(Symposium "Computercorpus der serbokroatischen Sprache", Belgrad 14.-18.12.1981)

LEGAL DOCUMENTATION WITH THE COMPUTER AIDED INDEXING SYSTEM CTX

Harald H. Zimmermann

Introduction

During the last twenty years, linguistic data processing mainly has been seen as a tool to develop linguistic regularities (or detect irregularities) of a given natural language, especially to handle large textual databases ("Corpora") . A second motivation to use a computer was to test some theories or models of a language system (or a part of it) using a simulation program.

As a result of both strategies, the "Saarbrücken Text Analysis System" has been implemented. At present, a very large lexical database is available to analyse written German texts morphologically and syntactically.

On the other hand, the development of large textual databases within different fields (e.g. law, patent specifications, medicine) is increasing rapidly. Therefore, a computer aided indexing system ('Computergestützte Texterschließung: CTX’) has been developed at Saarbrücken University to improve the natural language oriented access to textual data ("free text") applying linguistic strategies to information retrieval processes.

Main results of a feasibility study, in the field of legal documentation are presented.

Within the research project to be described an attempt was made to build a bridge between a computer information system and the casual user. The CTX System is a reference retrieval system: As an 'answer' to a 'query', the system refers (where possible) to one or several documents in which the question is dealt with.

In contrast to former systems, the 'query' can, however, also be formulated in the form of a 'natural-language problem description' - as a series of sentences or subsentences. These search units are processed by linguistic and partly statistical machine procedures in the same way as were the documents in the database. Thus, a mutual adaptation of document characterization (indexing) and document identification (retrieval) is achieved at the same time.

In addition, paraphrases are constructed and thesaurus relations are compiled. During the automatic text analysis an attempt is made to resolve all syntactic and - in the extended version CTX-II - semantic ambiguities.

I. Goals

Everybody knows that methodical linguistic features are also important in working processes which are not yet mechanized, for instance in the realization of artificial documentation languages, or in the development of terminologies or thesauri. For practical reasons, however, the use of computational linguistic procedures will be, in the following, the main point when discussing the use of linguistic methods in processes of information and documentation.

These procedures can be aimed at:

1. main developments in the direction of automatic "understanding" of natural language expressions by the computer (AI-systems);
2. automatic translation of texts/documents (on "any" level) so that either the "human translator" is aided in his work or that the computer (by intellectually developed dictionaries) furnishes sufficiently informative (raw) translations;
3. improvement of the methods of indexing in order to specify the text processing and to simplify the retrieval.

Considering the complexity of the documents, the first attempt (AI) seems to be the less suitable for furnishing an appreciable easing for the informational practice. It cannot be denied, however, that practically relevant results can be achieved by this method.

By now, the automatic translation has got a chance to be really applied. We refer to the automatic translation of the INSPEC database from English to Japanese (Nagao, Kyoto); the EC uses the SYSTRAN system for some special subjects for making raw translations in order to support the human translators (reduction of cost). The development of a translation system, however, strongly depends on the availability of extensive and highly qualified multilingual dictionaries. This shows that considerable efforts will be necessary before achieving the possibility of producing translations "just as on a production line" (which will not be "perfect" at all anyway). At Saarbrücken University, a system for "informative translation" (ITS) - i.e. raw, but understandable translation - of data bases is under development.

The procedure of automatic indexing - referred to a natural language – resembles automatic translation. Ideally, the procedure consists (as well as the automatic translation) of a linguistic analysis to find the basic forms (morphologic analysis), connection of words (syntactical analysis), semantic disambiguation (semantic analysis), and the construction of so-called "descriptors" in a certain "canonized" form (synthesis). The expenditure, however, is highly reduced compared to automatic translation. This is due to the fact of being monolingual. The component of synthesis is, on the surface, at most a “simple" transformation of deeper structures into an (artificial) language of documentation.

For this reason, automatic indexing procedures promise the most successful application in practical information services, especially in textual database systems.

In the Federal Republic of Germany, two main ways in automatic indexing are in the state of research and development:

1. According to intellectual procedures, few - extremely relevant - keywords are eliminated from texts by automatic statistical and linguistic methods. This procedure was developed at the technical University of Darmstadt on the basis of English texts. It is being tested on a larger, practical basis in collaboration with the FIZ energy, mathematics, and physics (FIZ 4) at Karlsruhe. First experimental results show that, with this method, one can obtain results which are almost as good as the results of an intellectual indexing /1/.
2. The way (described in (1)) of concentrating information implies, according to the intellectual indexing, considerable loss of information. As an alternative, one has to find methods which

   - represent the content of (specialized) texts completely;

   - supply results structured in a way so that by suitable retrieval systems according to the application of boolean operators a concrete search for information becomes possible.

