NATURAL LANGUAGE PROCESSING

Preeti¹ and BrahmaleenKaurSidhu²

¹ Research Scholar, University College of Engineering, Punjabi University, Patiala, Punjab, Pin Code-147002, India
  preetspal@yahoo.com

² Assistant Professor, University College of Engineering, Punjabi University, Patiala, Punjab, Pin Code-147002, India
  brahmaleen_sidhu@yahoo.co.in

ABSTRACT

Natural language processing (NLP) work began more than sixty years ago; it is a field of computer science and linguistics devoted to creating computer systems that use human (natural) language. Natural Language Processing holds great promise for making computer interfaces that are easier to use for people, since people will be able to talk to the computer in their own language, rather than learn a specialized language of computer commands. Natural Language processing techniques can make possible the use of natural language to express ideas.

1. 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 [1]. In recent years, the natural language text interpretation and processing technologies have also gained an increasing level of sophistication. 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 [2].

2. GOAL

The goal of natural language processing (NLP) is to build computational models of natural language for its analysis and generation. First, there is technological motivation of building intelligent computer systems such as machine translation systems, natural language interfaces to databases, man-machine interfaces to computers in general, speech understanding systems, text analysis and understanding systems. Second, there is a cognitive and linguistics motivation to gain a better insight into how humans communicate using natural language (NL) [3].

The goal of the Natural Language Processing (NLP) is to accomplish human-like language processing, design and build software that will analyse, understand, and generate languages that humans use naturally, so that eventually you will be able to address your computer as though you were addressing another person.
3. LEVELS

In all applications of NLP the natural languages are analyzed at different levels. These levels exist because of the regularities and the properties exhibited by a language at a number of somewhat separable levels [4]. These are:

i. **Phonology:** This level deals with the interpretation of speech sounds within and across words. There are, in fact, three types of rules used in phonological analysis:
   1) phonetic rules – for sounds within words; 2) phonemic rules – for variations of pronunciation when words are spoken together, and; 3) prosodic rules – for fluctuation in stress and intonation across a sentence.

ii. **Morphology:** The structure of words in the language, which are composed of morphemes – the smallest units of meaning that deals with the componential nature of words.

iii. **Syntactic:** How the sequences are structured. Syntactic requires both a grammar and a parser, which focuses on analyzing the words in a sentence so as to uncover the grammatical structure of the sentence. The output of this level of processing is a representation of the sentence that reveals the structural dependency relationships between the words.

iv. **Lexical:** Several types of processing contribute to word-level understanding – the first of these being assignment of a single part-of-speech tag to each word. NLP systems and human interpret the meaning of individual words. In this processing, words that can function as more than one part-of-speech are assigned the most probable part of speech tag based on the context in which they occur. Words that have only one possible sense or meaning can be replaced by a semantic representation of that meaning.

v. **Semantics:** Related to the meaning of the strings. Semantic focusing on the interactions among word-level meanings in the sentence. An analogous way to how syntactic disambiguation of words that can function as multiple parts-of-speech is accomplished at the syntactic level, include the semantic disambiguation of words with multiple senses. Semantic disambiguation permits one and only one sense of polysemous words to be selected and included in the semantic representation of the sentence.

vi. **Pragmatics:** is concerned with the use of language in situations and utilizes contents of the text for understanding. Without actually being encoded into text is the goal to explain how extra meaning is read. Some NLP applications may utilize knowledge bases and inferencing modules which requires much world knowledge, included the understanding of intentions, plans and goals.

vii. **Discourse:** The discourse level of NLP works with units of text longer than a sentence while syntax and semantics work with sentence-length units. It interprets multi sentence texts as singly. Discourse focuses on the properties of the text as a whole that convey meaning by making connections between component sentences. At this level several types of discourse processing can occur at two of the most common being anaphora resolution and discourse/text structure recognition.

