

Natural Language Processing: Current state and future directions

Ranjit Bose

Anderson School of Management
University of New Mexico
Albuquerque, NM 87131
E-mail: bose@mgt.unm.edu

Abstract

The value in being able to communicate with computers by speaking or writing via “natural language” cannot be overstated. Computational linguistics, or work on “natural language processing” (NLP) began more than sixty years ago. A recent study by the Kelsey Group reports that increasing numbers of companies are investing in and deploying voice or speech recognition and processing technologies at an alarming rate to save money by replacing operators and to improve service to their customers. In recent years, the natural language text interpretation and processing technologies have also gained an increasing level of sophistication. For example, generic engines are now available which can deliver semantic representations for sentences, or deliver sentences from representations. NLP technologies are becoming extremely important in the creation of user-friendly decision-support systems for everyday non-expert users, particularly in the areas of knowledge acquisition, information retrieval and language translation. The purpose of this research is to survey and report the

current state and the future directions of the use of NLP technologies and systems in the corporate world. The research is intended to assist business managers to stay abreast with the NLP technologies and applications.

Keywords: *Natural language recognition and processing system*

1. Introduction

Over thirty years ago, “2001: A Space Odyssey,” made predictions for computers used at the turn of the century. One of the HALs was able to have meaningful dialogues with the astronauts. Speech recognition and understanding together with psychological advice were packaged into a friendly chat.

Over the years, the HAL’s dream was followed and NLP research concentrated on “designing and building a computer system that would analyze, understand and generate languages that humans use naturally, so that

eventually you could address your computer as though you were addressing another person,” which is Microsoft’s NLP research definition (Andolsen, 2002).

Market researcher Datamonitor Technology reports that more than one-fourth of the *Fortune 500* companies invested in speech systems last year, up 60% from a year ago. Hundreds of companies are replacing some service reps with voice software. For example, AT&T recently replaced 200 operators with a voice-recognition system to handle night and weekend toll-free directory assistance calls. Operators are still reachable other times. BellSouth and Verizon Communications already use voice software to solicit city and listing during directory assistance calls. An operator often delivers the number. Qwest Communications is considering replacing operators with voice-recognition systems for more services.

A recent study by the Kelsey Group reports that increasing number of companies are investing in and deploying voice or speech recognition and processing technologies at an alarming rate to save money by replacing operators and to improve service to their customers. A slew of companies, including United Airlines, Charles Schwab, E-Trade and Amazon.com have added voice systems to handle general calls. AirTran Airways cut customer service costs by 20% by shifting some flight information calls from operators to voice systems. It is considering the system for reservations. Most companies still use operators for complex tasks, such as correcting financial information and retrieving passwords. U.S. firms have spent \$7.4 billion last year to improve voice and touch-tone systems. A call handled by a worker costs, on average, \$5.50 or 10 times as much as an automated call, says researcher Cahners In-Stat Group.

