on cultural contexts. We refer to this as *cultural commonsense knowledge*, or *CCSK* for short. Given a set of *cultures/cultural groups* (or *subjects*, which will be used interchangeably), our goal is to acquire CCSK assertions for those subjects that concern several cultural facets of interest. For example, we collect CCSK assertions such as:

- `CULTURAL-GROUP:East Asia, CULTURAL-FACET:food, Tofu is a major ingredient in many East Asian cuisines`, or
- `CULTURAL-GROUP:firefighter, CULTURAL-FACET:behavior, Firefighters use ladders to reach fires`.

The value of having a knowledge base with this information lies in making AI applications more situative and more robust.

Collecting CCSK is a challenging task, especially at scale. The few prior works with data that have specifically addressed the socio-cultural dimension of CSK are the projects Quasimodo (Romero et al. 2019), StereoKG (Deshpande et al. 2022), and the work of Acharya et al. (2021). The latter, which employed human annotations from Amazon MTurk, only produced a few hundred assertions. StereoKG used a specialized way of automatically extracting stereotypes from QA forums; its resulting resource is still small in size and suffers from high noise. On the other hand, Quasimodo aimed to collect general CSK and its KB only contains a small fraction of culturally relevant assertions.

Our goal is to collect CCSK of wide coverage while assuring high precision.

Approach. CCSK is expressed in text form on web pages and social media, but this is often very noisy and difficult to extract. We devised an end-to-end methodology and system, called Candle (extra**c**ting cultur**a**l commo**n**sense knowle**d**ge at sca**le**), to automatically extract and systematically organize a large collection of CCSK assertions. For scale, similarly to

Colored annotations: **cultural domains**, **subjects (cultural groups)**, **cultural facets**, **concepts**.

geography>country	Germany	drinks
German beer festivals in October are a celebration of beer drinking.		
geography>region	East Asia	food
Tofu is a major ingredient in many East Asian cuisines.		
geography>region	South Asia	traditions
In South Asia, henna is often used in bridal makeup or to celebrate festivals.		
occupation	lawyer	clothing
Lawyers wear suits to look professional.		
occupation	firefighter	behaviors
Firefighters run into burning buildings to save lives.		

Figure 4.1: Example assertions from CANDLE.

ASCENT++ (Chapter 3), we tap into the C4 web crawl (Raffel et al. 2020), a huge collection of web pages. This provides an opportunity to construct a sizable CCSK collection, but also a challenge in terms of scale and noise.

The output of CANDLE is a set of 1.1M CCSK assertions, organized into 60K coherent clusters. The set is organized by three domains of interest – geography, religion, occupation – with a total of 386 instances, referred to as *subjects* (or *cultures/cultural groups*). Per subject, the assertions cover five *facets of culture*: food, drinks, clothing, rituals, traditions (for geography and religion subjects) or behaviors (for occupations). In addition, we annotate each assertion with its salient *concepts*. Examples for the computed CCSK are shown in Figure 4.1.

Given a set of cultural subjects and facets, CANDLE operates in 6 steps. First and second, we identify candidate assertions using simple techniques for *subject detection* (named entity recognition and string matching), and *generic rule-based filtering*. Third, we *classify assertions* into specific cultural facets, which is challenging because we have several combinations of cultural groups and cultural facets, making it very expensive to create specialized training data. Instead, we creatively leverage LMs pre-trained on the natural language inference (NLI) task to perform zero-shot classification on our data, with judicious techniques to enhance the accuracy. Fourth we use state-of-the-art techniques for *assertion clustering*, and fifth a simple but effective method to *extract concepts* in assertions. Lastly, we combine several features to *score* the interestingness of assertions, such as frequency, specificity, distinctiveness. This way, we steer away from overly generic assertions (which LLMs like GPT-3 tend to generate) and favor assertions that set their subjects apart from others.

Contributions. The project's key contributions are:

1. *Methodology* (Section 4.3): We propose Candle, an end-to-end methodology to extract high-quality CCSK from very large text corpora. Candle contains new techniques for judiciously classifying and filtering CCSK-relevant text snippets, and for scoring assertions by their interestingness.
2. *Resource* (Section 4.4): We construct and publicly release a large collection of CCSK assertions for 386 subjects covering three domains (geography, religion, occupation) and several facets (food, drinks, clothing, traditions, rituals, behaviors).

The evaluation with human judgments shows that the assertions in Candle are of significantly higher quality than those from prior works (Section 4.5). An extrinsic use case demonstrates that our CCSK can improve performance of GPT-3 in question answering (Section 4.6). Code and data can be accessed at https://candle.mpi-inf.mpg.de.

4.2 Knowledge Representation

Our representation of CCSK is based on the notions of *subjects* (from three major domains: geography, religion and occupation) and *cultural facets*. These are the key labels for CCSK *assertions*, which are informative sentences with salient *concepts* marked up.

Input. We assume two sets to be given:

- $\mathcal{G}$: A set of *subjects (cultural groups)* $g_1, ..., g_n$ from a *cultural domain*, e.g., based on geo-locations (China, Middle East, California), religious groups (Christians, Muslims, Buddhists) or occupations (taxi driver, professor, web developer);
- $\mathcal{F}$: A set of *cultural facets* $F_1, ..., F_m$, e.g., food, drinks, clothing, traditions, rituals, behaviors.

Note that the cultural facets need not be mutually exclusive, e.g., food assertions sometimes overlap with traditions.

Format. Our objective is to collect a set of CCSK assertions for a given subject and a cultural facet. Existing commonsense resources store assertions in triple format (e.g. (Speer et al. 2017, Romero et al. 2019)), semantic frames (Ascent++ in Chapter 3) or generic sentences (Bhakthavatsalam et al. 2020). Although the traditional triple-based and frame-based data models are convenient for structured querying, and well suited for regular assertions like birth dates, citizenships, etc., they often falls short of capturing nuanced natural-language assertions, as essential for CSK. Moreover, recent advances in pre-trained LMs have made it easier to feed downstream tasks with less structured knowledge.

With CANDLE, we thus follow the approach of GenericsKB (Bhakthavatsalam et al. 2020), and use natural-language sentences to represent assertions.

In principle, an assertion could comprise even several sentences. The longer the assertions are, however, the harder it is to discern their core. In this work, for higher precision and simplicity of computations, we only consider single sentences.

Organization. Since natural language often allows to express similar assertions in many different ways, and web harvesting naturally leads to discovering similar assertions multiple times, we employ clustering as an essential component in our approach.

A *cluster* of CCSK assertions for a subject and a cultural facet contains assertions of identical meaning, and for presentation purposes, is summarized by a single summary sentence. Each cluster also comes with a score denoting its interestingness.

To further organize assertions, we also identify salient *concepts*, i.e., important terms inside assertions, that can be used for concept-centric browsing of assertion sets.

Several examples of CCSK assertions produced by CANDLE are shown in Figure 4.1.

4.3 Methodology

We propose an end-to-end system, called CANDLE, to extract and organize CCSK assertions based on the proposed CCSK representation. Notably, our system does not require annotating new training data, but only leverages pre-trained models with judicious techniques to enhance the accuracy. The system takes in three inputs:

- an English text corpus (e.g., a large web crawl);
- a set of *subjects* (cultural groups);
- a set of *facets* of culture.

CANDLE consists of six modules (see Figure 4.2). Throughout the system, step by step, we reduce a large input corpus (which could contain billions of documents, mostly noisy) into high-quality clusters of CCSK assertions for the given subjects and facets. Each cluster in the output is also accompanied by a representative sentence and an interestingness score. A summary of techniques and models applied in each module is presented in Table 4.1. We next elaborate on each module.

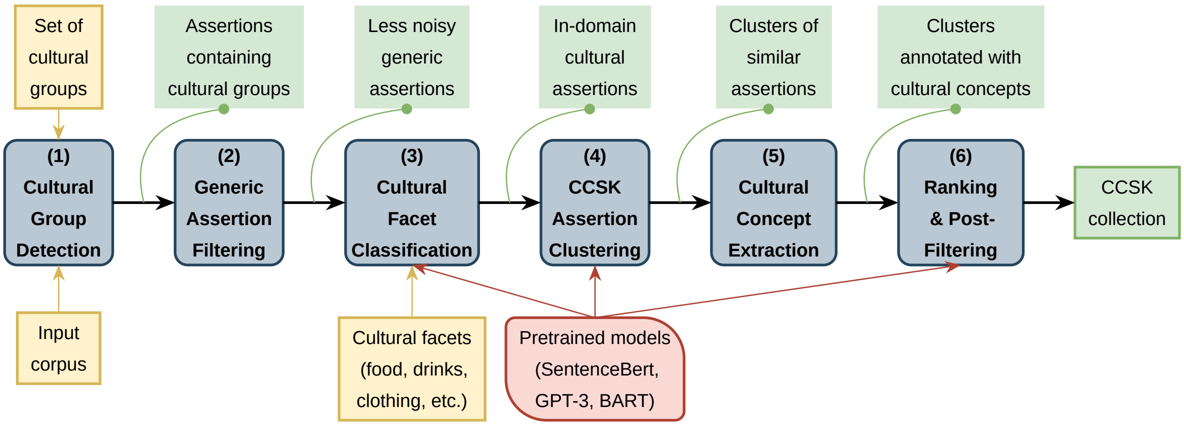


Figure 4.2: Architecture of the Candle system.

Table 4.1: Models and techniques used in Candle.

#	Module	Techniques/models used
1	Cultural group detection	String matching, named entity recognition
2	Generic assertion filtering	Hand-crafted lexico-syntactic rules
3	Cultural facet classification	The *bart-large-mnli* model
4	Assertion clustering	SentenceBert + HAC
	Cluster summarization	The GPT-3 *curie-001* model
5	Concept extraction	Common n-gram extraction
6	Cluster ranking	Ad-hoc ranking features
	Post-filtering	Rule-based filtering

4.3.1 Cultural Group Detection

We start the extraction by searching for sentences that contain mentions of the given subjects (cultural groups). These will be the candidate sentences used in the subsequent modules. To achieve high recall, we utilize generous techniques such as string matching and named entity recognition (NER). We will use more advanced filtering techniques in later modules, to ensure high precision.

For the *geography* and *religion* domains, in which subjects are named entities, we use spaCy's NER module to detect subjects. Specifically, geo-locations are detected with the GPE tag (geopolitical entities), and religions are detected with the NORP tag (nationalities or religious or political groups). For each subject, we also utilize a list of aliases for string matching, which can be the location's alternate names (e.g., United States, the U.S., the States), or demonyms (e.g., Colombians, Chinese, New Yorker), or names for religious adherents (e.g., Christians, Buddhists, Muslims) - which can be detected with the NORP tag as well.

For the *occupation* domain, we simply use exact-phrase matching to detect candidates. Each occupation subject is enriched with its alternate names and its plural form to enhance coverage.

4.3.2 Generic Assertion Filtering

CSK aims at covering *generic assertions*, not episodic or personal experiences. For example, *"Germans like their currywurst"* is a generic assertion, but *"I visited Germany to eat currywurst"* or *"This restaurant serves German currywurst"* are not.

GenericsKB (Bhakthavatsalam et al. 2020) is arguably the most popular work on automatically identifying generic sentences in texts. GenericsKB used a set of 27 hand-crafted lexico-syntactic rules to extract high-quality generic sentences from different text corpora (the ARC corpus, SimpleWikipedia, and the Waterloo crawl of education websites). For example, the lexical rules look for sentences of short length, starting with a capitalized character, having no bad first words (e.g., determiners), ending with a period, having no URL-like snippets, etc. The syntactic rules only accept a sentence if its root (in the dependency tree) is a verb and not the first word, and if there is a noun before the root verb, etc.

CANDLE adopts the GenericsKB rules. However, as GenericsKB only deals with general concepts (e.g., `tree`, `bird`, `car`), some of the rules are not applicable for the cultural subjects, which can be named entities. Hence, depending on the subjects and facets, we adaptively modify the rules (by dropping some of them) so that we will not miss out valuable assertions. For instance, for geography subjects, the *has-no-determiners-as-first-word* rule will filter out valuable assertions such as "The Chinese use chopsticks to eat their food" or "The currywurst is a traditional German fast food dish", and it must be dropped. In another situation, when exploring `CULTURAL-FACET:traditions`, the *remove-past-tense-verb-roots* rule would be too aggressive as it rejects assertions about past traditions. The rule that rejects sentences with PERSON entities can be used for the geography and occupation subjects, but must not be used for religions, because it will filter out sentences about, e.g., Buddha or Jesus Christ. Full details are in the published code base[2].

[2]https://github.com/cultural-csk/candle/blob/main/candle/pipeline/component_generic_sentence_filter.py

4.3.3 Cultural Facet Classification

To organize CCSK and filter out irrelevant assertions, we classify candidate sentences into several facets of culture. Traditional methods for this classification task would require a substantial amount of annotated data to train a supervised model. The costs of data annotation are often a critical bottleneck in large-scale settings. In Candle, we aim to minimize the degree of human supervision by leveraging pre-trained models for zero-shot classification.

A family of pre-trained models that is suitable for our setting is *textual entailment*, a.k.a. *natural language inference* (NLI): given two sentences, does one entail the other (or are they contradictory or unrelated)? Our approach to adopting such a model for cultural facet classification is inspired by the zero-shot inference method of Yin et al. (2019). Given a sentence sent and a facet F, we construct the NLI test as follows:

Input: $\text{Premise} \leftarrow \text{sent}, \text{Hypothesis} \leftarrow$ "This text is about F"
Output: $P[\text{sent} \in F] \leftarrow P[\text{Premise} \Rightarrow \text{Hypothesis}]$

The probability of Premise entailing Hypothesis will be taken as the probability of sent being labeled as F, denoted as $P[\text{sent} \in F]$. For example, with the sentence *"German October festivals are a celebration of beer and fun"*, the candidate entailments will be *"This text is about drinks"*, *"... about food"*, *"... about traditions"*, and so on. Multiple of these facets may yield high scores in these NLI tests.

To enhance precision, we introduce a set of *counter-labels* for topics that are completely outside the scope of CCSK, for example, politics or business. A sentence sent will be accepted as a good candidate for facet F if

$$\begin{cases} P[\text{sent} \in F] \geq \rho_+ & \text{and} \\ P\left[\text{sent} \in \tilde{F}\right] \leq \rho_- & \text{for all counter-labels } \tilde{F} \end{cases} \tag{4.1}$$

where ρ_+ and ρ_- are hyperparameters in the range $[0, 1]$, giving us the flexibility to tune for either precision or recall.

In our experiments, we use the BART model (Lewis et al. 2020) finetuned on the MultiNLI dataset (Williams et al. 2018) for NLI tests. Our crowdsourcing evaluations show that the zero-shot classifiers with the enhanced techniques achieved high precision (see Section 4.5.3).

> Given the following sentences:
> (1) The basic color for a Chinese funeral is all white.
> (2) In China, white is reserved for funerals.
> (3) At a traditional Chinese funeral, guests are expected to wear somber colors.
> (4) The Chinese wear white at funerals.
> (5) The Chinese color for mourning and funerals is white rather than black.
> Summarize them using one short sentence:
>
> In China, white is the traditional color for funerals and mourning.

Figure 4.3: A screenshot of using GPT-3 for cluster summarization.

4.3.4 Assertion Clustering

The same assertion can be expressed in many ways in natural language. For example, *"Fried rice is a popular Chinese dish"* can also be written as *"Fried rice is a famous dish from China"* or *"One of the most popular Chinese food is fried rice"*. Clustering is used to group such assertions, which reduces redundancies, and allows to obtain frequency signals on assertions.

Clustering. Following Ascent++, we leverage SentenceBert (Reimers and Gurevych 2019), to compute vector representations for all assertions, and use the hierarchical agglomerative clustering (HAC) algorithm for clustering. Clustering is performed on assertions of each subject-facet pair.

Cluster Summarization. Since each cluster can have from a few to hundreds of sentences, it is important to identify what those sentences convey, in a concise way.

One way to compute a representative assertion for a cluster is to compute the centroid of the cluster, then take its closest assertion as the representative. Yet for natural-language data, this does not work particularly well.

In Candle, we therefore approach cluster summarization as a generative task by using GPT-3, specifically the *curie-001* model (see Figure 4.3 for an example). Annotator-based evaluations show that GPT-generated representatives received significantly better scores than the base sentences in the clusters (see Section 4.5).

4.3.5 Concept Extraction

While the cultural groups are regarded as subjects, *concepts* are akin to objects of the assertions. Identifying these concepts enables concept-focused browsing (e.g., browsing Japan assertions only about the Miso soup, etc.).

We postulate that main concepts of an assertion cluster are terms shared by many members. To this end, we extract all n-grams ($n = 1..3$) of all assertions in a cluster (excluding subjects themselves, and stop words); and retain the ones that occur in more than 60% of the assertions. If both a phrase and its sub-phrase appear, we only keep the longer phrase in the final output. Noun-phrase concepts are normalized by singularization.

4.3.6 Cluster Ranking and Post-Filtering

Ranking commonsense assertions is a crucial task. Unlike encyclopedic knowledge, which is normally either true or false, precision of CSK is usually not a binary concept, as it generalizes over many groups. With CANDLE, we aim to pull out the most interesting assertions for each subject, and avoid overly generic assertions such as *"Chinese food is good"* or *"Firefighters work hard"*, which are very common in text.

Extracting and clustering assertions from large corpora gives us an important signal of an assertion, its *frequency*. However, ranking based on frequency alone may lead to reporting bias. As we compile a CCSK collection at large scale, it also enables us to compute the *distinctiveness* of an assertion against others in the collection. The notion of these two metrics can be thought of as term frequency and inverse document frequency in the established TF-IDF technique for IR document ranking (Sparck Jones 1988). Besides *frequency* and *distinctiveness*, we score the interestingness of assertion clusters based on two other custom metrics: *specificity* (how many objects are mentioned in the assertion?) and *facet relevance* (how relevant is the assertion to the cultural facet?).

Frequency. For each subject-facet pair, we normalize cluster sizes into the range $[0, 1]$, using min-max normalization.

Distinctiveness. We compute the IDF of a cluster cls as follows:

$$IDF(cls) = \frac{\sum_{cls' \in CLS} size(cls')}{\sum_{cls' \in CLS} size(cls') \times \sigma(cls, cls')} \tag{4.2}$$

where CLS is the set of all clusters for a given facet (e.g., food) and domain (e.g., geography>country), and

$$\sigma(cls, cls') = \begin{cases} 1 & \text{if } sim(cls, cls') \geq \theta \\ 0 & \text{otherwise} \end{cases} \tag{4.3}$$

Here, $sim(cls, cls')$ is the semantic similarity between the two clusters cls and cls', and θ is a predefined threshold. In CANDLE, to reduce computation, we approximate $sim(cls, cls')$ as the similarity between their summary sentences, which can be computed as the cosine similarity between their embedding vectors. When computing these embeddings, the subjects in the sentences are replaced with the same [MASK] tokens so that we only compare the expressed properties. Then, we normalize the logarithmic IDF values into the range $[0, 1]$ to get the distinctiveness scores of clusters.

Specificity. We compute the specificity of an assertion based on the fraction of nouns in it. Concretely, in CANDLE, the specificity of a cluster is computed as the specificity of its summary sentence.

Facet Relevance. For each facet, we compute the facet relevance of a cluster by taking the average of the probability scores given to its members by the cultural facet classifier.

Combined Score. The final interestingness score for cluster cls is the average of the four feature scores. A higher score means higher interestingness.

Post-Filtering. Lastly, to eliminate redundancies and noise, and further improve the final output quality, we employ a few hand-crafted rules:

- At most 500 clusters per subject-facet pair are retained, as further clusters mostly represent redundancies or noise.
- We remove clusters that have no concepts extracted, or that are based on too few distinct sentences (more than 2/3 same sentences) or web source domains.
- We remove any cluster if either its summary sentence or many of its member sentences match a bad pattern. We compile a set of about 200 regular expression patterns, which were written by a knowledge engineer in one day. For e.g., we reject assertions that contain “the menu”, “the restaurant” (likely advertisements for specific restaurants), or animal and plant breeds named after locations, such as “American bison”, “German Shepherd”, etc.

4.4 Implementation

4.4.1 Input

Corpus. In Candle, we tap into the broad web as knowledge source, because of its diversity and coverage, which are important for long-tail subjects. Besides the benefits, the most challenging problem when processing web contents is the tremendous amount of noise, offensive materials, incorrect information etc., hence, choosing a corpus that has been chiefly cleaned is beneficial. We reuse the processed C4 dataset (Raffel et al. 2020) from the Ascent++ project (see Section 3.4.1), which contains 365M English articles, each with text content and source URL.

Cultural Groups. We collect CCSK for subjects from three cultural domains: geography (272 subjects), religions (14 subjects) and occupations (100 subjects). For geography, we split into four sub-domains: countries, continents, geopolitical regions (e.g., Middle East, Southeast Asia, etc.) and US states, which were collected from the GeoNames database (http://www.geonames.org/), which also provides alias names. We further enriched these aliases with demonyms from Wikipedia (https://en.wikipedia.org/wiki/Demonym).

Cultural Facets. We consider five cultural facets four each subject: *food*, *drinks*, *clothing*, *rituals*, and *traditions* (for geography/religion) or *behaviors* (for occupation), selected based on an article on facets of culture (https://en.wikipedia.org/wiki/Outline_of_culture).

Specially for religion and occupation subjects, we first ask the NLI model (cf. Section 4.3.3) to check if the input sentence is *"... about religions"* or *"... about professions"* before categorizing it into one of the five cultural facets. If the sentence is in the general domain (religions/professions) but it does not fall into one of the predefined facets, we classify its cultural facet as "Other" (see Table 4.3).

4.4.2 Hyperparameters

Based on tuning on small withheld data, we select the following values for hyperparameters.

For *cultural facet classification*, we fix ρ_+ to 0.5 and ρ_- to 0.3 in Equation (4.1).