The research works at the University of the Sarre followed this way. In the following, the procedure will be described as a summary.

2. Computer-aided text processing system (CTX)

The computer-aided text processing system CTX represents the result of many years of research work at the university concerning information software. The models for language analysis developed by basic research work have been further developed - for practical application - into system components for text processing. building up upon a text processing system for words and sentences, key words (descriptors) with regard to the form and the contents are supplied for German texts/documents.
Legal documents of the subject "data protection" were the base for the first labtest of CTX. Presently, the system in its standard version CTX-I (without semantic disambiguation) is tested – near to a practical application - in different areas. CTX has a modular structure and interfaces which are independent from the type of the computer. Therefore, it may be integrated into different information retrieval systems or can be added as an indexing component.

For the processing of texts in natural language, CTX has the following tasks:

- The efficient word forms extracted from the text are automatically reduced (by means of a general dictionary with more than 140 000 stems) to their basic form.

  Examples

text word	basic form
Vorzüge	VORZUG
trat	TRETEN
trifft ... zu	ZUTREFFEN

- In addition, compounds are decomposed into efficient words and disambiguated, if possible. (only CTX-II)

  Example: Persönlichkeitssphäre

  part: PERSÖNLICHKEIT
  part: SPHÄRE

- The ambiguity of text words is shown; ambiguous words can be disambiguated (computer-aided) by means of the context.
  Examples.

  ... in der Praxis der Datenverarbeitung ...

PRAXIS (prakt. Vorgehen)
(in contrary to "Arztpraxis")

- Expressions of two or more words are dentified by means of a suitable system of rules. (only CTX II)

  Examples:
  tritt in Kraft IN KRAFT TRETEN
  <u>personenbezogene</u>
  durch das Gesetz
  geschützte <u>Daten</u> PERSONENBEZOGENE DATEN

- Special simple and complex descriptors are identified by using identified syntactic structures.

  Examples: .. modernen Industriestaaten …

  (simple) INDUSTRIESTAAT
  (complex) MODERNER INDUSTRIESTAAT

- Words without sense ("'functional" words: THAT, BUT, AND) are eliminated.

The realization of an efficient dictionary component and the (mostly) automatic improvement of the dictionary was of extreme importance. Common words are registered in a morpho-syntactical and in a semantic dictionary. Special words are identified and described in a special dictionary and in a dictionary of semantic relations ("thesaurus"). For special fields, the user may determine the specific structure and the use of words.

The SUSY translation system (developed in Saarbrücken) is the basis of the text processing component; the analysis component of SUSY was integrated in CTX. The linguistic analysis is mainly working on sentence level .

If necessary, the system can be integrated into retrieval components of a special computer. Thus, the analysis and the processing of the description of a problem in natural language during the retrieval becomes possible by using the same rules as for the text analysis (indexing of the documents). This allows a simple formal adaption of the words used in the question to the indexed words of the texts/documents.

In the following, a short description may explain the procedure of the text exploitation in its components "text analysis" and "supply of descriptors".

INPUT OF THE TEXT

- Input of the (special) text in computer-compatible form (perhaps adaption to the input interface of the system). Preparation for the following processing, sentence after sentence.

TEXT ANALYSIS

- Determination of the possible basic forms (if necessary, with decomposition of the compounds) by means of a morpho-syntactical dictionary.
- Analysis of syntactical ambiguities.
- Decomposition of the sentence in potential segments (for instance subordinate clauses, co-ordinate sentences).
- Determination and analysis of complex syntactical structures (for instance groups of nouns, verbal groups).

SEMANTIC ANALYSIS (only CTX-II)

- Reduction of ambiguities of words by means of a system of semantic rules (the semantic dictionary still under development)

The text word forms are reduced to basic forms which are partly already definite, the structure of the sentences is determinated.

SUPPLY OF DESCRIPTORS

- Key words, complex expressions, and information concerning their syntactical structures are made available. By means of a specialized dictionary, the ambiguities which have not yet been reduced are analyzed by a specialized weighting procedure.

Result of the supply of descriptors: Descriptors with regard to the form and the contents with additional information concerning the structure, independent of the system.

3. Practical applications of the CTX system

By now, the system has been applied in several large areas. First, at the university an information system was developed and implemented for legal documentation within the area of “data protection”. This model served mainly for the clarification of principal procedures, for the construction of a test database (according to the JURIS implementation) and the principal anticipation of potential problems of the users. The database constructed for this purpose covers by now a quantity of more than 150.000 running words /2/.

The legal documents used during the development of the JUDO system were primarily regulations and legislative drafts concerning data security and privacy. The procedures are, however, equally applicable to court rulings, technical literature (e.g.. abstracts), as well as reports, executive orders, etc.. and even newspaper reports ("semi-technical literature"). In each case, the main focus is on the processing of running plain-text. The structural elements of these documents (e.g.., effective date of laws) is not explicitly handled within the present framework, but such information is nevertheless at least partially integrated into the retrieval system in order to achieve the proper design, with a view to the system user, characterized by a combination of formatted and unformatted (i.e. plain-textual) data.

