CHAPTER 2

Natural Language Processing

Gobinda G. Chowdhury
University of Strathclyde

Introduction

Natural Language Processing (NLP) is an area of research and application that explores how computers can be used to understand and manipulate natural language text or speech to do useful things. NLP researchers aim to gather knowledge on how human beings understand and use language so that appropriate tools and techniques can be developed to make computer systems understand and manipulate natural languages to perform desired tasks. The foundations of NLP lie in a number of disciplines, namely, computer and information sciences, linguistics, mathematics, electrical and electronic engineering, artificial intelligence and robotics, and psychology. Applications of NLP include a number of fields of study, such as machine translation, natural language text processing and summarization, user interfaces, multilingual and cross-language information retrieval (CLIR), speech recognition, artificial intelligence, and expert systems.

One important application area that is relatively new and has not been covered in previous *ARIST* chapters on NLP relates to the proliferation of the World Wide Web and digital libraries. Several researchers have pointed out the need for appropriate research in facilitating multi- or cross-lingual information retrieval, including multilingual text processing

and multilingual user interface systems, in order to exploit the full benefit of the Web and digital libraries (see for example, Borgman, 1997; Peters & Picchi, 1997).

Scope

Several *ARIST* chapters have reviewed the field of NLP (Haas, 1996; Warner, 1987). Literature reviews on large-scale NLP systems, as well as related theoretical issues, have also appeared in a number of publications (see for example, Jurafsky & Martin, 2000; Mani & Maybury, 1999; Manning & Schutze, 1999; Sparck Jones, 1999; Wilks, 1996). Smeaton (1999) provides a good overview of past research on the applications of NLP in various information retrieval tasks. Several *ARIST* chapters have appeared on areas related to NLP, such as machine-readable dictionaries (Amsler, 1984; Evans, 1989), speech synthesis and recognition (Lange, 1993), and cross-language information retrieval (Oard & Diekema, 1998). Research on NLP is regularly published in the annual proceedings of the Association of Computational Linguistics (ACL) and its European counterpart EACL, biennial proceedings of the International Conference on Computational Linguistics (COLING), annual proceedings of the Message Understanding Conferences (MUCs), Text REtrieval Conferences (TRECs), and ACM SIGIR (Association for Computing Machinery, Special Interest Group on Information Retrieval) conferences. The most prominent journals reporting NLP research are *Computational Linguistics* and *Natural Language Engineering*. Articles reporting NLP research also appear in a number of information science journals such as *Information Processing & Management, Journal of the American Society for Information Science and Technology,* and *Journal of Documentation.* Several researchers have also conducted domain-specific NLP studies and reported them in journals specifically dealing with the field in question, such as the *International Journal of Medical Informatics* and *Journal of Chemical Information and Computer Science.*

Beginning with the basic issues of NLP, this chapter aims to chart the major research activities in this area since the last *ARIST* chapter in 1996 (Haas, 1996): natural language text processing systems—text summarization, information extraction, information retrieval, and so on, including domain-specific applications; natural language interfaces;

NLP in the context of the Web and digital libraries; and evaluation of NLP systems.

Linguistic research in information retrieval is not covered; this huge area is dealt with separately in this volume by David Blair. Similarly, NLP issues related to information retrieval tools (e.g., search engines) for Web search are not covered here because indexing and retrieval for the Web are the subjects of Edie Rasmussen's chapter, also in this volume.

Tools and techniques developed for building NLP systems are discussed in this chapter along with the specific areas of application for which they are built. Although machine translation (MT) is an important part, and in fact the origin, of NLP research, this topic is not addressed in great detail because it is a large area and demands separate treatment. Similarly, cross-language information retrieval, although a very important area in NLP research, is not covered in depth. A chapter on CLIR research appeared recently in *ARIST* (Oard & Diekema, 1998). However, MT and CLIR have become two important areas of research in the context of digital libraries. This chapter reviews some works on MT and CLIR in the context of NLP and information retrieval (IR) in digital libraries and the Web. Artificial intelligence techniques, including neural networks, are excluded.

Some Theoretical Developments

Previous *ARIST* chapters (Haas, 1996; Warner, 1987) described a number of theoretical developments that have influenced research in NLP. The most recent can be grouped into four classes: (1) statistical and corpus-based methods in NLP, (2) efforts to use WordNet for NLP research, (3) the resurgence of interest in finite-state and other computationally lean approaches to NLP, and (4) the initiation of collaborative projects to create large grammar and NLP tools.

Statistical methods are used in NLP for a number of purposes, e.g., for word sense disambiguation, for generating grammars and parsing, for determining stylistic evidence of authors and speakers, and so on. Charniak (1995) points out that 90 percent accuracy can be obtained in assigning part-of-speech tags to words by applying simple statistical measures. Jelinek (1999) is a widely cited source on the use of statistical methods in NLP, especially in speech processing. Rosenfield (2000)

reviews statistical language models for speech processing and argues for a Bayesian approach to the integration of linguistic theories of data.

Mihalcea and Moldovan (1999) mention that, although statistical approaches have thus far been considered best for word sense disambiguation, they are useful in only a small set of texts. They propose the use of WordNet to improve the results of statistical analyses of natural language texts. WordNet is an online lexical reference system developed at Princeton University. This excellent NLP tool contains English nouns, verbs, adjectives, and adverbs organized into synonym sets, each representing one underlying lexical concept. Details of WordNet are available in Fellbaum (1998) and on the Web (http://www.cogsci.princeton.edu/~wn). WordNet is now used in a number of NLP research and application areas. One of the major applications of WordNet in NLP has been in Europe with the formation of EuroWordNet in 1996. EuroWordNet is a multilingual database with WordNets for several European languages including Dutch, Italian, Spanish, German, French, Czech, and Estonian, structured in the same way as the WordNet for English (http://www.hum.uva.nl/~ewn).

Finite-state automation is the mathematical approach used to implement regular expressions—the standard notation for characterizing text sequences. Variations of automata such as finite-state transducers, Hidden Markov models, and n-gram grammars are important components of speech recognition and speech synthesis, spell checking, and information extraction, which are the principal applications of NLP. Different applications of the finite state methods in NLP have been discussed by Jurafsky and Martin (2000), Kornai (1999), and Roche and Shabes (1997).

The work of NLP researchers has been greatly facilitated by the availability of large-scale grammars for parsing and generation. Researchers can gain access to large-scale grammars and related tools through several Web sites; for example, Lingo (http://lingo.stanford.edu/), Computational Linguistics and Phonetics (http://www.coli.uni-sb.de/software.phtml), and the Parallel Grammar Project (http://www.parc.xerox.com/istl/groups/nltt/pargram/). Another significant recent development is the formation of various national and international consortia and research groups that can help share expertise and facilitate research in NLP. The Linguistic Data Consortium (LDC) (http://www.ldc.upenn.edu/)

at the University of Pennsylvania is a typical example that creates, collects, and distributes speech and text databases, lexicons, and other resources for research and development among universities, companies, and government research laboratories. The Parallel Grammar project is another example of international cooperation. This project is a collaborative effort involving researchers from Xerox PARC in California, the University of Stuttgart and the University of Konstanz in Germany, the University of Bergen in Norway, and Fuji Xerox in Japan. This project aims to produce wide-coverage grammars for English, French, German, Norwegian, Japanese, and Urdu, which are written collaboratively with a commonly agreed upon set of grammatical features (http://www.parc.xerox.com/istl/groups/nltt/pargram/). The recently formed Global WordNet Association is yet another example of cooperation. It is a non-commercial organization that provides a platform for discussing, sharing, and connecting WordNets for all languages in the world. The first international WordNet conference, held in India in January 2002, addressed various problems of NLP faced by researchers from different parts of the world (http://www.ciil.org/gwn/report.html).

Natural Language Understanding

At the core of any NLP task is the important issue of natural language understanding. The process of building computer programs that understand natural language involves three major problems: The first relates to thought processes, the second to the representation and meaning of the linguistic input, and the third to world knowledge. Thus, an NLP system may begin at the word level to determine the morphological structure and nature (such as part-of-speech or meaning) of the word; and then may move on to the sentence level to determine the word order, grammar, and meaning of the entire sentence; and then to the context and the overall environment or domain. A given word or sentence may have a specific meaning or connotation in a given context or domain, and may be related to many other words and/or sentences in the given context.

Liddy (1998) and Feldman (1999) suggest that in order to understand natural languages, it is important to be able to distinguish among the following seven interdependent levels that people use to extract meaning from text or spoken languages:

- Phonetic or phonological level that deals with pronunciation
- Morphological level that deals with the smallest parts of words that carry meaning, and suffixes and prefixes
- Lexical level that deals with lexical meaning of words and parts of speech analyses
- Syntactic level that deals with grammar and structure of sentences
- Semantic level that deals with the meaning of words and sentences
- Discourse level that deals with the structure of different kinds of text using document structures
- Pragmatic level that deals with the knowledge that comes from the outside world, i.e., from outside the content of the document

A natural language processing system may involve all or some of these levels of analysis.

NLP Tools and Techniques

A number of researchers have attempted to improve the technology for performing various activities that form important parts of NLP work. These activities may be categorized as follows:

- Lexical and morphological analysis, noun phrase generation, word segmentation, and so forth (Bangalore & Joshi, 1999; Barker & Cornacchia, 2000; Chen & Chang, 1998; Dogru & Slagle, 1999; Kazakov, Manandhar, & Erjavec, 1999; Lovis, Baud, Rassinoux, Michel, & Scherter, 1998; Tolle & Chen, 2000; Zweigenbaum & Grabar, 1999).
- Semantic and discourse analysis, word meaning, and knowledge representation (Kehler, 1997; Meyer & Dale, 1999; Mihalcea & Moldovan, 1999; Pedersen & Bruce, 1998; Poesio & Vieira, 1998; Tsuda & Nakamura, 1999).
- Knowledge-based approaches and tools for NLP (Argamon, Dagan, & Krymolowski, 1998; Fernandez & Garcia-Serrano, 2000; Martinez

& Garcia-Serrano, 1998; Martinez, de Miguel, Cuadra, Nieto, & Castro, 2000).