For *assertion clustering*, we use the SentenceBert model *all-MiniLM-L6-v2* for computing sentence embeddings. For the HAC algorithm, we measure point-wise Euclidean distance of the normalized embeddings. Then, we use the Ward's linkage (Ward 1963), with the maximal distance threshold set to 1.5. In the few cases where input sets are larger, we truncate them at 50K most frequent sentences per subject-facet pair, since larger inputs only contain further redundancies, that are not worth the cubic effort of clustering. This concerns only 15 out of 386 subjects.

For *cluster summarization*, we consider the 500 most populated clusters for each subject-facet pair with a minimum size of 3 sentences.

Table 4.2: Processing time and output size of each step in CANDLE.

#	Step	Time	Output/Data size
-	Input	1.5 days for NLP preprocessing (cf. Section 3.4.1)	C4 corpus: 8B sentences, 196 countries, 705 alternate names
1	Subject detection	2 hours	367M subject matches 300M sentences (-96%)
2	Generic assertion filtering	2 hours	13M generic sentences (-96%)
3	Cultural facet classification	4 hours	769K positive sentences (-94%)
4	Assertion clustering	4 hours	42K clusters (-93%)
5	Concept extraction	$<$ 5 minutes	12.4K concepts
6	Cluster ranking and post-filtering	$<$ 5 minutes	8.8K clusters (-80%)
-	*Total*	*~ 12 hours*	

For domain geography>country and `FACET:food`*.*

Table 4.3: Statistics of the CANDLE CCSK collection.

Facet	Geo-Locations		Religions		Occupations	
	#Assertions	#Clusters	#Assertions	#Clusters	#Assertions	#Clusters
Food	240,459	12,981	9,750	680	9,837	511
Drinks	95,394	5,923	3,079	218	3,321	227
Clothing	14,170	1,237	1,695	141	4,367	278
Rituals	116,839	8,007	74,651	3,026	22,581	1,253
Traditions	214,931	13,606	68,202	2,798	-	-
Behaviors	-	-	-	-	25,152	1,495
Other	-	-	60,483	2,292	159,239	5,461
All	681,793	41,754	217,860	9,155	224,497	9,225

For *cluster ranking*, we fix θ in Equation (4.3) to 0.8.

4.4.3 Execution

We executed CANDLE on a cluster of 6K CPU cores (AMD EPYC 7702) and 40 GPUs (a mix of NVIDIA RTX 8000, Tesla A100 and A40 GPUs).

Regarding processing time, for the domain country (196 subjects), it took a total of 12 hours to complete the extraction, resulting in 8.4K clusters for the facet food (cf. Table 4.2). Occupations and religions took 8 and 6 hours each.

4.4.4 Result Statistics

We provide statistics of the output in Table 4.3. In total, the resulting collection has 1.1M CCSK assertions (i.e., base sentences) which form 60K clusters for the given cultural groups and facets.

Experiment Overview

We perform the following evaluations:

1. A comparison of quality of Candle's output and existing socio-cultural CSK resources: This analysis will show that our CCSK collection is of significantly higher quality than existing resources (Section 4.5.1), and even outperforms GPT-3-generated assertions (Section 4.5.2).
2. Per-domain and per-facet quality of Candle: We break down the Candle CCSK collection into domains and facets and analyze in details the assertion quality for each subcollection (Section 4.5.3).
3. Two extrinsic use cases for CCSK: In this evaluation, we perform two downstream applications, cultural question answering (QA) and GuessTheCountry game, showing that using CCSK assertions from Candle is beneficial for these tasks and outperform those generated by GPT-3 (Section 4.6).

4.5 Intrinsic Evaluation

We compare Candle with materialized resources of similar kind (Section 4.5.1) and with assertions generated by GPT-3 (Section 4.5.2). Then, we break down the collection into domains and facets to analyze assertion quality for each subcollection (Section 4.5.3).

4.5.1 Comparison with Other Resources

4.5.1.1 Setup

Compared Resources. We compare Candle with three prominent CSK resources: Quasimodo (Romero et al. 2019), AcharyaEtAl (Acharya et al. 2021), StereoKG (Deshpande et al. 2022). The former covers broad domains including assertions for countries and religions, while the others focus on cultural knowledge.

We evaluate two versions of our CCSK collection, one where each base assertion is retained independently (CANDLE^{raw}), the other containing only the cluster representatives (CANDLE^{cluster}).

Other popular CSK resources such as ConceptNet (Speer et al. 2017), GenericsKB (Bhakthavatsalam et al. 2020), ATOMIC (Sap et al. 2019a), ASER (Zhang et al. 2020b), TransOMCS (Zhang et al. 2020a), and ASCENT++ (Chapter 3) do not have their focus on cultural knowledge and contain very little to zero assertions for geography or religion subjects, hence, they are not qualified for this comparison.

Evaluation Metrics. Following previous works (Romero et al. 2019, Deshpande et al. 2022), we analyze assertion quality along several complementary metrics, annotated by Amazon MTurk crowdsourcing.

- *Commonality:* This dimension measures whether annotators have heard of the assertion before.
- *Plausibility:* This dimension measures whether assertions are considered to be generally true, a CCSK-softened variant of correctness/precision.
- *Distinctiveness:* This dimension measures discriminative informativeness of assertions, i.e., whether the assertion differentiates the subject from others.

Each metric is evaluated on a 3-point Likert scale for *negation* (0), *ambiguity* (1) and *affirmation* (2). We present the crowdsourcing questions and answer options in Table 4.4.

The distinctiveness metric is only applicable if the answer to the plausibility question is either 1 or 2. In case the annotators are not familiar with the assertion, we advise them to perform a quick search on the web to find out the answers for the plausibility and distinctiveness questions. Additionally, we ask if the annotator would consider the assertion as an inappropriate or offensive material, measuring *offensiveness*.

Evaluation Scheme. For comparability, all resources are compared on 100 random assertions of the same five country subjects covered in StereoKG (Deshpande et al. 2022), namely United States, China, India, Germany and France. We note that among all compared resources, ACHARYAETAL (Acharya et al. 2021) only contain two subjects (United States and India), so for that resource, we only sample from those. For StereoKG, we use their natural-language assertions. For Quasimodo and ACHARYAETAL, we verbalize their triples using crafted rules. Each MTurk task consists of five assertions evaluated by three different annotators.

Table 4.4: Crowdsourcing questions for CCSK evaluation in Candle.

Dimension	**Question and answer options**
Commonality	*Have you heard of the assertion before?*
	0 - Never | No, I have never heard of it. *1 - Sometimes* | I have heard of it once or twice. *2 - Often* | Yes, I have heard of it many times.
Plausibility	*Do you consider the assertion to be true?*
	0 - No | No, not at all. *1 - Somewhat* | Somewhat. *2 - Yes* | Yes, absolutely.
Distinctiveness	*Does the assertion set the culture apart from others?*
	0 - No | This assertion applies to many cultures. *1 - Somewhat* | This assertion applies to several cultures. *2 - Yes* | This assertion is quite unique and applies to only a few or this one culture.

Details of MTurk Tasks. Workers are compensated $0.50 per task. We select Master workers with lifetime's acceptance rate more than 99% (this rate is provided by the platform). We obtain fair inter-annotator agreements given by Fleiss' κ (Fleiss and Cohen 1973): 0.26 for plausibility and 0.25 for distinctiveness. This number for commonality (0.13) is lower than others because it is an objective question (has the annotator heard of the assertion?).

Table 4.5: CANDLE in comparison to other CSK resources.

Resource	Format	#Assertions			Quality [0..2]			Offensiveness (%)	Average length (#chars)
		196 countries	10 religions	100 occupations	Plausibility	Commonality	Distinctiveness		
Crowdsourcing									
ACHARYAETAL	Fixed relations	225	0	0	1.32	**1.22**	0.25	2	102
Extractive									
StereoKG	OpenIE triples	2,181	1,810	0	0.54	0.46	0.21	18	37
Quasimodo	OpenIE triples	22,588	10,628	51,124	0.68	0.65	0.31	13	32
CANDLE^{raw}	Sentences	520,971	226,807	238,057	1.21	0.93	0.76	1	69
CANDLE^{cluster}	Sentences	28,711	8,823	9,826	**1.50**	1.15	**1.03**	1	73

Quality evaluated on assertions of five countries in StereoKG (United States, China, India, Germany, France).

4.5.1.2 Results

A summary of comparison with other resources is shown in Table 4.5.

Resource Size and Assertion Length. Candle outperforms all other resources on the number of base sentences. When turning to clusters, our resource still has significantly more assertions than AcharyaEtAl (which was constructed manually at small scale) and StereoKG (extracted from Reddit/Twitter questions). Quasimodo has comparable size with Candle^{cluster} for the country and religion domains and has more for the occupation domain.

The OpenIE-based methods, Quasimodo and StereoKG, produce the shortest assertion (32 and 37 characters on average, respectively). The manually-constructed KG (AcharyaEtAl) has the longest assertions (102 characters). Candle, having average assertion lengths (69 and 73), stands between those two approaches.

Assertion Quality. In general, Candle^{cluster} considerably outperforms all other baselines on two of the three metrics (*plausibility* and *distinctiveness*). Our resource only comes behind AcharyaEtAl on the *commonality* metric (1.15 and 1.22 respectively), which is expected because AcharyaEtAl only covers a few relations about common rituals (e.g., birthday, wedding) in two countries, USA and India, and their assertions are naturally known by many workers on Amazon MTurk, who are mostly from these two countries (Ross et al. 2010). Importantly, AcharyaEtAl is based on crowdsourcing and only contains a small set of 225 assertions for a few rituals.

Candle^{cluster} even outperforms the manually-constructed KG (AcharyaEtAl) on the *plausibility* metric. This could be caused by an annotation task design that was geared towards abnormalities, or lack of annotation quality assurance.

Candle also has the highest scores on the *distinctiveness* metric, while most of the assertions in other resources were marked as not distinguishing by the annotators.

Between the two versions of Candle, the cluster representatives consistently outperform the base sentences on all evaluated metrics. This indicates that still some of the raw sentences in the collection are noisy, on the other hand, the computed cluster representatives are more coherent and generally of better quality.

We also measured the *offensiveness* of each resource, i.e., the percentage of assertions that were marked as inappropriate or offensive materials by at least one of the human-annotators. Quasimodo and StereoKG, extracted from raw social media contents, have the highest number of assertions considered offensive (18% and 13%). Meanwhile, Candle's judicious filters only miss a small fraction (1% of final assertions).

In summary, our CANDLE CCSK collection has the highest quality by a large margin compared to other resources. Our resource provides assertions of high plausibility and distinctiveness. The clustering and cluster summarization steps also help to improve the presentation quality of the CCSK.

4.5.2 Comparison with LLM Generations

Knowledge extraction directly from pre-trained LMs is recently popular, e.g., the LAMA probe (Petroni et al. 2019) or ATOMIC-10x (West et al. 2022). There are major pragmatic challenges to this approach, in particular, that assertions cannot be contextualized with truly observed surrounding sentences, and that errors cannot be traced back to specific sources. Nonetheless, it is intrinsically interesting to compare assertion quality between extractive and generative approaches. In this section, we compare CANDLE with assertions generated by GPT-3 (Brown et al. 2020).

Assertion Generation. We query the GPT-3 model *davinci-002* with the following prompt template: *"Please write 20 short sentences about notable <facet> in <subject>."* We run each prompt ten times and set the randomness (temperature) to 0.7, so as to obtain a larger resource. We run the query for five facets and 210 subjects (196 countries and 14 religions), resulting in 188,061 unique sentences. Henceforth we call this dataset GPT3RES (GPT-3 resource), and reuse it in the extrinsic use cases (Section 4.6).

Evaluation Setup. For each resource, we sample 100 assertions for each of the five facets (hence, 500 assertions in total) and perform human evaluation on the three metrics: commonality, plausibility, and distinctiveness.

Results. The quality comparison between assertions of CANDLE and GPT3RES is shown in Table 4.6. While plausibility scores are the same, and CANDLE performs better in commonality, the difference that stands out is in distinctiveness: GPT3RES performs significantly worse, reconfirming a known problem of language models, evasiveness and over-generality

Table 4.6: Quality comparison between CANDLE and GPT3RES.

Resource	Quality [0..2]			Offensiveness	Length
	Plausibility	Commonality	Distinctiveness	(%)	(#chars)
GPT3RES	1.26	0.80	0.73	1	81
CANDLE	1.25	**0.89**	**0.89**	1	75

Evaluated on assertions of 196 countries.

Table 4.7: Example assertions of Candle and GPT3Res.

#	Candle	GPT3Res
1	The bride usually wears red in a traditional Chinese wedding.	Chinese people also like to wear modern clothes such as jeans and t-shirts.
2	The Chinese wear white at funerals because it is associated with mourning in Chinese culture.	Shoes are also very important in Chinese culture.
3	The Chinese wear new clothes for the New Year to symbolize new beginnings.	Chinese people also like to dress their children in very cute clothes.
4	The costumes in Chinese opera are very colorful and important.	In China, you will often see little girls wearing dresses and boys wearing shorts.
5	In ancient China, only the emperor was allowed to wear the color yellow.	In the winter, people in China wear coats and scarves to keep warm.

For `SUBJECT:China` and `FACET:clothing`.

(Li et al. 2016). We illustrate this with anecdotal evidence in Table 4.7, for `SUBJECT:China` and `FACET:clothing`. None of the listed GPT-3 examples is specific for China.

4.5.3 Per-Domain and Per-Facet Quality Evaluation

We break down the Candle CCSK collection into domains and facets and evaluate the assertion quality for each of these subcollections and get more insights into the constructed data.

Per-Domain Quality. Candle contains three subject domains, namely geography, religion and occupation. For each domain, we sample 100 assertions and perform crowdsourcing evaluation with the three metrics: plausibility, commonality and distinctiveness. We present the evaluation results in Table 4.8. Besides the raw scores (0, 1, 2), we binarize and denote them as acceptance rates, i.e., a score greater than zero means "accepted".

Candle achieves a high plausibility score of 1.54 on average. Performance on this metric is relatively consistent across all domains. Meanwhile, the commonality metric is highest for the occupation domain and lowest for geography domain.

More than 80% of plausible assertions are annotated as distinctive. Religion and occupation assertions perform significantly better than geography's on this metric. That could be caused by several assertions for geography subjects being correct but too generic (e.g., "Japanese food is enjoyed by many people", or "German beer is good"). In fact, religions and occupations are more distinguishing from one another, while countries or geo-regions usually have cultural overlaps.

Table 4.8: Quality of CANDLE assertions per domain.

Domain	**Quality [0..2]**			**Acceptance rate (%)**		
	PLA	**COM**	**DIS**	**PLA $\geq$ 1**	**COM $\geq$ 1**	**DIS $\geq$ 1**
Geography	1.52	1.19	1.03	84.00	66.00	61.33
Religion	1.51	1.29	1.22	85.76	74.67	72.00
Occupation	1.59	1.50	1.25	86.67	82.67	73.67
Average	1.54	1.33	1.17	85.44	74.44	69.00

Metrics: ***PLA*** *- plausibility,* ***COM*** *- commonality,* ***DIS*** *- distinctiveness.*

Table 4.9: Quality of CANDLE assertions per facet.

Cultural facet	**Quality [0..2]**			
	Facet relevance	**Plausibility**	**Commonality**	**Distinctiveness**
Food	1.42	1.23	0.94	0.97
Drinks	1.51	1.40	1.14	1.19
Clothing	1.49	1.30	1.04	1.07
Rituals	1.45	1.27	1.06	1.20
Traditions	1.42	1.27	1.02	1.11
Average	1.46	1.29	1.04	1.11

Evaluated on assertions of domain geography>country.

Per-Facet Quality. We select the assertions for the domain country, and for each facet (food, drinks, clothing, traditions, rituals), we sample 100 assertions for crowdsourcing evaluation. Besides commonality, plausibility and distinctiveness, here we introduce one more evaluation metric, *facet relevance*, measuring if an assertion conveys information about the cultural facet of interest:

- Question: *Is the assertion about <facet> in <culture>?*
- Answer options:
  - *0: No* | No, not at all.
  - *1: Somewhat* | It is partially in the domain of interest.
  - *2: Yes* | Yes, I totally see it is.

Only when the facet-relevance score is greater than zero, the other metrics will be evaluated. We present the evaluation results in Table 4.9.

It can be seen that CANDLE maintains good quality on all evaluation metrics. Notably, scores for the facet relevance metric are consistently high for all facets, suggesting that the enhanced techniques for zero-shot classification work well on our data. Interestingly, the facet *drinks* outperforms all other facets on three of the four metrics (facet relevance, plausibility and commonality), especially for plausibility, its score is significantly higher than others. Assertions for *drinks* and *rituals* are also more distinctive than for other facets.

4.6 Extrinsic Evaluation

We present two extrinsic tasks concerning cultural knowledge: cultural question answering (Section 4.6.1), and GuessTheCountry game (Section 4.6.2).

4.6.1 Cultural Question Answering

Similar to the extrinsic evaluation in Ascent++ (cf. Section 3.6), we employ retrieval-augmented generation (RAG) (Guu et al. 2020, Lewis et al. 2020) to show that CCSK assertions from resources like Candle can help LLMs perform better in cultural question answering (QA) tasks.

Dataset. For *questions*, we collected cultural knowledge quizzes from multiple websites, which resulted in 500 multiple-choice questions, each with two to five answer options (only one of them is correct). An example question is: *"What is the appropriate color to wear at a Hindu funeral?"*, with four answer options: *"white"*, *"black"*, *"gold"*, *"blue"*; the correct answer is *"white"*.

Models and Settings. We use the GPT-3 model *davinci-002* as QA agent (with temperature=0 and max_length=16), and compare its performance in three settings: (1) when only the questions are given, and when questions and their related contexts retrieved from (2) Candle or (3) GPT3Res are given to the LLM.

For *context retrieval*, we use the the SentenceBert model *all-mpnet-base-v2*, and for each question, retrieve the top-2 assertions from Candle^{cluster} and GPT3Res based on cosine similarity of their computed embeddings. We provide example input and output of this task in Figure 4.4.

Results. We measure the precision of the answers and present the results in Table 4.10. It can be seen that with Candle context, the performance is consistently better than when no context is given on all facets of culture, and better than GPT3Res context on three out of four facets. This shows that GPT-3, despite its hundred billions of parameters, still struggles with question answering tasks that require socio-cultural knowledge, and external resources such as Candle can help to alleviate this problem.

Without context ↓

What is the traditionally appropriate color for mourners to wear to a Hindu funeral?
A. White
B. Black
C. Gold
D. Blue
Answer:

B. Black

With a Candle assertion as context ↓

Context: Hindus wear white clothing to indicate mourning, while Christians wear white to weddings.
What is the traditionally appropriate color for mourners to wear to a Hindu funeral?
A. White
B. Black
C. Gold
D. Blue
Answer:

A. White

Figure 4.4: GPT-3 generations in the cultural QA task, without and with CCSK.

Table 4.10: Precision (%) of cultural QA.

Cultural facet	#Questions	Augmented context		
		None	GPT3Res	Candle
Food & Drinks	88	92.05	**94.32**	93.18
Behaviors	125	60.80	57.60	**63.20**
Rituals	135	87.41	85.93	**92.59**
Traditions	152	72.37	69.74	**79.61**
All	500	77.00	75.40	**81.40**

4.6.2 GuessTheCountry

The rule of GuessTheCountry game is as follows: Given five CCSK assertions about a country, a player has to guess the name of that country.

Dataset. As *input*, we select a random set of 100 countries, and take assertions from either Candle or GPT3Res. The game has five rounds, each is associated with a facet of culture.

CULTURAL-GROUP:Vietnam. **CULTURAL-FACET:drinks**. Assertions from GPT3Res ↓

Given the following sentences, guess the name of the hidden country?
- Drinking culture in […] is often seen as a way to relax and unwind.
- Drinking culture in […] is often considered to be very healthy, as many of the traditional drinks are made with natural ingredients.
- Drinking culture in […] is often considered to be very refreshing, as many of the traditional drinks are made with fresh ingredients.
- There are many bars and nightclubs in […].
- Beer is the most popular type of alcohol in […].

The correct answer is: Germany

CULTURAL-GROUP:Vietnam. **CULTURAL-FACET:drinks**. Assertions from Candle ↓

Given the following sentences, guess the name of the hidden country?
- […] iced coffee is a delicious, refreshing drink that is perfect for hot summer days.
- […] has a strong coffee culture, with coffee being a very popular drink among locals.
- Snake wine is a popular drink in […] that is made with rice wine and a snake.
- The […] like to drink beer with ice cubes.
- […] cuisine uses lime juice in many dishes, as well as a pickled lime called chanh muối.

The correct answer is: Vietnam

Figure 4.5: GPT-3 generations for GuessTheCountry game.

In each round, for each country, we draw the top-5 assertions from each resource (sorted by interestingness in Candle or by frequency in GPT3Res). All mentions of the countries in the input sentences are replaced with "[...]", before being revealed to the player.

Method. This is a game that requires a player who possesses a wide range of knowledge across many cultures. Instead of human players, we choose the GPT-3 model *davinci-002* as our player, which has been shown to be excellent at many QA tasks (Brown et al. 2020). We set temperature=0 and max_length=8. An example of input and output is shown in Figure 4.5.

Results. We measure the precision of the answers and present the results in Table 4.11. It can be seen that the player got significantly more correct answers when given assertions from Candle than from GPT3Res (i.e., assertions previously written by the player itself!). This confirms that assertions in Candle are more informative than GPT3Res assertions.

Table 4.11: Precision (%) of GuessTheCountry game.