It has become increasingly clear in the course of the research (especially in examining the documents on data privacy regulation) that, beyond the evaluation of sentence structures in indexing, computerized linguistic analysis affords an additional opportunity to aid in the simplification and optimization of the retrieval process. So far, it has not been possible to implement all the possibilities of evaluating the results of linguistic analysis. A tentative investigation of terms of more than one word (in particular, the Standard Expressions or terminological concretions), as ascertained by (legal) experts with the help of technical dictionaries and the indices of various commentaries, has nevertheless shown that these composite terms occur as certain specific syntactic surface structures, structures which are produced during machine analysis as sub-structures:

Examples:

(<u>Technical Term</u>)	(<u>Syntactic Structure</u>)
criminal act	adjective + noun
right to privacy	noun + prepositional phrase
concurrence of persons concerned	noun + possessive (in German: Genitive NP)

Since the legal experts were given no fixed instructions (which they can hardly be given in practice anyway) as to which descriptors should be directly considered as "lexicalized" units of more than one word, there were a number of border-line cases -- one must also consider that the area of data privacy regulation is relatively new. This problem area can at least be bounded, however, if the possibility is opened up of including syntactic (sub-)structures produced by the automatic sentence analysis in the indexing and (thus also) the retrieval procedures.

Initially, then, all syntactic relations of the form

adjective-noun	
noun-genitive NP	(relator: G)
noun-prepositional phrase	(relator: P)
noun-noun sequence	(relator: K)

are placed at the disposal of the retrieval procedure as "complex descriptors", marked with the appropriate relator (G, K, P), in each case where the automatic analysis recognizes the relations as such. In the semantic orientated system CTX-II, these structures then become retrievable when a Standard Expression was recognized using the semantic lexicon. A corresponding manually assigned relation of synonymy in the thesaurus insures that the user can formulate his inquiry using either variant (Standard Expression or complex descriptor). The system expert (as information distributor) and the technical expert in the particular field of inquiry will prefer the more comfortable Standard Expressions; in the case of an unsuccessful retrieval the complex descriptor would be in order. It is conceivable, of course, to go beyond these "mini-sequences" and to offer (and process) more complex (nominal) structures, for indexing or for automatic extraction.

Of particular interest here is the output interface for the further processing of the results of linguistic analysis in the CTX system.

As has already been touched upon, CTX utilizes only a part of the results produced by the SUSY System. In particular, nominal structures and the resolution of syntactic and semantic ambiguities are focussed upon. Since SUSY is capable of processing German texts with or without capitalization of words (nouns or words at the beginning of sentences), texts can thus be handled in which the nouns (always capitalized in Standard German) within sentences are not marked (where they are not marked, an increasing in processing time results at most -- hardly any increase in the proportion of erroneous analyses can be detected).

The structural (syntactic) analysis produced by SUSY provides, among other things, nominal structures (as mentioned above). Examples from the Federal Data Privacy Act, being focussed upon in the laboratory test (cf. e.g. § 1 of this code) are nominal structures of the following type (Original: German):

-- GOALS OF DATA PRIVACY	(genitive attribute)
-- PERSON-RELATED DATA	("Standard Expression")
-- STORAGE TRANSFER	(nominal sequence)
-- AUTOMATIC PROCEDURES	(adjectival attribute)
-- PROTECTION AGAINST MISUSE	(prepositional attribute)

Such structures are assigned to documents by CTX as complex descriptors. At the same time, single words become accessible as (simple) descriptors.

The constituents of compound words and Standard Expressions which are considered relevant are also assigned as descriptors (via the corresponding lexical entries). The basis for any assignment is the decision of the legal expert: for PERSONENVEREINIGUNG, both PERSON and VEREINIGUNG are assigned, whereby it is marked that these simple terms or constituents stem from the division of a more complex unit.

Contextual disambiguation of semantically ambiguous words (only in the extended version CTX-II) is attempted in two ways:

Primarily linguistic methods are applied. Syntactic rules (e.g. "jemand HANDELT" vs. "es HANDELT sich") and the detailed semantic context (sometimes the presence of individual words as well) are of use (e.g. die ÜBERMITTELNDE STELLE = "institution" vs. "die VERMITTELTE STELLE" = "job").

In a number of cases such rules (as are to be included in the basic system SUSY) are insufficient, where the context -- at present primarily the sentence context -- is too neutral for disambiguation. At best, certain semantic variants can be excluded. In view of all the linguistically possible readings of words, it appears that, in many cases, a resolution of semantic ambiguity cannot be achieved by the sole use of formal rules.

