Chowdhury, G. (2003) *Natural language processing.* Annual Review of Information Science and Technology, 37. pp. 51-89. ISSN 0066-4200

Natural Language Processing

Gobinda G. Chowdhury
Dept. of Computer and Information Sciences
University of Strathclyde, Glasgow G1 1XH, UK
e-mail: gobinda@dis.strath.ac.uk

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 the desired tasks. The foundations of NLP lie in a number of disciplines, viz. computer and information sciences, linguistics, mathematics, electrical and electronic engineering, artificial intelligence and robotics, psychology, etc. Applications of NLP include a number of fields of studies, 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, and so on.

One important area of application of NLP that is relatively new and has not been covered in the previous ARIST chapters on NLP has become quite prominent due 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 www and digital libraries (see for example, Borgman, 1997; Peters & Picchi, 1997)

Scope

Several ARIST chapters have reviewed the field of NLP. The most recent ones include that by Warner in 1987, and Haas in 1996. Reviews of literature on large-scale NLP systems, as well as the various theoretical issues have also appeared in a number of publications (see for example, Jurafsky & Martin, 2000; Manning & Schutze, 1999; Mani & Maybury, 1999; Sparck Jones, 1999; Wilks, 1996). Smeaton (1999) provides a good overview of the past research on the applications of NLP in various information retrieval tasks. Several ARIST chapters have appeared on areas related to NLP, such as on 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 a number of conferences such as the annual proceedings of ACL (Association of Computational Linguistics) 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 of 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 and 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 have reported them in journals specifically dealing with the domain concerned, 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), including:

(i) natural language text processing systems – text summarization, information extraction, information retrieval, etc., including domain-specific applications;
(ii) natural language interfaces;
(iii) NLP in the context of www and digital libraries ; and
(iv) evaluation of NLP systems.

Linguistic research in information retrieval has not been covered in this review, since this is a huge area and has been dealt with separately in this volume by David Blair. Similarly, NLP issues related to the information retrieval tools (search engines, etc.) for web search are not covered in this chapter since a separate chapter on indexing and retrieval for the Web has been written in this volume by Edie Rasmussen.

Tools and techniques developed for building NLP systems have been discussed in this chapter along with the specific areas of applications for which they have been built. Although machine translation (MT) is an important part, and in fact the origin, of NLP research, this paper does not cover this topic with sufficient detail since this is a huge area and demands a separate chapter on its own. Similarly, cross-language information retrieval (CLIR), although is a very important

and big area in NLP research, is not covered in great detail in this chapter. A separate chapter on CLIR research appeared in ARIST (Oard & Diekema, 1998). However, MT and CLIR have become two important areas of research in the context of the www digital libraries. This chapter reviews some works on MT and CLIR in the context of NLP and IR in digital libraries and www. Artificial Intelligence techniques, including neural networks etc., used in NLP have not been included in this chapter.

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 theoretical developments can be grouped into four classes: (i) statistical and corpus-based methods in NLP, (ii) recent efforts to use WordNet for NLP research, (iii) the resurgence of interest in finite-state and other computationally lean approaches to NLP, and (iv) 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 evidences of authors and speakers, and so on. Charniak (1995) points out that 90% accuracy can be obtained in assigning part-of-speech tag to a word 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 & Moldovan (1999) mention that although thus far statistical approaches have been considered the best for word sense disambiguation, they are useful only in 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 is an excellent NLP tool containing English nouns, verbs, adjectives and adverbs organized into synonym sets, each representing one underlying lexical concept. Details of WordNet is 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 applications. One of the major applications of WordNet in NLP has been in Europe with the formation 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/). The finite-state automation is the mathematical device 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 important applications of NLP. Different applications of the Finite State methods in NLP have been discussed by Jurafsky & Martin (2000), Kornai (1999) and Roche & Shabes (1997). The work of NLP researchers has been greatly facilitated by the availability of large-scale grammar for parsing and generation. Researchers can get access to large-scale grammars and tools through several websites, for example Lingo (http://lingo.stanford.edu), Computational Linguistics & Phonetics (http://www.coli.uni-sb.de/software.phtml), and Parallel grammar project (http://www.parc.xerox.com/istl/groups/nltt/pargram/). Another significant development in recent years is the formation of various national and international consortia and research groups that can facilitate, and help share expertise, research in NLP. LDC (Linguistic Data Consortium) (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 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, Fuji Xerox in Japan. The aim of this project is 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 to be held in India in early 2002 is expected to address various problems of NLP by researchers from different parts of the world.

Natural Language Understanding

At the core of any NLP task there is the important issue of natural language understanding. The process of building computer programs that understand natural language involves three major problems: the first one relates to the thought process, the second one to the representation and meaning of the linguistic input, and the third one to the world knowledge. Thus, an NLP system may begin at the word level – to determine the morphological structure, nature (such as part-of-speech, meaning) etc. of the word – and then may move on to the sentence level – to determine

the word order, grammar, meaning of the entire sentence, etc.— and then to the context and the overall environment or domain. A given word or a 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 a 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 and
- pragmatic level that deals with the knowledge that comes from the outside world, i.e., from outside the contents 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 come up with improved technology for performing various activities that form important parts of NLP works. These works may be categorized as follows:

- Lexical and morphological analysis, noun phrase generation, word segmentation, etc. (Bangalore & Joshi, 1999; Barker & Cornacchia,2000; Chen & Chang, 1998; Dogru & Slagle, 1999; Kam-Fai et al.. 1998; Kazakov et al.. , 1999; Lovis et al.. 1998; Tolle & Chen, 2000; Zweigenbaum & Grabar, 1999)

- Semantic and discourse analysis, word meaning and knowledge representation (Kehler, 1997; Mihalcea & Moldovan,1999; Meyer & Dale, 1999; Pedersen & Bruce, 1998; Poesio & Vieira,1998; Tsuda & Nakamura, 1999)

- Knowledge-based approaches and tools for NLP (Argamon et al.., 1998; Fernandez & Garcia-Serrano, 2000; Martinez et al.., 2000, 1998).