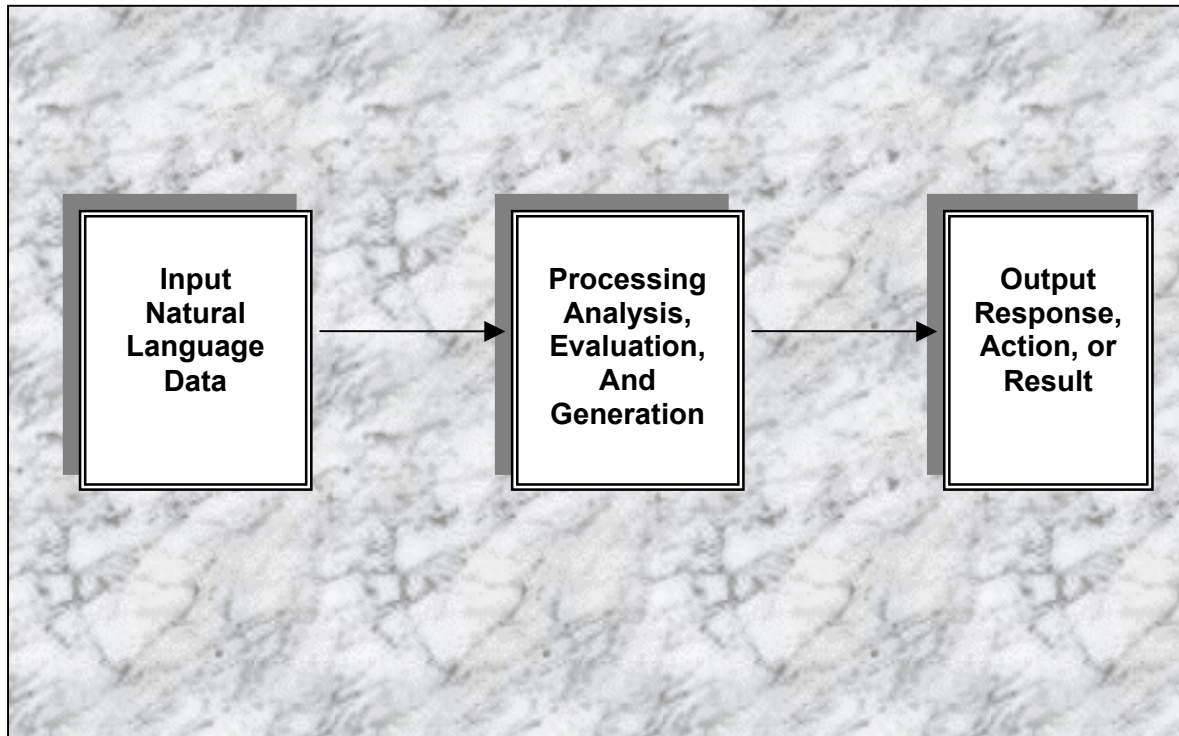
In recent years, the natural language text interpretation and processing technologies have also gained an increasing level of sophistication. For example, generic engines are now available which can deliver semantic representations for sentences, or deliver sentences from representations. It is now possible to build very-targeted systems for specific purposes, for example, finding index terms in open text, and also the ability to judge what level of syntax analysis is appropriate. NLP technologies are becoming extremely important in the creation of user-friendly decision-support systems for everyday non-expert users, particularly in the areas of knowledge acquisition, information retrieval and language translation. The purpose of this research is to survey and report the current state and the future directions of the use of NLP technologies and systems in the corporate world. The research is intended to assist business managers to stay abreast with the NLP technologies and applications.

2. Background Knowledge

The research and development in NLP over the last sixty years can be categorized into the following five areas:

- Natural Language Understanding
- Natural Language Generation
- Speech or Voice recognition
- Machine Translation
- Spelling Correction and Grammar Checking (Biermann, *et al.*; 1992)

Language is more than transfer of information. Language is a set of resources to enable us to share meanings, but isn’t best thought of as a means for “encoding” meanings. The following graphic depicts the flow of information in NLP:



An NLP system must possess considerable knowledge about the structure of the language itself, including what the words are, how to combine the words into sentences, what the words mean, how these word meanings contribute to the sentence meaning, and so on. The system would need methods of encoding and using this knowledge in ways that will produce the appropriate behavior. Furthermore, the knowledge of the current situation (or context) plays a crucial role in determining how the system interprets a particular sentence.

2.1 Categories of Knowledge

The different forms of knowledge have traditionally been defined into the following six categories (Allen, 1987):

- *Phonetic and Phonological (speech recognition) knowledge* – concerns how words are realized as sounds.

- *Morphological knowledge* – concerns how words are constructed from basic meaning units called phonemes.
- *Syntactic knowledge* – concerns how words can be put together to form sentences.
- *Semantic knowledge* – concerns what words mean and how these meanings combine in sentences to form sentence meanings.
- *Pragmatic and Discourse knowledge* – concerns how sentences are used in different contexts and how context affects sentence interpretation. Language is analyzed in more than a single utterance.
- *World knowledge* – include general knowledge about the structure of the world that the users must have in order to maintain a conversation (Wohleb, 2001).

2.2 Issue of Ambiguity

All tasks in NLP have to resolve ambiguity at one or more of these above six categories (Jurafsky and Martin, 2000). One can say that input is ambiguous if there are alternative linguistic structures that can be built for it. For example, the sentence “I made her duck” can be interpreted as ambiguous on the following categories:

- Morphological and syntactical ambiguity (Duck could be either a verb or a noun)
- Semantic ambiguity (Make, a verb, can mean either “create” or “cook.”)