Resource	Cultural facet					Average
	Food	**Drinks**	**Clothing**	**Rituals**	**Traditions**	
GPT3Res	63.0	30.0	44.0	70.0	**84.0**	58.2
Candle	**85.0**	**74.0**	**62.0**	**76.0**	80.0	**75.4**

4.7 Summary

In this chapter, we acquired refined commonsense knowledge for the second type of entry points: cultures. We presented Candle, an end-to-end methodology for automatically collecting cultural commonsense knowledge (CCSK) from broad web contents at scale. We executed Candle on several cultural groups and facets and produced CCSK of high quality. Our experiments showed the superiority of the resulting CCSK collection over existing resources, which have limited coverage for this kind of knowledge, and also over methods based on prompting LLMs. This work expands CSKB construction into a domain that has been largely ignored so far. Code, data, and a web interface are accessible at https://candle.mpi-inf.mpg.de.

5

COMBINING CONCEPTS AND CULTURES

In this chapter, we acquire commonsense knowledge for *both* types of entry points: concepts and cultures.

Our proposed method, called MANGO, distills high-accuracy, high-recall assertions of *culture-specific knowledge* from LLMs. Our LLM prompts are constructed judiciously and iteratively from the two entry points, covering a wide range of cultural subjects and concepts. Outputs are consolidated via clustering and generative summarization. Running the MANGO method with GPT-3.5 as underlying LLM yields 167K high-accuracy assertions for 30K concepts and 11K cultures, surpassing prior resources, including the CANDLE collection, by a large margin in quality and size. In an extrinsic evaluation for intercultural dialogues, we explore augmenting dialogue systems with cultural knowledge assertions. Notably, despite LLMs inherently possessing cultural knowledge, we find that adding knowledge from MANGO improves the overall quality, specificity, and cultural sensitivity of dialogue responses, as judged by human annotators. Interestingly, GPT-3.5, the LLM that generates MANGO assertions, also benefits from this resource in the extrinsic use case.

The project website is hosted at https://mango.mpi-inf.mpg.de, including downloadable code and data.

5.1 Introduction

Motivation and Research Questions. In the Candle project (Chapter 4), we investigated the acquisition of *cultural commonsense knowledge* (CCSK) from large-scale web contents. Candle extracted knowledge for a given set of cultural groups. The resulting knowledge base comprises ca. 60K assertions automatically distilled from nearly 1M sentences mined from the C4 crawl. An important limitation of Candle is that it only considered a handful of general cultural facets, hence potentially missed many culturally relevant concepts. Moreover, the resource is still insufficient in coverage of cultural groups, as cultures are not just geo-regions, but should ideally consider also demographic and social traits of the respective groups (e.g., female teenagers in Korea, French people of North-African descent, etc.).

The work presented in this chapter aims to expand the *coverage* of cultural groups and culture-specific assertions, while maintaining or even improving the *quality* of the assertions. This poses four research questions:

- *RQ1:* How can we substantially enlarge the amount of captured assertions, going beyond single-sentence retrieval from the web?
- *RQ2:* How can we systematically capture more and diverse cultural groups?
- *RQ3:* How can we ensure the specificity of assertions?
- *RQ4:* How can we avoid stereotypes and limit redundancy due to frequency bias?

Approach. The methodology devised in this work, called Mango (**m**ulti-cultur**a**l commo**n**sense knowled**g**e distillati**o**n), addresses these research questions as follows.

- To enlarge the pool of candidates (*RQ1*), we leverage an LLM, specifically GPT-3.5 (Ouyang et al. 2022), to generate assertions. Unlike Ascent++ and Candle, which tapped into web crawls, this is an implicit way of tapping the LLM's pre-training collection. Compared to web retrieval, the scales are similar but the advantage is that LLM training data involves efforts to remove spam and offensive content. Moreover, with properly designed prompts, LLMs can inherently incorporate information from different parts in their training data into coherent sentences, meanwhile single sentences extracted directly from web texts often need surrounding contexts to be meaningful.
- For high coverage of diverse cultures (*RQ2*), an important novelty in this method is to construct prompts for both concepts (incl. human activities) from an existing (culturally agnostic) CSKB and cultural groups from a large pool of cultures.
- Asking an LLM to generate assertions requires judicious prompting. As we have seen previously in the comparison between Candle assertions and those generated by

GPT-3 (Section 4.5.2), using a simple prompt template (*"Please write 20 short sentences about notable <facet> in <subject>."*) was insufficient and often led to evasive and overly generic answers. To ensure the informativeness of assertions (*RQ3*), our prompts contain example assertions and detailed instructions to direct the LLM to generate salient culture-specific knowledge.

- To tame the redundancy of generated candidates (*RQ4*), our method consists of steps for assertion consolidation. This is carried out by clustering the pool of assertions into topically and culturally coherent groups, considering both the key concept in an assertion and the culture to which it refers. Similarly to CANDLE, the LLM is leveraged again to generate a concise summary statement for each cluster. This exploits the LLM's language skills, but does not rely on actual knowledge by the LLM.

Contributions. This project's key contributions are:

1. *Methodology* (Section 5.3): We propose the MANGO methodology for efficiently distilling CCSK from LLMs, at high precision and recall, from two entry points: concepts and cultures.
2. *Resource* (Section 5.4): We construct and publicly release a CCSK collection of 167K assertions for 30K concepts and 11K cultures by running the MANGO method with GPT-3.5 as underlying LLM.

The evaluation with human judgements shows that the MANGO CCSK collection substantially surpasses prior CCSK resources in size and quality (Section 5.5), including CANDLE. An extrinsic evaluation for intercultural dialogues shows that the injection of MANGO assertions significantly improves the specificity and cultural sensitivity of LLM responses (Section 5.6). Code and data can be accessed at https://mango.mpi-inf.mpg.de.

5.2 Knowledge Representation

Given a concept c and a culture/cultural group g, our method generates statements that represent cultural beliefs, norms, or common practices around the concept c that apply to the cultural group g. Following CANDLE (cf. Section 4.2), we opt for using concise natural-language sentences to represent CCSK. However, instead of using general cultural facets (e.g., food, drinks, rituals, etc.), in this project, we use more specific concepts (e.g., `tipping`, `drinking beer`, `meditation`, etc.) in order to acquire more informative assertions.

For example, we collect CCSK assertions such as:

- `CULTURAL-GROUP:Japanese, CONCEPT:tipping, Tipping is not customary in Japan`, or
- `CULTURAL-GROUP:Buddhist, CONCEPT:meditation, Meditation is widely practiced for spiritual and mental well-being in Buddhist culture`.

5.3 Methodology

We propose Mango, a workflow for distilling and consolidating CCSK using LLMs. The goal of Mango is to generate CCSK for diverse cultures, covering a wide variety of concepts. Moreover, by clustering assertions, we obtain ranking signals for the assertions, which can be useful for downstream applications.

Our workflow consists of two phases, each consisting of two steps:

- *Phase 1: Assertion Generation* (Section 5.3.1)
  - Step 1a: Generating CCSK for a given concept.
  - Step 1b: Generating CCSK involving a given culture.
- *Phase 2: Assertion Consolidation* (Section 5.3.2)
  - Step 2a: Clustering CCSK assertions.
  - Step 2b: Generating cluster representatives.

An overview of the Mango workflow is depicted in Figure 5.1 alongside its examples.

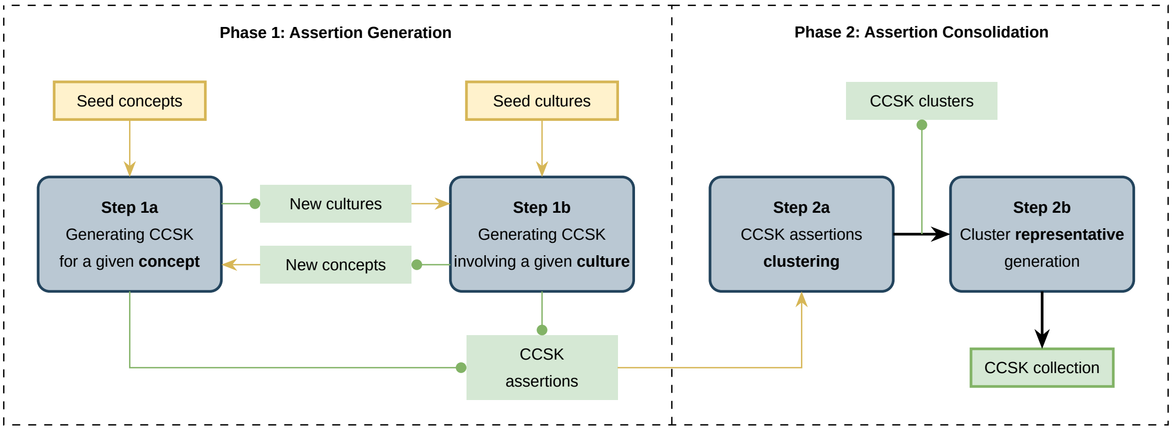
5.3.1 Phase 1: Assertion Generation

People of different cultures may have different perspectives on certain concepts. We are interested in collecting these cultural differences, as they are crucial in situations where cross-cultural knowledge is required to understand one another, as opposed to "universal CSK" captured by prominent CSK projects. Our previous project Candle could not capture these interesting differences directly since it only extracted single sentences from web data that contain a target culture and express a general cultural facet (e.g., food, drinks, rituals, etc.).

In the present Mango project, we address this issue by asking LLMs to generate pairs of CCSK assertions that represent different perspectives on the same concept in different cultures. To this end, one might ask LLMs to generate CCSK for a given pair of concept *and* culture (e.g., prompting for perceptions of `tipping` in `Japan`). However, the main drawback of this approach is that it requires concept and culture pairings, which is problematic because of the large number of possible combinations, and the fact that some combinations may not make sense. Instead, we propose to use LLMs to generate CCSK for a given concept or a given cultural group *separately*. This way we let the model decide which concepts are relevant for a given cultural group and vice versa, which will reduce the chance of nonsensical concept-culture combinations, hence reducing costs.

For each prompt, we provide the LLM with five pairs of example assertions randomly drawn from a set of human-written CCSK assertions. Each example consists of a concept, and two different perspectives on that concept associated with two or more different cultures.

Figure 5.1: The MANGO distillation workflow and its examples.



Step 1a: Generating CCSK for a given concept

Input concept: chopsticks
Prompt: You are a helpful assistant that writes culture-specific commonsense assertions. Some examples assertions are listed below:
* car | Important in US, Germany | Considered luxury item in poorer countries
* <4 more examples...>
Please write assertions for the concept: chopsticks.
Parsed output:

- Concept: *chopsticks*. Culture: *Japan*. Statement: *Standard eating utensils.*
- Concept: *chopsticks*. Culture: *Western countries*. Statement: *Considered exotic and less commonly used for everyday meals.*

Step 1b: Generating CCSK involving a given culture

Input culture: Japan
Prompt: You are a helpful assistant that writes culture-specific commonsense assertions. Some examples assertions are listed below:
* rice | Staple food in East Asia | Side dish in European countries
* <4 more examples...>
Please write assertions where one of the cultures is: Japan.
Parsed output:

- Concept: *tipping*. Culture: *Japan*. Statement: *Not a common practice.*
- Concept: *tipping*. Culture: *USA*. Statement: *Common and expected practice in the service industry.*

Step 2a: CCSK assertion clustering

Input: more than 500K CCSK assertions
Sample output cluster:

- Concept: *tipping*. Culture: *Japanese*. Statement: *Not a common practice.* (Frequency: 5)
- Concept: *leaving tip*. Culture: *Japanese culture*. Statement: *Not a common practice and may even be seen as rude.* (Frequency: 2)
- Concept: *tipping at restaurants*. Culture: *Japan*. Statement: *Tipping is not commonly practiced and can even be considered rude as it implies that the service is not already included in the price.* (Frequency: 1)
- Concept: *tipping service staff*. Culture: *Japan*. Statement: *Not a common practice and can even be considered rude or disrespectful.* (Frequency: 1)

Step 2b: Cluster representative generation

Prompt: Please generate a representative sentence for the following assertions: `<the sample cluster above>`
Output: Concept: *tipping*. Culture: *Japan*. Statement: *Tipping is not a common practice in Japan and can be considered rude or impolite.* (Frequency: 9)

We implement our workflow using the *gpt-3.5-turbo-1106* model. However, in practice, other LLMs could be used. We also run each prompt several times at a high temperature (i.e., high creativity), which can hopefully lead to different output assertions given the same input.

Step 1a: Generating CCSK for a Given Concept. Following instructions and example assertions in the prompt, we ask the LLM: *"Write culture-specific commonsense assertions for the concept: <concept>"*. For example, given the concept `chopsticks`, we expect the LLM to generate assertions like: "Chopsticks | Standard eating utensils in Japan | Considered exotic and less commonly used for everyday meals in Western countries".

We seed this step with everyday concepts from ConceptNet (Speer et al. 2017) and cultural concepts from Candle (Chapter 4).

Step 1b: Generating CCSK Involving a Given Culture. Following instructions and example assertions in the prompt, we ask the LLM: *"Write culture-specific commonsense assertions where one of the cultures is: <culture>"*. For example, given the culture `Japan`, we expect the LLM to generate assertions like: "Tipping at restaurants | Not a common practice in Japan | Common and expected practice in USA".

Note that the goal is not generating assertions only for the given culture, but we ask the LLM to generate pairs of assertions that express different perspectives in two or more cultures. This way we let the LLM come up with a distinctive cultural concept in the given cultre.

The seed cultures used for this step are taken from Candle, including geo-locations (countries, continents, geo-regions), and religions.

Iterative Generation. Step 1a and Step 1b are processed independently, and each step can generate new cultural groups and concepts, respectively. These new concepts and cultural groups can be fed back to the corresponding approach in the next iteration.

Output Format. For convenience, we ask the LLM to structure its responses into JSON objects with keys "concept", "culture", and "assertion". A sample of full prompts and outputs is presented in Figure 5.2.

5.3.2 Phase 2: Assertion Consolidation

It is natural that our CCSK generation approaches produce similar assertions that have very little or zero semantic differences. The repetitions can come from the fact that there exist concepts with close meanings (e.g., `tea`, `drinking tea`, and `tea drinking`), and terms referring to the same groups (e.g., `East Asia`, `East Asian cultures`, and `East Asian countries`) in the

<table>
<tr><th>Role</th><th>Message</th></tr>
<tr><td>System</td><td>You are a helpful assistant who writes culture-specific commonsense assertions. You always respond in JSON format with the following fields:
- “concept”: string
- “commonsense”: list of two objects representing <u>significantly different common sense</u> about the concept in different cultures:
  - “culture”: list of strings
  - “assertion”: string

Some examples of culture-specific commonsense assertions are listed below, in the format: * concept | commonsense assertion in culture 1 | commonsense assertion in culture 2.

* car | Important in US, Germany | Considered luxury item in poorer countries
* pig | Important farm animal in Europe, China | Considered dirty/shunned in Middle East
* bread | Dark/full-grain in Germany | Fluffy/toast-bread in Indonesia
* chopsticks | Standard eating utensils in East Asia | An exception in Europe
* window | Used to keep heat inside in northern countries | Used to keep heat outside in tropical countries</td></tr>
<tr><td colspan="2">Example 1: Generating CCSK for a given <u>concept</u></td></tr>
<tr><td>User</td><td>Write culture-specific commonsense assertions for the concept: feeding dogs.</td></tr>
<tr><td>Assistant</td><td><pre>{
  "concept": "feeding dogs",
  "commonsense": [
    {
      "culture": ["United States", "Europe"],
      "assertion": "Feeding dogs high-quality, specialized pet food is common practice."
    },
    {
      "culture": ["India", "Nepal"],
      "assertion": "Feeding dogs leftovers or homemade food is a common and acceptable practice."
    }
  ]
}</pre></td></tr>
<tr><td colspan="2">Example 2: Generating CCSK involving a given <u>culture</u></td></tr>
<tr><td>User</td><td>Write culture-specific commonsense assertions where one of the cultures is: Vietnam.</td></tr>
<tr><td>Assistant</td><td><pre>{
  "concept": "motorbike",
  "commonsense": [
    {
      "culture": ["Vietnam"],
      "assertion": "Common mode of transportation for individuals and families"
    },
    {
      "culture": ["USA"],
      "assertion": "Often used as a recreational vehicle rather than a primary mode of transportation"
    }
  ]
}</pre></td></tr>
</table>

Figure 5.2: Examples of prompting GPT-3.5 to generate CCSK assertions.

input. Furthermore, as we run each prompt several times, even at high output randomness (temperature), repetitions are inevitable.

While untreated redundancy is undesirable, we propose to take advantage of these redundancies in order to rank our assertions. Specifically, by grouping together assertions of identical meaning, we create frequency signals, which can be useful for downstream applications that only look for a subset of highest-significance assertions.

Step 2a: Assertion Clustering. As we generate hundreds of thousands of assertions in Phase 1, it is prohibitively expensive to run a clustering algorithm on all assertions at once. It would even be infeasible for algorithms such as the hierarchical agglomerative clustering (HAC) algorithm to process such a large amount of data points. Instead, we propose a divide-and-conquer approach to clustering this large set of assertions. Our approach consists of three substeps:

1. *Clustering concepts:* The concepts in the assertions generated in Phase 1 are clustered into groups of semantically similar concepts (e.g., `tea`, `drinking tea`, and `tea drinking`).
2. *Clustering cultures:* We group together different expressions of the same cultural groups (e.g., `East Asia`, `East Asian cultures`, and `East Asian countries`).
3. *Clustering subsets of assertions:* For each pair of concept cluster and culture cluster, we only cluster the subset of assertions associated with any of the concepts and any of the cultures in those clusters.

As the sizes of the entire concept and culture sets and the corresponding assertion subsets are substantially smaller than that of the entire set of generated assertions, standard clustering algorithms can process them efficiently. In our experiment, we use SentenceBert embeddings (Reimers and Gurevych 2019) and the HAC algorithm for all three substeps. Nevertheless, in practice, other text embedding models and clustering algorithms could be used.

An example assertion cluster can be found in Figure 5.1 (row *Step 2a: CCSK assertion clustering*), in which the respective concept cluster includes `tipping`, `leaving tip`, `tipping at restaurants`, `tipping service staff`, and the culture cluster includes `Japan`, `Japanese`, `Japanese culture`.

Step 2b: Representative Generation. For each cluster of CCSK assertions, we generate a *full sentence* as the representative of the cluster. We also obtain the frequency of each cluster by adding up all the frequencies of its member assertions.

For the final output, we store a set of CCSK cluster instances, each containing the following information: *concept* (the representative concept of the CCSK cluster), *culture* (the representative culture of the CCSK cluster), *statement* (the representative sentence of the CCSK cluster), *similar statements* (the set of member assertions), and *frequency* (the frequency sum of the cluster).

5.4 Implementation

We implement our workflow using the GPT-3.5 model *gpt-3.5-turbo-1106*. In the assertion generation phase, we run each prompt five times with a temperature of 1.0 (high creativity). We summarize the input and output of our implementation in Table 5.1.

Table 5.1: Input and output of each step in our implementation of MANGO.

Step	Input	Output
1a (2 runs)	31,196 concepts	468,543 assertions
1b (2 runs)	10,796 cultures	115,220 assertions
2a	507,780 filtered assertions	167,396 assertion clusters
2b	167,396 assertion clusters	167,396 full CCSK sentences

5.4.1 Input Concepts and Cultures

We run each step in Phase 1 twice, whereas the first run is fed with seed concepts and seed cultures from prior datasets, and the second run is fed with new concepts and cultures generated in the first run.

- *Seed concepts:* We take 8,357 concepts from CANDLE, which have more than one assertion, and 16,480 concepts from ConceptNet (Speer et al. 2017), which have more than three assertions. We clean this set of concepts by filtering out incomprehensible phrases (which mostly come from CANDLE) using GPT-3.5, retaining a seed set of 19,940 concepts for the first run of Step 1a.
- *Seed cultures:* We take 286 groups (272 geo-locations and 14 religions) from CANDLE as seed cultures for the first run of Step 1b.

After the first run of Step 1a with the seed concepts, we obtain 10,510 new cultures. After the first run of Step 1b with the 286 seed cultures, and the second run with the new 10,510 cultures, we obtain 11,256 new concepts, which are used for the second run of Step 1a.

5.4.2 Assertion Filtering

After Phase 1, we obtain 581,563 distinct CCSK assertions. We perform various simple filters to clean these assertions:

- Removing too long or too short assertions: We restrict the length of assertion to be between 2 and 25 words.
- Removing assertions that have more than one sentence.
- Removing assertions whose cultures are too general or noisy: We construct a dictionary of words that a valid culture should not contain. For example, we filter out cultures such as “Other cultures” (context-dependent), “Non-European countries”, “Some parts of Asia” (not specific enough), etc. More specifically, if the culture of an assertion contains any of the following punctuations, words and phrases, it will be rejected: *other*, *general*,

Table 5.2: Popular concepts and cultures in the Mango collection.

Concept	**#Assertions**	**Culture**	**#Assertions**
family	372	*Countries*	
tea	352	United States	9,097
hospitality	349	Japan	4,597
personal space	338	India	3,670
marriage	285	*Regions*	
family structure	234	Western countries	5,788
education	228	Eastern countries	1,849
gender roles	200	Middle Eastern countries	1,059
public transportation	193	*Religions*	
music	183	Christianity	142
traditional clothing	180	Islamic countries	121
time	174	Hinduism	106
food	149	*Ethnic groups*	
spicy food	148	Amazonian tribes	143
communication	147	Inuit communities	101
fishing	132	Maori culture	73

1, *2*, *(*, *)*, *and*, *,*, *some*, *unknown*, *parts of*, *few*, */*, *non-*, *many*, *outside*, *part of*, *various*, *elsewhere*, *rest of*, *certain*.

After filtering, we obtain 507,780 assertions (i.e., 87% of the original set), 14,298 cultural groups, and 32,126 concepts.

5.4.3 Clustering

For both concept, culture and assertion clustering, we use the HAC algorithm combined with the SentenceBert embedding model *all-MiniLM-L6-v2*. We use point-wise Euclidean distance of normalized embeddings, the Ward's linkage (Ward 1963), and the distance threshold set to 1.5, which are adopted from the Candle method.