Dogru & Slagle (1999) propose a model of 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. Kam-Fai et al.. (1998) report the features of an NLP tool called *Chicon* used for word segmentation in Chinese text. Zweigenbaum & Grabar (1999) propose a method for acquiring morphological knowledge about words in medical literature. It 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 since 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 the NLP of medical texts.

Noun phrasing is considered to be an important NLP technique used in information retrieval. One of the major goals of noun phrasing research is to investigate the possibility of 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 only requires

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 their 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 & 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 & 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 Internet using existing search engines.

Martinez & 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 modelling heuristics (Martinez et al.. 2000). Fernandez & Garcia-Serrano (2000) comment that knowledge engineering is increasingly regarded as a means to complement traditional formal NLP models by adding symbolic modelling 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 that integrates different formalisms, reuses existing language resources and supports 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 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 some 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 get the morphological variants of the words involved. The lexical and syntactic processing involve the utilization of lexicons for determining the characteristics of the words, recognition of their parts-of-speech, determining the words and phrases, and for parsing of the sentences.

Past research concentrating on natural language text processing systems has been reviewed by Haas (1986), Mani & 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 patent texts.

Abstracting

Automatic abstracting and text summarization are now used synonymously that aim to generate abstracts or summaries of texts. This area of NLP research is becoming more common in the web and digital library environment. 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 interests in automatic abstracting and text summarization are reflected by the huge number of research papers appearing in a number of international conferences and workshops including ACL, ACM, AAAI, SIGIR, and various national and regional chapters of the Associations. Several techniques are used for automatic abstracting and text summarization. Goldstein et al.. (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 US patent abstracts.

Roux and Ledoray (2000) report a project, called *Aristotle*, that aims to build an automatic medical data system that is capable of producing a semantic representation of the 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 in a corpus. Jin and Dong-Yan (2000) propose a methodology for generating automatic abstracts that provides an integration of the advantages of methods based on linguistic analysis and those based on statistics.

Moens and Uyttendaele (1997) describe the SALOMON (Summary and Analysis of Legal texts FOR Managing Online Needs) project that automatically summarizes legal texts written in Dutch. The system extracts relevant information from the full texts of Belgian criminal cases and uses it to summarize each decision. A text grammar represented as a semantic network is used to determine the category of each case. The system extracts relevant information about each case, such as the name of the court that issues the decision, the decision date, the offences charged, the relevant statutory provisions disclosed by the court, as well as the legal principles applied in the case. RAFI (resume automatique a 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 technique 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, viz. previous knowledge, content, method and new knowledge.

Most of the 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 an abstractor's assistance software called TEXNET. However, recent experiments on the usefulness of the automatically extracted keywords and phrases from full texts by TEXNET in the actual process of abstracting by human abstractors showed some considerable variation among subjects, and only 37% of the subjects found the keywords and phrases to be 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 a number of information science journals have published special issues reporting research on these topics (see for example, Benoit, 2001;Qin and Norton, 1999; Raghavan et al.., 1998; Trybula, 1997; Vickery, 1997). Knowledge discovery and data mining research use a variety of techniques in order to extract useful information from source documents. Information extraction (IE) is a subset of knowledge discovery and data mining research that aims to extract useful bits of textual information from natural language texts (Gaizauskas & Wilks, 1998). A variety of information extraction (IE) techniques are used and the extracted information can be used for a number of purposes, for example to prepare a summary of texts, to populate databases, fill-in slots in frames, identify keywords and phrase for information retrieval, and so on. IE techniques are also used for classifying text items according to some pre-defined categories. An earlier example of text categorization system is CONSTRUE, developed for Reuters, that classifies news stories (Hayes, 1992). The CONSTRUE software was subsequently generalized into a commercial product called TCS (Text Categorization Shell). An evaluation of five text categorization systems has been reported by Yang and Liu (1999).

Morin (1999) suggests that although many IE systems can successfully extract terms from documents, acquisition of relations between terms is still a difficulty. PROMETHEE is a system that extracts lexico-syntactic patterns relative to a specific conceptual relation from technical

corpora (Morin,1999). Bondale et al.. (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) approach for the analysis of a real life natural language corpus that consists of responses to open-ended questionnaires in the field of advertising.

Glasgow et al.. (1998) report a system called MITA (Metlife's Intelligent Text Analyzer) that extracts information from life insurance applications. Ahonen et al.. (1998) propose a general framework for text mining that uses pragmatic and discourse level analyses of text. Sokol et al.. (2000) report research that uses visualization and NLP technologies to perform text mining. Heng-Hsou et al.. (2000) argue that IE systems are usually event-driven (i.e., are usually based on domain knowledge built on various events) and propose an event detection driven intelligent information extraction by using the neural network paradigm. They use the backpropagation (BP) learning algorithm to train the event detector, and apply NLP technology to aid the selection of nouns as feature words which are supposed to characterize documents appropriately. These nouns are stored in ontology as a knowledge base, and are used for the extraction of useful information from e-mail messages.