Dogru and Slagle (1999) propose a model lexicon that involves automatic acquisition of the words as well as representation of the semantic content of individual lexical entries. Kazakov et al. (1999) report research on word segmentation based on an automatically generated, annotated lexicon of word-tag pairs. Wong et al. (1998) report the features of an NLP tool called Chicon used for word segmentation in Chinese text. Zweigenbaum and Grabar (1999) propose a method for acquiring morphological knowledge about words in the medical literature, which takes advantage of commonly available lists of synonym terms to bootstrap the acquisition process. Although the authors experimented with the method on the SNOMED International Microglossary for Pathology in its French version, they claim that because the method does not rely on a priori linguistic knowledge, it is applicable to other languages such as English. Lovis et al. (1998) propose the design of a lexicon for use in processing medical texts.

Noun phrasing is an important NLP technique used in information retrieval. One of the major research areas is combining traditional keyword and syntactic approaches with semantic approaches to text processing in order to improve the quality of information retrieval. Tolle and Chen (2000) compared four noun phrase generation tools in order to assess their ability to isolate noun phrases from medical journal abstracts databases. The NLP tools evaluated were: Chopper, developed by the Machine Understanding Group at the MIT Media Laboratory; Automatic Indexer and AZ Noun Phraser, developed at the University of Arizona; and NPTool, a commercial NLP tool from LingSoft, a Finnish Company. The National Library of Medicine's SPECIALIST Lexicon was used along with the AZ Noun Phraser. This experiment used a reasonably large test set of 1.1 gigabytes of text, comprising 714,451 abstracts from the CANCERLIT database. This study showed that with the exception of Chopper, the NLP tools were fairly comparable in their performance, measured in terms of recall and precision. The study also showed that the SPECIALIST Lexicon increased the ability of the AZ Noun Phraser to generate relevant noun phrases. Pedersen and Bruce (1998) propose a corpus-based approach to word-sense disambiguation that

requires only information that can be automatically extracted from untagged text. Barker and Cornacchia (2000) describe a simple system for choosing noun phrases from a document based on their length, their frequency of occurrence, and the frequency of the head noun, using a base noun phrase skimmer and an off-the-shelf online dictionary. This research revealed some interesting findings: (1) the simple noun-phrase-based system performs roughly as well as a state-of-the-art, corpus-trained keyphrase extractor, (2) ratings for individual keyphrases do not necessarily correlate with ratings for sets of keyphrases for a document, and (3) agreement among unbiased judges on the keyphrase rating task is poor. Silber and McCoy (2000) report research that uses a linear time algorithm for calculating lexical chains, which is a method of capturing the "aboutness" of a document.

Mihalcea and Moldovan (1999) argue that the reduced applicability of statistical methods in word sense disambiguation is due basically to the lack of widely available, semantically tagged corpora. They report research that enables the automatic acquisition of sense tagged corpora, and is based on (1) the information provided in WordNet, and (2) the information gathered from the Internet using existing search engines.

Martinez and Garcia-Serrano (1998) and Martinez et al. (2000) propose a method for the design of structured knowledge models for NLP. The key features of their method comprise the decomposition of linguistic knowledge sources in specialized sub-areas to tackle the complexity problem and a focus on cognitive architectures that allow for modularity, scalability, and reusability. The authors claim that their approach profits from NLP techniques, first-order logic, and some modeling heuristics (Martinez et al. 2000). Fernandez and Garcia-Serrano (2000) comment that knowledge engineering is increasingly regarded as a means to complement traditional formal NLP models by adding symbolic modeling and inference capabilities in a way that facilitates the introduction and maintenance of linguistic experience. They propose an approach that allows the design of linguistic applications to integrate different formalisms, reuse existing language resources, and support the implementation of the required control in a flexible way. Costantino (1999) argues that qualitative data, particularly articles from online news agencies, are not yet successfully processed, and as a result, financial operators, notably traders, suffer from qualitative data

overload. IE-Expert is a system that combines the techniques of NLP, information extraction, and expert systems in order to be able to suggest investment decisions from a large volume of texts (Constantino, 1999).

Natural Language Text Processing Systems

Manipulation of texts for knowledge extraction, for automatic indexing and abstracting, or for producing text in a desired format, has been recognized as an important area of research in NLP. This is broadly classified as the area of natural language text processing that allows structuring of large bodies of textual information with a view to retrieving particular information or to deriving knowledge structures that may be used for a specific purpose. Automatic text processing systems generally take some form of text input and transform it into an output of a different form. The central task for natural language text processing systems is the translation of potentially ambiguous natural language queries and texts into unambiguous internal representations on which matching and retrieval can take place (Liddy, 1998). A natural language text processing system may begin with morphological analyses. Stemming of terms, in both the queries and documents, is done in order to derive the morphological variants of the words involved. The lexical and syntactic processing involves the utilization of lexicons for determining the characteristics of the words, recognizing their parts of speech, determining the words and phrases, and parsing of the sentences.

Past research concentrating on natural language text processing systems has been reviewed by Haas (1986), Mani and Maybury (1999), Smeaton (1999), and Warner (1987). Some NLP systems have been built to process texts using particular small sublanguages to reduce the size of the operations and the nature of the complexities. These domain-specific studies are largely known as "sublanguage analyses" (Grishman & Kittredge, 1986). Some of these studies are limited to a particular subject area such as medical science, whereas others deal with a specific type of document, such as patents.

Abstracting

The terms "automatic abstracting" and "text summarization" are now used synonymously. This area of NLP research is becoming more common in the Web and digital library environments. In simple abstracting or summarization systems, parts of text—sentences or paragraphs—are selected automatically, based on some linguistic and/or statistical criteria, to produce the abstract or summary. More sophisticated systems may merge two or more sentences, or parts thereof, to generate one coherent sentence, or may generate simple summaries from discrete items of data.

Recent interest in automatic abstracting and text summarization is reflected in the large volume of research appearing in a number of international, national, and regional conferences and workshops presented by the ACL, the American Association for Artificial Intelligence (AAAI), and the ACM SIGIR. Several techniques are used for automatic abstracting and text summarization. Goldstein, Kantrowitz, Mittal, and Carbonell (1999) use conventional IR methods and linguistic cues for extracting and ranking sentences for generating news article summaries. A number of studies on text summarization have been reported recently. Silber and McCoy (2000) claim that their linear time algorithm for calculating lexical chains is an efficient method for preparing automatic summarization of documents. Chuang and Yang (2000) report a text summarization technique using cue phrases appearing in the texts of U.S. patent abstracts.

Roux and Ledoray (2000) report a project, Aristotle, that aims to build an automatic medical data system capable of producing a semantic representation of a text in a canonical form. Song and Zhao (2000) propose a method of automatic abstracting that integrates the advantages of both linguistic and statistical analysis of a corpus.

Moens and Uyttendaele (1997) describe the SALOMON (Summary and Analysis of Legal Texts for Managing Online Needs) project that automatically summarizes Belgian criminal court cases. The system extracts relevant information from the full texts, such as the name of the court issuing the decision, the decision date, the offenses charged, the relevant statutory provisions disclosed by the court, and the legal principles applied in the case. A text grammar represented as a semantic

network is used to determine the category of each case. RAFI (Résumé Automatique à Fragments Indicateurs) is an automatic text summarization system that transforms full-text scientific and technical documents into condensed texts (Lehmam, 1999). RAFI adopts discourse analysis techniques, using a thesaurus for recognition and selection of the most pertinent elements of texts. The system assumes a typical structure of areas from each scientific document, namely, previous knowledge, content, method, and new knowledge.

Most automatic abstracting and text summarization systems work satisfactorily within a small text collection or within a restricted domain. Building robust and domain-independent systems is a complex and resource-intensive task. Arguing that purely automatic abstracting systems do not always produce useful results, Craven (1988, 1993, 2000) proposes a hybrid abstracting system in which some tasks are performed by human abstractors and others by assistance software called TEXNET. However, recent experiments on the usefulness of the automatically extracted keywords and phrases in actual abstracting by human abstractors showed considerable variation among subjects, with only 37 percent of the abstractors finding keywords and phrases useful in writing their abstracts (Craven, 2000).

Information Extraction

Knowledge discovery and data mining have become important areas of research over the past few years, and in addition to *ARIST*, several information science journals have published special issues reporting research on these topics (see, for example, Benoît, 2002; Qin & Norton, 1999; Raghavan, Deogun, & Server, 1998; Trybula, 1997; Vickery, 1997). Knowledge discovery and data mining use a variety of techniques to extract information from source documents. Information extraction (IE) is a subset of knowledge discovery and data mining that aims to extract useful bits of textual information from natural language texts (Gaizauskas & Wilks, 1998). A variety of IE techniques is used and the extracted information can serve a number of purposes: for example, to prepare a summary of texts, populate databases, fill in slots in frames, and identify keywords and phrases for information retrieval. IE techniques are also used for classifying text items according to predefined

categories. CONSTRUE, an early categorization system developed for Reuters, classifies news stories (Hayes, 1992). The CONSTRUE software was subsequently generalized into a commercial product called Text Categorization Shell (TCS). Yang and Liu (1999) report an evaluation of five text categorization systems.