4. APPROACHES TO NATURAL LANGUAGE PROCESSING

Natural language processing approaches fall roughly into four categories: symbolic, statistical, connectionist, and hybrid [4].

i. **Symbolic Approach**

Symbolic approaches perform deep analysis of linguistic phenomena and are based on explicit representation of
facts about language through well-understood knowledge representation schemes and associated algorithms.

ii. Statistical Approach
Statistical approaches describe various mathematical techniques and often use large text corpora to develop approximate generalized models of linguistic phenomena based on actual examples of these phenomena provided by the text corpora without adding significant linguistic or world knowledge. In contrast to symbolic approaches, statistical approaches use observable data as the primary source of evidence. Statistical approaches have typically been used in tasks such as speech recognition, lexical acquisition, parsing, part-of-speech tagging, collocations, statistical machine translation, statistical grammar learning, and so on.

iii. Connectionist Approach
Generally speaking, a connectionist model is a network of interconnected simple processing units with knowledge stored in the weights of the connections between units (32). Local interactions among units can result in dynamic global behavior, which, in turn, leads to computation. Some connectionist models are called localist models, assuming that each unit represents a particular concept. Similar to the statistical approaches, connectionist approaches also develop generalized models from examples of linguistic phenomena. What separates connectionism from other statistical methods is that connectionist models combine statistical learning with various theories of representation - thus the connectionist representations allow transformation, inference, and manipulation of logic formulae. In addition, in connectionist systems, linguistic models are harder to observe due to the fact that connectionist architectures are less constrained than statistical ones.

iv. Comparison Among Approaches
We have seen that similarities and differences exist between approaches in terms of their assumptions, philosophical foundations, and source of evidence. In addition to that, the similarities and differences can also be reflected in the processes each approach follows, as well as in system aspects, robustness, flexibility, and suitable tasks.

Process: Research using these different approaches follows a general set of steps, namely, data collection, data analysis/model building, rule/data construction and application of rules/data in system. The data collection stage is critical to all three approaches although statistical and connectionist approaches typically require much more data than symbolic approaches. In the data analysis/model building stage, symbolic approaches rely on human analysis of the data in order to form a theory while statistical approaches manually define a statistical model that is an approximate generalization of the collected data. Connectionist approaches build a connectionist model from the data. In the rule/data construction stage, manual efforts are typical for symbolic approaches and the theory formed in the previous step may evolve when new cases are encountered. In contrast, statistical and connectionist approaches use the statistical or connectionist model as guidance and build rules or data items automatically, usually in relatively large quantity. After building rules or data items, all approaches then automatically apply them to specific tasks in the system. For instance, connectionist approaches may apply the rules to train the weights of links between units.

System aspects: By system aspects, we mean source of data, theory or model formed from data analysis, rules, and basis for evaluation.
- Data: As mentioned earlier, symbolic approaches use human introspective data, which are usually not directly observable. Statistical and connectionist approaches are built on the basis of machine observable facets of data, usually from text corpora.
Theory or model based on data analysis: As the outcome of data analysis, a theory is formed for symbolic approaches whereas a parametric model is formed for statistical approaches and a connectionist model is formed for connectionist approaches.

Rules: For symbolic approaches, the rule construction stage usually results in rules with detailed criteria of rule application. For statistical approaches, the criteria of rule application are usually at the surface level or under-specified. For connectionist approaches, individual rules typically cannot be recognized.

Basis for Evaluation: Evaluation of symbolic systems is typically based on intuitive judgments of unaffiliated subjects and may use system-internal measures of growth such as the number of new rules. In contrast, the basis for evaluation of statistical and connectionist systems are usually in the form of scores computed from some evaluation function. However, if all approaches are utilized for the same task, then the results of the task can be evaluated both quantitatively and qualitatively and compared.