2.3 Models and Algorithms for Disambiguation

The ambiguity problem gets resolved via disambiguation. The syntactic and morphological ambiguity in this case calls for the use of *parts-of-speech tagging* to resolve it. The semantic ambiguity can be solved via *word sense disambiguation*. Speech and language technology relies on the various categories of linguistic knowledge, which can be captured and used for the purpose of disambiguation in the following models:

- *State machines* – formal models that consist of states, transition among states and input representations
- *Formal rule systems* – regular grammars, context-free grammars
- *Logic* – first order logic, predicate calculus
- *Probability theory* – solving ambiguity, machine-learning models.

The algorithms associated with the models typically involve a search through a space-representing hypothesis about an input, such as:

- *State space search systems, and*
- *Dynamic programming algorithms.*

3. Analysis of NLP Knowledge

3.1 Phonetic & Phonological Knowledge

Phonological rules are captured through machine learning on training sets. Pronunciation dictionaries are also used for both text-to-speech and automatic speech recognition. Sounds (phonemes), as well as words can be predicted by using the *conditional probability* theory. There are many models of word prediction, among them N-Grams, which are evaluated by separating the corpus into a training test and test set (just like in the neural network). The input to a speech recognizer is a series of acoustic waves. The waves are then sampled, quantified and converted to spectral representation. Conditional probability is then used to evaluate each vector of the spectral representation with stored phonetic representation. Decoding or search is the process of finding the optimal sequence of input observations. Each successful match is later used in *embedded training* – a method for training speech recognizers.

3.2 Syntactic Knowledge

Syntax is a study of formal relationships between words. Computational models of this NLP knowledge category include parts-of-speech tagging, context-free grammars, lexicalized grammars or Chomsky’s hierarchy.

Parts-of-speech tagging, mentioned earlier, is the process of assigning a part of speech label to each sequence of words. Taggers, as they are called, are often evaluated by comparing their output from a test set to human labels for that test set.

A context-free grammar and its cousin, generative grammar – consist of a set of rules used to model a natural language. Any context-free grammar can be converted to Chomsky's normal form. In 1956, Noam Chomsky, the famous linguist, first created the context-free grammar parse trees. Since then, syntactic analysis of sentences has never been the same. Syntactic parsing began to be known as the task of recognizing a sentence and assigning a context-free tree to the input sentence. The following are the most common methods of parsing:

- *Top-down parsing* – searches for a parse tree by trying to build from a root node *S* (representing the sentence) down to the leaves via
- *Bottom-up parsing* – parsing starts with the words of the input and tries to build trees from the words up by applying the rules of grammar one at a time
- *Depth-first parsing* – expands the search space incrementally by exploring one state at a time. The state chosen for expansion is the most recently generated one.
- *Repeated parsing of subtrees* – designed to help with resolving ambiguity, and deals with the inefficiency of other parsing algorithms. Parser often backtracks to fix successive failures in previous parsing attempts.
- *Dynamic programming parsing algorithms* – use partial parsing to resolve ambiguity.

See Figure 2 below for a parsing tree used in the above approaches:

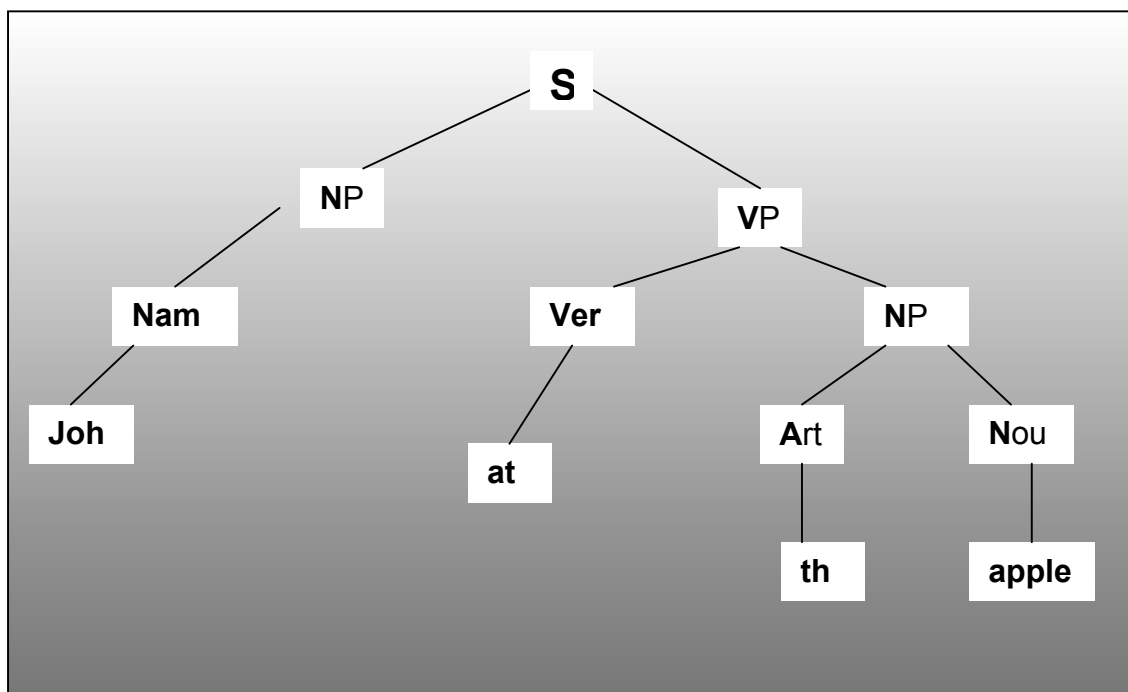


Figure 2. Syntactic model of parsing in a context-free grammar

Just as in the phonological component, the probability theory also plays a role in syntactic analysis. Probabilistic grammars assign a probability to a sentence while attempting to capture more sophisticated syntactic information. Every rule defined in the grammar is assigned a probability of being chosen.

3.3 Semantic Knowledge

Semantics is the study of the meaning of linguistic utterances. One of the key issues in semantics is modeling how the meaning of an utterance is related to meaning of phrases, words and morphemes that constitute it. The issues with knowledge base representation of semantic knowledge include:

- *Verifiability* – system must be able to relate the meaning of the sentence to world knowledge represented in the knowledge base
- *Unambiguous representation* – linguistic inputs have to be represented unambiguously based on the circumstances in which they occur
- *Vagueness* – system must be able to draw conclusions about the truth of propositions that are not explicitly represented in the knowledge base, but are nevertheless logically derivable from other propositions.

Semantic analysis can be performed with such models as First Order predicate Calculus or Semantic Networks. A semantic analysis system has to be able to *infer* or draw valid conclusions based on the meaning representation of input and its match in the knowledge base, or if the knowledge base does not have an exact match, the system has to be able to derive the meaning from other representations in the knowledge base.

First Order predicate Calculus, heavily used in semantic knowledge representation, is a very flexible model, which provides a computational basis for verifiability, inference and expressiveness requirements. The represented semantic knowledge contains objects, object properties and relations among objects. Inference is the ability to add valid propositions to the knowledge base or to determine their truth is based on algorithms, which can be found in other components of a typical DSS system, such as forward or backward chaining. The Prolog (programming in logic) programming language, which is very popular in NLP implementations, is implemented using backward chaining design strategy. An example of semantic/lexical representation at work is WordNet – a widely used database of lexical relations in English (Allen, 1987).

4. Notable NLP Systems/Prototypes

ELIZA

Eliza (Weizenbaum) is an extremely simple computer program intended as a study in natural language communication. It plays the part of a Rogerian psychoanalyst, and basically repeats everything the user says in the form of a question. To the shock of its programmer it was received with enthusiasm by the psychiatric community and was recommended for eventual therapeutic use. Here's an example of a dialog with Eliza:

E: GOOD DAY WHAT IS YOUR PROBLEM

I feel a bit depressed

E: MACHINES CAN MAKE PEOPLE HAPPIER

Can you make me happier?

E: PERHAPS YOU ALREADY KNOW THE ANSWER TO THAT?

You sound very philosophical

E: SOME PEOPLE USE LONG WORDS
TO IMPRESS OTHERS

Now you sound stupid

E: TELL ME MORE ABOUT YOURSELF

I've told you enough already

E: DO YOU THINK YOU ARE UNIQUE?

Bye

E: BYE FOR NOW

Bye

LUNAR

The LUNAR prototype, developed by NASA, allowed the geologists to ask questions about the chemical composition of lunar rock and soil samples. For example, one could ask LUNAR questions of the following nature: "What is the average modal plagioclase concentration for lunar samples that contain rubidium?"

The system was never put into real use but in tests it achieved a 78% success rate.

5. Most Successful NLP Uses

The most successful applications of NLP, beyond in the *voice recognition and processing* arena mentioned earlier, are in the *text interpretation* arena. Text Interpretation programs retrieve, categorize, filter and extract information from texts. Text Interpretation NLP uses can be subdivided into three types:

- *Information Retrieval* (most Web search engines)
- *Text Categorization* (sorting into fixed categories – new wire services)
- *Data Extraction* – derives from text assertions that can be stored in a structured database.