We obtain 4,571 concept clusters and 1,610 culture clusters, with the largest candidate set for assertion clustering containing 463 assertions. Finally, we obtain 167,396 assertion clusters, where the largest cluster contains 141 assertions. Table 5.2 lists the top cultures and concepts in our dataset.

5.4.4 Costs

We ran the Mango method with the *gpt-3.5-turbo-1106* model in January 2024. The assertion generation steps costed $71. On average, each thousand generated assertions costed 12 cents. The cluster representative generation costed $41. On average, each thousand cluster representative generations costed 25 cents.

The LLM's API had limits at 1M input tokens per minute, and 10K requests per minute. By making concurrent requests, our workflow could be executed in less than five hours, including the clustering steps which took less than 30 minutes.

Experiment Overview

We perform intrinsic evaluation with human judgements and show that the Mango CCSK collection has higher quality than Candle and other resources (Section 5.5). Our extrinsic use case of intercultural dialogues shows that injecting Mango assertions into LLM prompts improves their responses on specificity and cultural sensitivity (Section 5.6).

5.5 Intrinsic Evaluation

5.5.1 Setup

We evaluate the assertion quality of Mango by comparison to several resources of similar kind via human annotations.

Compared Resources. Our baseline resources are Quasimodo (Romero et al. 2019), StereoKG (Deshpande et al. 2022) and Candle (Chapter 4). For this comparison, we only consider Quasimodo assertions that contain geo-locations or religions (those present in Candle) in their subjects or objects. We did not include the manually-built resource by Acharya et al. (2021) in this evaluation due to cost limits (see Section 4.5 for a comparison of this resource and others).

For Mango, in addition to the full assertion set (Mango^{full}), we also evaluate Mango^{top}, which contains the top-ranked assertions up to the same size as Candle (i.e., 60K assertions), for compatibility.

Metrics. We employ three complementary dimensions (Bhatia and Shwartz 2023) to evaluate the intrinsic quality of CCSK resources, which extend the metrics used in Section 4.5 (i.e., plausibility, distinctiveness, offensiveness).

- *Cultural relevance* measures the assertion's relevance to the cultural context, and its plausibility in that cultural context.
- *Stereotype avoidance* assesses if the assertion avoids reinforcing cultural stereotypes and presenting offensive materials.
- *Linguistic accuracy* checks if the assertion is grammatical, meaningful and understandable as a standalone sentence.

We also ask annotators to specify how familiar they are with the cultures mentioned in the assertions (akin to *commonality* in Section 4.5). The three options for this question are "very familiar", "somewhat familiar", and "not familiar". In the latter case, our guidelines advise the annotators to search and inspect additional sources on the Internet.

Evaluation Scheme. Our evaluation is conducted on the Amazon MTurk platform. Specifically, we ask human annotators on MTurk to evaluate each assertion along the three dimensions, each on a 3-point Likert scale: *negation* (0), *ambiguity* (1) and *affirmation* (2). We present the MTurk questions and answer options in Table 5.3. For each resource, we randomly sample 300 assertions for evaluation, each assessed by three different workers.

Details of MTurk Tasks. We estimated each HIT would take 1.5 minutes to complete, hence an hourly wage of $12. The annotators were required to be Master Workers with a lifetime approval rate greater than 99% (this rate is provided by the platform). There were totally 33 different workers who participated in this task. We observed the following Krippendorff's alpha values for each of the metrics: cultural relevance: 0.33, stereotype avoidance: 0.29, linguistic accuracy: 0.39.

In these MTurk tasks, 41.8% of the annotators reported being "very familiar" with the cultures presented in the tasks, 37.4% being "somewhat familiar", and 20.8% being "not familiar".

5.5.2 Results

We report average scores given by the annotators in Table 5.4. It can be seen that Mango assertions have the highest quality compared to prior resources. Notably, despite being nearly three times larger than the largest baseline, Mango^{full} has better quality than all baselines, on all metrics.

Table 5.3: Crowdsourcing questions for CCSK evaluation in MANGO.

Dimension	**Question and answer options**
Linguistic accuracy	*Is the assertion meaningful and understandable without further context? Is it grammatical? Are the vocabulary and idiomatic expressions appropriate in the mentioned socio-cultural group?*
	0 - Never | No, not at all. *1 - Somewhat* | I think I understand what it means but not fully sure. *2 - Yes* | Yes, absolutely.
Cultural relevance	*Is the assertion factually accurate? Does it reflect the values, customs, traditions, and societal norms associated with the mentioned socio-cultural group?*
	0 - No | No, not at all. *1 - Somewhat* | Somewhat. *2 - Yes* | Yes, absolutely.
Stereotype avoidance	*Is the assertion neutral? Does it avoid perpetuating stereotypes about the mentioned socio-cultural group?*
	0 - No | No, it is offensive, prejudiced, or inappropriate. *1 - Somewhat* | Somewhat, it is neutral but may perpetuate stereotypes. *2 - Yes* | Yes, it is neutral and does not perpetuate stereotypes.

Table 5.4: Results of *intrinsic evaluation* via human annotations.

Resource	**Quality [0..2]**			**Size**	
	Cultural relevance	**Stereotype avoidance**	**Linguistic accuracy**	**#Cultures**	**#Assertions**
Extractive					
StereoKG	0.79	0.85	1.11	10	4K
Quasimodo	0.85	1.22	1.12	0.4K	131K
CANDLE	1.42	1.54	1.67	0.4K	60K
Generative					
MANGO^{full}	1.53*	1.61*	1.79*	**11.1K**	**167K**
MANGO^{top}	**1.59***	**1.65***	**1.82***	8.1K	60K

(*) indicates statistically significant gains over CANDLE *(*$p < 0.05$ *in the Student's t-test).*

Moreover, the top-ranked assertion subset, MANGO^{top}, has the same size as CANDLE, but it outperforms this resource on all three metrics by a large margin. This affirms that

<table>
<tr><th colspan="2">Input</th></tr>
<tr><th colspan="2">Narrative</th></tr>
<tr><td colspan="2">John, an American, is visiting his friend Kenji, who lives in Tokyo. They are paying their bill for dinner at a restaurant.</td></tr>
<tr><th colspan="2">Ongoing dialogue</th></tr>
<tr><td colspan="2">John: That’s a great meal, Kenji. I really liked the sushi.
Kenji: My pleasure, John. I’m glad you enjoyed it.
John: Let me see the bill. It is 8,000 yen. I’m gonna leave 10,000 yen.</td></tr>
<tr><th colspan="2">Methods</th></tr>
<tr><th>Method 1: Vanilla GPT-3.5</th><th>Method 2: GPT-3.5 with explicit CCSK</th></tr>
<tr><td>-</td><td>Injected CCSK: Tipping is not a common practice in Japan and can be considered rude or impolite.</td></tr>
<tr><th colspan="2">Output</th></tr>
<tr><td>Kenji: Thank you, John. You’re too kind. Next time, dinner is on me. It’s a very generous tip too.</td><td>Kenji: Oh, no, John. You don’t need to leave a tip here in Japan. Just 8,000 yen is fine. Thank you for offering though.</td></tr>
</table>

Figure 5.3: Example input and output for the next utterance generation task.

the frequency signals we obtain in the clustering step are helpful for pulling out the most significant assertions.

5.6 Extrinsic Evaluation

Explicit knowledge has been used to improve LLM performance in downstream tasks such as social dialogue generation (Kim et al. 2023) and intra-cultural dialogue synthesis (Li et al. 2023). However, other important scenarios, which involve *intercultural* interactions, remain under-explored. To address this gap, we introduce two new tasks concerning intercultural conversations. We will show that explicit CCSK assertions from resources like Mango can improve LLM responses in dialogue applications requiring cross-cultural knowledge.

5.6.1 Next Utterance Generation

The input of this task includes: (1) a narrative that involves interactions between two people of different cultures, and (2) their ongoing dialogue. The task is to generate the utterance of the next speaker. To excel in this task, LLMs need to leverage cultural knowledge such that the generated responses are natural, consistent, culturally sensitive and specific. We provide an example of input and output for this task in Figure 5.3.

Input Construction. First, using GPT-3.5 and a small set of in-context examples, we generate short narratives involving two people of different cultures, e.g., *“Liz from England is*

Role	Message
User	You are a narrative generator. Your task is to generate short narratives of less than 5 sentences around a cultural concept that involves two people from two different cultures. <u>The narrative should lead to an intercultural interaction where cultural differences play a significant role.</u> You must not include the cultural differences, or cultural knowledge, or the resolution, or the consequences of the situation in the narrative. Some examples: - Anna, an American, is visiting a remote village in Vietnam where Minh is a local. Anna asks Minh where she can get food for her dog. - Erling from Norway is visiting Seoul. He and his new Korean friend, Heungmin, are picking foods for their dinner at a traditional restaurant. - Liz and Qiang are two friends, who are currently in England. Qiang is from China who is visiting the country. Liz is a local. They are preparing tea together. Please write 3 more narratives:
Assistant	- Maria, a Mexican, is visiting Japan and is shopping in a local market. She meets Yuki, a Japanese woman, and asks for help in choosing a traditional Japanese outfit for a festival. - Pablo, a Spaniard, is traveling in India and meets Rajesh, a local, at a temple. They both want to participate in a religious ceremony, and Pablo asks Rajesh for guidance on the proper etiquette. - Fatima, a Saudi Arabian, is studying in France and meets Pierre, a French student, at a party. They both want to dance, and Fatima asks Pierre to teach her a traditional French dance.

Figure 5.4: An example prompt used to generate intercultural narratives.

visiting her friend Qiang in China. They are preparing tea together." (see in Figure 5.4 for a sample prompt).

For each narrative, we ask the LLM to generate a possible dialogue between the two participants in that situation. The first three turns of each generated dialogue are used to construct test samples.

Methods. We compare two methods, without and with explicit CCSK assertions in the prompts (see Figure 5.3).

1. *Method 1 (Vanilla LLM)* prompts an LLM with the task description, the input narrative, and the dialogue history.
2. *Method 2 (LLM with explicit CCSK)* alters the prompts of *Method 1* by augmenting relevant CCSK, following the idea of retrieval-augmented generation, a.k.a. RAG (Guu et al. 2020, Lewis et al. 2020).

For *Method 2*, we use a simple dense embeddings retrieval approach. First, we compute the embeddings of all assertions in our dataset using the SentenceBert model *all-MiniLM-L6-v2*.

Retrieval example 1	
Narrative	Carlos from Argentina is visiting Korea. He greets his new Korean friend, Jihoon, by giving him a friendly pat on the back.
Query	X from Argentina is visiting Korea. He greets his new Korean friend, Y, by giving him a friendly pat on the back.
Results	(1) In South Korea, beckoning with an open hand or palm facing downwards is considered polite. *(Similarity: 0.5216)* (2) In South Korean culture, it is common to gently pat someone on the shoulder or back as a sign of encouragement or reassurance. *(Similarity: 0.5201)*
Retrieval example 2	
Narrative	Maria, a woman from Spain, is visiting a Bedouin tribe in Jordan upon an invitation from her new friend, Ahmed. They are preparing to have dinner under the star-lit desert sky.
Query	X, a woman from Spain, is visiting a Bedouin tribe in Jordan upon an invitation from her new friend, Y. They are preparing to have dinner under the star-lit desert sky.
Results	(1) Bedouins, a Middle Eastern culture, are known for their nomadic lifestyle, hospitality, and expertise in desert survival. *(Similarity: 0.5217)* (2) Desert cultures highly value hospitality and express it through offering food, drinks, and shelter to guests. *(Similarity: 0.5207)*

Figure 5.5: Examples of retrieving Mango's assertions relevant to given narratives.

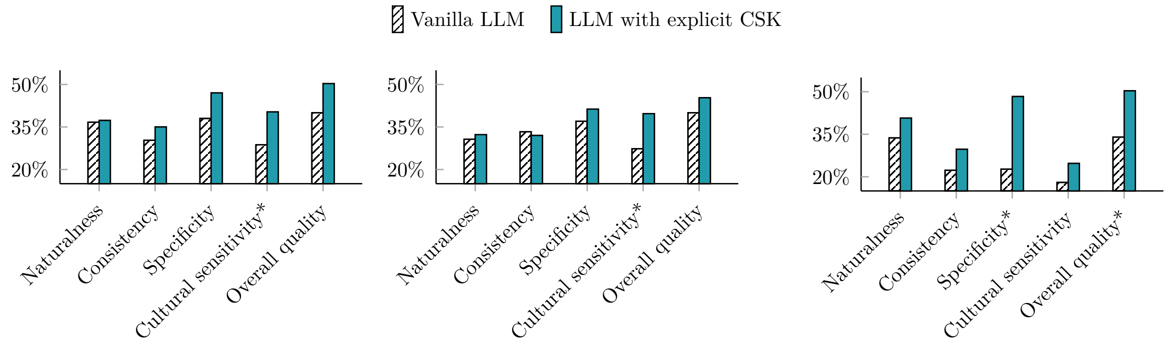
Then, given a narrative, we replace the names of the people with "X" and "Y" in order to reduce the distractions for the embedding model. For each given narrative, we retrieve the top-2 most similar CCSK assertions from based on cosine similarity of the embeddings. We also restrict to assertions with similarity scores better than 0.5. We only consider narratives that have relevant CCSK assertions retrieved from in our evaluation. Examples of CCSK retrieval can be seen in Figure 5.5.

Models. We experiment with three different LLMs from some of the most popular model families: (i) GPT-3.5[3], also known as ChatGPT, by OpenAI, (ii) the strongest model of the Mistral family by MistralAI[4], and (iii) Google's largest globally accessible language model as of January 2024, Bard[5].

Evaluation Scheme and Metrics. For each LLM, we randomly draw 100 test samples and use the two methods to generate 100 pairs of responses. We employ human annotators on MTurk to evaluate the quality of the generated utterances. Specifically, we show side-by-side

[3]Model name: *gpt-3.5-turbo-1106*
[4]Model name: *mistral-medium* (https://docs.mistral.ai/platform/endpoints/)
[5]Accessed via https://bard.google.com/

The y-axis represents the percentage of samples preferred by the human evaluators (the "Tie" cases are omitted from the plots). The (*) marker indicates results with statistically significant differences ($p <$ 0.05 in the Wilcoxon T-test).



(a) GPT-3.5 (b) Mistral (c) Bard

Figure 5.6: Results of head-to-head comparison for next utterance generation.

the outputs of the two methods given the same input narrative, and ask the evaluators to choose the better one based on the following dimensions:

1. *Naturalness:* The response does not sound awkward or unnatural.
2. *Consistency:* The response does not contradict the narrative or previous utterances.
3. *Specificity:* The response contains specific details rather than vague/generic information.
4. *Cultural sensitivity:* The response shows respect and understanding of the cultures.
5. *Overall quality:* The overall satisfaction with the response.

Dimensions 1-3 and 5 are established criteria for evaluating dialogues (Mehri et al. 2022, Kim et al. 2023), dimension 4 is a dimension we introduce in this evaluation, as per our focus on cross-cultural knowledge.

Each sample is evaluated by three different MTurk workers. To alleviate positional bias, the orders of the outputs are randomly shuffled before being revealed to the annotators. The outputs are labeled Response A and Response B. For each evaluation metric, we provide the annotators with the following option set: "A is Better", "B is Better", and "Tie". The instruction and layout of our annotation page used for MTurk evaluations are influenced by the work of Kim et al. (2023).

Details of MTurk Tasks. For these HITs, we require the workers to be Master Workers who have a lifetime approval rate of more than 99%. There were 41 different MTurk workers who participated in this evaluation. Each HIT was compensated \$0.20. We estimated each HIT would take 1 minute to complete, hence an hourly wage of \$12. In these MTurk tasks, 40.9% of the annotators reported being "very familiar" with the presented cultures, 48.2% being "somewhat familiar", and 10.9% being "not familiar".

Table 5.5: Mango versus Candle in the next utterance generation task.

Metric	Candle	Mango
Naturalness	30.7%	**32.7%**
Consistency	28.0%	**30.7%**
Specificity	37.0%	**39.7%**
Cultural sensitivity	**34.3%**	**34.3%**
Overall quality	**46.0%**	**46.0%**

The numbers present the percentage of samples preferred by the human evaluators.

Results. The human evaluation results are demonstrated in Figure 5.6. It can be seen that explicit CCSK assertions help all three LLMs to perform significantly better in this task, as the *overall quality* of the generated utterances are preferred by the human annotators compared to the outputs of the vanilla LLMs. This is attributed by the fact that the CCSK-enhanced utterances are more *specific* and *culturally sensitive* (for all three LLMs), and even more *natural* and *consistent* (for Bard). This shows that even though LLMs inherently possess cultural knowledge, they often fail to incorporate such knowledge into the generated utterances, and that explicit CCSK injected in the prompts can mitigate this problem. We provide some evaluated samples in Figure 5.7.

Mango vs. Candle. Additionally, we also compared using Mango vs. Candle assertions for augmenting prompts to GPT-3.5. We sample 100 test narratives and use the same retrieval method for both resources in this comparison. The results are presented in Table 5.5. It can be seen that human annotators preferred utterances generated with Mango assertions injected in terms of naturalness, consistency and specificity.

5.6.2 Full Dialogue Generation

In this task, instead of generating a single utterance, we ask LLMs to generate a full dialogue between the two people in a given narrative. We anticipate that this task is more difficult both for the LLMs to solve and for humans to evaluate.

We randomly sample 100 test narratives and ask GPT-3.5 to generate dialogues. We use the same prompting methods and evaluation setup as in the previous task. The evaluation results can be found in Table 5.6. It can be seen that injecting explicit CCSK in the prompts does not diminish the natural flow and consistency of the generated dialogues, on the other hand, it provides more specificity and cultural sensitivity into the intercultural interactions.

<table>
<tr><th colspan="6">Sample 1</th></tr>
<tr><th colspan="6">Narrative</th></tr>
<tr><td colspan="6">Ali from Tunisia is visiting Tokyo for the first time. He and his newly made Japanese friend, Hiroshi, are deciding how to celebrate Ali's birthday in one of Tokyo's iconic districts.</td></tr>
<tr><th colspan="6">Ongoing dialogue</th></tr>
<tr><td colspan="6">Ali: Hiroshi, I'm so excited to celebrate my birthday in Tokyo! What do you think we should do?
Hiroshi: Happy birthday, Ali! I think we should start with a traditional Japanese meal at a local izakaya.
Ali: That sounds great! I've been wanting to try authentic Japanese cuisine. And then maybe we could check out the vibrant nightlife in Shibuya?</td></tr>
<tr><th colspan="3">Vanilla GPT</th><th colspan="3">GPT + CCSK</th></tr>
<tr><td colspan="3">-</td><td colspan="3">Injected CCSK:
• In Japan, birthdays are celebrated with symbolic rituals, special meals, and a focus on traditional ceremonies and gift-giving.</td></tr>
<tr><td colspan="3">Hiroshi: Absolutely, Ali! Shibuya is the perfect place to experience Tokyo's bustling nightlife. We could start with a few drinks at a traditional Japanese izakaya and then explore the neon-lit streets and vibrant energy of Shibuya. It's going to be an unforgettable birthday celebration!</td><td colspan="3">Hiroshi: That sounds like a fantastic plan, Ali! But I also think it would be wonderful to incorporate some Japanese birthday traditions into our celebration. How about we visit a shrine to make a prayer for good luck and then participate in the Omikuji fortune-telling ritual? It's a special way to mark your birthday and immerse yourself in Japanese culture. And of course, we can still enjoy the bustling nightlife in Shibuya afterwards.</td></tr>
<tr><td>Metric</td><td>Naturalness</td><td>Consistency</td><td>Specificity</td><td>Cult. sensitivity</td><td>Overall quality</td></tr>
<tr><td>Annotator 1</td><td>GPT + CCSK</td><td>Tie</td><td>GPT + CCSK</td><td>GPT + CCSK</td><td>GPT + CCSK</td></tr>
<tr><td>Annotator 2</td><td>Tie</td><td>Tie</td><td>GPT + CCSK</td><td>Tie</td><td>GPT + CCSK</td></tr>
<tr><td>Annotator 3</td><td>Vanilla GPT</td><td>GPT + CCSK</td><td>GPT + CCSK</td><td>GPT + CCSK</td><td>GPT + CCSK</td></tr>
<tr><th colspan="6">Sample 2</th></tr>
<tr><th colspan="6">Narrative</th></tr>
<tr><td colspan="6">Fiona, a woman from Scotland, is in Nairobi visiting her online friend Makena. Makena is taking her to a traditional Kenyan hair salon for the first time.</td></tr>
<tr><th colspan="6">Ongoing dialogue</th></tr>
<tr><td colspan="6">Makena: Welcome to the salon, Fiona! Are you excited to try out some Kenyan hair styles?
Fiona: Oh, I am a bit nervous, but also curious to see what kind of styles are popular here. It's quite different from what I'm used to in Scotland.
Makena: Don't worry, you're going to love it! Kenyan hair stylists are really talented and they have some unique techniques.</td></tr>
<tr><th colspan="3">Vanilla GPT</th><th colspan="3">GPT + CCSK</th></tr>
<tr><td colspan="3">-</td><td colspan="3">Injected CCSK:
• Hairdressing education in Kenya focuses on intricate braiding and natural hair care techniques for diverse African hair types and textures.
• In Kenya, elaborate and intricate braided hairstyles are a significant aspect of cultural identity.</td></tr>
<tr><td colspan="3">Fiona: That's good to hear! I can't wait to see what they can do. I've always admired African hairstyles from afar, so I'm really looking forward to experiencing it firsthand.</td><td colspan="3">Fiona: That sounds amazing! I've always admired the intricate braided hairstyles I've seen here. I've never had the chance to experience it myself, so I'm really looking forward to it.</td></tr>
<tr><td>Metric</td><td>Naturalness</td><td>Consistency</td><td>Specificity</td><td>Cult. sensitivity</td><td>Overall quality</td></tr>
<tr><td>Annotator 1</td><td>Tie</td><td>GPT + CCSK</td><td>GPT + CCSK</td><td>Tie</td><td>GPT + CCSK</td></tr>
<tr><td>Annotator 2</td><td>GPT + CCSK</td><td>GPT + CCSK</td><td>GPT + CCSK</td><td>GPT + CCSK</td><td>GPT + CCSK</td></tr>
<tr><td>Annotator 3</td><td>GPT + CCSK</td><td>GPT + CCSK</td><td>GPT + CCSK</td><td>GPT + CCSK</td><td>GPT + CCSK</td></tr>
</table>

Figure 5.7: Examples of generated responses and annotations of MTurk workers.