Cowie and Lehnert (1996) reviewed the earlier research on IE and commented that the NLP research community is ill-prepared to tackle the difficult problems of semantic feature-tagging, co-reference resolution, and discourse analysis, all of which are important issues of IE research. Gaizauskas and Wilks (1998) reviewed the IE research from its origin in the Artificial Intelligence world in the sixties and seventies through to the modern days. They discussed the major IE projects undertaken in different sectors, viz., 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: the extraction of proper names from full text documents, extraction of 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. He recommends that if each web author is given a template to fill-in in order to characterize his/her document, then eventually a more controlled and systematic method of creating document surrogates can be achieved. However, he warns that a single all-

purpose metadata format will not be applicable for all authors in all the domains, and further research is necessary to come up with appropriate formats for each.

Arguing that IR has been the subject of research and development and has been delivering working solutions for many decades whereas IE is a more recent and emerging technology, Smeaton (1997) comments that it is of interest to the IE community to see how a related task, perhaps the most-related task, IR, has managed to use the NLP base technology in its development so far. Commenting on the future challenges of IE researchers, Gaizauskas and Wilks (1998) mention that the performance levels of the common IE systems, which stand in the range of 50% for combined recall and precision, should improve significantly to satisfy information analysts. A major stumbling block of IE systems development is the cost of 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. Since they depend heavily on the domain knowledge, a given IE system may work satisfactorily in a relatively smaller 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 the NLP and the wrapper induction technology in information extraction from the web documents. In contrast to NLP, wrapper induction operates independently of specific domain knowledge. Instead of analysing the meaning of discourse at the sentence level, the wrapper technology identifies relevant content based on the textual qualities that surround desired data. Wrappers operate on the surface features of document texts that characterize 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 has been a major area of application of NLP, and consequently a number of research projects, dealing with the various applications on NLP in IR, have taken place throughout the world resulting in a large volume of publications. Lewis and Sparck Jones (1996) comment that the generic challenge for NLP in the field of IR is whether the necessary NLP of texts and queries is doable, and the specific challenges are whether non-statistical and statistical data can be combined and whether data about individual documents and whole files can be combined. They further comment that there are major challenges in making the NLP technology

operate effectively and efficiently and also in conducting appropriate evaluation tests to assess whether and how far the approach works in an environment of interactive searching of large text files. Feldman (1999) suggests 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 are more meaningful than statistically obtained word pairs, and thus are more powerful for discriminating among documents, Narita and Ogawa (2000) use a shallow syntactic processing instead of statistical processing to automatically 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 as the best for every query. In other words, their conclusion was that different queries demand different techniques.

Pirkola (2001) shows that languages vary significantly in their morphological properties. However, for each language there are two variables that describe the morphological complexity, viz., 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 index of fusion (IF) that describes the ease with which two morphemes can be separated in a language. Pirkola (2001) shows that calculation of the ISs and IFs in a language is a relatively simple task, and once they have been established, they could be utilized fruitfully in empirical IR research and system development.

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

Khoo et al.. (2001) report an experiment that investigates whether information obtained by matching cause-effect relations expressed in documents with the cause-effect relations expressed in user queries can be used to improve results in document retrieval compared with the use of only the keywords without considering the relations. Their experiment with the Wall Street Journal full text database revealed that causal relations matching where either the cause or the effect is a wildcard can be used to improve information retrieval effectiveness if the appropriate weight for each type of matching can be determined for each query. However, the authors stress that the results of this study were not as strong as they had expected it to be.

Chandrasekar & Srinivas (1998) propose that coherent text contains significant latent information, such as syntactic structure and patterns of language use, and this information could be used to improve the performance of information retrieval systems. They describe a system, called *Glean*, that uses syntactic information for effectively filtering irrelevant documents, and thereby improving the precision of information retrieval systems.

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 track, filtering track, interactive track, question-answering track, and the web track. Reports of progress of the NLIR (Natural Language Information Retrieval) project are available in the TREC reports (Perez-Carballo & Strzalkowski, 2000; Strzalkowski. et al.., 1997, 1998, 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 compared to the simple keyword and string-based methods used in statistical full-text retrieval (Strzalkowski, T. et al.., 1999). However, results indicate that simple linguistically motivated indexing (LMI) did not prove to be 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 compared to terse one-sentence search queries. Thus, it was concluded that query expansion, using NLP techniques, leads to a 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 results in

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 review of such systems see Chowdhury, 1999b, Chapter 19;Haas, 1996; Stock, 2000).

Much of the efforts in natural language interface design to date have focused on handling rather simple natural language queries. A number of 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 using NLP tools and techniques to get the correct answer from the source texts. Breck et al. (1999) report a question answering system that uses techniques from knowledge representation, information retrieval, and NLP. The authors claim that this combination enables 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 (Text Retrieval Conferences) show some interesting results. The basic technology used by the participants in the QA track included several steps. First, cue words/phrase like 'who' (as in 'who is the prime minister of Japan'), '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 for identifying the entities required for an answer. If no appropriate answer type was found then 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 2/3 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% 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 documents will be handled by QA track researchers.

Owei (2000) argues that the drawbacks of most natural language interfaces to database systems stem 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 argues that the difficulty with NL database query languages (DBQLs) can be overcome by combining concept based DBQL paradigms with NL approaches to enhance the overall ease-of-use of the query interface.

Zadrozny et al. (2000) suggest 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 then should be able to provide intelligent answers or ask relevant questions. However, they comment 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 (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 is using with the user), and
- Lack of middleware and toolkits.

NLP Software

A number of specific NLP software products have been developed over the past decades, some of which are available for free, while others are available commercially. Many such NLP software packages and tools have already been mentioned in the discussions throughout this chapter. Some more NLP tools and software are mentioned in this section.