The following is a text retrieval engine example. EBSCO's Psychology & Behavioral Sciences Collection is a subject database with full-text articles. The collection includes 350 peer-reviewed journals as well as about 50 popular magazines. The earliest articles are from 1984, and most titles actually do stretch that far back. Some titles have a 12-month full-text embargo so that articles published within the last year only have the citation and abstract included. When available, links to places where the full text of embargoed articles are located online are provided.

The text retrieval engine's interface is clear, easy to use, and offers several different search options. The default search is by keyword; Boolean operators can be employed. Other search options include natural language, advanced, and expert. Advanced search offers multiple input boxes with Boolean connectors available in dropdown menus. In advanced searching, users can also specify which fields to search. Expert searching allows the use of extra limiters (type of document, number of pages, whether or not the article is a cover story) and saves those searches for further manipulation.

Another example of a successful text retrieval system is Lexus-Nexis iPhrase (Quint, 2001). The iPhrase natural language search and navigation technology allows users to pose typical questions that the system can interpret to locate precise results. Its software can even recognize follow-up questions from users and answer them in the context of what has gone before. It can tap into all the fields available in a database structure and present the results in a variety of attractive, useful formats. In the course of adopting the technology, the content provider works extensively with iPhrase to customize the knowledge base behind the one step system, helping it learn the types of

questions users will ask, the jargon of the trade, synonyms, taxonomies, and such.

Typical, text search and retrieval tasks are performed by intelligent software agents, for example, a network information software agent. This agent is a software robot type of application that would make the agent go out on the Internet, look for and find information, and produce natural language answers to your questions with references or links to primary information sources (Perez, 2001).

6. Future of NLP

NLP's future will be redefined as it faces new technological challenges and a push from the market to create more user-friendly systems. Market's influence is prompting fiercer competition among existing NLP based companies. It is also pushing NLP more towards Open Source Development. If the NLP community embraces Open Source Development, it will make NLP systems less proprietary and therefore less expensive. The systems will also be built as easily replaceable components, which take less time to build and more user-friendly. Many companies, T Rowe Price for example, are looking into creating more user-friendly systems.

Frappaola (2000) notes that for the coming years, the emphasis will be on customer relationship management in all facets of the business such as better phone service, new call center systems, voice response systems, and development of a "natural language" technology which can understand and respond to plainly spoken customer requests. The customer should be able to say, "Hi, I'm calling about my position with IBM, where do I stand?, what's my account balance?," and those type of things.

Web portal services interface are becoming increasingly user-friendly. NLP will increasingly play a critical role in the design and development of successful Web portals. As the universal platform of the Web broadens the user audience for portals, the search tool must be appealing to many types of users. Searching must not require an education in SQL, Boolean logic, lexical analysis, or the underlying structures of information repositories. Users overwhelmingly accept search functionality that is natural language-based and intuitive. Searches of all types of data are expected to interpret and expand queries lexically, while simultaneously delivering precise results focused on the essence of the search. These results should be ranked by perceived relevancy to the query. Queries, whether of structured data records or documents, should deliver answers – not database records or collection of documents. In this manner, a search tool may also support a portal's presentation and personalization features, giving users control over the level of detail and presentation of the answer set.

Ultimately the search tool should function against both structured and unstructured types of data repositories with a single query, delivering a single, combined answer set that is data neutral – be able to return streaming video resources as well as database fields or relevant segments of text documents. To meet these expanded demands, a new market for search technology is emerging, one in which established vendors are seeking to broaden their functionality and new technology is coming to market with innovative approaches against new Web-based engines.

Another example of a future NLP system, illustrating consumer's use of NLP via a wireless PDA, is depicted below in Figure 3.

