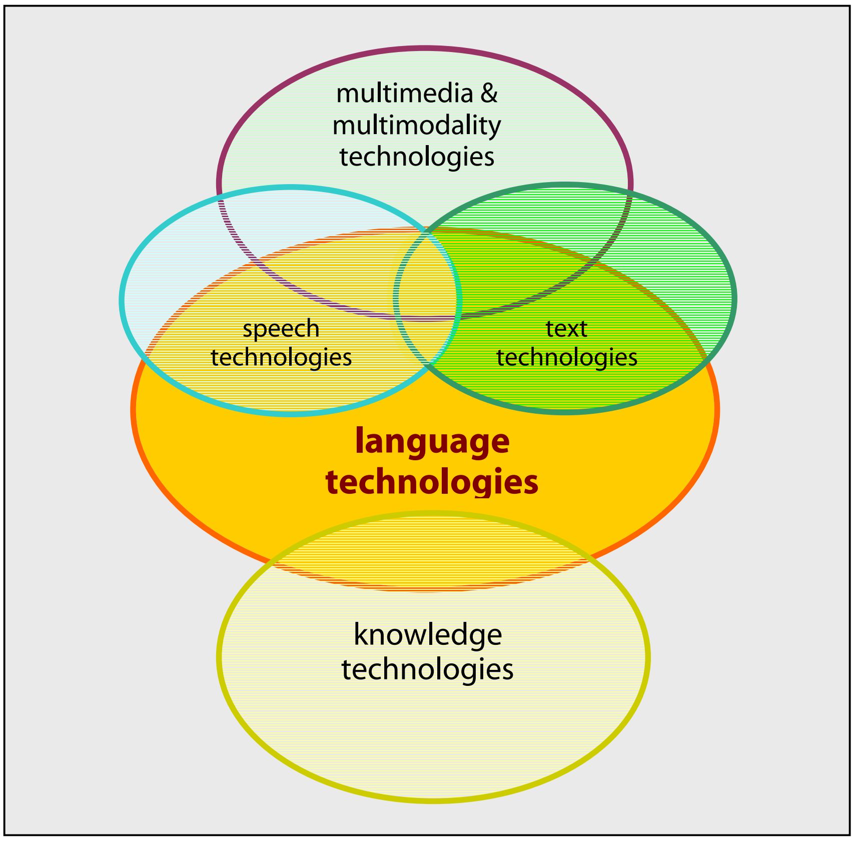
Language Technology
A First Overview

Hans Uszkoreit

1. Scope

Language technologies are information technologies that are specialized for dealing with the most complex information medium in our world: human language. Therefore these technologies are also often subsumed under the term Human Language Technology. Human language occurs in spoken and written form. Whereas speech is the oldest and most natural mode of language communication, complex information and most of human knowledge is maintained and transmitted in written texts. Speech and text technologies process or produce language in these two modes of realization. But language also has aspects that are shared between speech and text such as dictionaries, most of grammar and the meaning of sentences. Thus large parts of language technology cannot be subsumed under speech and text technologies. Among those are technologies that link language to knowledge. We do not know how language, knowledge and thought are represented in the human brain. Nevertheless, language technology had to create formal representation systems that link language to concepts and tasks in the real world. This provides the interface to the fast growing area of knowledge technologies.

In our communication we mix language with other modes of communication and other information media. We combine speech with gesture and facial expressions. Digital texts are combined with pictures and sounds. Movies may contain language and spoken and written form. Thus speech and text technologies overlap and interact with many other technologies that facilitate processing of multimodal communication and multimedia documents.



For a comprehensive introduction to the field, the reader is referred to: Cole R.A., J. Mariani, H. Uszkoreit, G. Varile, A. Zaenen, V. Zue, A. Zampolli (Eds.) (1997) Survey of the State of the Art in Human Language Technology, Cambridge University Press and Giardini. (http://www.dfki.de/~hansu/HLT-Survey.pdf)

2. Applications

Although existing LT systems are far from achieving human ability, they have numerous possible applications. The goal is to create software products that have some knowledge of human language. Such products are going to change our lives. They are urgently needed for improving human-machine interaction since the main obstacle in the interaction between human and computer is merely a communication problem. Today's computers do not understand our language but computer languages are difficult to learn and do not correspond to the structure of human thought. Even if the language the machine understands and its domain of discourse are very restricted, the use of human language can increase the acceptance of software and the productivity of its users.