Pasero & Sabatier (1998) describe principles underlying ILLICO, a generic natural-language software tool for building larger applications for performing specific linguistic tasks such as analysis, synthesis, and guided composition. Liddy (1998) and Liddy et al. (2000) discuss the commercial use of NLP in IR with the example of DR-LINK (**D**ocument **R**etrieval Using **LIN**guistic **K**nowledge) system demonstrating the capabilities of NLP for IR. Detailed product information and a demo of DR-LINK are now available online (http://www.textwise.com/dr-

link.html). Nerbonne et al. (1998) report on GLOSSER, an intelligent assistant for Dutch students for learning to read French. Scott (1999) describes the *Kana Customer Messaging System* that can categorize inbound e-mails, 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, called Brightware, that uses NLP techniques to elicit meaning from groups of words or phrases and reply to some e-mails automatically. *NLPWin* is an NLP system from Microsoft that accepts sentences and delivers detailed syntactic analysis, 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, 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 (Natural Language …, n.d.)

- *MARMOT Text Bracketting Module*, a text file translator which 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*, that analyzes bracketed text and produces case frame instantiations according to application-specific domain guidelines.
- *CRYSTAL Dictionary Induction Module*, that learns text extraction rules, suitable for use by BADGER, from annotated training texts.
- *ID3-S Inductive Learning Module*, a variant on ID3 which induces decision trees on the basis of training examples.

Waldrop (2001) briefly describes the features of three NLP software packages, viz.

- *Jupiter*, a product of the MIT research Lab that works in the field of weather forecast
- *Movieline*, a product of Carnegie Mellon that talks about local movie schedules, and

- *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, such as

- *ConQuest*, a part of *Excalibur*, that incorporates a lexicon that is 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 centres in computational linguistics and NLP, carried out under the auspices of a Socrates Working Group on Advanced Computing in the Humanities, revealed that Java has already reached the status of second most commonly taught programming language (Black et al., 2000). In addition, Java based programs are being used to develop interactive instructional materials. Black et al. (2000) review some Java-based courseware in use 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 improvements 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 some problems related to information processing and retrieval more prominent. According to a recent Survey (Global Reach, 2001), 55% of the Internet users are non-English speakers and this is increasing rapidly, thereby reducing the percentage of net users who are native English speakers. However, about 80% of the Internet and digital library resources available today are in English (Bian, Guo-Wei & Chen, 2000). This calls for the urgent need for the establishment of multilingual information systems and CLIR facilities. How to manipulate the large volume of multilingual data has become a major research question. In fact, several issues are involved here. At the user interface level, there has to be a query translation system that should translate the query 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 is a major issue here. The corpus-based approach uses parallel corpora for word selection, where the problem lies with the domain and scale of the corpora. Bian & Chen (2000) propose a Chinese-English CLIR system on www, called MTIR, that integrates the query translation and document translation. They also address a number of issues of machine translation on the web, viz., the role played by the 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) describe the features of an intelligent information agent called GETESS that uses semantic methods and NLP capabilities in order to gather tourist information from the web and present it to the human user in an intuitive, user-friendly way. Ceric (2000) reviews the advancements of the web search technology and mentions that, among others, NLP technologies will have very good impact on the success of the search engines. Mock and Vemuri (1997) describe the Intelligent News Filtering Organizational System (INFOS) that is designed to filter out unwanted news items from a Usenet. *INFOS* builds a profile of user interests based on the user feedback. After the user browses each article, INFOS asks the user to rate the article, and uses this as a criterion for selection (or rejection) of similar articles next time round. News articles are classified by a simple keyword method, called the Global Hill Climbing (GHC), that is used as a simple quick-pass method. Articles that cannot be classified by GHC are passed through a *WordNet* knowledgebase through a Case based reasoning (CBR) module which is a slower but more accurate method. Very small-scale evaluation of *INFOS* suggests 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% of the articles; the major reasons for errors being the weakness of the system to disambiguate pronouns.

One of the major stumbling blocks of providing personalized news delivery to users over the Internet is the problem involved in the automatic association of related items of different media type. Carrick and Watters (1997) describe a system that aims 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 in course of this research was tested against a test data set as well as new data sets. The pair of news items and

photos generated by the system were checked by human experts. The system performed, in terms of recall, precision and time, similarly on the new data sets as it did on the training set.

Because of the volume of text available on the web, many researchers have proposed to use the web as the testbed for NLP research. Grefenstette (1999) argues 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. There are two sets of issues here: (1) recognition, manipulation and display of multiple languages, and (2) cross-language information search and retrieval (Peter & Picchi, 1997). The first set of issues relate 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 (discussed below).

Oard (1997) comments 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 is not available in a language that the user can read, some form of translation will be needed. Multilingual thesauri such as EUROVOC help to address this challenge by facilitating controlled vocabulary search using terms from several languages, and services such as INSPEC produce English abstracts for documents in other languages (Oard, 1997). However, as Oard mentions, fully automatic MT is presently neither sufficiently fast nor sufficiently accurate to adequately support interactive cross-language information seeking in the web and digital libraries. Fortunately, an active and rapidly growing research community has coalesced around these and

other related issues, applying techniques drawn from several fields - notably IR and NLP - to provide access to large multilingual collections.

Borgman (1997) comments 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 illustrates the multi-language DL challenge with examples drawn from the research library community, which typically handles collections of materials in about 400 different languages.