Morin (1999) suggests that, although many IE systems can successfully extract terms from documents, revealing relations between terms is still difficult. PROMETHEE is a system that extracts lexico-syntactic patterns relative to a specific conceptual relation from technical corpora (Morin, 1999). Bondale, Maloor, Vaidyanathan, Sengupta, and Rao (1999) suggest that IE systems must operate at many levels, from word recognition to discourse analysis at the level of the complete document. They report an application of the Blank Slate Language Processor (BSLP) for the analysis of a real-life natural language corpus of responses to open-ended questionnaires in the field of advertising.

Glasgow, Mandell, Binney, Ghemri, and Fisher (1998) describe a system called Metlife's Intelligent Text Analyzer (MITA) that extracts information from life insurance applications. Ahonen, Heinonen, Klemettinen, and Verkamo (1998) propose a general framework for text mining using pragmatic and discourse level analyses of text. Sokol, Murphy, Brooks, and Mattox (2000) combine visualization and NLP technologies to perform text mining. Chang, Ko, and Hsu (2000) argue that IE systems are usually event-driven (i.e., based on domain knowledge built on various events) and propose using the neural network paradigm for event detection to drive intelligent information extraction. They employ the back propagation (BP) learning algorithm to train the event detector, and apply NLP technology to aid the selection of nouns as feature words to characterize documents. These nouns are stored in an ontology as a knowledge base and are used for the extraction of useful information from e-mail messages.

Cowie and Lehnert (1996) reviewed research on IE and observed that the NLP research community was ill-prepared to tackle the difficult problems of semantic feature tagging, co-reference resolution, and discourse analysis, all of which are important aspects of IE research. Gaizauskas and Wilks (1998) reviewed IE research from its origin in the artificial intelligence world in the 1960s and 1970s through to the present. They discussed major IE projects undertaken in different sectors,

namely, academic research, employment, fault diagnosis, finance, law, medicine, military intelligence, police, software system requirements specification, and technology/product tracking.

Chowdhury (1999a) reviewed research that used template mining techniques in a range of contexts: extracting proper names from full-text documents, extracting facts from press releases, abstracting scientific papers, summarizing new product information, extracting specific information from chemical texts, and so on. He also discussed how some Web search engines use templates to facilitate information retrieval, and recommended that each Web author complete a template to characterize his/her document, to regularize the creation of document surrogates. However, he warns that a single, all-purpose metadata format will not be applicable for all authors in all domains, and further research is necessary to develop appropriate formats for each.

Smeaton (1997) argues that IE has been the subject of much research and development and has been delivering working solutions for many decades, whereas IE is a more recent and emerging technology. He urges the IE community to see how a related task—perhaps the most-related task, information retrieval—has managed to use basic NLP technology in its development. Commenting on the future challenges of IE researchers, Gaizauskas and Wilks (1998) mention that the performance levels of the common IE systems, which lie in the 50 percent range for combined recall and precision, need to improve significantly to satisfy information analysts. Cost is a major stumbling block of IE systems development. CONSTRUE, for example, required 9.5 person years of effort (Hayes & Weinstein, 1991). Portability and scalability are also two big issues for IE systems, which depend heavily on domain knowledge. A given IE system may work satisfactorily in a relatively small text collection, but it may not perform well in a larger collection or in a different domain. Alternative technologies are now being used to overcome these problems. Adams (2001) discusses the merits of NLP and the wrapper induction technology in extracting information from Web documents. In contrast to NLP, wrapper induction operates independently of specific domain knowledge. Instead of analyzing the meaning of discourse at the sentence level, wrapper technology identifies relevant content based on the textual qualities that surround the desired data. Wrappers operate on the surface features that characterize texts of

training examples. A number of vendors, such as Jango (purchased by Excite), Junglee (purchased by Amazon), and Mohomine employ wrapper induction technology (Adams, 2001).

Information Retrieval

Information retrieval (IR) has been a major application area of NLP, resulting in a large number of publications. Lewis and Sparck Jones (1996) commented that the generic challenge for NLP in information retrieval was whether the necessary processing of texts and queries was doable; and the specific challenges were whether non-statistical and statistical data could be combined, and whether data about individual documents and whole files could be combined. They further suggested that there were major challenges in making NLP technology operate effectively and efficiently and also in conducting appropriate evaluation tests to assess whether and how far the approach worked in the case of interactive searching of large text files. Feldman (1999) suggested that in order to achieve success in IR, NLP techniques should be applied in conjunction with other technologies such as visualization, intelligent agents, and speech recognition.

Arguing that syntactic phrases were more meaningful than statistically obtained word pairs, and thus more powerful for discriminating among documents, Narita and Ogawa (2000) used shallow syntactic processing instead of statistical processing to identify candidate phrasal terms from query texts. Comparing the performance of Boolean and natural language searches, Paris and Tibbo (1998) found that, in their experiment, Boolean searches had better results than freestyle (natural language) searches. However, they concluded that neither could be considered superior for every query. In other words, different queries demand different techniques.

Pirkola (2001) showed that languages vary significantly in their morphological properties. However, for each language there are two variables that describe the morphological complexity, namely, an index of synthesis (IS) that describes the amount of affixation in an individual language, i.e., the average number of morphemes per word in the language; and an index of fusion (IF) that describes the ease with which two morphemes can be separated in a language. Pirkola (2001) found that

calculation of the ISs and IFs in a language is a relatively simple task, and once established, they could be utilized fruitfully in empirical IR research and system development.

Variations in presenting subject matter greatly affect IR, and, hence, the linguistic variation of document texts is one of the field's greatest challenges. In order to investigate how consistently newspapers choose words and concepts to describe an event, Lehtokangas and Järvelin (2001) chose articles on the same news stories from three Finnish newspapers. Their experiment revealed that for short newswire items the consistency was 83 percent and for long articles, 47 percent. The newspapers were very consistent in using concepts to represent events, with a level of consistency varying between 92 and 97 percent.

Khoo, Myaeng, and Oddy (2001) investigated whether information obtained by matching cause–effect relations expressed in documents with the cause–effect relations expressed in user queries improves results in document retrieval when compared with the keywords only, without considering the relations. Their experiment with the *Wall Street Journal* full-text database revealed that searching either the cause or the effect as a wildcard can improve information retrieval effectiveness if the appropriate weight for the type of match can be determined. However, the authors stressed that the results of this study were not as strong as they had expected.

Chandrasekar and Srinivas (1998) suggested that coherent text contains significant latent information, such as syntactic structure and patterns of language use, and that this information could be used to improve the performance of information retrieval systems. They described Glean, a system that uses syntactic information to filter irrelevant documents, thereby improving the precision of information retrieval.

A number of tracks (research groups or themes) in the TREC series of experiments deal directly or indirectly with NLP and information retrieval, such as the cross-language, filtering, interactive, question-answering, and Web tracks. Reports of progress of the Natural Language Information Retrieval (NLIR) project are available in the TREC reports (Perez-Carballo & Strzalkowski, 2000; Strzalkowski, Fang, Perez-Carballo, & Jin, 1997; Strzalkowski et al., 1998; Strzalkowski et al., 1999). The major goal of this project has been to demonstrate that robust NLP techniques used for indexing and searching of text documents perform

better than the simple keyword and string-based methods used in statistical full-text retrieval (Strzalkowski et al., 1999). However, results indicated that simple linguistically motivated indexing (LMI) was no more effective than well-executed statistical approaches in English language texts. Nevertheless, it was noted that more detailed search topic statements responded well to LMI, when compared to terse one-sentence search queries. Thus, it was concluded that query expansion, using NLP techniques, leads to sustainable advances in IR effectiveness (Strzalkowski et al., 1999).

Natural Language Interfaces

A natural language interface is one that accepts query statements or commands in natural language and sends data to some system, typically a retrieval system, which then provides appropriate responses to the commands or query statements. A natural language interface should be able to translate the natural language statements into appropriate actions for the system. A large number of natural language interfaces that work reasonably well in narrow domains have been reported in the literature (for a review of such systems see Chowdhury, 1999b, Chapter 19; Haas, 1996; Stock, 2000).

Many of the efforts in natural language interface design to date have focused on handling rather simple natural language queries. Several question-answering systems are now being developed that aim to provide answers to natural language questions, as opposed to documents containing information related to the question. Such systems often use a variety of IE and IR operations employing NLP tools and techniques to derive the correct answer from the source texts. Breck, Burger, House, Light, and Mani (1999) report a question-answering system that uses techniques from knowledge representation, information retrieval, and NLP. The authors claim that this combination promotes domain independence and robustness in the face of text variability, both in the question and in the raw text documents used as knowledge sources. Research reported in the Question Answering (QA) track of TREC shows some interesting results. The basic technology used by the participants in the QA track included several steps. First, cue words and phrases like "who" (as in "Who is the prime minister of Japan?") and "when" (as in "When

did the Jurassic period end?") were identified to guess what was needed; and then a small portion of the document collection was retrieved using standard text retrieval technology. This was followed by a shallow parsing of the returned documents, to identify the entities required for an answer. If no appropriate answer type was found, the best matching passage was retrieved. This approach works well as long as the query types recognized by the system have broad coverage, and the system can classify questions reasonably accurately (Voorhees, 1999). In TREC-8, the first QA track of TREC, the most accurate QA systems could answer more than two thirds of the questions correctly. In the second QA track (TREC-9), the best performing QA system, the Falcon system from Southern Methodist University, was able to answer 65 percent of the questions (Voorhees, 2000). These results are quite impressive in a domain-independent, question-answering environment. However, the questions were still simple in the first two QA tracks. In the future more complex questions, requiring answers to be obtained from more than one document, will be handled by QA track researchers.

Owei (2000) argued that the drawbacks of most natural language interfaces to database systems result primarily from their weak interpretative power, which is caused by their inability to deal with the nuances in human use of natural language. The author further argued that the difficulty with NL database query languages (DBQLs) could be overcome by combining concept-based DBQL paradigms with NL approaches to enhance the overall ease-of-use of the query interface.