Friendly technology should listen and speak

Natural language interfaces enable the user to communicate with the computer in French, English, German, or another human language. Some applications of such interfaces are database queries, information retrieval from texts, so-called expert systems, and robot control. Current advances in the recognition of spoken language improve the usability of many types of natural language systems. Communication with computers using spoken language will have a lasting impact upon the work environment; completely new areas of application for information technology will open up.

However, spoken language needs to be combined with other modes of communication such as pointing with mouse or finger. If such multimodal communication is finally embedded in an effective general model of cooperation, we have succeeded in turning the machine into a partner. The ultimate goal of research is the omnipresent access to all kinds of technology and to the global information structure by natural interaction. In an ambitious but not too far-fetched scenario, language technology provides the interface to an ambient intelligence providing assistance at work and in many situations of daily life.

Machines can also help people communicate with each other

Language technologies can also help people communicate with each other. Much older than communication problems between human beings and machines are those between people with different mother tongues. One of the original aims of language technology has always been fully automatic translation between human languages. From bitter experience scientists have realized that they are still far away from achieving the ambitious goal of translating unrestricted texts. Nevertheless, they have been able to create software systems that simplify the work of human translators and clearly improve their productivity. Less than perfect automatic translations can also be of great help to information seekers who have to search through large amounts of texts in foreign languages.

The most serious bottleneck for e-commerce is the volume of communication between business and customers or among businesses. Language technology can help to sort, filter and route incoming email. It can also assist the customer relationship agent to look up information and to compose a response. In cases where questions have been answered before, language technology can find appropriate earlier replies and automatically respond.

Language is the fabric of the web

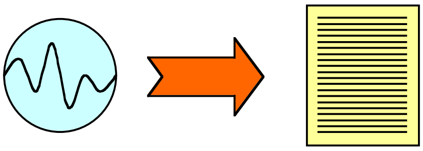
The rapid growth of the Internet/WWW and the emergence of the information society pose exciting new challenges to language technology. Although the new media combine text, graphics, sound and movies, the whole world of multimedia information can only be structured, indexed and navigated through language. For browsing, navigating, filtering and processing the information on the web, we need software that can get at the contents of documents. Language technology for content management is a necessary precondition for turning the wealth of digital information into collective knowledge. The increasing multilinguality of the web constitutes an additional challenge for language technology. The global web can only be mastered with the help of multilingual tools for indexing and navigating. Systems for crosslingual information and knowledge management will surmount language barriers for e-commerce, education and international cooperation.

3. Technologies

In the following a selection of the most relevant language technologies will be summarized. By clicking on the names of the technologies, you can access additional information.

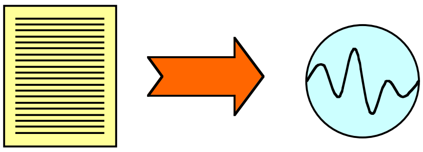
Speech recognition

Spoken language is recognized and transformed in into text as in dictation systems, into commands as in robot control systems, or into some other internal representation.



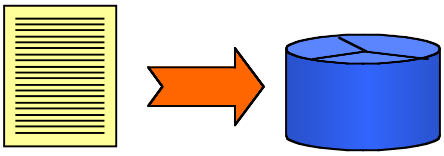
Speech synthesis

Utterances in spoken language are produced from text (text-to-speech systems) or from internal representations of words or sentences (concept-to-speech systems)



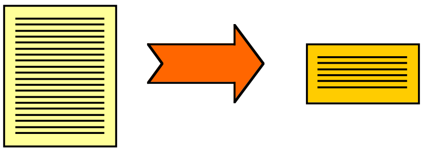
Text categorization

This technology assigns texts to categories. Texts may belong to more than one category, categories may contain other categories. Filtering is a special case of categorization with just two categories.



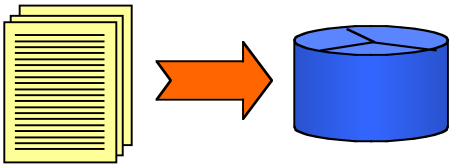
Text Summarization

The most relevant portions of a text are extracted as a summary. The task depends on the needed lengths of the summaries. Summarization is harder if the summary has to be specific to a certain query.