Ruiz and Srinivasan (1998) investigate an automatic method for CLIR that utilizes the multilingual Unified Medical Language System (UMLS) Metathesaurus to translate Spanish natural-language queries into English. They conclude that the UMLS Metathesaurus-based CLIR method is at least equivalent to, if not better, than multilingual dictionary based approaches. Dan-Hee et al. (2000), comment that there is no reliable guideline as to how large machine readable corpus resources should be compiled to develop practical NLP software package and/or complete dictionaries for humans and computational use. They propose a new mathematical tool: a piecewise curve-fitting algorithm, and suggest 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 sector, TRANSLIB and CANAL/LS, were active between 1995 and 1997 (Oard,1997). Both these 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, 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 which facilitates finding and evaluating the desired information quickly and accurately (MULINEX, n.d.). *TwentyOne*, started in 1996, is a EU funded project which has the target to develop a tool for efficient dissemination of multimedia information in the field of sustainable development (TwentyOne, n.d.). Details of these and CLIR research projects in the US and other parts of the world have been reviewed by Oard & Diekama (1998).

Magnini et al. (2000) report two projects where NLP has been used for improving the performance in the public administration sector. The first project, GIST, is concerned with automatic multilingual generation of instructional texts for form-filling. The second project,

TAMIC, aims at providing an interface for interactive access to information, centered on NLP and supposed to be used by the clerk but with the active participation of the citizen.

Powell and Fox (1998) describe a federated search system, called *SearchDB-ML Lite*, for searching heterogeneous multilingual theses and dissertations collections on the World Wide Web NDLTD: Networked Digital Library of Theses and Dissertations (NDLTD, n.d.). 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 querying over 50 sites simultaneously revealed that in some cases more sophisticated query mapping is necessary to retrieve results sets that truly correspond to the original query. The authors report that an extended version of the *SearchDB* markup language is being developed that can reflect the default and available query modifiers for each search engine; work is also underway to implement a mapping system that uses this information

A number of companies now provide machine translation service, for example (McMurchie, 1998):

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

A large number of research papers are available that discuss various research projects dealing with MT and CLIR with reference to specific languages, for example in Chinese (Kwok et al. 2000; Lee et al., 1999), Japanese (Jie & Akahori , 2000; Ma, et al. 2000; Ogura et al. 2000) , Portugese (Barahona & Alferes, 1999), Sinhalese (Herath & Herath, 1999), Spanish (Weigard & Hoppenbrouwers,1998; Marquez et al., 2000), Thai (Isahara et al.,2000), Turkish (Say, 1999), and so on. Some studies have considered more than two languages; see for example Ide, 2000. These papers address various issues of MT, 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 (Jie & 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, et al., 1999)

Weigard & Hoppenbrouwers (1998) report the way an English/Spanish lexicon, including an ontology, is constructed for NLP tasks in an ESPRIT project called TREVI. Emphasizing the point that there has not been any study of natural language information retrieval in Swedish, Hedlund et al. (2001) describe the features of Swedish language and point out a number of research problems. They further stress that separate research in NLP in Swedish is required because the research results and tools for other languages do not quite apply to Swedish because of the unique features of the language.

Commenting on the progress of MT research, Jurafsky & Martin (2000; p. 825) comment that "machine translation system design is hard work, requiring careful selection of models and algorithms and combination into a useful system." They further comment that "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 come up with appropriate evaluation mechanisms for

large-scale information systems. Consequently, 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 for all NLP tasks. Different evaluation parameters may be required for each task, such as IE and automatic abstracting which are significantly different in nature compared to some other NLP tasks such as MT, CLIT or natural language user interfaces.

The ELSE (Evaluation in Language and Speech Engineering) project under the contract from the European Commission aimed to study the possible implementation of comparative evaluation in NLP systems. Comparative evaluation in Language Engineering has been used since 1984 as a basic paradigm in the DARPA research program in the US on human language technology since 1984. Comparative evaluation consists of a set of participants that compare the results of their systems using similar tasks and related data with metrics that were agreed upon. Usually this evaluation is performed in a number of successive evaluation campaigns with more complex task to perform at every campaign. ELSE proposition 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 to assess the amount of improvement it brings over 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.

EAGLES (The Expert Advisory Group on Language Engineering Standards – Evaluation Workgroup) (Centre for .., 2000), phase one (EAGLES-I: 1993—1995) and phase two

(EAGLES-II:1997—1998), is an European Initiative that proposed a user-centred evaluation of NLP systems. The EAGLES work takes as its starting point an existing Standard, viz. ISO 9126, which is concerned primarily with the definition of quality characteristics to be used in the evaluation of software products.

The 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 webpage of the DiET project (DiET, 1997) says 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.", and that 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".

MUC, the Message Understanding Conferences, which have now ceased, was the pioneer in opening 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 had focused on analyzing free text, identifying events of a specified type, and filling a data base template with information about each such event (MUC-6, 1996). After MUC-5, a broad set of objectives was defined for the forthcoming MUCs, such as: to push information extraction systems towards greater portability to new domains, and to encourage evaluations of some basic language analysis technologies. In MUC-7 (the last MUC), the multilingual NE (named entities) evaluation was run using training and test articles from comparable domains for all languages (Chinchor, n.d.). The papers in the MUC-7 conference report some interesting observations by system developers who were non-native speakers of the language of their system and system developers who were native speakers of the language of their system. Results of MUC-3 through MUC-7 have been summarized by Chinchor (n.d.).

Conclusion

Results of some NLP experiments reported in this paper show encouraging results. However, one should not forget that most of these experimental systems end 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 re-usable 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 re-use large test collections and experimental tools and techniques. Since the origin of the Message Understanding Conferences, group research efforts have proliferated with the regular conferences and workshops, for example, the TREC series and other conferences organized by NAACL (North American Chapter of the Association for Computational Linguistics), EACL (European ACL), and so on. These group research efforts help researchers share their expertise by building re-usable NLP tools, test collections, and experimental methodologies. References to some re-usable NLP tools and cooperative research groups have been made earlier in this paper (see under the heading *Some Theoretical Developments*).

