ISSN 2278-3091

International Journal of Advanced Trends in Computer Science and Engineering, Vol.2 , No.6, Pages : 175-179 (2013)
Special Issue of ICETEM 2013 - Held on 29-30 November, 2013 in Sree Visvesvaraya Institute of Technology and Science, Mahabubnagar – 204, AP, India

Role of NLP in Production and Comprehension to Communicate and Understand Human Language

G.Pratibha[1], Dr. Nagaratna Hegde[2]
[1]Asst.Professor, Matrusri Engineering College, Saidabad, Hyderabad, Andhra Pradesh, India, pratibhareddy19@gmail.com
[2]AProfessor, Vasavi Engineering College, Ibrahimbagh, Hyderabad, Andhra Pradesh, India, nagaratnaph@gmail.com

Abstract: Natural Language Processing (NLP) is an interdisciplinary research area at developing computer programs capable of human-like activities related to understanding or producing texts or speech in a natural language, such as English.

Natural language processing has been in existence for more than fifty years. During this time, it has significantly contributed to the field of human-computer interaction in terms of theoretical results and practical applications. As computers continue to become more affordable and accessible, the importance of user interfaces that are effective, robust, unobtrusive, and user-friendly regardless of user expertise or impediment becomes more pronounced. Since natural language usually provides for effortless and effective communication in human-human interaction, its significance and potential in human-computer interaction should not be overlooked – either spoken or typewritten, it may effectively complement other available modalities, such as windows, icons, and menus, and pointing; in some cases, such as in users with disabilities, natural language may even be the only applicable modality.

In this paper, we examines the field of natural language processing as it relates to human computer interaction by focusing on its history, interactive application areas, and how natural language programming contributes a lot for natural language processing.

Keywords : NLP, User Interfaces, Human Computer Interaction, Modality, Interactive Application Areas.

I. INTRODUCTION

The field of natural language processing has entered its sixth decade. During its relatively short lifetime, it has made significant contributions to the fields of human-computer interaction and linguistics. It has also influenced other scientific fields such as computer science, philosophy, mathematics, statistics, psychology, biology, and engineering by providing the motivation for new ideas, as well as a computational framework for testing and refining existing theoretical assumptions, models, and techniques. Finally, it has impacted society through applications that have shaped and continue to shape the way we work and live our lives.

Developing a program that understands natural language is a difficult problem. Number of natural languages is large, they contain infinitely many sentences. Also there is much ambiguity in natural language. Many words have several meanings, such as can, bear, fly, orange, and sentences have meanings different in different contexts. This makes creation of programs that understands a natural language, a challenging task.

Natural Language Processing (NLP) is the computerized approach to analyzing text that is based on both a set of theories and a set of technologies. And, being a very active area of research and development, there is not a single agreed-upon definition that would satisfy everyone, but there are some aspects, which would be part of any knowledgeable person's definition. Natural Language Processing is a theoretically motivated range of computational techniques for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of tasks or applications.

There are two motivations for NLP, one scientific and one technological (Allen, 1994a). The *scientific motivation* is to understand the nature of language. Other traditional disciplines, such as linguistics, psycholinguistics, and philosophy, do not have tools to evaluate extensive theories and models of language comprehension and production. It is only through the tools provided by computer science that one may construct implementations of such theories and models. These implementations are indispensable in exploring the significance and improving the accuracy (through iterative refinement) of the original theories and models. The *technological motivation* is to improve communication between humans and machines.

Computers equipped with effective natural language models and processes could access all human knowledge recorded in linguistic form; considering the revolution in information dissemination and communication infrastructure that has been introduced by the World-Wide-Web, one could easily see the importance and potential of such systems. User interfaces with natural language modalities (either input or output, spoken or typewritten) would enhance human-computer interaction by facilitating access to computers by unsophisticated computer users, users in hands-busy/eyes-busy situations (such as car driving, space walking, and air traffic control tasks), and users with disabilities. Actually, the development of this technology for the latter group is motivated by federal legislation and guidelines, such as (a) the US Public Laws 99-506 and 100-542 which mandate the establishment of accessible environments to citizens with disabilities, (b) the 1989 US General Services Administration's guide, *Managing End User Computing for Users with Disabilities*, which describes accommodations for disabled computer users (Shneiderman, 1993), and (c) the 1996