Zadrozny, Budzikowska, Chai, and Kambhatla (2000) suggested that in an ideal information retrieval environment, users should be able to express their interests or queries directly and naturally, by speaking, typing, and/or pointing; the computer system should then be able to provide intelligent answers or ask relevant questions. However, they commented that even though we build natural language systems, this goal cannot be fully achieved due to limitations of science, technology, business knowledge, and programming environments. The specific problems include the following (Zadrozny et al., 2000):

- Limitations of NL understanding
- Managing the complexities of interaction (for example, when using NL on devices with differing bandwidth)

- Lack of precise user models (for example, knowing how demographics and personal characteristics of a person should be reflected in the type of language and dialogue the system employs to interact with the user)
- Lack of middleware and toolkits

NLP Software

A number of specific NLP software products have been developed over the years, some of which are free, while others are available commercially. Many such NLP software packages and tools have already been mentioned in this chapter. More NLP tools and software are introduced in this section.

Pasero and Sabatier (1998) describe the principles underlying ILLICO, a generic natural language software tool for building larger applications that perform specific linguistic tasks such as analysis, synthesis, and guided composition. Liddy (1998) and Liddy, Diamond, and McKenna (2000) discuss the commercial use of NLP in IR with the example of DR-LINK (Document Retrieval Using LINguistic Knowledge) system demonstrating the capabilities of NLP for IR. Detailed product information and a demo of DR-LINK are available online (http://www.textwise.com/dr-link.html). Nerbonne, Dokter, and Smit (1998) report on GLOSSER, an intelligent assistant for Dutch students learning to read French. Scott (1999) describes the Kana Customer Messaging System that can categorize inbound e-mails and messages, forward them to the right department, and generally streamline the response process. Kana also has an auto-suggestion function that helps a customer service representative answer questions on unfamiliar territory. Scott (1999) describes another system, Brightware, which uses NLP techniques to elicit meaning from groups of words or phrases and reply to some e-mails and messages automatically. NLPWin is an NLP system from Microsoft that accepts sentences and delivers detailed syntactic analyses, together with a logical form representing an abstraction of the meaning (Elworthy, 2000). Scarlett and Szpakowicz (2000) report a diagnostic evaluation of DIPETT, a broad-coverage parser of English sentences.

The Natural Language Processing Laboratory, Center for Intelligent Information Retrieval at the University of Massachusetts

(http://www-nlp.cs.umass.edu/nlplic.html), distributes source codes and executables to support IE system development efforts at other sites. Each module is designed to be used in a domain-specific and task-specific customizable IE system. Available software includes:

- MARMOT Text Bracketting Module, a text file translator that segments arbitrary text blocks into sentences, applies low-level specialists such as date recognizers, associates words with part-of-speech tags, and brackets the text into annotated noun phrases, prepositional phrases, and verb phrases
- BADGER Extraction Module, analyzes bracketed text and produces case frame instantiations according to application-specific domain guidelines
- CRYSTAL Dictionary Induction Module, learns text extraction rules, suitable for use by BADGER, from annotated training texts
- ID3-S Inductive Learning Module, a variant on ID3 that induces decision trees on the basis of training examples

Waldrop (2001) briefly describes the features of three NLP software packages:

- Jupiter, a product of the MIT Research Lab that works in the field of weather forecasting
- Movieline, a product of Carnegie Mellon University that talks about local movie schedules
- MindNet from Microsoft Research, a system for automatically extracting a massively hyperlinked Web of concepts, from, say, a standard dictionary

Feldman (1999) mentions a number of NLP software packages:

- ConQuest, a part of Excalibur, that incorporates a lexicon implemented as a semantic network
- InQuery that parses sentences, stems words, and recognizes proper nouns and concepts based on term co-occurrence
- The LinguistX parser from Xerox PARC that extracts syntactic information and is used in InfoSeek
- Text mining systems like NetOwl from SRA and KNOW-IT from TextWise

A recent survey of 68 European university centers in computational linguistics and NLP, carried out under the auspices of the Socrates Working Group on Advanced Computing in the Humanities, revealed that Java has already become the second most commonly taught programming language (Black, Rinaldi, & McNaught, 2000). In addition, Java-based programs are being used to develop interactive instructional materials. Black et al. (2000) review some Java-based courseware and discuss the issues involved in more complex natural language processing applications that use Java.

Internet, Web and Digital Library Applications of NLP

The Internet and the Web have brought significant advances in the way we create, look for, and use information. A huge volume of information is now available through the Internet and digital libraries. However, these developments have made problems related to information processing and retrieval more prominent. According to a recent survey (Global Reach, 2001), 55 percent of Internet users are non-English speakers; this is increasing rapidly, thereby reducing the percentage of Net users who are native English speakers. However, about 80 percent of the Internet and digital library resources available today are in English (Bian & Chen, 2000). This dramatizes the urgent need for multilingual information systems and CLIR facilities. How to manipulate the large volume of multilingual data has become a major research question. At the user interface level, a query translation system must translate from the user's native language to the language of the system. Several approaches have been proposed for query translation. The dictionary-based approach uses a bilingual dictionary to convert terms from the source language to the target language. Coverage and up-to-dateness of the bilingual dictionary are major issues here. The corpus-based approach uses parallel corpora for word selection, where the problem lies with the domain and scale of the corpora. Bian and Chen (2000) proposed MTIR, a Chinese-English CLIR system on the Web that integrated query and document translation. They also addressed a number of issues regarding machine translation on the Web; specifically, the role played by HTML tags in translation, the

trade-off between the speed and performance of the translation system, and the form in which the translated material is presented.

Staab et al. (1999) described the features of an intelligent information agent called GETESS that used semantic methods and NLP capabilities to gather tourist information from the Web and present it to the human user in an intuitive, user-friendly way. Ceric (2000) reviewed advances in Web search technology and mentioned that NLP technologies would have a positive impact on the success of search engines. Mock and Vemuri (1997) described the Intelligent News Filtering Organizational System (INFOS) that was designed to filter unwanted news items from a Usenet news group. INFOS built a profile of user interests based on user feedback. After the user browsed each article, INFOS asked the user to rate the article, and used this as a criterion for selection (or rejection) of similar articles next time around. News articles were classified by a simple, quick-pass keyword method, called Global Hill Climbing (GHC). Articles that could not be classified by GHC were passed through a WordNet knowledge base through a case-based reasoning (CBR) module, a slower but more accurate method. Very small-scale evaluation of INFOS suggested that the indexing pattern method, i.e., mapping of the words from the input text into the correct concepts in the WordNet abstraction hierarchy, correctly classified 80 percent of the articles; the major reason for errors was the weakness of the system in disambiguating pronouns.

One of the major stumbling blocks to providing personalized news delivery over the Internet is the difficulty of automatically associating related items from different media. Carrick and Watters (1997) described a system that aimed to determine to what degree any two news items refer to the same news event. This research focused on determining the association between photographs and stories by using names. The algorithm developed was tested against a test data set as well as new data sets, with human experts checking the pairs of news items and photos generated by the system. In terms of recall, precision, and time, the system performed comparably for the new and training sets.

Because of the volume of text available on the Web, many researchers have proposed using it as the testbed for NLP research. Grefenstette (1999) argued that, although noisy, Web text presents language as it is used, and statistics derived from the Web can have practical uses in many NLP applications.

Machine Translation and CLIR

With the proliferation of the Web and digital libraries, multilingual information retrieval has become a major challenge. Two sets of issues are considered here: (1) recognition, manipulation, and display of multiple languages; and (2) cross-language information search and retrieval (Peters & Picchi, 1997). The first set of issues relates to the enabling technology that will allow users to access information in whatever language it is stored, while the second set implies permitting users to specify their information needs in their preferred language while retrieving information in whatever language it is stored. Text translation can take place at two levels: (1) translation of the full text from one language to another for the purpose of search and retrieval; and (2) translation of queries from one language to one or more different languages. The first option is feasible for small collections or for specific applications, as in meteorological reports (Oudet, 1997). Translation of queries is a more practicable approach, and promising results have been reported in the literature.

Oard (1997) commented that seeking information from a digital library could benefit from the ability to query large collections once using a single language. Furthermore, if the retrieved information was not available in a language the user can read, some form of translation would be needed. Multilingual thesauri such as EUROVOC help to address this challenge by facilitating controlled vocabulary searches using terms from several languages, whereas services such as INSPEC produce English abstracts for documents in other languages (Oard, 1997). However, as Oard noted, fully automatic MT was neither sufficiently fast nor sufficiently accurate to adequately support interactive cross-language information seeking in the Web and digital library environments. Fortunately, an active and rapidly growing research community has coalesced around these and related issues, applying techniques drawn from several fields, notably IR and NLP, to provide access to large multilingual collections.

Borgman (1997) commented that we have hundreds (and sometimes thousands) of years' worth of textual materials in hundreds of languages, created long before data encoding standards existed. She illustrated the multilanguage digital language challenge with examples drawn from the research library community, which typically handles collections of materials in about 400 different languages.

Ruiz and Srinivasan (1998) investigated an automatic method for CLIR that utilized the multilingual Unified Medical Language System (UMLS) Metathesaurus to translate Spanish natural language queries into English. They concluded that the UMLS Metathesaurus-based CLIR method was at least equivalent to, if not better than, multilingual-dictionary-based approaches. Yang, Gomez, and Song (2000) observed that there was no reliable guideline as to how large, machine readable corpora should be compiled to develop practical NLP software packages and/or complete dictionaries for humans and computational use. They proposed a new mathematical tool—a piecewise curve-fitting algorithm—and suggested how to determine the tolerance error of the algorithm for good prediction, using a specific corpus.