Some recent studies on evaluation also show promising results. Very small-scale evaluation of INFOS suggests 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% of the articles (Mock and Vemuri, 1997). Some large-scale experiments with NLP also show encouraging results. For example, Kwok et al. (2000,1999) report that their PIRCS system can perform the tasks of English-Chinese query translation with an effectiveness of over 80%. Strzalkowski et al. (TREC-8;1998) report that by using the algorithm of automatic expansion of queries, using NLP techniques, they obtained a 37% improvement of average precision over a baseline where no expansion was used. There are conflicting results too. For example, Elworthy (2000) reports that the NLP system, using the Microsoft product NLPWin, performed much poorer 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 for the future work may be looking at how to build a system that merges definitive, pre-encoded knowledge, and ad-hoc documents of unknown reliability.

As already mentioned earlier (in the section on Abstracting), Craven's study with TEXNET (Craven, 1996) shows a limited success (only 37%). Gaizauskas and Wilks mention that the performance levels of the common IE systems, stand in the range of 50% 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 the 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 statement 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, Haas, 1996; Chowdhury, 1999b, Chapters 18-21), natural language user interfaces are not still very common. The impediments to progress to the natural language interfaces 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 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 can have any piece of information about a person, we do not know how could we 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 quite ambiguous and full of special constructions, and exceptions to rules. Despite that there has been a steady development, and MT research has now reached a stage where the benefits can be enjoyed by people. A number of web search tools, viz. Altavista, Google, Lycos and AOL offer free MT facilities of web 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 US software company began to offer MT services to its European customers at a cost which is 50% less compared to the human translation services (Schenker, 2001). Though machine translations are not always perfect and do not produce as good translations as human translators would produce, the results, and evidences of interests in improving the performance level of MT systems, are very encouraging.

One area of application of NLP that has drawn much research attention, but where the results are yet to reach the general public with an acceptable level of performance, is the natural language question-answering system. While some systems, as reported in this chapter, produce acceptable results, there are still many failures and surprises. Results of systems reported under the QA track of TREC (reported under the heading of natural language interfaces in this paper) show promising results with some simple type of natural language queries. However, these systems are still at experimental stages, 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 retrieved form a number of 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 the user changes, entirely new patterns need to 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 high degree of sophistication may largely be, not the inefficiency of the systems or researchers, but the complexities and idiosyncrasies of human behaviour and communication patterns.

References

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,* 22-24 April 1998, Santa Barbara, CA. Los Alamitos, CA: IEEE Computer Society, pp. 2-11

Amsler, R.A.(1984). Machine-readable dictionaries. In: M. E. Williams, (ed.) *Annual Review of Information Science and Technology (ARIST:* Volume 19, White Plains, NY: Knowledge Industry Publications Inc. for the American Society for Information Science. pp.161-209.

Argamon, S.; Dagan, I. & Krymolowski, Y. (1998). A memory-based approach to learning shallow natural language patterns. In 17th International Conference on Computational Linguistics (COLING '98), August 10-14, 1998, Université de Montréal, Montréal, Québec, Canada , Montreal: ACL. pp. 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. Proceedings*, 21-24 Sept. 1999 Evora, Portugal. 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*. 14-17 May 2000, Montreal, Berlin: Springer-Verlag. pp. 40-52

Benoit, G. (2001) Data mining. In: Cronin, B. (ed.). *Annual Review of Information Science and Technology* (ARIST): Volume 36. Medford, NJ: Information today for ASIS, pp.

Bian, Guo-Wei & Chen, Hsin-Hsi (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.* 12-14 April 2000, Manchester. Practical Application Company: Blackpool. pp. 157-70

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*. [Online] Available 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,* 5-7 Nov. 1999, North Falmouth, MA. Menlo Park, CA: AAAI Press. pp. 26-31

Carrick, C. and 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 . [Online] Available: 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,* 13-16 June 2000, Pula, Croatia. SRCE University Computer Centre, Univ. Zagreb, pp. 211-20

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

Charniak, E. (1995). Natural language learning. ACM Computing Surveys, 27, 317-3319.

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. Overview of MUC-7/MET-2. [Online] Available: 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. In: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval, ACM, pp. 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*, Pula, Croatia, 15-18 June, 1999. Univ. Zagreb , Zagreb, Croatia, pp. 63-9

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

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

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. (1996). An experiment in the use of tools for computer-assisted abstracting. In: ASIS'96: Proceedings of the 59th ASIS Annual Meeting 1996. Baltimore, MD, October 21-24, 1996. Vol. 33, Medford, NJ: Information Today, pp. 203-208.

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

Dan-Hee, Y.; Gomez, P.C. & Song, M. (2000). An algorithm for predicting the relationship between lemmas and corpus size. *ETRI Journal,* 22, 20-31

DiET: Diagnostoc and Evaluation Tools for natural language applications (1997). [Online] Available: http://www.dfki.de/lt/projects/diet-e.html

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, ICCS'99 Proceedings,* 12-15 July 1999, Blacksburg, VA. Berlin: Springer-Verlag pp. 154-67

Elworthy, D. (2000). Question answering using a large NLP system. The Ninth Text REtrievalConference (TREC 9) [Online] Available: http://trec.nist.gov/pubs/trec9/papers/msrc-qa.pdf

Evans, M. (1989). Computer-readable Dictionaries. . In: M.E. Williams (Ed). Annual Review of Information Science and Technology (ARIST): Volume 24. Amsterdam, The Netherlands: Elsevier Science Publishers B.V. for the American Society for Information Science. 85-117.