ISSN 2278-3091

International Journal of Advanced Trends in Computer Science and Engineering, Vol.2 , No.6, Pages : 175-179 (2013)
Special Issue of ICETEM 2013 - Held on 29-30 November, 2013 in Sree Visvesvaraya Institute of Technology and Science, Mahabubnagar – 204, AP, India

Telecommunication Act. In this context, it does not matter how closely the model captures the complexity of natural language communication; it only matters that the resultant tool performs satisfactorily in a given domain of discourse, or complements/outperforms any alternative solutions. This article adheres to this perspective in presenting and discussing various NLP theories, models, and applications. In this context, and given the state-of-the-art, NLP could be defined as the discipline that studies the linguistic aspects of human-human and human-machine communication, develops models of linguistic competence and performance, employs computational frameworks to implement processes incorporating such models, identifies methodologies for iterative refinement of such processes/models, and investigates techniques for evaluating the resultant systems.

NLP is an interdisciplinary area based on many fields of study. These fields include computer science, which provides techniques for model representation, and algorithm design and implementation; linguistics, which identifies linguistic models and processes; mathematics, which contributes formal models and methods; psychology, which studies models and theories of human behavior; philosophy, which provides theories and questions regarding the underlying principles of thought, linguistic knowledge, and phenomena; statistics, which provides techniques for predicting events based on sample data; electrical engineering, which contributes information theory and techniques for signal processing; and biology, which explores the underlying architecture of linguistic processes in the brain.

II. BACKGROUND STUDY

Research in natural language processing has been going on for several decades dating back to the late 1940s. Machine translation (MT) was the first computer-based application related to natural language. While Weaver and Booth started one of the earliest MT projects in 1946 on computer translation based on expertise in breaking enemy codes during World War II, it was generally agreed that it was Weaver's memorandum of 1949 that brought the idea of MT to general notice and inspired many projects. He suggested using ideas from cryptography and information theory for language translation. Research began at various research Institutions in the United States within a few years.

Early work in MT took the simplistic view that the only differences between languages resided in their vocabularies and the permitted word orders. Systems developed from this perspective simply used dictionary-lookup for appropriate words for translation and reordered the words after translation to fit the word-order rules of the target language, without taking into account the lexical ambiguity inherent in natural language. This produced poor results. The apparent failure made researchers realize that the task was a lot harder than anticipated, and they needed a more adequate theory of language.

However, it was not until 1957 when Chomsky published Syntactic Structures introducing the idea of generative grammar, did the field gain better insight into whether or how mainstream linguistics could help MT. During this period, other NLP application areas began to emerge, such as speech recognition. The language processing Community and the speech community then was split into two camps with the language processing community dominated by the theoretical perspective of generative grammar and hostile to statistical methods, and the speech community dominated by statistical information theory and hostile to theoretical linguistics.

Due to the developments of the syntactic theory of language and parsing algorithms, there was over-enthusiasm in the 1950s that people believed that fully automatic high quality translation systems would be able to produce results indistinguishable from those of human translators, and such systems should be in operation within a few years. It was not only unrealistic given the then-available linguistic knowledge and computer systems, but also impossible in principle. The inadequacies of then-existing systems, and perhaps accompanied by the over enthusiasm, led to the ALPAC (Automatic Language Processing Advisory Committee of the National Academy of Science - National Research Council) report of 1966. The report concluded that MT was not immediately achievable and recommended it not be funded. This had the effect of halting MT and most work in other applications of NLP at least within the United States.

Although there was a substantial decrease in NLP work during the years after the ALPAC report, there were some significant developments, both in theoretical issues and in construction of prototype systems. Theoretical work in the late 1960's and early 1970's focused on the issue of how to represent meaning and developing computationally tractable solutions that the then-existing theories of grammar were not able to produce. In 1965, Chomsky introduced the transformational model of linguistic competence. However, the transformational generative grammars were too syntactically oriented to allow for semantic concerns. They also did not lend themselves easily to computational implementation. As a reaction to Chomsky's theories and the work of other transformational generativists, case grammar of Fillmore, semantic networks of Quillian, and conceptual dependency theory of Schank, were developed to explain syntactic anomalies, and provide semantic representations. Augmented transition networks of Woods, extended the power of phrase-structure grammar by incorporating mechanisms from programming languages such as LISP. Other representation formalisms included Wilks' preference semantics, and Kay's functional grammar.