Investigation of sub-disciplines (in this case data privacy law), however, has shown that the majority of potential (i.e. lexical) meanings of a word are, in a particular jargon and from a statistical point of view, never or hardly ever observed. Rather, a small number (often only one, possibly two or three) of topically relevant preferences can be given. In each case where linguistic procedures appear to be less productive or too extensive, this probability criterion can be applied, even if not absolutely reliably, to the process of semantic disambiguation.

Initial experiences with this procedure (supplementary to linguistic methods) have produced promising results. At present, percentage probabilities of occurrence of the various semantic variants are specified. The specification "0% " indicates that this variant occurs so infrequently in texts of this sub-discipline that it may be neglected for purposes of disambiguation.

This method to be used in CTX-II is presently still at a low state of sophistication. A more valuable aid in the disambiguation of descriptors would probably be a network of concepts pertaining to a given subject, even though more extensive manual work may be involved in the construction and maintenance of such a network.

A special procedure (NATURA) in the CTX system allows occasional users to have descriptors determined automatically from a natural-language problem description. These descriptors are produced in accordance with the conventions of the underlying information-retrieval system. The same procedures are used here as were used in processing the documents in the data base. The simple or complex descriptors thus ascertained (and in general, semantically disambiguated) are then applied to a Boolean retrieval procedure. Even at this early stage, (the procedure is not yet integrated in a retrieval system) a number of advantages of this procedure have become apparent: The user is automatically made aware of correct spelling of terms, terms of more than one word – complex descriptors - are identified and terms involving syntactic relations are marked appropriately. Inquiry and document profiles thus closely approximate each other.

The descriptors produced by the analysis of a document, the text of the document itself and, when necessary, additional manually assigned (structural) descriptors are made accessible in a largely implementation-independent (datafile) interface for further processing in an information retrieval system. A GOLEM version has been thus far implemented.

In conceiving the CTX system it was assumed from the very first that particular attention would have to be paid to the maintenance of the descriptor file. Since that time, initial results have indicated that the integration of semantic relations among descriptors can also contribute to syntactic (!) and semantic disambiguation. In particular, CTX has emphasized the use of subject-specific relations among terms. Here, too, a data base structure was selected independent of any particular information-retrieval systems, so as not to be limited by the various restrictions of such information systems (e.g. descriptor length, number of semantic relations allowed, etc.).

The following semantic relations are presently in use in the CTX systems variant for legal documentation:

Association: freely associated
Synonymy: "strictly synonymous"
Orthographical Synonymy: Orthographical variants, plural forms
Quasi-synonymy: "loosely synonymous"
Relation of Abbreviation: Abbreviation/Full form
Part-Whole Relation
Hypernym-Hyponym Relation
Antonymy
Subject-specific Relation: Legal Control/Object of Legal Control

In addition, relations of derivation (e.g. noun-verb) are compiled; special relations of synonymy have been developed for Standard Expressions: Genitive-Attribute Synonymy, Prepositional Synonymy, Adjective-Noun Synonymy and Noun-Sequence Synonymy. Further there are various relations of constituency, e.g. between a compound word and its parts.

In conclusion, the retrieval capabilities of CTX are to be briefly illustrated.

A retrieval request can be made using "simple descriptors"; meaning-numbers at the end of descriptors indicate the semantic variant intended; Standard Expressions (of more than one word) are treated as simple descriptors.

Ambiguous terms occurring in the request without an indication of which semantic reading is meant first leads the user to an "information document" in which the various meanings (together with their identifying numbers) are explained with examples. Complex descriptors (i.e. those based on syntactic relations) are included, along with their relator (G, K or P) as determined in the text; the same applies to the descriptors derived from compound constituents and Standard Expressions. Requests using thesaural relations (S for SYN, ABK, DER; F for QUA, ASS) are possible.

Tests of CTX using the present initial corpus (Federal Data Privacy Act and several regional Data Privacy Regulations) have already shown the possibilities opened up by automatic language analysis. A number of linguistic (but also technical) problems must still be overcome, however; improved analysis systems for the processing of mass textual data -- as already seen in the initial conception of the multilingual, computer-aided European translation system EUROTRA -- will enable further steps to be taken towards improved characterization of documents.

REFERENCES

KNORZ, Gerhard: Die Darmstädter Projekte zur Automatischen Indexierung: WAI and AIR. In: Das Inforum 11 (1981), p. 38-49. (ISSN 0720-3950)

ZIMMERMANN, H., KROUPA, E., KEIL, G.: CTX - Ein Verfahren zur computergestützten Texterschließung. In BMFT-Reihe "Information and Dokumentation": BMFT-FB-ID-003-006 (ISSN 0170-8996)