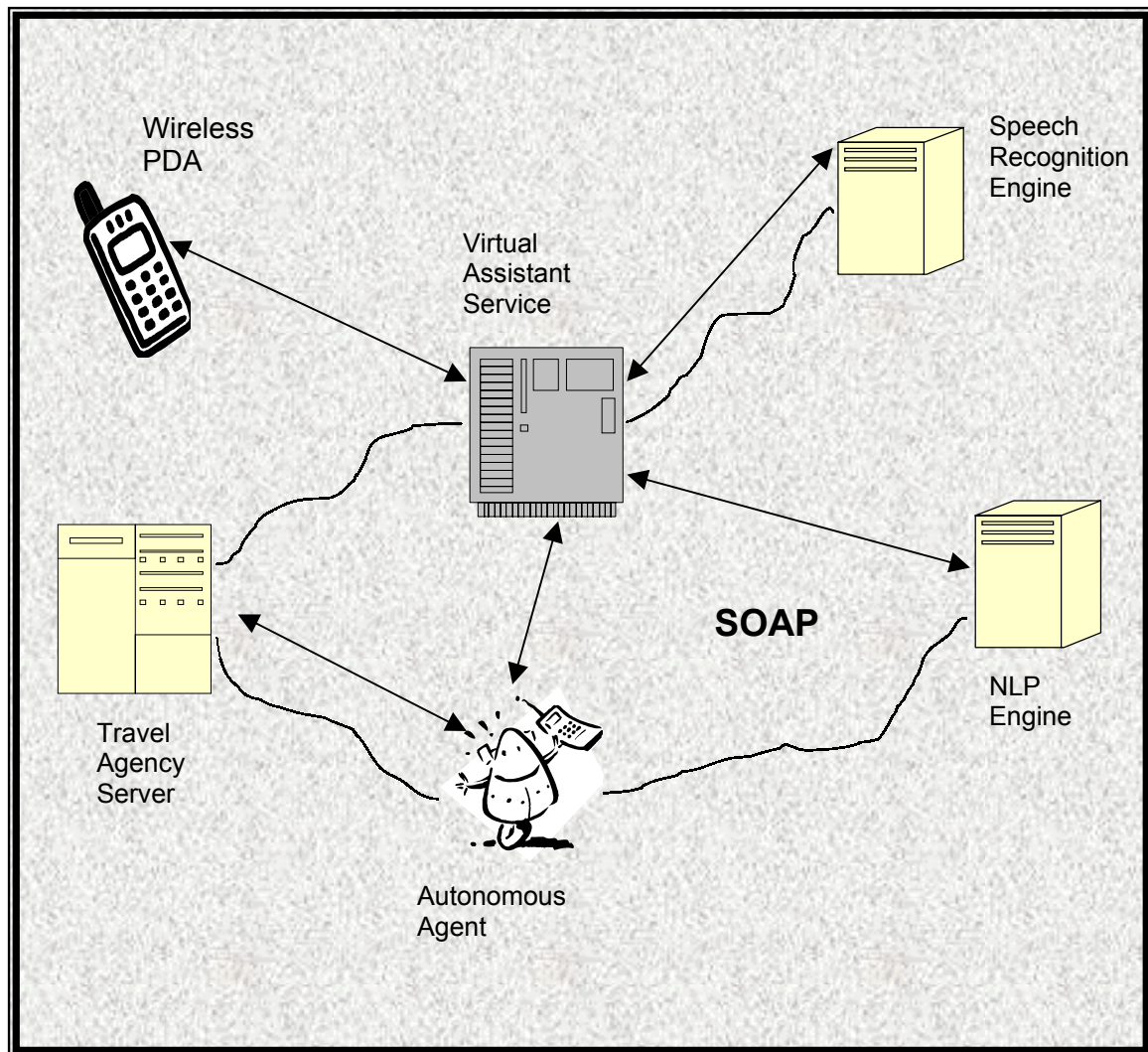


Figure 3. Consumer use of NLP via wireless PDA

A consumer could book a flight using a Web service as well as perform a variety of other functions. The agent service, receiving the verbal request via a wireless Internet connection could use centralized speech recognition and NLP to process the users utterance and activate the appropriate autonomous agent. This in turn would book the customer's flight as requested and return a confirmation message.

Several other future applications of NLP, most of them currently under development, are as follows:

- *Conversational systems.* The University of Colorado at Boulder has developed systems that are designed to assist with inquiries related to airline schedules, hotel reservations, the times of movies and their locations, or sports scores (Andolsen, 2002). The first challenge for a speech recognition system used in these systems still remains to be proper recognition of what is being spoken by a wide variety of people with differing vocabularies and accents.