As a result, the annotators prefer the CCSK-enhanced dialogues on overall quality, similar to the results of the next utterance generation task.

Table 5.6: Results of the full dialogue generation task.

Metric	Without CCSK	With CCSK
Natural flow	**39.3%**	38.7%
Consistency	21.0%	**22.7%**
Specificity	36.7%	**43.0%**
Cultural sensitivity	22.7%	**28.3%**
Overall quality	40.7%	**44.7%**

The numbers present the percentage of samples preferred by the human evaluators.

In these MTurk tasks, 33.7% of the annotators reported being "very familiar" with the presented cultures, 54.0% being "somewhat familiar", and 12.3% being "not familiar".

5.7 Summary

In this chapter, we acquired refined commonsense knowledge for both types of entry points: concepts and cultures. We presented Mango, a methodology to distill and consolidate culture-specific commonsense knowledge from LLMs. We executed the Mango workflow using GPT-3.5, obtaining a collection of 167K assertions covering 11K cultures and 30K concepts, which surpasses existing resources, including the Candle collection, in quality and coverage by a large margin. In the extrinsic evaluation, we showed that by augmenting explicit assertions from Mango into prompts, LLMs can perform better in intercultural dialogue generation tasks as their responses were judged more specific, more culturally sensitive, and better overall quality by human evaluators. Code, data, and a web interface are accessible at https://mango.mpi-inf.mpg.de.

6 DISCUSSION

Commonsense knowledge (CSK) is crucial for human-centric AI. Most prior CSK acquisition methods have significant shortcomings of expressiveness, as they rely on subject-predicate-object (SPO) triples. In addition, prominent CSK projects focus only on universal CSK, overlooking culture-specific CSK conditioned on socio-cultural contexts. This dissertation has addressed these gaps by introducing expressive knowledge models and developing methodologies to automatically acquire refined CSK at high precision and wide coverage, via text extraction or LLM-based knowledge distillation. In Table 6.1, we provide a summary of lessons learned and open issues across the three projects contributing to this dissertation.

We detail the lessons learned in Section 6.1, followed by a discussion of the projects' shortcomings and future research opportunities in Section 6.2. We take an outlook for CSK acquisition and application amidst the rise of large language models in Section 6.3.

6.1 Lessons Learned

Major lessons learned in the course of this dissertation, which generalize over the individual projects, are:

1. It is feasible to achieve high precision and wide coverage with automated CSKB construction if given sufficient thoughts about knowledge organization, source selection, and method design (Section 6.1.1).
2. High-quality CSKBs are beneficial to various downstream applications (Section 6.1.2).
3. CSKBs using more expressive knowledge models outperform those using standard models intrinsically and extrinsically (Section 6.1.3).
4. Clustering is important for frequency signals, and dealing with the heterogeneity of natural language (Section 6.1.4).

Table 6.1: A summary of lessons learned and open issues.

	Lessons learned	**Open issues**
Knowledge representation	• Expressive knowledge models outperform standard models	• Combining factual and cultural knowledge into one model
Methodology	• Automated CSKB construction is feasible with high precision and wide coverage • LLM-based distillation produces assertions of higher quality and coverage than text extraction • Clustering is important for frequency signals and dealing with heterogeneity	• Source tracing and ranking for LLM-generated assertions • Expanding refined subjects and further cultural groups • Expanding sources to more diverse languages, cultures, and modalities • Detecting and mitigating biases and stereotypes
Extrinsic value	• High-quality CSKBs are beneficial to various downstream applications, via using RAG	• Developing better methods for combining non-parametric and parametric knowledge
Evaluation	• Evaluating the quality of CSK resources relies heavily on human annotations	• Developing more comprehensive test suites for coverage and quality evaluation

6.1.1 Advancing Automated CSKB Construction

Previous CSK projects such as (Tandon et al. 2014a), (Romero et al. 2019), and (Zhang et al. 2020a) extracted CSK automatically from web contents, which allowed for high coverage but also introduced a substantial amount of noise. Studies (e.g., (Hwang et al. 2021), (Nguyen et al. 2021a)) reported that these resources suffer from high noise.

In the Ascent++ and Candle projects, we showed that it is feasible to extract high-quality CSK from web sources. The key to success includes the judicious design of filtering, aggregation, and consolidation techniques in our pipelines. In terms of filtering, in Ascent++, we employed filters both on the document level and the assertion level to select high-quality candidates; meanwhile in Candle, we used rule-based generic assertion filtering and classification-based filtering for culture-specific assertions. In addition, crafted dictionaries were used by both methods in order to clean the extracted assertions. Finally, we used advanced semantic clustering techniques to aggregate and consolidate the extracted assertions, creating frequency signals. These signals were used to rank the assertions, enabling downstream applications to select the most significant subset of knowledge for their tasks. Evaluations also showed that the top-ranked assertions in our resources consistently outperformed other assertions, confirming the effectiveness of our ranking models.

6.1.2 Extrinsic Values of CSKBs

Research on augmenting LLMs with commonsense knowledge has been limited, while retrieval-augmented generation (RAG) approaches have shown mixed results for tasks requiring rich world knowledge (Mallen et al. 2023). RAG has been used in all of our projects for question answering (in Ascent++ and Candle), and dialogue generation (in Mango). We showed that adding relevant CSK assertions from our resources to the prompts of LLMs significantly improved their performance in these tasks. Moreover, while LLMs also benefited from assertions of other resources, pulling assertions from higher-quality resources like ours led to better performance. This is an interesting result as LLMs inherently posses such knowledge, but often fail to use it in downstream tasks. Nevertheless, our extrinsic evaluations were rather simple. More realistic use cases of commonsense knowledge should be explored in future research.

6.1.3 Values of Expressive Knowledge Models

Standard data models that rely on subject-predicate-object (SPO) triples have been widely used in CSK projects. Although that enables structured representations and graph-based reasoning, these models have major limitations in expressiveness. In Ascent++, we introduced an advanced knowledge model for commonsense, which captures more informative assertions with refined subjects and semantic facets for SPO triples. Our CSKB built on this model outperformed existing CSK collections both intrinsically and extrinsically, as our assertions are more informative and specific. In Candle and Mango, where we used natural-language sentences to capture CCSK, our resulting resources also outperformed existing resources such as Quasimodo (Romero et al. 2019) and StereoKG (Deshpande et al. 2022), which are based on SPO triples, in terms of both quality and coverage.

6.1.4 Importance of Clustering

Collecting CSK at scale requires dealing with the heterogeneity of natural language and redundancy in the extracted assertions. To this end, across all three projects, we used clustering techniques to identify and group assertions of identical meaning. Besides reducing redundancy, clustering also helped to identify frequency signals. Although frequency can sometimes be a misleading signal (e.g., sensational content is often repeated many times), we showed in the Ascent++ and Candle projects that it is still useful for ranking the assertions when combined with other signals (e.g., specificity, distinctiveness, relevance).

6.2 Limitations

The limitations of our work come from the knowledge representations, the sources of CSK, the acquisition methods, and the evaluation settings. In the following, we discuss the main shortcomings of our methods, and future research opportunities.

6.2.1 Knowledge Representations

Choosing the Right Model. In Ascent++, we introduced an expressive knowledge model for commonsense, capturing more informative assertions with refined subjects and semantic facets for SPO triples. While structured representations like this are useful for many applications such as graph-based reasoning, they often fail to capture the nuances and complexities of natural language, especially in the commonsense domain.

In Candle and Mango, we dropped the structured representation, and used natural-language sentences to capture CCSK. This approach is capable of capturing more nuanced knowledge, and it can be directly leveraged by LLMs in downstream applications. However, natural-language assertions introduce redundancy, with differences often limited to small but important details. For example, *"Phở is a popular traditional noodle soup dish in Vietnamese cuisine"* and *"Phở is a traditional Vietnamese noodle soup typically eaten for breakfast"* are two assertions which are very similar, having a large overlap in their content but not identical as the added detail in the second assertion is very valuable (typical eating time).

While each of these representations has its own strengths and weaknesses, choosing the right representation depends on the applications and the trade-offs between expressiveness and efficiency, source faithfulness and structure. Applications such as dialogue systems might benefit more from natural-language assertions, while applications that depend on multi-hop reasoning might benefit more from structured representations.

Combining CSK Models. Most prior CSK projects and our first project Ascent++ focused on acquiring "universal CSK", while our other projects Candle and Mango aimed at collecting culture-specific CSK. Combining these two types of knowledge in one system could be an interesting research direction. For example, the combined knowledge base should capture elephants being animals as universal CSK, and elephants being *sacred* animals in *Thailand*, *India*, and *Sri Lanka* as culture-specific CSK. This could be useful for various applications, such as dialogue systems, where understanding both universal and culture-specific knowledge would be helpful to generate appropriate responses.

6.2.2 Sources

Biases and Stereotypes. As sources, we directly tapped into a huge web crawl in Ascent++ and Candle. This gave us access to a vast amount of information, but also introduced challenges such as noise, biases, and stereotypes. Our methods for filtering, aggregating, and consolidating assertions have shown that it is feasible to extract high-quality CSK from such sources. Nonetheless, the inherent limitation of this approach is that the information from these sources does not always reflect what is true, but rather what people believe, or what the (English-speaking) Internet users say. This issue is particularly relevant for cultural knowledge, where stereotypes and misinformation can be perpetuated. Our other approach that used LLMs to generate CCSK, Mango, produced higher-quality assertions. However, as LLMs are also trained on web data, they are not immune to the challenges of biases and stereotypes.

A possible research direction is to develop models that can detect stereotypes in the collected assertions. As generative AI is getting increasingly popular, stereotype detection has become a notable research area. There have been several efforts to develop datasets for stereotype classification (e.g., (Dinan et al. 2020, Nangia et al. 2020, Nadeem et al. 2021, Felkner et al. 2023)). Such datasets can be used to train stereotype detection models which can be helpful for downstream applications to select which assertions should be incorporated into their systems.

Another possibility is to rank the source websites by their trustworthiness, giving signals for stereotype detection, besides frequency signals already collected via clustering. However, this is only applicable to the extraction-based methods, since the LLM-based method does not provide the source of the assertions. To this end, future research could focus on developing methods to trace the source of LLM-generated assertions, for example, via web search, bridging the gap between the two approaches.

Source Expansion. Another interesting research direction is to acquire knowledge from more diverse sources, including sources of different languages, cultures, and modalities (e.g., images, videos), though it has received little attention. For example, (Chen et al. 2013) and (Xu et al. 2018) are the few works that extracted CSK from images, with focus on specific relations such as `IsA`, `AtLocation`, `PartOf`, and `LocatedNear`. This was done by looking at the position and size of objects in images relative to each other. More recently, Yao et al. (2023)

used pretrained vision-language models to find commonsense interactions between objects in images.

Expanding commonsense knowledge extraction into diverse sources could help to mitigate the bias in the extracted knowledge, and to provide a more comprehensive view of the world, not just the English-speaking Internet, as well as to capture CSK that is not expressed in text and CSK in long-tail cultural groups. A major challenge of this direction is that by increasing the diversity of sources, the noise and misinformation in the collected assertions may also increase. Furthermore, using multilingual corpora might not fully solve the problem of cultural biases, as English is still the dominant language on the web. Nonetheless, this direction could be promising for future research.

6.2.3 Acquisition Methods

Text Extraction and LLM-Based Knowledge Distillation. We proposed two approaches for acquiring CSK: text extraction and knowledge distillation using LLMs. While text extraction is more transparent and traceable, LLM-based distillation produces assertions of higher quality and coverage.

Our text extraction methods, Ascent++ and Candle, consist of several modules which required judicious design and tuning, and they often relied on manual efforts for crafting rules and heuristics. While those are not the bottleneck for the extraction process, they might be time consuming and error-prone.

Distilling knowledge from LLMs, on the other hand, is more straightforward, and requires significantly less manual labor. However, CSK distillation from LLMs has its own limitations. Compared to text extraction, the usage of LLMs for knowledge distillation means losing the ability to trace assertions to their source, which is an important signal to verify and rank the assertions. Moreover, specifically to GPT-3.5 which we used in Mango, this LLM is provided via a commercial API only, hence it cannot be guaranteed that the proposed pipeline is long-term reproducible. The training process of GPT-3.5 is also undocumented, so we do not know what corpus it has been trained on, and in particular, whether there might be systematic gaps or biases in its training corpus (know what you don't know).

Future research could aim to combine the strengths of both approaches. For example, one could perform knowledge extraction from texts with the helps from LLMs to reduce the manual efforts in filtering and cleaning the extracted assertions. Another possibility is to use LLMs to generate assertions, and then develop methods to trace the source of these assertions, as discussed above.

Subject Expansion using LLMs. Prior work focused on acquiring CSK for simple concepts, which are usually single nouns. This missed out on knowledge of fine-grained concepts. Our Ascent++ project tackled this limitation by introducing subgroups and aspects of concepts, and acquiring CSK assertions for them. These refined subjects were collected using simple lexico-syntactic heuristics on web texts. However, the coverage of these subgroups and aspects is still limited, and there are still noisy extractions.

LLMs could be used to generate more of these fine-grained subjects. We expect that LLMs can generate more diverse and specific subgroups and aspects, via judicious prompting and filtering, than using the simple heuristics as in Ascent++. LLMs could also be used to generate assertions for these subgroups and aspects for potentially better coverage and precision, but the same limitations of traceability and source verification discussed above apply.

Similarly for CCSK, one could also aim to expand the domain to more cultural groups, such as interest groups (e.g., punks, skaters, movie buffs), age-related groups (e.g., babies, toddlers, kids, teenagers, Gen Y, Gen Z), and other groups beyond geo-based cultures.

The challenge here is that for long-tail cultural groups and concepts, neither the training data for LLMs nor the web data might have sufficient coverage.

6.2.4 Evaluation Settings

Human Evaluation. Across the three projects, we conducted human evaluations to assess the quality of the acquired CSK, as well as the outputs in the extrinsic use cases. Although human evaluations have been widely used in most CSK projects, they are not without limitations: Besides the cost and time constraints, human evaluations are subject to the quality of the annotators, and the biases they might have.

The biases in the annotators can be particularly problematic in cross-cultural tasks, where the annotators need to evaluate assertions from different cultures. For example, in the Mango project, we recruited annotators from Amazon MTurk to evaluate CCSK assertions and responses of intercultural dialogues. It has been reported by Ross et al. (2010) that the majority of workers in this platform are from only two cultures (USA and India), which makes it challenging to recruit truly diverse annotators. Although the majority of the workers reported being somewhat or very familiar with the cultures, it is not a guarantee for the quality of the annotations. Conducting evaluations with more culturally diverse annotators would further reinforce our results, but it can be challenging due to the cost and time constraints.

Coverage Measurement. In the Ascent++ project, we used relative recall computed against a small reference set, the CSLB property norm dataset (Devereux et al. 2014) consisting of 22.6K human-written sentences expressing properties of 638 concepts (see Section 3.5.2). While this is a common practice in the CSK literature, it is not a perfect measure of coverage, as the reference set is far from being comprehensive (e.g., it is impossible to write down all possible things that elephants can do). However, this approach is still useful for comparing the coverage of CSK resources in a relative manner.

Creating a test suite that covers a wide range of concepts and relations is a challenging task, and it is even more challenging for culture-specific knowledge. In the Candle and Mango projects, we used the number of assertions and cultures for comparing coverage of CCSK resources. Although both Candle and Mango significantly improved upon existing resources on these metrics, we acknowledge that sizes are only a crude proxy for coverage. For example, an extreme case would be a resource with many cultures, but each culture has only a single assertion; or a resource with many assertions, but they are all about a facet of few cultures. Future research could aim to develop evaluation metrics which can better capture the quality and coverage of CSK resources.

6.2.5 Ethical Considerations

Adapting AI systems to different cultures is crucial for their successful deployment in diverse societies. However, classified as discriminating or not, knowledge about cultural groups is always to some degree imperfect, and even positive biases can disturb some people (e.g., "playing football is a Brazilian thing" might be a positive stereotype to many people, but some Brazilians might not like hearing it because they don't play football at all).

Collecting CCSK represents a dual-use technology. On the one hand, CCSK can be used to improve the performance of AI systems (as our evaluations in Chapter 4 and Chapter 5 showed), and to help fighting biases that are present in the training data of those systems. On the other hand, we may also codify and reinforce biases, as the assertions we collected can be used by others to train future AI models, perpetuating the biases.

Nevertheless, our data represents the outputs of a research prototype. We recommend against using our data in production systems without a careful evaluation of benefits (of having a system that is more culturally adept) and associated challenges and risks (of stereotype perpetuation etc.).

6.3 Outlook

Large language models (LLMs) have revolutionized the field of natural language processing (NLP), and have had significant impacts on many text-based tasks. As these models continue to grow in size and performance, we expect a shift towards using them for knowledge acquisition, replacing the traditional methods of extracting knowledge from raw or (semi-)structured texts, and crowdsourcing annotation. In the Mango project, we developed a relatively simple method for distilling culture-specific knowledge from LLMs. That resulted in a resource of unprecedented coverage with even higher quality than Candle, but also required significantly less manual labor, showing the great potential of using LLMs for knowledge base construction. Although LLM-based knowledge acquisition methods have their own limitations (discussed in Section 6.2.2 and Section 6.2.3), we expect that future models will be even more knowledgable, as they see more data and are trained on more advanced architectures, which will further improve the quality and coverage of the generated assertions. However, the trade-off between low bias and specificity in model outputs will likely remain. That is unavoidable because of the biased nature of the training data, and efforts to mitigate these biases in the fine-tuning process likely lead to a loss in specificity.

On the other hand, despite LLMs being increasingly capable, we have shown that high-quality CSK resources are crucial to improve their performance in various downstream applications, such as question answering and dialogue generation, via retrieval-augmented generation (RAG) approaches. We anticipate that incorporating external knowledge, including CSK, into LLM prompts will continue to be a major research direction, and it will still be beneficial even for future models of much larger sizes. The reason for that is, unlike parametric knowledge in LLMs, non-parametric knowledge like commonsense knowledge bases has the advantage of being scrutable and verifiable, which is highly relevant in applications where errors are costly. These two types of knowledge are complementary, and future research should focus on how to best combine them, e.g., developing better retrieval methods, and how to decide when to use non-parametric knowledge and when to use parametric knowledge.

List of Figures

List of Tables

Bibliography

Acharya A, Talamadupula K, Finlayson M A (2021) Towards an Atlas of Cultural Commonsense for Machine Reasoning. In: Workshop on Common Sense Knowledge Graphs (CSKGs). Available from: https://usc-isi-i2.github.io/AAAI21workshop/papers/Acharya_CSKGsAAAI-21.pdf (Accessed on: September 11, 2024).

Agostini A, Aein M J, Szedmák S, Aksoy E E, Piater J H, Wörgötter F (2015) Using structural bootstrapping for object substitution in robotic executions of human-like manipulation tasks. In: 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2015, Hamburg, Germany, September 28 - October 2, 2015. IEEE, 6479–6486. https://doi.org/10.1109/IROS.2015.7354303

Ahn L von, Kedia M, Blum M (2006) Verbosity: a game for collecting common-sense facts. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. CHI '06. Association for Computing Machinery, Montréal, Québec, Canada, 75–78. https://doi.org/10.1145/1124772.1124784

Al-Moadhen A, Packianather M, Qiu R, Setchi R, Ji Z (2015) Improving the Efficiency of Robot Task Planning by Automatically Integrating Its Planner and Common-Sense Knowledge Base. In: Tweedale J W, Jain L C, Watada J, Howlett R J (Eds.), Knowledge-Based Information Systems in Practice. Springer International Publishing, Cham, 185–199. https://doi.org/10.1007/978-3-319-13545-8_11

Al-Moadhen A, Qiu R, Packianather M S, Ji Z, Setchi R (2013) Integrating Robot Task Planner with Common-sense Knowledge Base to Improve the Efficiency of Planning. In: Watada J, Jain L C, Howlett R J, Mukai N, Asakura K (Eds.), 17th International Conference in Knowledge Based and Intelligent Information and Engineering Systems, KES 2013, Kitakyushu, Japan, 9-11 September 2013. Procedia Computer Science. Elsevier, 211–220. https://doi.org/10.1016/J.PROCS.2013.09.097

Anacleto J, Lieberman H, Tsutsumi M, Neris V, Carvalho A, Espinosa J, Godoi M, Zem-Mascarenhas S (2006) Can Common Sense uncover cultural differences in computer applications?. In: Bramer M (Ed.), Artificial Intelligence in Theory and Practice. Springer US, Boston, MA, 1–10. https://doi.org/10.1007/978-0-387-34747-9_1

Arnaout H, Nguyen T-P, Razniewski S, Weikum G (2023) UnCommonSense in Action! Informative Negations for Commonsense Knowledge Bases. In: Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining. WSDM '23. Association for Computing Machinery, New York, NY, USA, 1120–1123. https://doi.org/10.1145/3539597.3573027

Auer S, Bizer C, Kobilarov G, Lehmann J, Cyganiak R, Ives Z G (2007) DBpedia: A Nucleus for a Web of Open Data. In: Aberer K, Choi K-S, Noy N F, Allemang D, Lee K-I, Nixon L J B, Golbeck J, Mika P, Maynard D, Mizoguchi R, Schreiber G, Cudré-Mauroux

P (Eds.), The Semantic Web, 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007, Busan, Korea, November 11-15, 2007. Lecture Notes in Computer Science. Springer, 722–735. https://doi.org/10.1007/978-3-540-76298-0_52

Ayari N, Chibani A, Amirat Y, Matson E T (2015) A novel approach based on commonsense knowledge representation and reasoning in open world for intelligent ambient assisted living services. In: 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2015, Hamburg, Germany, September 28 - October 2, 2015. IEEE, 6007–6013. https://doi.org/10.1109/IROS.2015.7354232

Bang Y, Cahyawijaya S, Lee N, Dai W, Su D, Wilie B, Lovenia H, Ji Z, Yu T, Chung W, Do Q V, Xu Y, Fung P (2023) A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity. In: Park J C, Arase Y, Hu B, Lu W, Wijaya D, Purwarianti A, Krisnadhi A A (Eds.), Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Nusa Dua, Bali, 675–718. https://doi.org/10.18653/v1/2023.ijcnlp-main.45

Barbieri F, Camacho-Collados J, Espinosa Anke L, Neves L (2020) TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification. In: Cohn T, He Y, Liu Y (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2020. Association for Computational Linguistics, Online, 1644–1650. https://doi.org/10.18653/v1/2020.findings-emnlp.148

Bhakthavatsalam S, Anastasiades C, Clark P (2020) GenericsKB: A Knowledge Base of Generic Statements. CoRR. Available from: https://arxiv.org/abs/2005.00660 (Accessed on: September 13, 2024).