Fellbaum, C. (ed.) (1998). WordNet : an electronic lexical database. Cambridge, Mass : MIT Press

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

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, 59-71

Global Reach (2001). Global Internet Statistics (by language). [Online]. Available: http://www.euromktg.com/globstats/

Goldstein, J.; Kantrowitz, M.; Mittal, V. & Carbonell, J. (1999). Summarizing text documents: sentence selection and evaluation metrics. In: Proceeding of the 22nd Annual International Conference on Research and Development in Information Retrieval. ACM, pp. 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* 10-11 Nov. 1999, London: Aslib/IMI , pp. 12

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

Haas, S. W. (1996). Natural language processing: toward large-scale robust systems. In: M.E. Williams (Ed.). Annual Review of Information Science and Technology (ARIST): Volume 31. Medford, NJ: Learned Information Inc. for the American Society for Information Science. pp. 83-119.

Hayes, P. (1992) Intelligent high-volume text processing using shallow, domain-specific techniques. In: Jacobs, P.S., (ed.). *Text-based intelligent systems*, Hillsdale, NJ, Lawrence Erlbaum, pp. 227-241. Hayes, P. & Weinstein, S. (1991). Construe-TIS: a system for content-based indexing of a database of news stories. In: Rappaport, A. & Smith, R. (eds.), *Innovative applications of artificial intelligence 2*, Cambridge, MA, MIT Press, pp. 51-64.

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

Heng-Hsou Chang; Yau-Hwang Ko & Jang-Pong Hsu (2000). An event-driven and ontology-based approach for the delivery and information extraction of e-mails. *Proceedings International*

Symposium on Multimedia Software Engineering, 11-13 Dec. 2000, Taipei, Taiwan. Los Alamitos, CA: IEEE Computer Society, pp. 103-9

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, 223-34

Isahara, H.; Ma, Q.; Sornlertlamvanich, V. & Takahashi, N. (2000). ORCHID: building linguistic resources in Thai. *Literary & Linguistic Computing,*15, 465-78

Jelinek, F. (1999). Statistical Methods for Speech Recognition (Language, Speech, and Communication). 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.

Jie Chi Yang & Akahori, K. (2000). A discourse structure analysis of technical Japanese texts and its implementation on the WWW. *Computer Assisted Language Learning,* 13, 119-41

Jin, Song and Dong-Yan, Zhao (2000). Study of automatic abstracting based on corpus and hierarchical dictionary**,** *Journal of Software*, 11, 308-14

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.

Kam-Fai Wong; Lum, V.Y.& Wai-Ip Lam (1998). Chicon-a Chinese text manipulation language. *Software - Practice and Experience,* 28, 681-701

Kazakov, D.; Manandhar, S. & Erjavec, T. (1999)**.** Learning word segmentation rules for tag prediction. In: S. Dzeroski, S. & P. Flach (Eds.) *Inductive Logic Programming. 9th International Workshop, ILP-99 Proceedings,* 24-27 June 199, Bled, Slovenia. Berlin: Springer-Verlag , pp. 152-161

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, 73-80

Kornai, A. (ed.) (1999). Extended Finite State Models of Language (Studies in Natural Language Processing), Cambridge University Press.

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)*. [Online] Available: http://trec.nist.gov/pubs/trec9/t9_proceedings.html

Kwok, K.L.; Grunfield, L. & Chen, M. (1999). TREC-8 Ad-hoc, query filtering track experiments using PIRCS. The Eighth text retrieval Conference (TREC-8). [Online] Available: http://trec.nist.gov/pubs/trec8/papers/queenst8.pdf

Lange, H. (1993). Speech Synthesis and Speech Recognition: Tomorrow's Human-Computer Interfaces? In: M.E. Williams (Ed.). *Annual Review of Information Science and Technology* (ARIST): Volume 28. Medford, NJ: Learned Information Inc. for the American Society for Information Science. pp.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. & Jarvelin, K. (2001). Consistency of textual expression in newspaper articles: an argument for semantically base 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, 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. 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*. 24-27 July 2000. Como, Italy. Los Alamitos, CA: IEEE Comput. Soc , , pp. 91-96

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,* 21-24 May 2000, Anchorage, AK. Hershey, PA: Idea Group Publishing. pp. 500-504

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, 23-26 Feb. 1998, Berlin. London: Chapman & Hall pp. 166-179

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: Edwards, J. (ed.), Proceedings of the 23rd Australasian Computer Science Conference. ACSC 2000, IEEE Comput. Soc , Los Alamitos, CA, pp. 158-65

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,* 3-5 May 1999, Orlando, FL. Menlo Park, CA: AAAI Press , pp. 293-7

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

Moens, Marie-Francine & Uyttendaele, Caroline (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*. Innsbruck, Austria , 23-27 Aug. 1999. Vienna: TermNet pp. 268-78

MUC-6 (1996). [Online] Available : http://www.cs.nyu.edu/cs/faculty/grishman/muc6.html

MULINEX: Multilingual Indexing, Navigation and Editing Extensions for the World Wide Web. [Online]. Available: http://mulinex.dfki.de/

Narita, M.& Ogawa, Y. (2000). The use of phrases from query texts in information retrieval. *SIGIR Forum*, 34, 318-20

Natural Language Processing Laboratory, University of Massachusetts. [Online] Available: http://www-nlp.cs.umass.edu/nlplic.html

NDLTD: Networked Digital Library of Theses and Dissertations. [Online] Available: http://www.ndltd.org

Nerbonne, J.; Dokter, D. & Smit, P. (1998). Morphological Processing and Computer-Assisted Language Learning. *Computer Assisted Language Learning*, 11, 543-59

Oard, D. W. (1997). Serving users in many languages: cross-language information retrieval for digital libraries, *D-Lib Magazine.* [Online] Available: http://www.dlib.org/dlib/december97/oard/12oard.html