- Systems where a computer would be able to read a book, store the information about the book, and then answer questions about the book. These types of system would be dealing with advanced type of autoindexing.
- *Artificial Neural Networks*. One of the interesting products now being introduced on the market is DolphinSearch technology. Dolphins learn by recognizing the characteristics of objects off of which they bounce sonar waves. They learn by categorizing and remembering the various reflections that come back from the objects. In a similar manner, this approach relates words to one another so that, in ambiguous situations, their grammatical role becomes evident. For example, the word “force” can be either a noun or a verb. By analyzing the words around it, the system is able to determine whether it is being used as one or the other.
- *Microsoft MindNet* – combination of an extensive database and algorithms that can define relationships. The project is attempting to use dictionaries in seven languages and a variety of encyclopedias to create a system that recognizes relationships between simple words (from the dictionaries) and phrases or sentences (from the encyclopedias). The relationships are built and identified by simple questions directed at the system. MindNet also promises to be a powerful tool for machine translation. The idea is to have MindNet create separate conceptual webs for English and another language, Spanish, for example, and then align the Webs so that the English logical forms match their Spanish equivalents. MindNet then annotates these matched logical forms with data from the English-Spanish translator memory, so that translation can proceed smoothly in either direction (Waldrop, 2001).
- *Medication Assistant* – a medical DSS, which models the effects of therapy on patients with cardiovascular and other medical conditions. Prolog programming language, used in this DSS to control NLP links hierarchically linked data and grammatically corrects text (Temiroff *et al.*, 2001).
- *Chatterbots* – although they exist already, new generations of them are being constantly developed. Chatterbots use natural language processing to simulate conversations with users. Web sites are beginning to install chatterbots as Web guides and customer service agents (Anonymous, 2001).

7. Conclusions

With over sixty years of NLP research and development, the natural language systems are still very complicated to design. Multitude of models and algorithms exist today. Most of all, NLP systems are still not perfect because natural human language is complex, and it is difficult to capture the entire linguistic knowledge for hundred percent accuracy in processing. For example, even though hundreds of companies are replacing some service reps with voice software, emergency services like 911 will continue to be handled by humans for at least another decade or so because of their critical nature. The current voice systems still need adjustments -- some cannot understand

heavy accents, speech impediments or quiet voices.

Most NLP systems are currently proprietary – specific to the domain they serve – therefore expensive to build. If the information systems community responds to the challenge by building NLP systems with reusable components via Open Source programming, the future of NLP will start looking even brighter. Still, the possibility of *free* natural communication with a machine in the near future seems unlikely despite all the developments. There are still unresolved challenges for software programs to represent the entire knowledge, the different contexts and cultures of the world.

The Kelsey Group study reports that companies are using voice-recognition software more and worldwide consumers are likely to run into it. The report also states that there will be a fivefold increase in spending on voice-recognition software in the next three years. The currently used systems in text interpretation however, seem to offer more versatility to the users. These systems can offer real advantages in composing text, and online help such as dictionary support becomes very useful. In the longer term, cooperation between the learner and the system, where they both help each other with natural communication, will probably be the direction for further developments.

References

- Allen, J. *Natural Language Understanding*. The Benjamin/Cummings Publishing Company, Inc., 1987.
- Andolsen, A.A. "On the Horizon," *Information Management Journal*, Vol. 36, No. 2, 2002.
- Anonymous. "Just Talk to Me," *The Economist Technology Quarterly*, December 6, 2001.
- Biermann, A., *et al.* "A Voice- and Touch-Driven Natural Language Editor and its Performance," *International Journal of Man-Machine Studies*, Vol. 37, No. 1, 1992.
- Frappalo, C. "Now It's Personal," *Intelligent Enterprise*, Vol. 3, No. 17, 2000.
- Guerra, A. "T. Rowe Price to hone in on voice systems," *Wall Street and Technology*, Vol. 19, No. 3, 2000.
- Jurafsky, D. and Martin, J.H. *Speech and Language Processing*, Prentice Hall, New Jersey, 2000.
- Perez, E. "Finding needles in textstacks," *Online*, Vol. 25, No. 5, 2001.
- Quint, B. "LexisNexis applies iPhrase natural language search software to key directory files," *Information Today*, Vol. 18, No. 11, 2001.
- Temiroff, A., *et al.* "Predicting the effects of therapy on patients with multiple diseases: Prolog based medical decision support," *PC/AI*, November/December, 2001.
- Waldrop, M. "Natural Language Processing," *Technology Review*, Vol. 104, No. 1, 2001.
- Wohleb, R. "Natural Language Processing: Understanding Its Future," *PC/AI*, November/December, 2001.