Bhakthavatsalam S, Richardson K, Tandon N, Clark P (2020) Do Dogs have Whiskers? A New Knowledge Base of hasPart Relations. CoRR. Available from: https://arxiv.org/abs/2006.07510 (Accessed on: September 13, 2024).

Bhatia M, Shwartz V (2023) GD-COMET: A Geo-Diverse Commonsense Inference Model. In: Bouamor H, Pino J, Bali K (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Singapore, 7993–8001. https://doi.org/10.18653/v1/2023.emnlp-main.496

Bisk Y, Zellers R, Bras R L, Gao J, Choi Y (2020) PIQA: Reasoning about Physical Commonsense in Natural Language. In: The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020. AAAI Press, 7432–7439. https://doi.org/10.1609/AAAI.V34I05.6239

Bitton Y, Guetta N B, Yosef R, Elovici Y, Bansal M, Stanovsky G, Schwartz R (2022) WinoGAViL: Gamified Association Benchmark to Challenge Vision-and-Language Models. In: Koyejo S, Mohamed S, Agarwal A, Belgrave D, Cho K, Oh A (Eds.), Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022. Available from: http://papers.nips.cc/paper_files/paper/2022/hash/a96fe863f85c59789bba63588a9557b4-Abstract-Datasets_and_Benchmarks.html (Accessed on: September 13, 2024).

Bosselut A, Bras R L, Choi Y (2021) Dynamic Neuro-Symbolic Knowledge Graph Construction for Zero-shot Commonsense Question Answering. In: Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021. AAAI Press, 4923–4931. https://doi.org/10.1609/AAAI.V35I6.16625

Bosselut A, Rashkin H, Sap M, Malaviya C, Celikyilmaz A, Choi Y (2019) COMET: Commonsense Transformers for Automatic Knowledge Graph Construction. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Florence, Italy, 4762–4779. https://doi.org/10.18653/v1/P19-1470

Brachman R J, Levesque H J (2022) Toward a New Science of Common Sense. Proceedings of the AAAI Conference on Artificial Intelligence 36: 12245–12249. https://doi.org/10.1609/aaai.v36i11.21485

Brants T, Franz A (2006) Web 1T 5-gram Version 1 LDC2006T13. https://doi.org/10.35111/cqpa-a498

Brown T B, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S, Herbert-Voss A, Krueger G, Henighan T, Child R, Ramesh A, Ziegler D M, Wu J, Winter C, Hesse C, Chen M, Sigler E, Litwin M, Gray S, Chess B, Clark J, Berner C, McCandlish S, Radford A, Sutskever I, Amodei D (2020) Language Models Are Few-Shot Learners. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. NIPS'20. Curran Associates Inc., Vancouver, BC, Canada. Available from: https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf (Accessed on: September 13, 2024).

Cai H, Shen X, Xu Q, Shen W, Wang X, Ge W, Zheng X, Xue X (2023) Improving Empathetic Dialogue Generation by Dynamically Infusing Commonsense Knowledge. In: Rogers A, Boyd-Graber J, Okazaki N (Eds.), Findings of the Association for Computational Linguistics: ACL 2023. Association for Computational Linguistics, Toronto, Canada, 7858–7873. https://doi.org/10.18653/v1/2023.findings-acl.498

Cetto M, Niklaus C, Freitas A, Handschuh S (2018) Graphene: Semantically-Linked Propositions in Open Information Extraction. In: Bender E M, Derczynski L, Isabelle P (Eds.), Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018. Association for Computational Linguistics, 2300–2311. Available from: https://aclanthology.org/C18-1195/ (Accessed on: September 13, 2024).

CH-Wang S, Saakyan A, Li O, Yu Z, Muresan S (2023) Sociocultural Norm Similarities and Differences via Situational Alignment and Explainable Textual Entailment. In: Bouamor H, Pino J, Bali K (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Singapore, 3548–3564. https://doi.org/10.18653/v1/2023.emnlp-main.215

Chakrabarty T, Ghosh D, Muresan S, Peng N (2020) R^3: Reverse, Retrieve, and Rank for Sarcasm Generation with Commonsense Knowledge. In: Jurafsky D, Chai J, Schluter N, Tetreault J (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Online, 7976–7986. https://doi.org/10.18653/v1/2020.acl-main.711

Chalier Y, Razniewski S, Weikum G (2020) Joint Reasoning for Multi-Faceted Commonsense Knowledge. In: Das D, Hajishirzi H, McCallum A, Singh S (Eds.), Conference on Automated Knowledge Base Construction, AKBC 2020, Virtual, June 22-24, 2020. https://doi.org/10.24432/C58G6G

Chen H, Huang Y, Takamura H, Nakayama H (2021) Commonsense Knowledge Aware Concept Selection For Diverse and Informative Visual Storytelling. In: Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021. AAAI Press, 999–1008. https://doi.org/10.1609/AAAI.V35I2.16184

Chen Q, Ji F, Chen H, Zhang Y (2020) Improving Commonsense Question Answering by Graph-based Iterative Retrieval over Multiple Knowledge Sources. In: Scott D, Bel N, Zong C (Eds.), Proceedings of the 28th International Conference on Computational Linguistics. International Committee on Computational Linguistics, Barcelona, Spain (Online), 2583–2594. https://doi.org/10.18653/v1/2020.coling-main.232

Chen T, Jiang Z, Poliak A, Sakaguchi K, Van Durme B (2020) Uncertain Natural Language Inference. In: Jurafsky D, Chai J, Schluter N, Tetreault J (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Online, 8772–8779. https://doi.org/10.18653/v1/2020.acl-main.774

Chen X, Jia S, Xiang Y (2020) A review: Knowledge reasoning over knowledge graph. Expert Syst. Appl. 141. https://doi.org/10.1016/J.ESWA.2019.112948

Chen X, Shrivastava A, Gupta A (2013) NEIL: Extracting Visual Knowledge from Web Data. In: IEEE International Conference on Computer Vision, ICCV 2013, Sydney, Australia, December 1-8, 2013. IEEE Computer Society, 1409–1416. https://doi.org/10.1109/ICCV.2013.178

Chernova S, Chu V, Daruna A, Garrison H, Hahn M, Khante P, Liu W, Thomaz A (2020) Situated Bayesian Reasoning Framework for Robots Operating in Diverse Everyday Environments. In: Amato N M, Hager G, Thomas S, Torres-Torriti M (Eds.), Robotics Research. Springer International Publishing, Cham, 353–369. https://doi.org/10.1007/978-3-030-28619-4_29

Chiatti A, Motta E, Daga E (2022) Robots With Commonsense: Improving Object Recognition Through Size and Spatial Awareness. In: Martin A, Hinkelmann K, Fill H-G, Gerber A, Lenat D, Stolle R, Harmelen F van (Eds.), Proceedings of the AAAI 2022 Spring Symposium on Machine Learning and Knowledge Engineering for Hybrid Intelligence (AAAI-MAKE 2022), Stanford University, Palo Alto, California, USA, March 21-23, 2022. CEUR Workshop Proceedings. CEUR-WS.org. Available from: https://ceur-ws.org/Vol-3121/paper4.pdf (Accessed on: September 13, 2024).

Chiu Y Y, Jiang L, Antoniak M, Park C Y, Li S S, Bhatia M, Ravi S, Tsvetkov Y, Shwartz V, Choi Y (2024) CulturalTeaming: AI-Assisted Interactive Red-Teaming for Challenging LLMs' (Lack of) Multicultural Knowledge. CoRR. https://doi.org/10.48550/ARXIV.2404.06664

Clark P, Cowhey I, Etzioni O, Khot T, Sabharwal A, Schoenick C, Tafjord O (2018) Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge. CoRR. Available from: http://arxiv.org/abs/1803.05457 (Accessed on: September 13, 2024).

Clarke J, Srikumar V, Sammons M, Roth D (2012) An NLP Curator (or: How I Learned to Stop Worrying and Love NLP Pipelines). In: Calzolari N, Choukri K, Declerck T, Doğan M U, Maegaard B, Mariani J, Moreno A, Odijk J, Piperidis S (Eds.), Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12). European Language Resources Association (ELRA), Istanbul, Turkey, 3276–3283. Available from: http://www.lrec-conf.org/proceedings/lrec2012/pdf/664_Paper.pdf (Accessed on: September 13, 2024).

Conesa J, Storey V C, Sugumaran V (2008) Improving web-query processing through semantic knowledge. Data & Knowledge Engineering 66: 18–34. https://doi.org/https://doi.org/10.1016/j.datak.2007.07.009

Cycorp (2021) Cyc Technology Overview. Available from: https://cyc.com/wp-content/uploads/2021/04/Cyc-Technology-Overview.pdf (Accessed on: September 4, 2024).

Cycorp The OpenCyc Platform. Available from: https://web.archive.org/web/20160330064354/http://www.opencyc.org/ (Accessed on: May 16, 2022).

Dalvi Mishra B, Tandon N, Clark P (2017) Domain-Targeted, High Precision Knowledge Extraction. Transactions of the Association for Computational Linguistics 5: 233–246. https://doi.org/10.1162/tacl_a_00058

Daruna A A, Liu W, Kira Z, Chernova S (2019) RoboCSE: Robot Common Sense Embedding. In: International Conference on Robotics and Automation, ICRA 2019, Montreal, QC, Canada, May 20-24, 2019. IEEE, 9777–9783. https://doi.org/10.1109/ICRA.2019.8794070

Davis E (2023) Benchmarks for Automated Commonsense Reasoning: A Survey. ACM Comput. Surv. 56. https://doi.org/10.1145/3615355

Davis E, Marcus G (2015) Commonsense reasoning and commonsense knowledge in artificial intelligence. Commun. ACM 58: 92–103. https://doi.org/10.1145/2701413

Deshpande A, Ruiter D, Mosbach M, Klakow D (2022) StereoKG: Data-Driven Knowledge Graph Construction For Cultural Knowledge and Stereotypes. In: Narang K, Mostafazadeh Davani A, Mathias L, Vidgen B, Talat Z (Eds.), Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH). Association for Computational Linguistics, Seattle, Washington (Hybrid), 67–78. https://doi.org/10.18653/v1/2022.woah-1.7

Devereux B J, Tyler L K, Geertzen J, Randall B (2014) The Centre for Speech, Language and the Brain (CSLB) concept property norms. Behavior research methods. https://doi.org/10.3758/s13428-013-0420-4

Dinan E, Fan A, Wu L, Weston J, Kiela D, Williams A (2020) Multi-Dimensional Gender Bias Classification. In: Webber B, Cohn T, He Y, Liu Y (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, Online, 314–331. https://doi.org/10.18653/v1/2020.emnlp-main.23

Dragoni M, Poria S, Cambria E (2018) OntoSenticNet: A Commonsense Ontology for Sentiment Analysis. IEEE Intell. Syst. 33: 77–85. https://doi.org/10.1109/MIS.2018.033001419

Elazar Y, Mahabal A, Ramachandran D, Bedrax-Weiss T, Roth D (2019) How Large Are Lions? Inducing Distributions over Quantitative Attributes. In: Korhonen A, Traum D, Màrquez L (Eds.), Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Florence, Italy, 3973–3983. https://doi.org/10.18653/v1/P19-1388

Etzioni O, Banko M, Soderland S, Weld D S (2008) Open information extraction from the web. Commun. ACM 51: 68–74. https://doi.org/10.1145/1409360.1409378

Etzioni O, Cafarella M J, Downey D, Kok S, Popescu A-M, Shaked T, Soderland S, Weld D S, Yates A (2004) Web-scale information extraction in knowitall: (preliminary results). In: Feldman S I, Uretsky M, Najork M, Wills C E (Eds.), Proceedings of the 13th international conference on World Wide Web, WWW 2004, New York, NY, USA, May 17-20, 2004. ACM, 100–110. https://doi.org/10.1145/988672.988687

Felkner V, Chang H-C H, Jang E, May J (2023) WinoQueer: A Community-in-the-Loop Benchmark for Anti-LGBTQ+ Bias in Large Language Models. In: Rogers A, Boyd-Graber J, Okazaki N (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Toronto, Canada, 9126–9140. https://doi.org/10.18653/v1/2023.acl-long.507

Feng Y, Chen X, Lin B Y, Wang P, Yan J, Ren X (2020) Scalable Multi-Hop Relational Reasoning for Knowledge-Aware Question Answering. In: Webber B, Cohn T, He Y, Liu Y (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, Online, 1295–1309. https://doi.org/10.18653/v1/2020.emnlp-main.99

Fleiss J L, Cohen J (1973) The Equivalence of Weighted Kappa and the Intraclass Correlation Coefficient as Measures of Reliability. Educational and Psychological Measurement 33: 613–619. https://doi.org/10.1177/001316447303300309

Forbes M, Hwang J D, Shwartz V, Sap M, Choi Y (2020) Social Chemistry 101: Learning to Reason about Social and Moral Norms. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, Online, 653–670. https://doi.org/10.18653/v1/2020.emnlp-main.48

Fu X, He M, Lu Y, Wang W Y, Roth D (2024) Commonsense-T2I Challenge: Can Text-to-Image Generation Models Understand Commonsense?. CoRR. https://doi.org/10.48550/ARXIV.2406.07546

Fung Y, Chakrabarty T, Guo H, Rambow O, Muresan S, Ji H (2023) NORMSAGE: Multi-Lingual Multi-Cultural Norm Discovery from Conversations On-the-Fly. In: Bouamor H, Pino J, Bali K (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Singapore, 15217–15230. https://doi.org/10.18653/v1/2023.emnlp-main.941

Fung Y, Zhao R, Doo J, Sun C, Ji H (2024) Massively Multi-Cultural Knowledge Acquisition & LM Benchmarking. CoRR. https://doi.org/10.48550/ARXIV.2402.09369

Gabbay D M (2003) Many-Dimensional Modal Logics: Theory and Applications. Elsevier North Holland.

Galárraga L, Heitz G, Murphy K, Suchanek F M (2014) Canonicalizing Open Knowledge Bases. In: Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management. CIKM '14. Association for Computing Machinery, Shanghai, China, 1679–1688. https://doi.org/10.1145/2661829.2662073

Gao X, Chen W, Lu L, Cui Y, Dai X, Dai L, Wang K, Shen J, Wang Y, Wang S, Yu Z, Liu H (2023) An Interpretable Fake News Detection Method Based on Commonsense Knowledge Graph. Applied Sciences 13. https://doi.org/10.3390/app13116680

Gatti P, Penamakuri A S, Teotia R, Mishra A, Sengupta S, Ramnani R (2022) COFAR: Commonsense and Factual Reasoning in Image Search. In: He Y, Ji H, Li S, Liu Y, Chang C-H (Eds.), Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, Online only, 1185–1199. Available from: https://aclanthology.org/2022.aacl-main.87 (Accessed on: September 13, 2024).

Geertz C (1983) Local knowledge: Further essays in interpretive anthropology. Basic books.

Ghosal D, Majumder N, Gelbukh A, Mihalcea R, Poria S (2020) COSMIC: COmmonSense knowledge for eMotion Identification in Conversations. In: Cohn T, He Y, Liu Y (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2020. Association for Computational Linguistics, Online, 2470–2481. https://doi.org/10.18653/v1/2020.findings-emnlp.224

Ghosal D, Majumder N, Mihalcea R, Poria S (2021) STaCK: Sentence Ordering with Temporal Commonsense Knowledge. In: Moens M-F, Huang X, Specia L, Yih S W-t (Eds.), Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 8676–8686. https://doi.org/10.18653/v1/2021.emnlp-main.683

Girju R, Badulescu A, Moldovan D (2006) Automatic Discovery of Part-Whole Relations. Computational Linguistics 32: 83–135. https://doi.org/10.1162/coli.2006.32.1.83

Gordon J, Durme B V (2013) Reporting bias and knowledge acquisition. In: Suchanek F M, Riedel S, Singh S, Talukdar P P (Eds.), Proceedings of the 2013 workshop on Automated knowledge base construction, AKBC@CIKM 13, San Francisco, California, USA, October 27-28, 2013. ACM, 25–30. https://doi.org/10.1145/2509558.2509563

Gordon J, Durme B V, Schubert L K (2010) Learning from the Web: Extracting General World Knowledge from Noisy Text. In: Collaboratively-Built Knowledge Sources and Artificial Intelligence, Papers from the 2010 AAAI Workshop, Atlanta, Georgia, USA, July 11, 2010. AAAI Technical Report. AAAI. Available from: http://aaai.org/ocs/index.php/WS/AAAIW10/paper/view/2035 (Accessed on: September 13, 2024).

Gordon J, Schubert L K (2010) Quantificational Sharpening of Commonsense Knowledge. In: Commonsense Knowledge, Papers from the 2010 AAAI Fall Symposium, Arlington, Virginia, USA, November 11-13, 2010. AAAI Technical Report. AAAI. Available from: https://cdn.aaai.org/ocs/2324/2324-9487-1-PB.pdf (Accessed on: September 13, 2024).

Gu J, Zhao H, Lin Z, Li S, Cai J, Ling M (2019) Scene Graph Generation With External Knowledge and Image Reconstruction. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019. Computer Vision Foundation / IEEE, 1969–1978. https://doi.org/10.1109/CVPR.2019.00207

Guo Y, Song J, Gao L, Shen H T (2020) One-shot Scene Graph Generation. In: Chen C W, Cucchiara R, Hua X-S, Qi G-J, Ricci E, Zhang Z, Zimmermann R (Eds.), MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020. ACM, 3090–3098. https://doi.org/10.1145/3394171.3414025

Guu K, Lee K, Tung Z, Pasupat P, Chang M-W (2020) Retrieval Augmented Language Model Pre-Training. In: Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event. Proceedings of Machine Learning Research. PMLR, 3929–3938. Available from: http://proceedings.mlr.press/v119/guu20a.html (Accessed on: September 13, 2024).

Havasi C, Speer R, Alonso J (2007) ConceptNet 3: a flexible, multilingual semantic network for common sense knowledge. In: Recent advances in natural language processing. John Benjamins, 269–280. Available from: https://web.media.mit.edu/~lieber/Teaching/Common-Sense-Course/ConceptNet-3.pdf (Accessed on: September 13, 2024).

Hertling S, Paulheim H (2017) WebIsALOD: Providing Hypernymy Relations Extracted from the Web as Linked Open Data. In: d'Amato C, Fernández M, Tamma V A M, Lécué F, Cudré-Mauroux P, Sequeda J F, Lange C, Heflin J (Eds.), The Semantic Web - ISWC 2017 - 16th International Semantic Web Conference, Vienna, Austria, October

21-25, 2017, Proceedings, Part II. Lecture Notes in Computer Science. Springer, 111–119. https://doi.org/10.1007/978-3-319-68204-4_11

Hogan A, Blomqvist E, Cochez M, D'amato C, Melo G D, Gutierrez C, Kirrane S, Gayo J E L, Navigli R, Neumaier S, Ngomo A-C N, Polleres A, Rashid S M, Rula A, Schmelzeisen L, Sequeda J, Staab S, Zimmermann A (2021) Knowledge Graphs. ACM Comput. Surv. 54. https://doi.org/10.1145/3447772

Huang J, Yang D (2023) Culturally Aware Natural Language Inference. In: Bouamor H, Pino J, Bali K (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2023. Association for Computational Linguistics, Singapore, 7591–7609. https://doi.org/10.18653/v1/2023.findings-emnlp.509

Hwang J D, Bhagavatula C, Bras R L, Da J, Sakaguchi K, Bosselut A, Choi Y (2021) (Comet-) Atomic 2020: On Symbolic and Neural Commonsense Knowledge Graphs. In: Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021. AAAI Press, 6384–6392. https://doi.org/10.1609/AAAI.V35I7.16792

Ilievski F, Szekely P A, Schwabe D (2020) Commonsense Knowledge in Wikidata. In: Kaffee L-A, Tifrea-Marciuska O, Simperl E, Vrandecic D (Eds.), Proceedings of the 1st Wikidata Workshop (Wikidata 2020) co-located with 19th International Semantic Web Conference(OPub 2020), Virtual Conference, November 2-6, 2020. CEUR Workshop Proceedings. CEUR-WS.org. Available from: https://ceur-ws.org/Vol-2773/paper-10.pdf (Accessed on: September 13, 2024).