Two telematics application program projects in the Telematics for Libraries initiative, TRANSLIB and CANAL/LS, were active between 1995 and 1997 (Oard, 1997). Both projects investigated cross-language searching in library catalogs, and each included English, Spanish, and at least one other language; CANAL/LS added German and French, while TRANSLIB added Greek. MULINEX (http://mulinex.dfki.de), another European project, is concerned with the efficient use of multilingual online information. The project aims to process multilingual information and present it to the user in a way that facilitates finding and evaluating the desired information quickly and accurately. TwentyOne, started in 1996, is an EU-funded project aiming to develop a tool for efficient dissemination of multimedia information in the field of sustainable development (http://twentyone.tpd.tno.nl/twentyone). Details of these and CLIR research projects in the U.S. and other parts of the world have been reviewed by Oard and Diekema (1998).

Magnini, Not, Stock, and Strapparava (2000) described two projects where NLP had been used for improving performance in the public administration sector. The first project, GIST, was concerned with automatic multilingual generation of instructional texts for form filling. The second project, TAMIC, aimed to provide an interface for interactive access to information, centered on NLP and designed to be used by the clerk along with the active participation of the individual citizen.

Powell and Fox (1998) described SearchDB-ML Lite, a federated search system, for searching heterogeneous multilingual collections of theses and dissertations on the Web (the Networked Digital Library of

Theses and Dissertations, NDLTD [www.ndltd.org/]). A markup language, called SearchDB, was developed for describing the characteristics of a search engine and its interface, and a protocol was built for requesting word translations between languages. A review of the results generated from simultaneously querying over 50 sites revealed that in some cases more sophisticated query mapping was necessary to retrieve results sets that truly correspond to the original query. The authors reported that an extended version of the SearchDB markup language was being developed that could reflect the default and available query modifiers for each search engine; work was also underway to implement a mapping system that used this information.

Several companies now provide machine translation services, for example (McMurchie, 1998):

- Berlitz International, Inc. offers professional translation service in 20 countries
- Lernout and Hauspie has an Internet Translation Division
- Orange, California-based Language Force, Inc., has a product called Universal Translator Deluxe
- IBM MT services through its WebSphere Translation Server

Numerous research papers discuss MT and CLIR projects that deal with specific languages; for example, Chinese (Kwok, Grunfeld, Dinstl, & Chan, 2000; Lee, Ng, & Lu, 1999), Japanese (Ma et al. 2000; Ogura, Nakaiwa, Matsuo, Ooyama, & Bond, 2000; Yang & Akahori, 2000), Portugese (Barahona & Alferes, 1999), Sinhalese (Herath & Herath, 1999), Spanish (Marquez, Padro, & Rodriguez, 2000; Weigard & Hoppenbrouwers, 1998), Thai (Isahara, Ma, Sornlertlamvanich, & Takahashi, 2000), and Turkish (Say, 1999). Some studies have considered more than two languages; see, for example, Ide (2000). Various aspects of MT are considered, for example:

- Use of cue phrases in determining relationships among the lexical units in a discourse (Say, 1999)
- Generation of semantic maps of terms (Ma et al., 2000)

- Creation of language-specific semantic dictionaries (Ogura et al., 2000)
- Discourse analysis (Yang & Akahori, 2000)
- Lexical analysis (Ide, 2000; Lee et al., 1999)
- Part-of-speech tagging (Isahara et al., 2000; Marquez et al., 2000)
- Query translation (Kwok et al., 2000)
- Transliteration of foreign words for information retrieval (Jeong, Mayeng, Lee, & Choi, 1999)

Weigard and Hoppenbrouwers (1998) reported how an English/Spanish lexicon, including an ontology, is constructed for NLP tasks in an ESPRIT project called TREVI. Emphasizing the point that there had not been any study of natural language information retrieval in Swedish, Hedlund, Pirkola, and Järvelin (2001) described the features of the language and point out a number of research problems. They stressed that research in NLP in Swedish is required because the research findings and tools for other languages do not quite apply to Swedish because of the language's unique features.

Reviewing the progress of MT research, Jurafsky and Martin (2000, p. 825) commented "machine translation system design is hard work, requiring careful selection of models and algorithms and combination into a useful system." They continued, "despite half a century of research, machine translation is far from solved; human language is a rich and fascinating area whose treasures have only begun to be explored."

Evaluation

Evaluation is an important area in any system development activity, and information science researchers have long been struggling to determine appropriate evaluation mechanisms for large-scale information systems. Similarly, NLP researchers have also been trying to develop reliable methods for evaluating robust NLP systems. However, a single set of evaluation criteria will not be applicable to all NLP tasks. Different evaluation parameters may be required for each task, such as information extraction and automatic abstracting, which are significantly different in nature

when compared with other NLP tasks such, as MT, CLIR, or natural language user interfaces.

The Evaluation in Language and Speech Engineering (ELSE) project, under contract from the European Commission, studied the possible implementation of comparative evaluation in NLP systems. Comparative evaluation in language engineering has been used since 1984 in the U.S. Defense Advanced Research Projects Agency (DARPA) research program on human language technology. Comparative evaluation consists of a set of participants who compare the results of their systems using similar tasks and related data with agreed-upon metrics. Usually this evaluation is performed in a number of successive evaluation campaigns with more complex tasks performed at each stage. The ELSE approach departs from the DARPA research program in two ways: first by considering usability criteria in the evaluation, and second by trading competitive aspects for more contrastive and collaborative ones through the use of multidimensional results (Paroubek & Blasband, 1999). The ELSE consortium has identified the following five types of evaluation (Paroubek & Blasband, 1999):

- Basic research evaluation: tries to validate a new idea or assess the amount of improvement it creates in older methods
- Technology evaluation: tries to assess the performance and appropriateness of a technology for solving a problem that is well-defined, simplified, and abstracted
- Usage evaluation: tries to assess the usability of a technology for solving a real problem in the field. It involves the end-users in the environment intended for the deployment of the system under test
- Impact evaluation: tries to measure the socio-economic consequences of a technology
- Program evaluation: attempts to determine how worthwhile a funding program has been for a given technology

The Expert Advisory Group on Language Engineering Standards—Evaluation Workgroup (EAGLES) (Centre for Language Technology, 2000), phase one (1993–1995) and phase two (1997–1998), was a

European initiative that proposed a user-centered evaluation of NLP systems. The EAGLES work took as its starting point an existing standard, ISO 9126, which is concerned primarily with the definition of quality characteristics to be used in the evaluation of software products.

The Diagnostic and Evaluation Tools (DiET) project (1997–1999) was designed to develop data, methods, and tools for the glass-box evaluation of NLP components, building on the results of previous projects covering different aspects of assessment and evaluation. The Web page of the DiET project (http://www.dfki.de.lt/projects/diet-e.html) indicates that the project "will extend and develop test-suites with annotated test items for grammar, morphology, and discourse for English, French, and German. DiET will provide user support in terms of database technology, test-suite construction tools, and graphic interfaces." Further, it "will result in a tool-package for in-house and external quality assurance and evaluation, which will enable the commercial user to assess and compare Language Technology products."

The Message Understanding Conferences, now ceased, pioneered an international platform for sharing research on NLP systems. In particular, MUC researchers were involved in the evaluation of IE systems applied to a common task. The first five MUCs focused on analyzing free text, identifying events of a specified type, and filling a database template with information about each such event (http://www.cs.nyu.edu/cs/faculty/grishman/muc6.html). After MUC-5, a broad set of objectives was defined for subsequent MUCs, such as, to push information extraction systems toward greater portability to new domains, and to encourage evaluations of some basic language analysis technologies. In MUC-7 (the last MUC), the multilingual named entities evaluation was run using training and test articles from comparable domains for all languages (Chinchor, 2001). The MUC-7 papers report some interesting observations by system developers who were not native speakers of the languages of their systems and system developers who were. Results of MUC-3 through MUC-7 have been summarized by Chinchor (2001).

Conclusions

Some NLP experiments reported in this chapter show encouraging results. However, one should not forget that most of these experimental

systems remain in the lab; very few experimental systems are converted to real systems or products. One of the major stumbling blocks of NLP research, as in areas like information retrieval research, has been the absence of large test collections and reusable experimental methods and tools. Fortunately, the situation has changed over the past few years. Several national and international research groups are now working together to build and reuse large test collections and experimental tools and techniques. Since the origin of the Message Understanding Conferences, group research efforts have proliferated with regular conferences and workshops; for example, the TREC series and other conferences organized by the North American Chapter of the Association for Computational Linguistics, European ACL, and so on. These group efforts help researchers share their expertise by building reusable NLP tools, test collections, and experimental methodologies. References to some reusable NLP tools and cooperative research groups have already been made in this chapter (see the section "Some Theoretical Developments").

Some recent studies on evaluation are promising. Very small-scale evaluation of INFOS suggests that the indexing pattern method, i.e., mapping of the words from the input text into appropriate concepts in the WordNet abstraction hierarchy, correctly classified 80 percent of the articles (Mock & Vemuri, 1997). Some large-scale experiments with NLP also show encouraging results. For example, Kwok et al. (1999, 2000) report that their PIRCS system can perform the tasks of English-Chinese query translation with an effectiveness of over 80 percent. Strzalkowski et al. (1998) report that by automatically expanding queries using NLP techniques, they obtained an average 37 percent improvement over a baseline where no expansion was used. Conflicting results have arisen, too. For example, Elworthy (2000) reports that the NLP system, using the Microsoft product NLPWin, performed much more poorly in the TREC-9 test set compared with the TREC-8 test set. While trying to find out the reasons for this discrepancy, Elworthy (2000) comments that an important challenge may be figuring out how to build a system that merges definitive, pre-encoded knowledge, with ad hoc documents of unknown reliability.