Robustness: Symbolic systems may be fragile when presented with unusual or noisy input. To deal with anomalies, they can anticipate them by making the grammar more general to accommodate them. Compared to symbolic systems, statistical systems may be more robust in the face of unexpected input provided that training data is sufficient, which may be difficult to be assured of. Connectionist systems may also be robust and fault tolerant because knowledge in such systems is stored across the network. When presented with noisy input, they degrade gradually.

Flexibility: Since symbolic models are built by human analysis of well-formulated examples, symbolic systems may lack the flexibility to adapt dynamically to experience. In contrast, statistical systems allow broad coverage, and may be better able to deal with unrestricted text (21) for more effective handling of the task at hand. Connectionist systems exhibit flexibility by dynamically acquiring appropriate behavior based on the given input. For example, the weights of a connectionist network can be adapted in realtime to improve performance. However, such systems may have difficulty with the representation of structures needed to handle complex conceptual relationships, thus limiting their abilities to handle high-level NLP (36).

Suitable tasks: Symbolic approaches seem to be suited for phenomena that exhibit identifiable linguistic behavior. They can be used to model phenomena at all the various linguistic levels described in earlier sections. Statistical approaches have proven to be effective in modeling language phenomena based on frequent use of language as reflected in text corpora. Linguistic phenomena that are not well understood or do not exhibit clear regularity are candidates for statistical approaches. Similar to statistical approaches, connectionist approaches can also deal with linguistic phenomena that are not well understood. They are useful for low-level NLP tasks that are usually subtasks in a larger problem. To summarize, symbolic, statistical, and connectionist approaches have exhibited different characteristics, thus some problems may be better tackled with one approach while other problems by another. In some cases, for some specific tasks, one approach may prove adequate, while in other cases, the tasks can get so complex that it might not be possible to choose a single best approach. In addition, as Klavans and Resnik pointed out, there is no such thing as a “purely statistical” method. Every use of statistics is based upon a symbolic model and statistics alone is not adequate for NLP. Toward this end, statistical approaches are not at odds with symbolic approaches. In fact, they are rather complementary. As a result, researchers have begun developing hybrid techniques that utilize the strengths of each approach in an attempt to address NLP problems more effectively and in a more flexible manner.
5. APPLICATIONS

Natural language processing provides both theory and implementations for a range of applications. In fact, any application that utilizes text is a candidate for NLP. The most frequent applications utilizing NLP include the following [5][6].

i. Machine Translation

Machine translation refers to automatic translation of text from one human language to another. It is necessary to have an understanding of words and phrases, grammars of the two languages involved, semantics of the languages, and world knowledge. Various levels of NLP have been utilized in MT systems, ranging from the “word-based” approach to applications that include higher levels of analysis.

ii. Speech Recognition

Speech recognition is the process of mapping acoustic speech signals to a set of words. The difficulties arise due to wide variations in the pronunciation of words, homonym and acoustic ambiguities.

iii. Speech Synthesis

Speech synthesis refers to automatic production of speech (utterance of natural language sentences). In order to generate utterances, text has to be processed. So, NLP remains an important component of any speech synthesis system.

iv. Natural Language Interfaces to Databases

Natural language interfaces allow querying a structured database using natural language sentences. Computers have been widely used to store and manage large amounts of data. The data might pertain to railway reservation, library, banking, management information, and so on.

v. Information Retrieval

Information retrieval is concerned with identifying documents relevant to a user’s request. It is given the significant presence of text. Statistical approaches for accomplishing NLP have seen more utilization, few implementations utilize NLP. NLP techniques have found useful applications in information retrieval such as indexing, word sense disambiguation, query modification and knowledge bases have also been used in IR system.

vi. Information Extraction

System captures an information extraction and factual information output contained in a document. An information retrieval system responds to a user’s information need. The information need is not expressed as a keyword query in an information retrieval system. Instead it is specified as pre-defined database schemas or templates. An IR system identifies a subset of documents in a large repository of text database. IE focuses on the recognition, tagging, and extraction into a structured representation, certain key elements of information, e.g. persons, companies, locations, organizations, from large collections of text. These extractions can then be utilized for a range of applications including question-answering, visualization, and data mining.