Alongside theoretical development, many prototype systems were developed to demonstrate the effectiveness of particular principles. Weizenbaum's ELIZA was built to replicate the conversation between a psychologist and a patient, simply by permuting or echoing the user input. Winograd's SHRDLU simulated a robot that manipulated blocks on a tabletop. Despite its limitations, it showed that natural language understanding was indeed possible for the computer. PARRY attempted to embody a theory of paranoia in a system. Instead of single keywords, it used groups of keywords, and used synonyms if keywords were not found. LUNAR was developed by Woods as an interface system to a database that consisted of information about lunar rock samples using augmented transition network and procedural semantics.

In the late 1970's, attention shifted to semantic issues, discourse phenomena, and communicative goals and plans. Grosz analyzed task-oriented dialogues and proposed a

ISSN 2278-3091

International Journal of Advanced Trends in Computer Science and Engineering, Vol.2 , No.6, Pages : 175-179 (2013)
Special Issue of ICETEM 2013 - Held on 29-30 November, 2013 in Sree Visvesvaraya Institute of Technology and Science, Mahabubnagar – 204, AP, India

theory to partition the discourse into units based on her findings about the relation between the structure of a task and the structure of the task-oriented dialogue. Mann and Thompson developed Rhetorical Structure Theory, attributing hierarchical structure to discourse. Other researchers have also made significant contributions, including Hobbs and Rosensche in Polanyi and Scha, and Reichman.

This period also saw considerable work on natural language generation. McKeown's discourse planner TEXT and McDonald's response generator MUMMBLE used rhetorical predicates to produce declarative descriptions in the form of short texts, usually paragraphs. TEXT's ability to generate coherent responses online was considered a major achievement. In the early 1980s, motivated by the availability of critical computational resources, the growing awareness within each community of the limitations of isolated solutions to NLP problems, and a general push toward applications that worked with language in a broad, real-world context, researchers started re-examining non-symbolic approaches that had lost popularity in early days. By the end of 1980s, symbolic approaches had been used to address many significant problems in NLP and statistical approaches were shown to be complementary in many respects to symbolic approaches. In the last ten years of the millennium, the field was growing rapidly. This can be attributed to:

a) Increased availability of large amounts of electronic text;
b) Availability of computers with increased speed and memory; and
c) The advent of the Internet.

Statistical approaches succeeded in dealing with many generic problems in computational linguistics such as part-of-speech identification, word sense disambiguation, etc., and have become standard throughout NLP. NLP researchers are now developing next generation NLP systems that deal reasonably well with general text and account for a good portion of the variability and ambiguity of language.

III.KNOWLEDGE AND PROCESSING REQUIREMENTS

In this section, we studied certain processing requirements based on the knowledge which includes computational issues, understanding natural language, natural language knowledge levels, and classification of NLP systems.

It is important to remember that any model of natural language phenomena will eventually have to be communicated to and executed by a computing device. Turing machine will fall short of exploiting all the power of a computing device in its attempt to perform NLP; this might have considerable implications with respect to the utility of any NLP theory. On the other hand, Wegner (1997) discusses a thought-provoking alternative model of computation based on interaction, namely *interaction machines* that are more powerful than Turing machines. Specifically, he argues that any system that allows for interaction is capable of exhibiting richer behavior than a Turing machine. That is the "assertion that algorithms capture the intuitive notion of what computers compute is invalid". This supports claims of certain researchers that natural language could be effectively (if not completely) modeled by context - free, or even regular language frameworks (Blank, 1989; Marcus, 1980; Reich, 1969) – especially if such models can be trained through interaction.17 Actually, such results have contributed to empirical NLP applications in the late 1980s and 1990s based on text or speech corpora, finite-state-machine modeling frameworks, such as HMMs, and neural networks.

The success of any NLP system is highly dependent on its knowledge of the domain of discourse. Given the current state-of-the-art in NLP models, this knowledge may be subdivided into several levels. There exist different schools of thought, but, in general, researchers agree that linguistic knowledge can be subdivided into at least lexical, syntactic, semantic, and pragmatic levels. Each level conveys information in a different way. For example, the lexical level might deal with actual words (i.e., lexemes), their constituents (i.e., morphemes), and their inflected forms. The syntactic level might deal with the way words can be combined to form sentences in a given language. One way of expressing such rules is to assign words into different syntactic categories, such as noun, verb, and adjective, and specify legal combinations of these categories using a grammar. The semantic level might deal with the assignment of meaning to individual words and sentences. Finally, the pragmatic level might deal with monitoring of context/focus shifts within a dialog and with actual sentence interpretation in the given context.