As already mentioned in the section on "Abstracting," Craven's study with TEXNET (Craven, 1996) resulted in limited success (only

37 percent). Gaizauskas and Wilks (1998) mention that the performance levels of the common IE systems lie in the 50 percent range for combined recall and precision. Such low success rates are not acceptable in large-scale operational information systems.

Smith (1998) suggests that there are two possible scenarios for future relations between computers and humans: (1) in the user-friendliness scenario, computers become smart enough to communicate in natural language; and (2) in the computer friendliness scenario, humans adapt their practices in order to communicate with, and make use of, computers. He further argues that the use of computer-friendly encoding of natural language texts on the Web is symptomatic of a revolutionary trend toward the computerization of human knowledge. Petreley (2000, p. 102) raises a very pertinent question about natural language user interfaces: "Will the natural language interface have to wait until voice recognition becomes more commonplace?" This question appears to be quite legitimate when we see that, although a large number of natural language user interfaces were built—most at the laboratory level and a few at the commercial level (for details of these see Chowdhury, 1999b, [chapters 18–21]; Haas, 1996)—natural language user interfaces are still not common. The impediments to progress lie on several planes, including language issues. Zadrozny et al. (2000) mention that, except for very restricted domains, we do not know how to compute the meaning of a sentence based on the meanings of its words and its context. Another problem is caused by the lack of precise user models. Zadrozny et al. (2000) maintain that, even assuming that we could have any piece of information about a person, we do not know how to use this knowledge to make this person's interaction with a dialogue system most effective and pleasant.

MT involves a number of difficult problems, mainly because human language is at times highly ambiguous, full of special constructions, and replete with exceptions to rules. Nonetheless, there has been steady development, and MT research has now reached a stage where the benefits can be enjoyed. A number of Web search tools—AltaVista, Google, Lycos, and AOL—offer free MT facilities of Web-based information resources. A number of companies also provide MT services commercially. For example, the IBM WebSphere Translation Server for Multiplatforms is a machine translation service available commercially

for translating Web documents in a number of languages, such as English, French, Italian, Spanish, Chinese, Japanese, and Korean. In June 2001, Autodesk, a U.S. software company, began to offer MT services to its European customers at a cost that is 50 percent less than human translation services (Schenker, 2001). Although machine translations are not always perfect and do not produce translations as good as those produced by humans, the results and evidence of interest in improving the performance level of MT systems are very encouraging.

One application area that has drawn much research attention, but where the results have yet to provide the general public with an acceptable level of performance, is the natural language question-answering system. While some systems, as already noted, produce acceptable results, there are still many failures and surprises. Results from systems tested under the QA track of TREC (reported in the "Natural Language Interfaces" section) show promising results with some simple types of natural language queries. However, these systems are still in the experimental stage, and much research is needed before robust QA systems can be built that are capable of accepting user queries in any form of natural language and producing natural language answers from distributed information resources. Scalability and portability are the main challenges facing natural language text processing research. Adams (2001) argues that current NLP systems establish patterns that are valid for a specific domain and for a particular task only; as soon as the topic, context, or user changes, entirely new patterns must be established. Sparck Jones (1999) rightly warns that advanced NLP techniques, such as concept extraction, are too expensive for large-scale NLP applications. The research community, however, is making continuous efforts. The reason for not having reliable NLP systems that work at a high level of performance with a high degree of sophistication may have less to do with the inefficiency of the systems or researchers than with the complexities and idiosyncrasies of human behavior and communication patterns.

Bibliography

Adams, K. C. (2001). The Web as a database: New extraction technologies & content management, *Online*, *25*, 27–32.

Ahonen, H., Heinonen, O., Klemettinen, M., & Verkamo, A. I. (1998). Applying data mining techniques for descriptive phrase extraction in digital document

collections. *IEEE International Forum on Research and Technology. Advances in Digital Libraries—ADL '98*, 2–11.

Amsler, R. A. (1984). Machine-readable dictionaries. *Annual Review of Information Science and Technology, 19*, 161–209.

Argamon, S., Dagan, I., & Krymolowski, Y. (1998). A memory-based approach to learning shallow natural language patterns. *17th International Conference on Computational Linguistics (COLING '98)*, 67–73.

Bangalore, S. & Joshi, A. K. (1999). Supertagging: An approach to almost parsing. *Computational Linguistics, 25*, 237–265.

Barahona, P. & Alferes, J. J. (Eds.). (1999). Progress in artificial intelligence. *9th Portuguese Conference on Artificial Intelligence, EPIA '99*. Berlin: Springer-Verlag.

Barker, K. & Cornacchia, N. (2000). Using noun phrase heads to extract document keyphrases. In H. J. Hamilton (Ed.), *Advances in artificial intelligence. Proceedings of 13th Biennial Conference of the Canadian Society for Computational Studies of Intelligence, AI 2000*, 40–52.

Benoît, G. (2002). Data mining. *Annual Review of Information Science and Technology, 36*, 265–310.

Bian, G.-W. & Chen, H.-H. (2000). Cross-language information access to multilingual collections on the Internet. *Journal of the American Society for Information Science, 51*, 281–296.

Black, W. J., Rinaldi, F., & McNaught, J. (2000). Natural language processing in Java: Applications in education and knowledge management. *Proceedings of the Second International Conference on the Practical Application of Java*, 157–170.

Bondale, N., Maloor, P., Vaidyanathan, A., Sengupta, S. & Rao, P. V. S. (1999). Extraction of information from open-ended questionnaires using natural language processing techniques. *Computer Science and Informatics, 29*, 15–22.

Borgman, C. L. (1997). Multi-media, multi-cultural, and multi-lingual digital libraries: Or how do we exchange data in 400 languages? *D-Lib Magazine*. Retrieved December 5, 2001, from http://www.dlib.org/dlib/june97/06borgman.html.

Breck, E., Burger, J., House, D., Light, M., & Mani, I. (1999). Question answering from large document collections. *Question Answering Systems. Papers from the 1999 AAAI Fall Symposium*, 26–31.

Carrick, C. & Watters, C. (1997). Automatic association of news items. *Information Processing & Management, 33*, 615–632.

Centre for Language Technology. (2000). EAGLES-ll Information Page: Evaluation of NLP Systems. Retrieved December 5, 2001, from http://www.cst.ku.dk/projects/eagles2.html.

Ceric, V. (2000). Advancements and trends in the World Wide Web search. In D. Kalpic & V. H. Dobric (Eds.), *Proceedings of the 22nd International Conference on Information Technology Interfaces* (pp. 211–220). Zagreb: SRCE University Computer Centre.

Chandrasekar, R. & Srinivas, B. (1998). Glean: Using syntactic information in document filtering. *Information Processing & Management, 34*, 623–640.

Chang, H.-H., Ko, Y.-H., & Hsu, J.-P. (2000). An event-driven and ontology-based approach for the delivery and information extraction of e-mails. In *Proceedings/International Symposium on Multimedia Software Engineering* (pp.103–109). Los Alamitos, CA: IEEE Computer Society.

Charniak, E. (1995). Natural language learning. *ACM Computing Surveys, 27*, 317–319.

Chen, J. N. & Chang, J. S. (1998). Topical clustering of MRD senses based on information retrieval techniques. *Computational Linguistics, 24*, 61–96.

Chinchor, N. A. (2001.) Overview of MUC-7/MET-2. Retrieved December 5, 2001, from http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/muc_7_proceedings/overview.html.

Chowdhury, G. G. (1999a). Template mining for information extraction from digital documents. *Library Trends, 48*, 182–208.

Chowdhury, G. G. (1999b). *Introduction to modern information retrieval.* London: Library Association Publishing.

Chuang, W. & Yang, J. (2000). Extracting sentence segments for text summarization: A machine learning approach. *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 152–159.

Costantino, M. (1999). Natural language processing and expert system techniques for equity derivatives trading: The IE-Expert system. In D. Kalpic & V. H. Dobric (Eds.), *Proceedings of the 21st International Conference on Information Technology Interfaces*, (pp. 63–69). Zagreb, Croatia: University of Zagreb.

Cowie, J. & Lehnert, W. (1996). Information extraction. *Communications of the ACM, 39*(1), 80–91.

Craven, T. C. (1988). Text network display editing with special reference to the production of customized abstracts. *Canadian Journal of Information Science, 13*, 59–68.

Craven, T. C. (1993). A computer-aided abstracting tool kit. *Canadian Journal of Information Science, 18*, 19–31.

Craven, T. C. (1996). An experiment in the use of tools for computer-assisted abstracting. *Proceedings of the 59th ASIS Annual Meeting*, 203–208.

Craven, T. C. (2000). Abstracts produced using computer assistance. *Journal of the American Society for Information Science, 51*, 745–756.

Dogru, S., & Slagle, J. R. (1999). Implementing a semantic lexicon. In W. Tepfenhart & W. Cyre (Eds.), *Conceptual Structures: Standards and Practices. 7th International Conference on Conceptual Structures* (pp. 154–167). Berlin: Springer-Verlag.

Elworthy, D. (2000). Question answering using a large NLP system. *The Ninth Text REtrieval Conference (TREC 9)*. Retrieved December 5, 2001, from http://trec. nist.gov/pubs/trec9/papers/msrc-qa.pdf.

Evans, M. (1989). Computer-readable dictionaries. *Annual Review of Information Science and Technology, 24*, 85–117.

Feldman, S. (1999). NLP meets the jabberwocky. *Online*, *23*, 62–72.

Fellbaum, C. (Ed.). (1998). WordNet: An electronic lexical database. Cambridge, MA: MIT Press.

Fernandez, P. M. & Garcia-Serrano, A. M. (2000). The role of knowledge-based technology in language applications development. *Expert Systems With Applications*, *19*, 31–44.

Gaizauskas, R., & Wilks, Y. (1998). Information extraction: Beyond document retrieval. *Journal of Documentation*, *54*, 70–105.