Oard, D. W & Diekama, A.R. (1998). Cross-language Information Retrieval. In: M.E. Williams (Ed.). *Annual Review of Information Science and Technology (ARIST)*: Volume 33. Medford, NJ: Learned Information Inc. for the American Society for Information Science. pp. 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, 53-8

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, 439-92

Paris, L.A.H. & Tibbo, H.R. (1998). Freestyle vs. Boolean: a comparison of partial and exact match retrieval systems. *Information Processing & Management,* 34, 175-90

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. [Online] Available: http://www.limsi.fr/TLP/ELSE/PreambleXwhyXwhatXrev3.htm

Pasero, R. & Sabatier, P. (1998) Linguistic Games for Language Learning: A Special Use of the ILLICO Library. *Computer Assisted Language Learning*, 11, 561-85

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. 26-30 July 1998, Madison. Menlo Park, CA: WI AAAI Press/MIT Press pp. 800-5

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*. [Online] Available: 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*. [Online] Available: 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). Special topical issue: Knowledge discovery and data mining. *Journal of the American Society for Information Science*, 49(5).

Roche, E. and Shabes, Y. (eds.) (1997). Finite-State Language Processing (Language, Speech and Communication), MIT Press.

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

Roux, M.& Ledoray, V. (2000) Understanding of medico-technical reports. *Artificial Intelligence in Medicine,* 18, 149-72

Ruiz, M.E. & Srinivasan, P. (1998). Cross-Language Information Retrieval: an analysis of errors. *Proceedings of the 61st ASIS Annual Meeting*, Pittsburgh, PA, October 25-29, pp.153-65

Say, B (1999). Modeling cue phrases in Turkish: a case study. In: V. Matousek, V. et al (Eds.). *Text, Speech and Dialogue. Second International Workshop, TDS'99 Proceedings*, 13-17 Sept. 1999, Plzen, Czech Republic. Berlin: Springer-Verlag pp. 337-40

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, AI 2000 Proceedings,* 14-17 May 2000, Montreal. Berlin: Springer-Verlag pp. 138-50

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

Scott, J. (1999). E-mail Management: the key to regaining control. *Internet Business, December 1999,* 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,* 9-12 Jan. 2000, New Orleans, LA. New York: ACM pp. 252-5

Smeaton A.F. (1999). Using NLP or NLP Resources for Information Retrieval Tasks. In: T. Strzalkowski (Ed.), *Natural Language Information Retrieval*, Kluwer Academic Publishers, 99-111,

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*, 14-18 July 1997,

Frascati, Italy. Berlin: Springer-Verlag pp. 115-38

Smith, D. (1998). Computerizing Computer Science. *Communications of the ACM*, 41, 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, 11-13 April 2000, Manchester. Blackpool: Practical Application Company, pp. 57-61

Song Jin & Zhao Dong-Yan (2000). Study of automatic abstracting based on corpus and hierarchical dictionary. *Journal of Software,*11, 308-14

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

Staab, S.; Braun, C.; Bruder, I.; Dusterhoft, A.; Heuer, A.; Klettke, M.; Neumann, G.; Prager, B.; Pretzel, J.; Schnurr, H.-P.; Studer, R.; Uszkoreit, H.& Wrenger, B. (1999) GETESS-searching the Web exploiting German texts. *Cooperative Information Agents III. Third International Workshop, CIA'99 Proceedings,* 31 July-2 Aug. 1999, Uppsala, Sweden. Berlin: Springer-Verlag pp. 113-24

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 Publication500-240: The Sixth Text REtrieval Conference (TREC 6).* [Online] Available: 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:The Eighth Text REtrieval Conference (TREC 8)* [Online] Available: http://trec.nist.gov/pubs/trec8/papers/ge8adhoc2.pdf

Strzalkowski, T.; Stein, G.; Wise, G.B.; Perez-Carballo, J; Tapanainen, P.; Jarvinen, T.; Voutilainen, A. & Karlgren, J. (1998). *Natural languagee information retrieval: TREC-7 report. NIST Special Publication 500-242: The Seventh Text REtrieval Conference (TREC 7)* [Online] Available: 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. In: M.E. Williams (Ed.). *Annual Review of Information Science and Technology (ARIST)*: Volume 32. Medford, NJ: Learned Information Inc. for the American Society for Information Science, pp.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.* 31 Aug.-1 Sept. 1999, Adelaide, SA, Australia. Piscataway, NJ: IEEE , pp. 534-7

Twenty-One: development of a multimedia information dissemination and transaction tool. [Online] Available: http://twentyone.tpd.tno.nl/twentyone/

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. [Online] Available: http://trec.nist.gov/pubs/trec8/papers/qa-report.pdf

Voorhees, E. (2000). The TREC-9 question answering track report. [Online] Available: 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. In: Williams, Martha E. ed. Annual Review of Information Science and Technology (ARIST): Volume 22. Amsterdam, The Netherlands: Elsevier Science Publishers B.V. for the American Society for Information Science, 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 Ninth International Workshop on Database and Expert Systems Applications,* 26-28 Aug. 1998, Vienna. Los Alamitos, CA: IEEE Computer Society pp. 160-5

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

Yang, Y. & Liu, X (1999). A re-examination of text categorization methods. In: SIGIR '99 Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, pp. 42-49.

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

Zweigenbaum, P.& Grabar, N. (1999) Automatic acquisition of morphological knowledge for medical language processing. In: W. Horn, et al (Eds.). *Artificial Intelligence in Medicine. Joint European Conference on Artificial Intelligence in Medicine and Medical Decision Making, AIMDM'99 Proceedings,* 20-24 June 1999, Aalborg, Denmark. Berlin: Springer-Verlag pp. 416-20