1. *Acoustic/prosodic knowledge*: What are rhythm and intonation of language; how to form phonemes.
2. *Phonologic knowledge*: What are spoken sounds; how to form morphemes.
3. *Morphologic knowledge*: What are sub-word units; how to form words.
4. *Lexical knowledge*: What are words; how to derive units of meaning.
5. *Syntactic knowledge*: What are structural roles of words (or collection of words); how to form sentences.
6. *Semantic knowledge*: What is context-independent meaning; how to derive sentence meanings.
7. *Discourse knowledge*: What are structural roles of sentences (or collections of sentences); how to form dialogs.
8. *Pragmatic knowledge*: What is context-dependent meaning; how to derive sentence meanings relative to surrounding discourse.
9. *World knowledge:* What is generally known by the language user and the environment, such as user beliefs and goals; how to derive belief and goal structures. Currently, this is a catchall category for linguistic processes and phenomena that are not well understood yet. Based on past evolutionary trends, this knowledge level may be further subdivided in the future to account for new linguistic/cognitive theories and models.

The above list shows one commonly used classification which attempts to be as thorough as possible (given our current understanding of the language phenomenon) by accounting for acoustic, as well as general world knowledge (Akmajian *et al.*, 1990; Allen, 1994b; Manaris and Slator,

ISSN 2278-3091

International Journal of Advanced Trends in Computer Science and Engineering, Vol.2 , No.6, Pages : 175-179 (2013)
Special Issue of ICETEM 2013 - Held on 29-30 November, 2013 in Sree Visvesvaraya Institute of Technology and Science, Mahabubnagar – 204, AP, India

1996; Sowa, 1984). In this classification, each level is defined in terms of the declarative and procedural characteristics of knowledge that it encompasses.

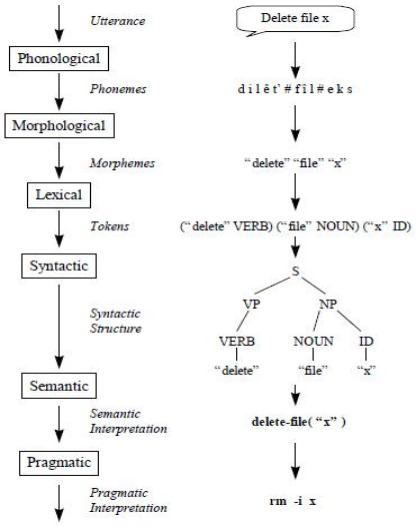

Fig 1: *Knowledge Levels in NLP Systems*

IV. NLP for NLP

Natural Language Processing and Programming Languages are both established areas in the field of Computer Science, each of them with a long research tradition. Although they are both centered on a common theme – "languages" – over the years, there has been only little interaction (if any) between them1. This paper tries to address this gap by proposing a system that attempts to convert natural language text into computer programs. While we overview the features of a natural language programming system that attempts to tackle both the descriptive and procedural programming paradigms, in this paper we focus on the aspects related to procedural programming. Starting with an English text, we show how a natural language programming system can automatically identify steps, loops, and comments, and convert them into a program skeleton that can be used as a starting point for writing a computer program, expected to be particularly useful for those who begin learning how to program.

We start by over viewing the main features of a descriptive natural language programming system METAFOR. We then describe in detail the main components of a procedural programming system as introduced in this paper. We show how some of the most difficult aspects of procedural programming, namely steps and loops, can be handled effectively using techniques that map natural language onto program structures. We demonstrate the applicability of this approach on a set of programming assignments automatically mined from the Web.

When storytellers speak fairy tales, they first describe the fantasy world – its characters, places, and situations – and then relate how events unfold in this world. Programming, resembling storytelling, can likewise be distinguished into the complementary tasks of description and proceduralization. While this paper tackles primarily the basics of building procedures out of steps and loops, it would be fruitful to also contextualize procedural rendition by discussing the architecture of the descriptive world that procedures animate.