Ilievski F, Szekely P, Zhang B (2021) CSKG: The CommonSense Knowledge Graph. In: The Semantic Web: 18th International Conference, ESWC 2021, Virtual Event, June 6–10, 2021, Proceedings. Springer-Verlag, Berlin, Heidelberg, 680–696. https://doi.org/10.1007/978-3-030-77385-4_41

Javia B, Cimiano P (2016) A knowledge-based architecture supporting declarative action representation for manipulation of everyday objects. In: Aßmann U, Brugali D, Piechnick C (Eds.), Proceedings of the 3rd Workshop on Model-Driven Robot Software Engineering, MORSE@RoboCup 2016, Leipzig, Germany, July 1, 2016. ACM, 40–46. https://doi.org/10.1145/3022099.3022105

Kan X, Cui H, Yang C (2021) Zero-Shot Scene Graph Relation Prediction Through Commonsense Knowledge Integration. In: Machine Learning and Knowledge Discovery in Databases. Research Track: European Conference, ECML PKDD 2021, Bilbao, Spain, September 13–17, 2021, Proceedings, Part II. Springer-Verlag, Bilbao, Spain, 466–482. https://doi.org/10.1007/978-3-030-86520-7_29

Khan M J, Breslin J G, Curry E (2022) Expressive Scene Graph Generation Using Commonsense Knowledge Infusion for Visual Understanding and Reasoning. In: The Semantic Web: 19th International Conference, ESWC 2022, Hersonissos, Crete, Greece, May 29 – June 2, 2022, Proceedings. Springer-Verlag, Hersonissos, Greece, 93–112. https://doi.org/10.1007/978-3-031-06981-9_6

Kim H, Hessel J, Jiang L, West P, Lu X, Yu Y, Zhou P, Bras R, Alikhani M, Kim G, Sap M, Choi Y (2023) SODA: Million-scale Dialogue Distillation with Social Commonsense Contextualization. In: Bouamor H, Pino J, Bali K (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Singapore, 12930–12949. https://doi.org/10.18653/v1/2023.emnlp-main.799

Krishna R, Zhu Y, Groth O, Johnson J, Hata K, Kravitz J, Chen S, Kalantidis Y, Li L-J, Shamma D A, Bernstein M S, Fei-Fei L (2017) Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. Int. J. Comput. Vision 123: 32–73. https://doi.org/10.1007/s11263-016-0981-7

Kümpel M, Groot A de, Tiddi I, Beetz M (2020) Using Linked Data to Help Robots Understand Product-related Actions. In: Hammar K, Kutz O, Dimou A, Hahmann T, Hoehndorf R, Masolo C, Vita R, Abbès S B, Hantach R, Calvez P, Sales T P, Porello D, Beßler D, Borgo S, Diab M, Gangemi A, Alarcos A O, Pomarlan M, Porzel R, Skjæveland M G, Lupp D P, Horrocks I, Klüwer J W, Kindermann C, Bozzato L, Mossakowski T, Serafini L (Eds.), Proceedings of the Joint Ontology Workshops co-located with the Bolzano Summer of Knowledge (BOSK 2020), Virtual & Bozen-Bolzano, Italy, August 31st to October 7th, 2020. CEUR Workshop Proceedings. CEUR-WS.org. Available from: https://ceur-ws.org/Vol-2708/robontics2.pdf (Accessed on: September 13, 2024).

Lan Z, Chen M, Goodman S, Gimpel K, Sharma P, Soricut R (2020) ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net. Available from: https://openreview.net/forum?id=H1eA7AEtvS (Accessed on: September 13, 2024).

Lenat D B (1995) CYC: A Large-Scale Investment in Knowledge Infrastructure. Commun. ACM 38: 32–38. https://doi.org/10.1145/219717.219745

Levesque H J, Davis E, Morgenstern L (2012) The Winograd Schema Challenge. In: Brewka G, Eiter T, McIlraith S A (Eds.), Principles of Knowledge Representation and Reasoning: Proceedings of the Thirteenth International Conference, KR 2012, Rome, Italy, June 10-14, 2012. AAAI Press. Available from: http://www.aaai.org/ocs/index.php/KR/KR12/paper/view/4492 (Accessed on: September 13, 2024).

Lewis M, Liu Y, Goyal N, Ghazvininejad M, Mohamed A, Levy O, Stoyanov V, Zettlemoyer L (2020) BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Online, 7871–7880. https://doi.org/10.18653/v1/2020.acl-main.703

Lewis P S H, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, Küttler H, Lewis M, Yih W-t, Rocktäschel T, Riedel S, Kiela D (2020) Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In: Larochelle H, Ranzato M, Hadsell R, Balcan M-F, Lin H-T (Eds.), Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual. Available from: https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html (Accessed on: September 13, 2024).

Li D, Zhu X, Li Y, Wang S, Li D, Liao J, Zheng J (2021) Enhancing emotion inference in conversations with commonsense knowledge. Knowledge-Based Systems 232: 107449–107450. https://doi.org/https://doi.org/10.1016/j.knosys.2021.107449

Li J, Meng F, Lin Z, Liu R, Fu P, Cao Y, Wang W, Zhou J (2022) Neutral Utterances are Also Causes: Enhancing Conversational Causal Emotion Entailment with Social Commonsense Knowledge. In: Raedt L D (Ed.), Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22. International Joint Conferences on Artificial Intelligence Organization, 4209–4215. https://doi.org/10.24963/ijcai.2022/584

Li J, Pan H, Lin Z, Fu P, Wang W (2021) Sarcasm Detection with Commonsense Knowledge. IEEE ACM Trans. Audio Speech Lang. Process. 29: 3192–3201. https://doi.org/10.1109/TASLP.2021.3120601

Li J, Galley M, Brockett C, Gao J, Dolan B (2016) A Diversity-Promoting Objective Function for Neural Conversation Models. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, San Diego, California, 110–119. https://doi.org/10.18653/v1/N16-1014

Li L, Lei J, Gan Z, Liu J (2021) Adversarial VQA: A New Benchmark for Evaluating the Robustness of VQA Models. In: 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021. IEEE, 2022–2031. https://doi.org/10.1109/ICCV48922.2021.00205

Li O, Subramanian M, Saakyan A, CH-Wang S, Muresan S (2023) NormDial: A Comparable Bilingual Synthetic Dialog Dataset for Modeling Social Norm Adherence and Violation. In: Bouamor H, Pino J, Bali K (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Singapore, 15732–15744. https://doi.org/10.18653/v1/2023.emnlp-main.974

Lieberman H, Liu H, Singh P, Barry B (2004) Beating Common Sense into Interactive Applications. AI Mag. 25: 63–76. https://doi.org/10.1609/AIMAG.V25I4.1785

Lin B Y, Chen X, Chen J, Ren X (2019) KagNet: Knowledge-Aware Graph Networks for Commonsense Reasoning. In: Inui K, Jiang J, Ng V, Wan X (Eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Association for Computational Linguistics, Hong Kong, China, 2829–2839. https://doi.org/10.18653/v1/D19-1282

Lin B Y, Lee S, Khanna R, Ren X (2020) Birds have four legs?! NumerSense: Probing Numerical Commonsense Knowledge of Pre-Trained Language Models. In: Webber B, Cohn T, He Y, Liu Y (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, Online, 6862–6868. https://doi.org/10.18653/v1/2020.emnlp-main.557

Lin B Y, Wu Z, Yang Y, Lee D-H, Ren X (2021) RiddleSense: Reasoning about Riddle Questions Featuring Linguistic Creativity and Commonsense Knowledge. In: Zong C, Xia F, Li W, Navigli R (Eds.), Findings of the Association for Computational Linguistics: ACL-

IJCNLP 2021. Association for Computational Linguistics, Online, 1504–1515. https://doi.org/10.18653/v1/2021.findings-acl.131

Lin B Y, Zhou W, Shen M, Zhou P, Bhagavatula C, Choi Y, Ren X (2020) CommonGen: A Constrained Text Generation Challenge for Generative Commonsense Reasoning. In: Cohn T, He Y, Liu Y (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2020. Association for Computational Linguistics, Online, 1823–1840. https://doi.org/10.18653/v1/2020.findings-emnlp.165

Liu B, Guo W, Niu D, Luo J, Wang C, Wen Z, Xu Y (2020) GIANT: Scalable Creation of a Web-scale Ontology. In: Maier D, Pottinger R, Doan A, Tan W-C, Alawini A, Ngo H Q (Eds.), Proceedings of the 2020 International Conference on Management of Data, SIGMOD Conference 2020, online conference [Portland, OR, USA], June 14-19, 2020. ACM, 393–409. https://doi.org/10.1145/3318464.3386145

Liu F, Bugliarello E, Ponti E M, Reddy S, Collier N, Elliott D (2021) Visually Grounded Reasoning across Languages and Cultures. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 10467–10485. https://doi.org/10.18653/v1/2021.emnlp-main.818

Liu H, Singh P (2004) ConceptNet — A Practical Commonsense Reasoning Tool-Kit. BT Technology Journal 22: 211–226. https://doi.org/10.1023/B:BTTJ.0000047600.45421.6d

Liu R, Lin Z, Tan Y, Wang W (2021) Enhancing Zero-shot and Few-shot Stance Detection with Commonsense Knowledge Graph. In: Zong C, Xia F, Li W, Navigli R (Eds.), Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Association for Computational Linguistics, Online, 3152–3157. https://doi.org/10.18653/v1/2021.findings-acl.278

Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V (2019) RoBERTa: A Robustly Optimized BERT Pretraining Approach. CoRR. Available from: http://arxiv.org/abs/1907.11692 (Accessed on: September 13, 2024).

Loshchilov I, Hutter F (2019) Decoupled Weight Decay Regularization. In: 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019. OpenReview.net. Available from: https://openreview.net/forum?id=Bkg6RiCqY7 (Accessed on: September 13, 2024).

Lv S, Guo D, Xu J, Tang D, Duan N, Gong M, Shou L, Jiang D, Cao G, Hu S (2020) Graph-Based Reasoning over Heterogeneous External Knowledge for Commonsense Question Answering. Proceedings of the AAAI Conference on Artificial Intelligence 34: 8449–8456. https://doi.org/10.1609/aaai.v34i05.6364

Ma Y, Peng H, Cambria E (2018) Targeted Aspect-Based Sentiment Analysis via Embedding Commonsense Knowledge into an Attentive LSTM. Proceedings of the AAAI Conference on Artificial Intelligence 32. https://doi.org/10.1609/aaai.v32i1.12048

Mallen A, Asai A, Zhong V, Das R, Khashabi D, Hajishirzi H (2023) When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories. In: Rogers A, Boyd-Graber J, Okazaki N (Eds.), Proceedings of the 61st Annual

Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Toronto, Canada, 9802–9822. https://doi.org/10.18653/v1/2023.acl-long.546

Marneffe M-C de, Manning C D, Nivre J, Zeman D (2021) Universal Dependencies. Computational Linguistics 47: 255–308. https://doi.org/10.1162/coli_a_00402

McRae K, Cree G S, Seidenberg M S, Mcnorgan C (2005) Semantic feature production norms for a large set of living and nonliving things. Behavior Research Methods 37: 547–559. https://doi.org/10.3758/BF03192726

Mehrabi N, Zhou P, Morstatter F, Pujara J, Ren X, Galstyan A (2021) Lawyers are Dishonest? Quantifying Representational Harms in Commonsense Knowledge Resources. In: Moens M-F, Huang X, Specia L, Yih S W-t (Eds.), Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 5016–5033. https://doi.org/10.18653/v1/2021.emnlp-main.410

Mehri S, Choi J, D'Haro L F, Deriu J, Eskénazi M, Gasic M, Georgila K, Hakkani-Tur D, Li Z, Rieser V, Shaikh S, Traum D R, Yeh Y-T, Yu Z, Zhang Y, Zhang C (2022) Report from the NSF Future Directions Workshop on Automatic Evaluation of Dialog: Research Directions and Challenges. CoRR. https://doi.org/10.48550/ARXIV.2203.10012

Mikolov T, Sutskever I, Chen K, Corrado G, Dean J (2013) Distributed representations of words and phrases and their compositionality. In: Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2. NIPS'13. Curran Associates Inc., Lake Tahoe, Nevada, 3111–3119. Available from: https://proceedings.neurips.cc/paper_files/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf (Accessed on: September 13, 2024).

Miller G A (1995) WordNet: a lexical database for English. Communication of the ACM 38: 39–41. https://doi.org/10.1145/219717.219748

Mitrevski A, Plöger P G, Lakemeyer G (2021) Ontology-Assisted Generalisation of Robot Action Execution Knowledge. In: IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2021, Prague, Czech Republic, September 27 - Oct. 1, 2021. IEEE, 6763–6770. https://doi.org/10.1109/IROS51168.2021.9636791

Nadeem M, Bethke A, Reddy S (2021) StereoSet: Measuring stereotypical bias in pretrained language models. In: Zong C, Xia F, Li W, Navigli R (Eds.), Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, Online, 5356–5371. https://doi.org/10.18653/v1/2021.acl-long.416

Nangia N, Vania C, Bhalerao R, Bowman S R (2020) CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models. In: Webber B, Cohn T, He Y, Liu Y (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, Online, 1953–1967. https://doi.org/10.18653/v1/2020.emnlp-main.154

Navigli R (2009) Word sense disambiguation: A survey. ACM Computing Surveys 41. https://doi.org/10.1145/1459352.1459355

Navigli R, Ponzetto S P (2012) BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artificial Intelligence 193: 217–250. https://doi.org/10.1016/j.artint.2012.07.001

Nguyen T-P, Razniewski S (2022) Materialized Knowledge Bases from Commonsense Transformers. In: Proceedings of the First Workshop on Commonsense Representation and Reasoning (CSRR 2022). Association for Computational Linguistics, Dublin, Ireland, 36–42. https://doi.org/10.18653/v1/2022.csrr-1.5

Nguyen T-P, Razniewski S, Romero J, Weikum G (2023a) Refined Commonsense Knowledge From Large-Scale Web Contents. IEEE Trans. Knowl. Data Eng. 35: 8431–8447. https://doi.org/10.1109/TKDE.2022.3206505

Nguyen T-P, Razniewski S, Varde A, Weikum G (2023b) Extracting Cultural Commonsense Knowledge at Scale. In: Proceedings of the Web Conference 2023. WWW '23. Association for Computing Machinery, Austin, TX, USA. https://doi.org/10.1145/3543507.3583535

Nguyen T-P, Razniewski S, Weikum G (2021a) Advanced Semantics for Commonsense Knowledge Extraction. In: Proceedings of the Web Conference 2021. WWW '21. Association for Computing Machinery, Ljubljana, Slovenia, 2636–2647. https://doi.org/10.1145/3442381.3449827

Nguyen T-P, Razniewski S, Weikum G (2021b) Inside ASCENT: Exploring a Deep Commonsense Knowledge Base and its Usage in Question Answering. In: Ji H, Park J C, Xia R (Eds.), Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations. Association for Computational Linguistics, Online, 40–47. https://doi.org/10.18653/v1/2021.acl-demo.5

Nguyen T-P, Razniewski S, Weikum G (2024) Cultural Commonsense Knowledge for Intercultural Dialogues. In: Proceedings of the 33rd ACM International Conference on Information and Knowledge Management. CIKM '24. Association for Computing Machinery, Boise, ID, USA. https://doi.org/10.1145/3627673.3679768

Nyga D, Beetz M (2012) Everything robots always wanted to know about housework (but were afraid to ask). In: 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2012, Vilamoura, Algarve, Portugal, October 7-12, 2012. IEEE, 243–250. https://doi.org/10.1109/IROS.2012.6385923

Ofek N, Poria S, Rokach L, Cambria E, Hussain A, Shabtai A (2016) Unsupervised Commonsense Knowledge Enrichment for Domain-Specific Sentiment Analysis. Cogn. Comput. 8: 467–477. https://doi.org/10.1007/S12559-015-9375-3

OpenAI (2023) GPT-4 Technical Report. CoRR. https://doi.org/10.48550/ARXIV.2303.08774

Ouyang L, Wu J, Jiang X, Almeida D, Wainwright C L, Mishkin P, Zhang C, Agarwal S, Slama K, Ray A, Schulman J, Hilton J, Kelton F, Miller L, Simens M, Askell A, Welinder

P, Christiano P F, Leike J, Lowe R (2022) Training language models to follow instructions with human feedback. In: Koyejo S, Mohamed S, Agarwal A, Belgrave D, Cho K, Oh A (Eds.), Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022. Available from: http://papers.nips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html (Accessed on: September 13, 2024).

Palmer M, Gildea D, Xue N (2010) Semantic Role Labeling. Springer Cham. https://doi.org/10.1007/978-3-031-02135-0

Palta S, Rudinger R (2023) FORK: A Bite-Sized Test Set for Probing Culinary Cultural Biases in Commonsense Reasoning Models. In: Rogers A, Boyd-Graber J, Okazaki N (Eds.), Findings of the Association for Computational Linguistics: ACL 2023. Association for Computational Linguistics, Toronto, Canada, 9952–9962. https://doi.org/10.18653/v1/2023.findings-acl.631

Pantel P, Pennacchiotti M (2006) Espresso: Leveraging Generic Patterns for Automatically Harvesting Semantic Relations. In: Calzolari N, Cardie C, Isabelle P (Eds.), 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. The Association for Computer Linguistics. https://doi.org/10.3115/1220175.1220190

Panton K, Matuszek C, Lenat D, Schneider D, Witbrock M, Siegel N, Shepard B (2006) Common Sense Reasoning – From Cyc to Intelligent Assistant. In: Cai Y, Abascal J (Eds.), Ambient Intelligence in Everyday Life: Foreword by Emile Aarts. Springer Berlin Heidelberg, Berlin, Heidelberg, 1–31. https://doi.org/10.1007/11825890_1

Pasca M, Durme B V (2008) Weakly-Supervised Acquisition of Open-Domain Classes and Class Attributes from Web Documents and Query Logs. In: McKeown K R, Moore J D, Teufel S, Allan J, Furui S (Eds.), ACL 2008, Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics, June 15-20, 2008, Columbus, Ohio, USA. The Association for Computer Linguistics, 19–27. Available from: https://aclanthology.org/P08-1003/ (Accessed on: September 13, 2024).

Pellissier Tanon T, Weikum G, Suchanek F (2020) YAGO 4: A Reason-able Knowledge Base. In: Harth A, Kirrane S, Ngonga Ngomo A-C, Paulheim H, Rula A, Gentile A L, Haase P, Cochez M (Eds.), The Semantic Web. Springer International Publishing, Cham, 583–596. https://doi.org/10.1007/978-3-030-49461-2_34

Petroni F, Lewis P S H, Piktus A, Rocktäschel T, Wu Y, Miller A H, Riedel S (2020) How Context Affects Language Models' Factual Predictions. In: Das D, Hajishirzi H, McCallum A, Singh S (Eds.), Conference on Automated Knowledge Base Construction, AKBC 2020, Virtual, June 22-24, 2020. https://doi.org/10.24432/C5201W

Petroni F, Rocktäschel T, Riedel S, Lewis P, Bakhtin A, Wu Y, Miller A (2019) Language Models as Knowledge Bases?. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Association for Computational Linguistics, Hong Kong, China, 2463–2473. https://doi.org/10.18653/v1/D19-1250

Ponzetto S P, Strube M (2011) Taxonomy induction based on a collaboratively built knowledge repository. Artif. Intell. 175: 1737–1756. https://doi.org/10.1016/J.ARTINT.2011.01.003

Pradeepani M K T, Jayawardena C, Rajapaksha U U S (2022) Adding Commonsense to Robotic Application Using Ontology-Based Model Retraining. In: 2022 International Research Conference on Smart Computing and Systems Engineering (SCSE). IEEE, 157–164. https://doi.org/10.1109/SCSE56529.2022.9905090

Prasad R, Dinesh N, Lee A, Miltsakaki E, Robaldo L, Joshi A, Webber B (2008) The Penn Discourse TreeBank 2.0. In: Calzolari N, Choukri K, Maegaard B, Mariani J, Odijk J, Piperidis S, Tapias D (Eds.), Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08). European Language Resources Association (ELRA), Marrakech, Morocco. Available from: http://www.lrec-conf.org/proceedings/lrec2008/pdf/754_paper.pdf (Accessed on: September 13, 2024).

Prasojo R E, Kacimi M, Nutt W (2018) StuffIE: Semantic Tagging of Unlabeled Facets Using Fine-Grained Information Extraction. In: Proceedings of the 27th ACM International Conference on Information and Knowledge Management. CIKM '18. Association for Computing Machinery, Torino, Italy, 467–476. https://doi.org/10.1145/3269206.3271812

Pratama F A, Mastrogiovanni F, Chong N Y (2014) An integrated epigenetic robot architecture via context-influenced long-term memory. In: 4th International Conference on Development and Learning and on Epigenetic Robotics, ICDL-EPIROB 2014, Genoa, Italy, October 13-16, 2014. IEEE, 68–74. https://doi.org/10.1109/DEVLRN.2014.6982956

Putri R A, Hong G, Myaeng S-H (2019) Aligning Open IE Relations and KB Relations using a Siamese Network Based on Word Embedding. In: Dobnik S, Chatzikyriakidis S, Demberg V (Eds.), Proceedings of the 13th International Conference on Computational Semantics - Long Papers. Association for Computational Linguistics, Gothenburg, Sweden, 142–153. https://doi.org/10.18653/v1/W19-0412

Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I, others (2019) Language models are unsupervised multitask learners. OpenAI blog 1: 9–10. Available from: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf (Accessed on: September 8, 2024).

Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, Zhou Y, Li W, Liu P J (2020) Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. J. Mach. Learn. Res. 21: 1–67. Available from: https://jmlr.org/papers/v21/20-074.html (Accessed on: September 13, 2024).