vii. Question Answering

Question Answering provides a list of potentially relevant documents in response to a user’s query, question-answering provides the user with either just the text of the answer itself or answer-providing passages. Given a question and a set of documents, a question answering system attempts to find the precise answer, or at least the precise portion of text in which the answer appears. A question answering system is different from an information extraction system in that the content that is to be extracted is unknown. In general, a question answering system benefits from having an information extraction system to identify entities in the text. A question


Available online@www.ijcta.com

ISSN:2229-6093
answering system requires more NLP than an information retrieval system or an information extraction system.

viii. Text Summarization
The higher levels of NLP, particularly the discourse level, can empower an implementation that reduces a larger text into a shorter, yet richly constituted abbreviated narrative representation of the original document. Text summarization deals with the creation of summaries of documents and involves syntactic, semantic, and discourse level processing of text.

ix. Dialogue Systems
Systems envisioned by large providers of end-user applications. Dialogue systems, which usually focus on a narrowly defined application (e.g. your refrigerator or home sound system), currently utilize the phonetic and lexical levels of language. It is believed that utilization of all the levels of language processing explained above offer the potential for truly habitable dialogue systems.

6. USES OF NLP:
There are many applications of natural language processing developed over the years. They can be mainly divided into two parts as follows [7].

i. Text-based applications
This involves applications such as searching for a certain topic or a keyword in a data base, extracting information from a large document, translating one language to another or summarizing text for different purposes.

7. Techniques used in analyzing NLP
There are several main techniques used in analysing natural language processing. Some of them can be briefly described as follows [8].

i. Pattern matching
The idea here is an approach to natural language processing is to interpret input utterances as a whole rather than building up their interpretation by combining the structure and meaning of words or other lower level constituents. That means the interpretations are obtained by matching patterns of words against the input utterance. For a deep level of analysis in pattern matching a large number of patterns are required even for a restricted domain. This problem can be ameliorated by hierarchical pattern matching in which the input is gradually canonicalized through pattern matching against sub phrases. Another way to reduce the number of patterns is by matching with semantic primitives instead of words.

ii. Syntactically driven Parsing
Syntax means ways that words can fit together to form higher level units such as phrases, clauses and sentences. Therefore syntactically driven parsing means interpretation of larger groups of words are built up out of the interpretation of their syntactic constituent words or phrases. In a way this is the opposite of pattern matching as here the interpretation of the input is done as a whole. Syntactic analyses are obtained by application of a grammar that determines what sentences are legal in the language that is being parsed.

iii. Semantic Grammars
Natural language analysis based on semantic grammar is bit similar to syntactically driven parsing except that in semantic grammar the categories used are defined semantically and syntactically. There here semantic grammar is also involved.
iv. Case frame instantiation

Case frame instantiation is one of the major parsing techniques under active research today. It has some very useful computational properties such as its recursive nature and its ability to combine bottom-up recognition of key constituents with top-down instantiation of less structured constituents.

8. FUTURE OF NLP

NLP’s future will be redefined as it faces new technological challenges to create more user-friendly systems. It is also forcing 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. 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. 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. 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. 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. 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. 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.

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

- **Conversational systems.** 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 auto indexing.

- **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.

- **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. MindNet then annotates these matched logical forms with data from the English-Spanish translator memory, so that translation can proceed smoothly in either direction.

- **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.

- **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.

9. **CONCLUSION**

With over sixty years of NLP research and development, the natural language systems are still very complicated to design. 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. 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. 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. There are still unresolved challenges for software programs to represent the entire knowledge, the different contexts and cultures of the world.

References:

[1] Natural Language Processing Gobinda G. Chowdhury Dept. of Computer and Information Sciences University of Strathclyde, Glasgow G1 1XH, UK