Glasgow, B., Mandell, A., Binney, D., Ghemri, L., & Fisher, D. (1998). MITA: An information-extraction approach to the analysis of free-form text in life insurance applications. *AI Magazine*, *19*(1), 59–71.

Global Reach (2001). Global Internet Statistics (by language). Retrieved December 5, 2001, from http://www.euromktg.com/globstats.

Goldstein, J., Kantrowitz, M., Mittal, V., & Carbonell, J. (1999). Summarizing text documents: Sentence selection and evaluation metrics. *Proceeding of the ACM SIGIR 22nd Annual International Conference on Research and Development in Information Retrieval*, 121–128.

Grefenstette, G. (1999). The World Wide Web as a resource for example-based machine translation tasks. *Translating and the Computer 21. Proceedings of the Twenty-first International Conference on Translating and the Computer.* Retrieved April 29, 2002, from: http://www.xcre.xerox.com/competencies/content-analysis/publications/Documents/P49030/content/ggaslib.pdf.

Grishman, R. & Kittredge, R. (Eds.). (1986). *Analyzing language in restricted domains: Sublanguage descriptions and processing*. London: Lawrence Erlbaum.

Haas, S. W. (1996). Natural language processing: Toward large-scale robust systems. *Annual Review of Information Science and Technology*, *31*, 83–119.

Hayes, P. (1992). Intelligent high-volume text processing using shallow, domain-specific techniques. In P. S. Jacobs (Ed.), *Text-based intelligent systems*, (pp. 227–241). Hillsdale, NJ: Lawrence Erlbaum.

Hayes, P. & Weinstein, S. (1991). Construe-TIS: A system for content-based indexing of a database of news stories. In A. Rappaport, & R. Smith (Eds.), *Innovative applications of artificial intelligence 2* (pp. 51–64). Cambridge, MA: MIT Press.

Hedlund, T., Pirkola, A., & Järvelin, K. (2001). Aspects of Swedish morphology and semantics from the perspectives of mono- and cross-language information retrieval. *Information Processing & Management*, *37*, 147–161.

Herath, S. & Herath, A. (1999). Algorithm to determine the subject in flexible word order language based machine translations: A case study for Sinhalese. *Communications of COLIPS*, *9*, 1–17.

Ide, N. (2000). Cross-lingual sense determination: Can it work? *Computers and the Humanities*, *34*, 223–234.

Isahara, H., Ma, Q., Sornlertlamvanich, V., & Takahashi, N. (2000). ORCHID: Building linguistic resources in Thai. *Literary & Linguistic Computing*, *15*, 465–478.

Jelinek, F. (1999). Statistical methods for speech recognition, Cambridge, MA: MIT Press.

Jeong, K. S., Mayeng, S. H., Lee, J. S., & Choi, K. S. (1999). Automatic identification and back-transliteration of foreign words for information retrieval. *Information Processing & Management, 35*, 523–540.

Jurafsky, D. & Martin, J. H. (2000). *Speech and language processing: An introduction to natural language processing, computational linguistics and speech recognition*. Upper Saddle River, NJ: Prentice Hall.

Kazakov, D., Manandhar, S., & Erjavec, T. (1999). Learning word segmentation rules for tag prediction. In S. Dzeroski & P. Flach (Eds.), *Inductive Logic Programming. 9th International Workshop, ILP-99 Proceedings* (pp. 152–161). Berlin: Springer-Verlag.

Kehler, A. (1997). Current theories of centering for pronoun interpretation: A critical evaluation. *Computational Linguistics, 23*, 467–475.

Khoo, C. S. G., Myaeng, S. H., & Oddy, R. N. (2001). Using cause-effect relations in text to improve information retrieval precision. *Information Processing & Management, 37*, 119–145.

Kim, T., Sim, C., Sanghwa, Y., & Jung, H. (1999). From To-CLIR: Web-based natural language interface for cross-language information retrieval. *Information Processing & Management, 35*, 559–586.

King, M. (1996). Evaluating natural language processing systems. *Communications of the ACM, 39*(1), 73–80.

Kornai, A. (Ed.). (1999). *Extended finite state models of language*. Cambridge, UK: Cambridge University Press.

Kwok, K. L., Grunfeld, L., Dinstl, N., & Chan, M. (1999). TREC-8 ad-hoc, query filtering experiments using PIRCS. *The Eighth Text REtrieval Conference (TREC 8)*. Retrieved December 5, 2001, from http://trec.nist.gov/pubs/trec8/papers/ queenst8.pdf.

Kwok, K. L., Grunfeld, L., Dinstl, N., & Chan, M. (2000). TREC-9 cross language, Web and question-answering track experiments using PIRCS. *The Ninth Text REtrieval Conference (TREC 9)*. Retrieved December 5, 2001, from http://trec.nist.gov/pubs/trec9/t9_proceedings.html.

Lange, H. (1993). Speech synthesis and speech recognition: Tomorrow's human-computer interfaces? *Annual Review of Information Science and Technology, 28*, 153–185.

Lee, K. H., Ng, M. K. M., & Lu, Q. (1999). Text segmentation for Chinese spell checking. *Journal of the American Society for Information Science, 50*, 751–759.

Lehmam, A. (1999). Text structuration leading to an automatic summary system: RAFI. *Information Processing & Management, 35*, 181–191.

Lehtokangas, R. & Järvelin, K. (2001). Consistency of textual expression in newspaper articles: An argument for semantically based query expansion. *Journal of Documentation, 57*, 535–548.

Lewis, D. D. & Sparck Jones, K. (1996). Natural language processing for information retrieval. *Communications of the ACM, 39*(1), 92–101.

Liddy, E. (1998). Enhanced text retrieval using natural language processing. *Bulletin of the American Society for Information Science, 24*(4), 14–16.

Liddy, E., Diamond, T., & McKenna, M. (2000). DR-LINK in TIPSTER III. *Information Retrieval, 3*, 291–311.

Lovis, C., Baud, R., Rassinoux, A. M., Michel, P. A., & Scherter, J. R. (1998). Medical dictionaries for patient encoding systems: A methodology. *Artificial Intelligence in Medicine, 14*, 201–214.

Ma, Q., Kanzaki, K., Murata, M., Utiyama, M., Uchimoto, K., & Isahara, H. (2000). Self-organizing semantic maps of Japanese nouns in terms of adnominal constituents. In S. Herath & A. Herath (Eds.), *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium*, (pp. 91–96). Los Alamitos, CA: IEEE Computer Society.

Magnini, B., Not, E., Stock, O., & Strapparava, C. (2000). Natural language processing for transparent communication between public administration and citizens. *Artificial Intelligence and Law, 8*, 1–34.

Mani, I. & Maybury, M. T. (1999). *Advances in automatic text summarization*. Cambridge, MA: MIT Press.

Manning, C. D. & Schutze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.

Marquez, L., Padro, L., & Rodriguez, H. (2000). A machine learning approach to POS tagging. *Machine Learning, 39*, 59–91.

Martinez, P., De Miguel, A., Cuadra, D., Nieto, C., & Castro, E. (2000). Data conceptual modelling through natural language: Identification and validation of relationship cardinalities. *Challenges of Information Technology Management in the 21st Century. 2000 Information Resources Management Association International Conference* (pp. 500–504). Hershey, PA: Idea Group Publishing.

Martinez, P. & Garcia-Serrano, A. (1998). A knowledge-based methodology applied to linguistic engineering. In R. N. Horspool (Ed.), *Systems Implementation 2000. IFIP TC2 WG2.4 Working Conference on Systems Implementation 2000: Languages, Methods and Tools* (pp. 166–179). London: Chapman & Hall.

McMurchie, L. L (1998). Software speaks user's language. *Computing Canada, 24*, 19–21.

Meyer, J. & Dale, R. (1999). Building hybrid knowledge representations from text. In J. Edwards (Ed.), *Proceedings of the 23rd Australasian Computer Science Conference. ACSC 2000*, (pp. 158–165). Los Alamitos, CA: IEEE Computer Society.

Mihalcea, R. & Moldovan, D. I. (1999). Automatic acquisition of sense tagged corpora. In A. N. Kumar & I. Russell (Eds.), *Proceedings of the Twelfth International Florida AI Research Society Conference*, (pp. 293–297). Menlo Park, CA: AAAI Press.

Mock, K. J. & Vemuri, V. R. (1997). Information filtering via hill climbing, WordNet and index patterns. *Information Processing & Management, 33*, 633–644.

Moens, M.-F. & Uyttendaele, C. (1997). Automatic text structuring and categorization as a first step in summarizing legal cases. *Information Processing & Management, 33*, 727–737.

Morin, E. (1999). Automatic acquisition of semantic relations between terms from technical corpora. In P. Sandrini (Ed.), *TKE '99. Terminology and Knowledge Engineering. Proceedings, Fifth International Congress on Terminology and Knowledge Engineering*, (pp. 268–278). Vienna: TermNet.

Narita, M. & Ogawa, Y. (2000). The use of phrases from query texts in information retrieval. *SIGIR Forum, 34*, 318–320.

Nerbonne, J., Dokter, D., & Smit, P. (1998). Morphological processing and computer-assisted language learning. *Computer Assisted Language Learning, 11*, 543–559.

Oard, D. W. (1997). Serving users in many languages: Cross-language information retrieval for digital libraries. *D-Lib Magazine*. Retrieved December 5, 2001, from http://www.dlib.org/dlib/december97/oard/12oard.html.

Oard, D. W. & Diekema, A. R. (1998). Cross-language information retrieval. *Annual Review of Information Science and Technology, 33*, 223–256.