Rajpurkar P, Jia R, Liang P (2018) Know What You Don't Know: Unanswerable Questions for SQuAD. In: Gurevych I, Miyao Y (Eds.), Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers. Association for Computational Linguistics, 784–789. https://doi.org/10.18653/V1/P18-2124

Ravi S, Chinchure A, Sigal L, Liao R, Shwartz V (2023) VLC-BERT: Visual Question Answering with Contextualized Commonsense Knowledge. In: IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2023, Waikoloa, HI, USA, January 2-7, 2023. IEEE, 1155–1165. https://doi.org/10.1109/WACV56688.2023.00121

Reimers N, Gurevych I (2019) Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Association for Computational Linguistics, Hong Kong, China, 3982–3992. https://doi.org/10.18653/v1/D19-1410

Romero J, Razniewski S, Pal K, Z. Pan J, Sakhadeo A, Weikum G (2019) Commonsense Properties from Query Logs and Question Answering Forums. In: Proceedings of the 28th ACM International Conference on Information and Knowledge Management. CIKM '19. Association for Computing Machinery, Beijing, China, 1411–1420. https://doi.org/10.1145/3357384.3357955

Ross J, Irani L, Silberman M S, Zaldivar A, Tomlinson B (2010) Who Are the Crowdworkers? Shifting Demographics in Mechanical Turk. In: CHI '10 Extended Abstracts on Human Factors in Computing Systems. CHI EA '10. Association for Computing Machinery, Atlanta, Georgia, USA, 2863–2872. https://doi.org/10.1145/1753846.1753873

Russell B C, Torralba A, Murphy K P, Freeman W T (2008) LabelMe: A Database and Web-Based Tool for Image Annotation. Int. J. Comput. Vis. 77: 157–173. https://doi.org/10.1007/S11263-007-0090-8

Sakaguchi K, Bras R L, Bhagavatula C, Choi Y (2021) WinoGrande: an adversarial winograd schema challenge at scale. Commun. ACM 64: 99–106. https://doi.org/10.1145/3474381

Sap M, Le Bras R, Allaway E, Bhagavatula C, Lourie N, Rashkin H, Roof B, Smith N A, Choi Y (2019a) ATOMIC: An Atlas of Machine Commonsense for If-Then Reasoning. In: Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence. AAAI'19. AAAI Press, Honolulu, Hawaii, USA. https://doi.org/10.1609/aaai.v33i01.33013027

Sap M, Rashkin H, Chen D, Le Bras R, Choi Y (2019b) Social IQa: Commonsense Reasoning about Social Interactions. In: Inui K, Jiang J, Ng V, Wan X (Eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Association for Computational Linguistics, Hong Kong, China, 4463–4473. https://doi.org/10.18653/v1/D19-1454

Schubert L (2002) Can we derive general world knowledge from texts?. In: Proceedings of the Second International Conference on Human Language Technology Research. HLT '02. Morgan Kaufmann Publishers Inc., San Diego, California, 94–97. Available from: https://www.cs.rochester.edu/~schubert/papers/can-we-derive-hlt02.pdf (Accessed on: September 8, 2024).

Schütz A (1944) The Stranger: An Essay in Social Psychology. American Journal of Sociology 49: 499–507. Available from: http://www.jstor.org/stable/2771547 (Accessed on: September 5, 2024).

Seitner J, Bizer C, Eckert K, Faralli S, Meusel R, Paulheim H, Ponzetto S P (2016) A Large DataBase of Hypernymy Relations Extracted from the Web. In: Calzolari N, Choukri K, Declerck T, Goggi S, Grobelnik M, Maegaard B, Mariani J, Mazo H, Moreno A, Odijk J, Piperidis S (Eds.), Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016. European Language Resources Association (ELRA). Available from: http://www.lrec-conf.org/proceedings/lrec2016/summaries/204.html (Accessed on: September 13, 2024).

Shwartz V (2022) Good Night at 4 pm?! Time Expressions in Different Cultures. In: Findings of the Association for Computational Linguistics: ACL 2022. Association for Computational Linguistics, Dublin, Ireland, 2842–2853. https://doi.org/10.18653/v1/2022.findings-acl.224

Shwartz V, Waterson C (2018) Olive Oil is Made of Olives, Baby Oil is Made for Babies: Interpreting Noun Compounds Using Paraphrases in a Neural Model. In: Walker M, Ji H, Stent A (Eds.), Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). Association for Computational Linguistics, New Orleans, Louisiana, 218–224. https://doi.org/10.18653/v1/N18-2035

Siddique A, Jamour F, Xu L, Hristidis V (2021) Generalized Zero-shot Intent Detection via Commonsense Knowledge. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '21. Association for Computing Machinery, Virtual Event, Canada, 1925–1929. https://doi.org/10.1145/3404835.3462985

Singh P, Lin T, Mueller E T, Lim G, Perkins T, Li Zhu W (2002) Open Mind Common Sense: Knowledge Acquisition from the General Public. In: Meersman R, Tari Z (Eds.), On the Move to Meaningful Internet Systems 2002: CoopIS, DOA, and ODBASE. Springer Berlin Heidelberg, Berlin, Heidelberg, 1223–1237. https://doi.org/10.1007/3-540-36124-3_77

Singhania S, Nguyen T-P, Razniewski S (Eds.) (2022) 3274 Proceedings of the Semantic Web Challenge on Knowledge Base Construction from Pre-trained Language Models 2022 co-located with the 21st International Semantic Web Conference (ISWC2022), Virtual Event, Hanghzou, China, October 2022. CEUR-WS.org. Available from: http://ceur-ws.org/Vol-3274 (Accessed on: September 13, 2024).

Snow R, Jurafsky D, Ng A Y (2006) Semantic Taxonomy Induction from Heterogenous Evidence. In: Calzolari N, Cardie C, Isabelle P (Eds.), 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. The Association for Computer Linguistics. https://doi.org/10.3115/1220175.1220276

Soderland S, Gilmer J, Bart R, Etzioni O, Weld D S (2013) Open Information Extraction to KBP Relations in 3 Hours. In: Proceedings of the Sixth Text Analysis Conference, TAC 2013, Gaithersburg, Maryland, USA, November 18-19,

2013. NIST. Available from: https://tac.nist.gov/publications/2013/participant.papers/UWashington.TAC2013.proceedings.pdf (Accessed on: September 13, 2024).

Sparck Jones K (1988) A Statistical Interpretation of Term Specificity and Its Application in Retrieval. In: Document Retrieval Systems. Taylor Graham Publishing, GBR, 132–142. https://doi.org/10.1108/eb026526

Speer R (2007) Open Mind Commons: An Inquisitive Approach to Learning Common Sense. In: Proceedings of the Workshop on Common Sense and Interactive Applications. Available from: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=e82919ef4ba11314affdecfcc9a6d72170847e05 (Accessed on: September 13, 2024).

Speer R, Chin J, Havasi C (2017) ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. In: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence. AAAI'17. AAAI Press, San Francisco, California, USA, 4444–4451. https://doi.org/10.1609/aaai.v31i1.11164

Speer R, Havasi C (2013) ConceptNet 5: A Large Semantic Network for Relational Knowledge. In: Gurevych I, Kim J (Eds.), The People's Web Meets NLP: Collaboratively Constructed Language Resources. Springer Berlin Heidelberg, Berlin, Heidelberg, 161–176. https://doi.org/10.1007/978-3-642-35085-6_6

Stanovsky G, Michael J, Zettlemoyer L, Dagan I (2018) Supervised Open Information Extraction. In: Walker M, Ji H, Stent A (Eds.), Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). Association for Computational Linguistics, New Orleans, Louisiana, 885–895. https://doi.org/10.18653/v1/N18-1081

Suchanek F M, Kasneci G, Weikum G (2007) YAGO: A Core of Semantic Knowledge. In: Proceedings of the 16th International Conference on World Wide Web. WWW '07. Association for Computing Machinery, Banff, Alberta, Canada, 697–706. https://doi.org/10.1145/1242572.1242667

Talmor A, Herzig J, Lourie N, Berant J (2019) CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge. Burstein J, Doran C, Solorio T (Eds.). : 4149–4158. https://doi.org/10.18653/V1/N19-1421

Talmor A, Yoran O, Bras R L, Bhagavatula C, Goldberg Y, Choi Y, Berant J (2021) CommonsenseQA 2.0: Exposing the Limits of AI through Gamification. In: Vanschoren J, Yeung S-K (Eds.), Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual. Available from: https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/3ef815416f775098fe977004015c6193-Abstract-round1.html (Accessed on: September 13, 2024).

Tandon N (2016) Commonsense knowledge acquisition and applications. https://doi.org/10.22028/D291-26668

Tandon N, Hariman C, Urbani J, Rohrbach A, Rohrbach M, Weikum G (2016) Commonsense in parts: mining part-whole relations from the web and image tags. In: Proceedings of the

Thirtieth AAAI Conference on Artificial Intelligence. AAAI'16. AAAI Press, Phoenix, Arizona, 243–250. https://doi.org/10.1609/aaai.v30i1.9992

Tandon N, Melo G de, De A, Weikum G (2015a) Knowlywood: Mining Activity Knowledge From Hollywood Narratives. In: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. CIKM '15. Association for Computing Machinery, Melbourne, Australia, 223–232. https://doi.org/10.1145/2806416.2806583

Tandon N, Melo G de, Suchanek F, Weikum G (2014a) WebChild: Harvesting and Organizing Commonsense Knowledge from the Web. In: Proceedings of the 7th ACM International Conference on Web Search and Data Mining. WSDM '14. Association for Computing Machinery, New York, New York, USA, 523–532. https://doi.org/10.1145/2556195.2556245

Tandon N, Melo G de, Weikum G (2014b) Acquiring Comparative Commonsense Knowledge from the Web. In: Brodley C E, Stone P (Eds.), Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, July 27 -31, 2014, Québec City, Québec, Canada. AAAI Press, 166–172. https://doi.org/10.1609/AAAI.V28I1.8735

Tandon N, Melo G de, Weikum G (2017) WebChild 2.0 : Fine-Grained Commonsense Knowledge Distillation. In: Bansal M, Ji H (Eds.), Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, System Demonstrations. Association for Computational Linguistics, 115–120. https://doi.org/10.18653/V1/P17-4020

Tandon N, Weikum G, Melo G d, De A (2015b) Lights, Camera, Action: Knowledge Extraction from Movie Scripts. In: Proceedings of the 24th International Conference on World Wide Web. WWW '15 Companion. Association for Computing Machinery, Florence, Italy, 127–128. https://doi.org/10.1145/2740908.2742756

Tomasello M (2014) A Natural History of Human Thinking. Harvard University Press, Cambridge, MA and London, England. https://doi.org/doi:10.4159/9780674726369

Töberg J-P, Ngomo A-C N, Beetz M, Cimiano P (2024) Commonsense knowledge in cognitive robotics: a systematic literature review. Frontiers Robotics AI 11. https://doi.org/10.3389/FROBT.2024.1328934

Varshney D, Prabhakar A, Ekbal A (2022) Commonsense and Named Entity Aware Knowledge Grounded Dialogue Generation. Carpuat M, Marneffe M-C de, Meza Ruiz I V (Eds.). : 1322–1335. https://doi.org/10.18653/v1/2022.naacl-main.95

Vrandečić D, Krötzsch M (2014) Wikidata: a free collaborative knowledgebase. Commun. ACM 57: 78–85. https://doi.org/10.1145/2629489

Wang P, Peng N, Ilievski F, Szekely P, Ren X (2020) Connecting the Dots: A Knowledgeable Path Generator for Commonsense Question Answering. In: Cohn T, He Y, Liu Y (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2020. Association for Computational Linguistics, Online, 4129–4140. https://doi.org/10.18653/v1/2020.findings-emnlp.369

Wang P, Wu Q, Shen C, Dick A R, Hengel A van den (2017) Explicit Knowledge-based Reasoning for Visual Question Answering. In: Sierra C (Ed.), Proceedings of the Twenty-Sixth

International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017. ijcai.org, 1290–1296. https://doi.org/10.24963/IJCAI.2017/179

Wang S, Thompson L, Iyyer M (2021) Phrase-BERT: Improved Phrase Embeddings from BERT with an Application to Corpus Exploration. In: Moens M-F, Huang X, Specia L, Yih S W-t (Eds.), Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 10837–10851. https://doi.org/10.18653/v1/2021.emnlp-main.846

Wang Y, Zhang S, Lee J (2019) Bridging Commonsense Reasoning and Probabilistic Planning via a Probabilistic Action Language. Theory Pract. Log. Program. 19: 1090–1106. https://doi.org/10.1017/S1471068419000371

Ward J H (1963) Hierarchical Grouping to Optimize an Objective Function. Journal of the American Statistical Association 58: 236–244. https://doi.org/10.1080/01621459.1963.10500845

Weikum G, Dong X L, Razniewski S, Suchanek F M (2021) Machine Knowledge: Creation and Curation of Comprehensive Knowledge Bases. Found. Trends Databases 10: 108–490. https://doi.org/10.1561/1900000064

West P, Bhagavatula C, Hessel J, Hwang J D, Jiang L, Bras R L, Lu X, Welleck S, Choi Y (2022) Symbolic Knowledge Distillation: from General Language Models to Commonsense Models. In: Carpuat M, Marneffe M-C de, Ruíz I V M (Eds.), Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022. Association for Computational Linguistics, 4602–4625. https://doi.org/10.18653/V1/2022.NAACL-MAIN.341

Williams A, Nangia N, Bowman S (2018) A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). Association for Computational Linguistics, New Orleans, Louisiana, 1112–1122. https://doi.org/10.18653/v1/N18-1101

Wu S, Li Y, Zhang D, Zhou Y, Wu Z (2020) Diverse and Informative Dialogue Generation with Context-Specific Commonsense Knowledge Awareness. In: Jurafsky D, Chai J, Schluter N, Tetreault J (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Online, 5811–5820. https://doi.org/10.18653/v1/2020.acl-main.515

Wu S, Li Y, Zhang D, Zhou Y, Wu Z (2020) TopicKA: Generating Commonsense Knowledge-Aware Dialogue Responses Towards the Recommended Topic Fact. In: Bessiere C (Ed.), Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20. International Joint Conferences on Artificial Intelligence Organization, 3766–3772. https://doi.org/10.24963/ijcai.2020/521

Wu W, Li H, Wang H, Zhu K Q (2012) Probase: a probabilistic taxonomy for text understanding. In: Candan K S, Chen Y, Snodgrass R T, Gravano L, Fuxman A (Eds.), Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2012, Scottsdale, AZ, USA, May 20-24, 2012. ACM, 481–492. https://doi.org/10.1145/2213836.2213891

Xu F F, Lin B Y, Zhu K (2018) Automatic Extraction of Commonsense LocatedNear Knowledge. In: Gurevych I, Miyao Y (Eds.), Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Association for Computational Linguistics, Melbourne, Australia, 96–101. https://doi.org/10.18653/v1/P18-2016

Yang K, Zhang T, Ji S, Ananiadou S (2023) A Bipartite Graph is All We Need for Enhancing Emotional Reasoning with Commonsense Knowledge. In: Proceedings of the 32nd ACM International Conference on Information and Knowledge Management. CIKM '23. Association for Computing Machinery, Birmingham, United Kingdom, 2917–2927. https://doi.org/10.1145/3583780.3614758

Yang P, Li L, Luo F, Liu T, Sun X (2019) Enhancing Topic-to-Essay Generation with External Commonsense Knowledge. In: Korhonen A, Traum D, Màrquez L (Eds.), Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Florence, Italy, 2002–2012. https://doi.org/10.18653/v1/P19-1193

Yang X, Tang K, Zhang H, Cai J (2019) Auto-Encoding Scene Graphs for Image Captioning. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019. Computer Vision Foundation / IEEE, 10685–10694. https://doi.org/10.1109/CVPR.2019.01094

Yao Y, Yu T, Zhang A, Li M, Xie R, Weber C, Liu Z, Zheng H-T, Wermter S, Chua T-S, Sun M (2023) Visually Grounded Commonsense Knowledge Acquisition. In: Williams B, Chen Y, Neville J (Eds.), Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023. AAAI Press, 6583–6592. https://doi.org/10.1609/AAAI.V37I5.25809

Yin D, Bansal H, Monajatipoor M, Li L H, Chang K-W (2022) GeoMLAMA: Geo-Diverse Commonsense Probing on Multilingual Pre-Trained Language Models. In: Goldberg Y, Kozareva Z, Zhang Y (Eds.), Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022. Association for Computational Linguistics, 2039–2055. https://doi.org/10.18653/V1/2022.EMNLP-MAIN.132

Yin D, Li L H, Hu Z, Peng N, Chang K-W (2021) Broaden the Vision: Geo-Diverse Visual Commonsense Reasoning. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2115–2129. https://doi.org/10.18653/v1/2021.emnlp-main.162

Yin W, Hay J, Roth D (2019) Benchmarking Zero-shot Text Classification: Datasets, Evaluation and Entailment Approach. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Association for Computational Linguistics, Hong Kong, China, 3914–3923. https://doi.org/10.18653/v1/D19-1404

Zareian A, Karaman S, Chang S-F (2020) Bridging Knowledge Graphs to Generate Scene Graphs. In: Vedaldi A, Bischof H, Brox T, Frahm J-M (Eds.), Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXIII. Lecture Notes in Computer Science. Springer, 606–623. https://doi.org/10.1007/978-3-030-58592-1_36

Zellers R, Bisk Y, Farhadi A, Choi Y (2019) From Recognition to Cognition: Visual Commonsense Reasoning. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019. Computer Vision Foundation / IEEE, 6720–6731. https://doi.org/10.1109/CVPR.2019.00688

Zellers R, Bisk Y, Schwartz R, Choi Y (2018) SWAG: A Large-Scale Adversarial Dataset for Grounded Commonsense Inference. In: Riloff E, Chiang D, Hockenmaier J, Tsujii J (Eds.), Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018. Association for Computational Linguistics, 93–104. https://doi.org/10.18653/V1/D18-1009

Zellers R, Holtzman A, Bisk Y, Farhadi A, Choi Y (2019) HellaSwag: Can a Machine Really Finish Your Sentence?. In: Korhonen A, Traum D R, Màrquez L (Eds.), Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers. Association for Computational Linguistics, 4791–4800. https://doi.org/10.18653/V1/P19-1472

Zhang D, Mukherjee S, Lockard C, Dong L, McCallum A (2019) OpenKI: Integrating Open Information Extraction and Knowledge Bases with Relation Inference. In: Burstein J, Doran C, Solorio T (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Association for Computational Linguistics, Minneapolis, Minnesota, 762–772. https://doi.org/10.18653/v1/N19-1083

Zhang H, Khashabi D, Song Y, Roth D (2020a) TransOMCS: From Linguistic Graphs to Commonsense Knowledge. In: Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20. International Joint Conferences on Artificial Intelligence Organization, California, USA, 4004–4010. https://doi.org/10.24963/ijcai.2020/554

Zhang H, Liu X, Pan H, Song Y, Leung C W-K (2020b) ASER: A Large-scale Eventuality Knowledge Graph. In: Huang Y, King I, Liu T-Y, Steen M van (Eds.), WWW '20: The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020. ACM / IW3C2, 201–211. https://doi.org/10.1145/3366423.3380107

Zhang H, Liu Z, Xiong C, Liu Z (2020c) Grounded Conversation Generation as Guided Traverses in Commonsense Knowledge Graphs. In: Jurafsky D, Chai J, Schluter N, Tetreault

J (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Online, 2031–2043. https://doi.org/10.18653/v1/2020.acl-main.184

Zhang S, Rudinger R, Duh K, Van Durme B (2017) Ordinal Common-sense Inference. Lee L, Johnson M, Toutanova K (Eds.). Transactions of the Association for Computational Linguistics 5: 379–395. https://doi.org/10.1162/tacl_a_00068

Zhang S, Stone P (2015) CORPP: Commonsense Reasoning and Probabilistic Planning, as Applied to Dialog with a Mobile Robot. In: Bonet B, Koenig S (Eds.), Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA. AAAI Press, 1394–1400. https://doi.org/10.1609/AAAI.V29I1.9385

Zhang S, Jiang X, Yang Z, Wan T, Qin Z (2022) Reasoning with Multi-Structure Commonsense Knowledge in Visual Dialog. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2022, New Orleans, LA, USA, June 19-20, 2022. IEEE, 4599–4608. https://doi.org/10.1109/CVPRW56347.2022.00506

Zhang Y, Tian G, Lu J, Zhang M, Zhang S (2019) Efficient Dynamic Object Search in Home Environment by Mobile Robot: A Priori Knowledge-Based Approach. IEEE Trans. Veh. Technol. 68: 9466–9477. https://doi.org/10.1109/TVT.2019.2934509

Zhang Y, Li Y, Cui L, Cai D, Liu L, Fu T, Huang X, Zhao E, Zhang Y, Chen Y, Wang L, Luu A T, Bi W, Shi F, Shi S (2023) Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models. CoRR. https://doi.org/10.48550/ARXIV.2309.01219

Zhou H, Young T, Huang M, Zhao H, Xu J, Zhu X (2018) Commonsense Knowledge Aware Conversation Generation with Graph Attention. In: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18. International Joint Conferences on Artificial Intelligence Organization, 4623–4629. https://doi.org/10.24963/ijcai.2018/643

Zhou P, Gopalakrishnan K, Hedayatnia B, Kim S, Pujara J, Ren X, Liu Y, Hakkani-Tur D (2022) Think Before You Speak: Explicitly Generating Implicit Commonsense Knowledge for Response Generation. In: Muresan S, Nakov P, Villavicencio A (Eds.), Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Dublin, Ireland, 1237–1252. https://doi.org/10.18653/v1/2022.acl-long.88

Ziems C, Dwivedi-Yu J, Wang Y-C, Halevy A, Yang D (2023) NormBank: A Knowledge Bank of Situational Social Norms. In: Rogers A, Boyd-Graber J, Okazaki N (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Toronto, Canada, 7756–7776. https://doi.org/10.18653/v1/2023.acl-long.429