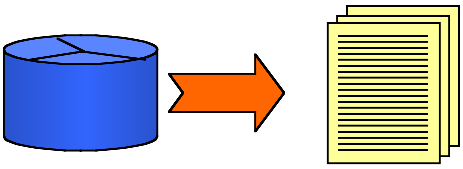
Text Indexing

As a precondition for document retrieval, texts are are stored in an indexed database. Usually a text is indexed for all word forms or – after lemmatization – for all lemmas. Sometimes indexing is combined with categorization and summarization.



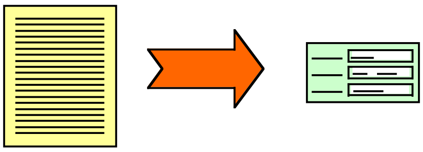
Text Retrieval

Texts are retrieved from a database that best match a given query or document. The candidate documents are ordered with respect to their expected relevance. Indexing, categorization, summarization and retrieval are often subsumed under the term information retrieval.



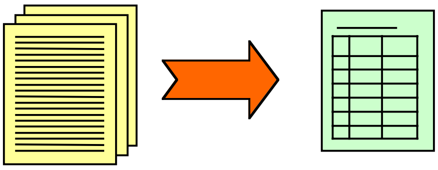
Information Extraction

Relevant information pieces of information are discovered and marked for extraction. The extracted pieces can be: the topic, named entities such as company, place or person names, simple relations such as prices, destinations, functions etc. or complex relations describing accidents, company mergers or football matches.



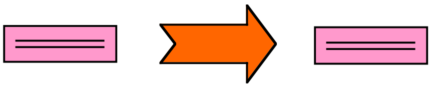
Data Fusion and Text Data Mining

Extracted pieces of information from several sources are combined in one database. Previously undetected relationships may be discovered.



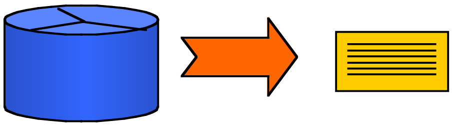
Question Answering

Natural language queries are used to access information in a database. The database may be a base of structured data or a repository of digital texts in which certain parts have been marked as potential answers.



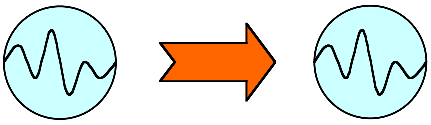
Report Generation

A report in natural language is produced that describes the essential contents or changes of a database. The report can contain accumulated numbers, maxima, minima and the most drastic changes.



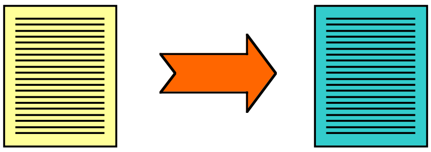
Spoken Dialogue Systems

The system can carry out a dialogue with a human user in which the user can solicit information or conduct purchases, reservations or other transactions.



Translation Technologies

Technologies that translate texts or assist human translators. Automatic translation is called machine translation. Translation memories use large amounts of texts together with existing translations for efficient look-up of possible translations for words, phrases and sentences.



4. Methods and Resources

As the investigation and modelling of human language is a truly interdisciplinary endeavor, the methods of language technology come from several disciplines: computer science, computational and theoretical linguistics, mathematics, electrical engineering and psychology.

Generic CS Methods

Programming languages, algorithms for generic data types, and software engineering methods for structuring and organizing software development and quality assurance.

Specialized Algorithms

Dedicated algorithms have been designed for parsing, generation and translation, for morphological and syntactic processing with finite state automata/transducers and many other tasks.

Nondiscrete Mathematical Methods

Statistical techniques have become especially successful in speech processing, information retrieval, and the automatic acquisition of language models. Other methods in this class are neural networks and powerful techniques for optimization and search.

Logical and Linguistic Formalisms

For deep linguistic processing, constraint based grammar formalisms are employed. Complex formalisms have been developed for the representation of semantic content and knowledge.

Linguistic Knowledge

Linguistic knowledge resources for many languages are utilized: dictionaries, morphological and syntactic grammars, rules for semantic interpretation, pronunciation and intonation.

Corpora and Corpus Tools

Large collections of application-specific or generic collections of spoken and written language are exploited for the acquisition and testing of statistical or rule-based language models.