Ogura, K., Nakaiwa, H., Matsuo, Y., Ooyama, Y., & Bond, F. (2000). The electronic dictionary. Goi-Taikei: A Japanese lexicon and its applications. *NTT Review*, 12, 53–58.

Oudet, B. (1997). Multilingualism on the Internet. *Scientific American, 276*(3), 77–78.

Owei, V. (2000). Natural language querying of databases: An information extraction approach in the conceptual query language. *International Journal of Human-Computer Studies, 53*, 439–492.

Paris, L. A. H. & Tibbo, H. R. (1998). Freestyle vs. Boolean: A comparison of partial and exact match retrieval systems. *Information Processing & Management, 34*, 175–190.

Paroubek, P. & Blasband, M. (1999). Executive summary of a blueprint for a general infrastructure for natural language processing systems evaluation using semi-automatic quantitative black box approach in a multilingual environment. Retrieved December 5, 2001, from http://www.limsi.fr/TLP/ELSE/Preamble XwhyXwhatXrev3.htm.

Pasero, R., & Sabatier, P. (1998). Linguistic games for language learning: A special use of the ILLICO library. *Computer Assisted Language Learning, 11*, 561–585.

Pedersen, T., & Bruce, R. (1998). Knowledge lean word-sense disambiguation. *Proceedings Fifteenth National Conference on Artificial Intelligence (AAAI-98). Tenth Conference on Innovative Applications of Artificial Intelligence* (pp. 800–805). Menlo Park, CA: AAAI Press/MIT Press.

Perez-Carballo, J. & Strzalkowski, T. (2000). Natural language information retrieval: Progress report. *Information Processing & Management, 36*, 155–178.

Peters, C. & Picchi, E. (1997). Across languages, across cultures: Issues in multilinguality and digital libraries, *D-Lib Magazine*. Retrieved December 5, 2001, from http://www.dlib.org/dlib/may97/peters/05peters.html.

Petreley, N. (2000). Waiting for innovations to hit the mainstream: What about natural language? *InfoWorld, 22*(4), 102.

Pirkola, A. (2001). Morphological typology of languages for IR. *Journal of Documentation, 57*, 330–348.

Poesio, M. & Vieira, R. (1998). A corpus-based investigation of definite description use. *Computational Linguistics, 24*, 183–216.

Powell, J. & Fox, E. A. (1998). Multilingual federated searching across heterogeneous collections. *D-Lib Magazine*. Retrieved December 5, 2001, from http://www.dlib.org/dlib/september98/powell/09powell.html.

Qin, J. & Norton, M. J. (Eds.). (1999). Introduction. [Special Issue] Knowledge discovery in bibliographic databases. *Library Trends, 48*, 1–8.

Raghavan, V. V., Deogun, J. S., & Server, H. (Eds.). (1998). Knowledge discovery and data mining [Special topics issue] *Journal of the American Society for Information Science, 49*(5).

Roche, E. & Shabes, Y. (Eds.). (1997). *Finite-state language processing*. Cambridge, MA: MIT Press.

Rosenfield, R. (2000). Two decades of statistical language modeling: Where do we go from here? *Proceedings of the IEEE. 88*(8), 1270–1278.

Roux, M. & Ledoray, V. (2000). Understanding of medico-technical reports. *Artificial Intelligence in Medicine, 18*, 149–172.

Ruiz, M. E., & Srinivasan, P. (1998). Cross-language information retrieval: An analysis of errors. *Proceedings of the 61st ASIS Annual Meeting*, 153–165.

Say, B. (1999). Modeling cue phrases in Turkish: A case study. In V. Matousek et al. (Eds.). *Text, speech and dialogue. Second International Workshop, TDS '99* (pp. 337–340). Berlin: Springer-Verlag.

Scarlett, E., & Szpakowicz, S. (2000). The power of the TSNLP: Lessons from a diagnostic evaluation of a broad-coverage parser. In H. J. Hamilton (Ed.), *Advances in Artificial Intelligence. 13th Biennial Conference of the Canadian Society for Computational Studies of Intelligence* (pp. 138–150). Berlin: Springer-Verlag.

Schenker, J. L., (2001, July 16). The gist of translation: How long will it be before machines make the Web multilingual? *Time*, 158, 54.

Scott, J. (1999, December). E-mail management: The key to regaining control. *Internet Business*, 60–65.

Silber, H. G., & McCoy, K. F. (2000). Efficient text summarization using lexical chains. In H. Lieberman (Ed.), *Proceedings of IUI 2000 International Conference on Intelligent User Interfaces* (pp. 252–255). New York: ACM.

Smeaton, A. F. (1999). Using NLP or NLP resources for information retrieval tasks. In T. Strzalkowski (Ed.), *Natural language information retrieval* (pp. 99–111). Dordrecht, Netherlands: Kluwer Academic.

Smeaton, A. F. (1997). Information retrieval: Still butting heads with natural language processing? In M. T. Pazienza (Ed.), *Information extraction. A multidisciplinary approach to an emerging information technology international summer school, SCIE '97* (pp. 115–138). Berlin: Springer-Verlag.

Smith, D. (1998). Computerizing computer science. *Communications of the ACM, 41* (9), 21–23.

Sokol, L., Murphy, K., Brooks, W., & Mattox, D. (2000). Visualizing text-based data mining. *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining* (pp. 57–61). Blackpool, UK: Practical Application Company.

Song J., & Zhao, D.-Y. (2000). Study of automatic abstracting based on corpus and hierarchical dictionary. *Journal of Software, 11*, 308–314.

Sparck Jones, K. (1999). What is the role for NLP in text retrieval? In T. Strzalkowski (Ed.), *Natural language information retrieval* (pp. 1–25). Dordrecht, Netherlands: Kluwer.

Staab, S., Braun, C., Bruder, I., Dusterhoft, A., Heuer, A., Klettke, M., et al. (1999). GETESS-searching the Web exploiting German texts. *Cooperative Information Agents III. Third International Workshop, CIA '99* (pp. 113–124). Berlin: Springer-Verlag.

Stock, O. (2000). Natural language processing and intelligent interfaces. *Annals of Mathematics and Artificial Intelligence, 28*, 39–41.

Strzalkowski, T., Fang, L., Perez-Carballo, J., & Jin, W. (1997). *Natural language information retrieval TREC-6 Report, NIST Special Publication 500-240.* Retrieved December 5, 2001, from http://trec.nist.gov/pubs/trec6/t6_proceedings.html

Strzalkowski, T., Perez-Carballo, J., Karlgren, J., Hulth, A., Tapanainen, P., & Lahtinen, T. (1999). *Natural language information retrieval: TREC-8 report. NIST Special Publication 500-246.* Retrieved December 5, 2001, from http://trec.nist. gov/pubs/trec8/papers/ge8adhoc2.pdf.

Strzalkowski, T., Stein, G., Wise, G. B., Perez-Carballo, J., Tapanainen, P., Jarvinen, et al. (1998). *Natural language information retrieval: TREC-7 report. NIST Special Publication 500-242.* Retrieved December 5, 2001, from http://trec.nist.gov/pubs/trec7/t7_proceedings.html.

Tolle, K. M. & Chen, H. (2000). Comparing noun phrasing techniques for use with medical digital library tools. *Journal of the American Society for Information Science, 51*, 352–370.

Trybula, W. J. (1997). Data mining and knowledge discovery. *Annual Review of Information Science and Technology, 32*, 197–229.

Tsuda, K., & Nakamura, M. (1999). The extraction method of the word meaning class. In L. C. Jain (Ed.), *Third International Conference on Knowledge-Based Intelligent Information Engineering Systems* (pp. 534–537). Piscataway, NJ: IEEE.

Vickery, B. (1997). Knowledge discovery from databases: An introductory review. *Journal of Documentation, 53*, 107–122.

Voorhees, E. (1999). The TREC-8 question answering track report. Retrieved December 5, 2001, from http://trec.nist.gov/pubs/trec8/papers/qa-report.pdf.

Voorhees, E. (2000). The TREC-9 question answering track report. Retrieved December 5, 2001, from http://trec.nist.gov/pubs/trec9/papers/qa-report.pdf.

Waldrop, M. M. (2001). Natural language processing. *Technology Review, 104*, 107–108.

Warner, A. J. (1987). Natural language processing. *Annual Review of Information Science and Technology, 22*, 79–108.

Weigard, H., & Hoppenbrouwers, S. (1998). Experiences with a multilingual ontology-based lexicon for news filtering. In A. M. Tjoa & R. R. Wagner (Eds.), *Proceedings of the Ninth International Workshop on Database and Expert Systems Applications* (pp. 160–165). Los Alamitos, CA: IEEE Computer Society.

Wilks, Y. (1996). Natural language processing. *Communications of the ACM, 39*(1), 60.

Wong, K.-F., Lum, V. Y. & Lam, W.-I.(1998). Chicon: A Chinese text manipulation language. *Software - Practice and Experience, 28*, 681–701.

Yang, D. H., Gomez, P. C., & Song, M. (2000). An algorithm for predicting the relationship between lemmas and corpus size. *ETRI Journal, 22*, 20–31.

Yang, J. C. & Akahori, K. (2000). A discourse structure analysis of technical Japanese texts and its implementation on the WWW. *Computer Assisted Language Learning, 13*, 119–141.

Yang, Y. & Liu, X. (1999). A re-examination of text categorization methods. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 42–49.

Zadrozny, W., Budzikowska, M., Chai, J., & Kambhatla, N. (2000). Natural language dialogue for personalized interaction. *Communications of the ACM, 43*(8), 116–120.

Zweigenbaum, P., & Grabar, N. (1999). Automatic acquisition of morphological knowledge for medical language processing. In W. Horn, et al. (Eds.), *Joint European Conference on Artificial Intelligence in Medicine and Medical Decision Making* (pp. 416–420). Berlin: Springer-Verlag.