In procedural programming, a computer program is typically composed of sequences of action statements that indicate the operations to be performed on various data structures. Correspondingly, procedural natural language programming is targeting the generation of computer programs following the procedural paradigm, starting with a natural language text.

For example, starting with the natural language text on the left side of figure 2, we would ideally like to generate a computer program as the one shown on the right side of the figure3. While this is still a long term goal, in this section we show how we can automatically generate computer program skeletons that can be used as a starting point for creating procedural computer programs. Specifically, we focus on the description of three main components of a system for natural language procedural programming:

– The step finder, which has the role of identifying in a natural language text the action statements to be converted into programming language statements.

– The loop finder, which identifies the natural language structures that indicate repetition.

– Finally, the comment identification components, which identifies the descriptive statements that can be turned into program comments.

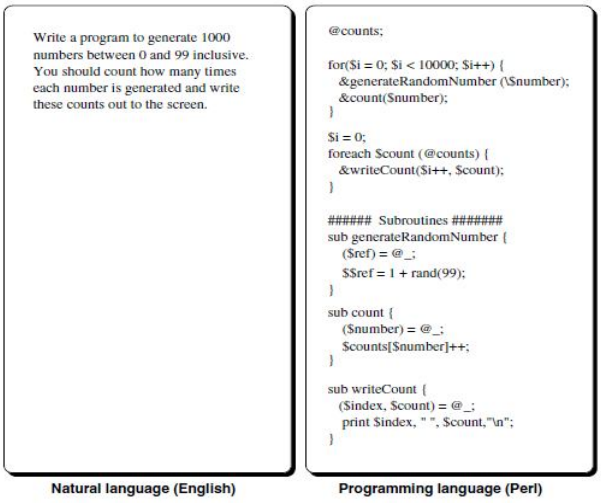

Fig 2*: Side by side: the natural language (English) and programming language (Perl) expressions for the same problem.*

V. APPLICATIONS

The most important applications of natural language processing include information retrieval and information organization, machine translation, and natural language interfaces, among others. However, as in any science, the activities of the researchers are mostly concentrated on its internal art and craft, that is, on the solution of the problems arising in analysis or generation of natural language text or speech, such as syntactic and semantic analysis, disambiguation, or compilation of dictionaries and grammars necessary for such analysis.

1. A Low-Complexity Constructive Learning Automaton Approach to Handwritten Character Recognition.

2. Utterances Assessment in Chat Conversations.
3. Punctuation Detection with Full Syntactic Parsing.
4. User Profile Modeling in eLearning using Sentiment Extraction from Text.
5. Predicting the Difficulty of Multiple - Choice Close Questions for Computer - Adaptive Testing.
6. Mathematical Text in a Controlled Natural Language.
7. Summarization
8. Machine Translation
9. Dialogue Systems
10. Information Retrieval and Extraction.

CONCLUSION

While NLP is a relatively recent area of research and application, as compared to other information technology approaches, there have been sufficient successes to date that suggest that NLP-based information access technologies will continue to be a major area of research and development in information systems now and far into the future.

ACKNOWLEDGEMENT

I would like to thank all the people who helped me in doing this research. Especially I thank my parents for their continuous support and my co-author for his contribution to this paper.

REFERENCES

[1] Alexander Gelbukh – "Natural Language Processing and Its Applications", ***Research in Computing Science*** *is published by the Center for Computing Research of IPN.* ***Volume 46****, March, 2010, p.p.no 311- 335.*
[2] Xiaoyong Liu – "Natural Language Processing", School of Information Studies at Syracuse University, Volume 2, p.p.no 1-14.
[3] Akshar Bharati, Vineet Chaitanya, Rajiv Sangal – "Natural Language Processing: A Paninian Prespective ", Prentice Hall of India, p.p.no 1-15.
[4] Bill Manaras – "Natural Language Processing: A Human–Computer Interaction Perspective", Appears in *Advances in Computers* (Marvin V. Zelkowitz, ed.), vol. 47, pp. 1-66, Academic Press, New York, 1998. p.p.no 1-35.
[5] R. Mihalcea, H. Liu, and H. Lieberman – "NLP (Natural Language Processing) for NLP (Natural Language Programming)", Springer Verlag Berlin Heidelberg 2006, p.p.no 319-330.