IRACST – Engineering Science and Technology: An International Journal (ESTIJ), ISSN: 2250-3498, Vol.3, No.1, February 2013

REVIEW ON NATURAL LANGUAGE PROCESSING

Prof. Alpa Reshamwala
Dept. Of computer Engineering
MPSTME, NMIMS University
Mumbai, India
alpa.reshamwala@nmims.edu

Prof. Dhirendra Mishra
Dept. Of computer Engineering
MPSTME, NMIMS University
Mumbai, India
dhirendra.mishra@nmims.edu

Prajakta Pawar
Dept. Of computer Engineering
MPSTME, NMIMS University
Mumbai, India
Prajakta.pawar@gmail.com

***Abstract*— Natural language processing is a branch of computer science and artificial intelligence which is concerned with interaction between computers and human languages. Natural language processing is the study of mathematical and computational modeling of various aspects of language and the development of a wide range of systems. These includes the spoken language systems that integrate speech and natural language. Natural language processing has a role in computer science because many aspects of the field deal with linguistic features of computation. Natural language processing is an area of research and application that explores how computers can be used to understand and manipulates natural language text or speech to do useful things. The applications of Natural language processing includes fields of study, such as machine translation, natural language text processing and summarization, user interfaces, multilingual and cross language information retrieval (CLIR), speech recognition, artificial intelligence(AI) and expert systems.**

Keywords: Natural language Proceesing (NLP), Cross Language Information Retrieval (CLIR), Artificial intelligence (AI).

I. INTRODUCTION

Natural Language processing is an branch of computer science, artificial intelligence and linguistics concerned with the interactions between computers and human (natural) language. Natural languages are languages spoken by humans. Natural language is any language that humans learn from their environment and use to communicate with each other. Whatever the form of the communication, natural languages are used to express our knowledge and emotions and to convey our responses to other people and to our surroundings. Natural languages are usually learned in early childhood from those around us. Currently we are not yet at the point where these languages in all of their unprocessed forms can be understood by computers. Natural language processing is the collection of techniques employed to try and accomplish that goal. The field of natural language processing (NLP) is deep and diverse. Natural language processing (NLP) is a collection of techniques used to extract grammatical structure and meaning from input in order to perform a useful task as a result, natural language generation builds output based on the rules of the target language and the task at hand. NLP is useful in the tutoring systems, duplicate detection, computer supported instruction and database interface fields as it provides a pathway for increased interactivity and productivity.

II. LITERATURE REVIEW

The research work in the natural language processing has been increasingly addressed in the recent years. The natural language processing is the computerized approach to analyzing text and being a very active area of research and development. The literature distinguishes the main application of natural language processing and the methods to describe it.

1) Natural language processing for Speech Synthesis:

This is based on the text to speech conversion i.e (TTS) in which the text data is the first input into the system. It uses high level modules for speech synthesis. It uses the sentence segmentation which deals with punctuation marks with a simple decision tree.

2) Natural language processing for Speech Recognition:

Automatic speech recognition system make use of natural language processing techniques based on grammars. It uses the context free grammars for representing syntax of that language presents a means of dealing with spontaneous through the spotlighting addition of automatic summarization including indexing, which extracts the gist of the speech transcriptions in order to deal with Information retrieval and dialogue system issues.

III. LEVELS OF NLP

The most explanatory method for presenting what actually happens within a Natural Language Processing system is by means of the ‘levels of language’ approach. This is also

referred to as the synchronic model of language and is distinguished from the earlier sequential model, which hypothesizes that the levels of human language processing follow one another in a strictly sequential manner. Psycholinguistic research suggests that language processing is much more dynamic, as the levels can interact in a variety of orders. Introspection reveals that we frequently use information we gain from what is typically thought of as a higher level of processing to assist in a lower level of analysis. For example, the pragmatic knowledge that the document you are reading is about biology will be used when a particular word that has several possible senses is encountered, and the word will be interpreted as having the biology sense. Of necessity, the following description of levels will be presented sequentially. The key point here is that meaning is conveyed by each and every level of language and that since humans have been shown to use all levels of language to gain understanding, the more capable an NLP system is, the more levels of language it will utilize.

A. 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 [11]:

1) Phonetic rules:

It is used for sound within words.

2) Phonemic rules :

It is used for variations of pronunciation when words are spoken together.

3) Prosodic rules :

It is used to check for fluctuation in stress and intonation across a sentence.

In an NLP system that accepts spoken input, the sound waves are analyzed and encoded into a digitized signal for interpretation by various rules or by comparison to the particular language model being utilized.

B. Morphology:

Morphology is the first stage of analysis once input has been received. It looks at the ways in which words break down into their components and how that affects their grammatical status. Morphology is mainly useful for identifying the parts of speech in a sentence and words that interact together. The following quote from Forsberg gives a little background on the field of morphology.

Morphology is a systematic description of words in a natural language. It describes a set of relations between words' surface forms and lexical forms. A word's surface form is its graphical or spoken form, and the lexical form is an analysis of the word into its lemma (also known as its dictionary form) and its grammatical description. This task is more precisely called inflectional morphology. Being able to identify the part of speech is essential to identifying the grammatical context a word belongs to. In English, regular verbs have a ground form with a limited set of modifications, however, irregular verbs do not follow these modification rules, and greatly increase the complexity of a language. The information gathered at the morphological stage prepares the data for the syntactical stage which looks more directly at the target language's grammatical structure.

1) Syntax:

Syntax involves applying the rules of the target language's grammar, its task is to determine the role of each word in a sentence and organize this data into a structure that is more easily manipulated for further analysis. Semantics are the examination of the meaning of words and sentences.

a) Grammar:

In English, a statement consists of a noun phrase, a verb phrase, and in some cases, a prepositional phrase. A noun phrase represents a subject that can be summarized or identified by a noun. This phrase may have articles and adjectives and/or an embedded verb phrase as well as the noun itself. A verb phrase represents an action and may include an imbedded noun phrase along with the verb. A prepositional phrase describes a noun or verb in the sentence. The majority of natural languages are made up of a number of parts of speech mainly: verbs, nouns, adjectives, adverbs, conjunctions, pronouns and articles.

b) Parsing:

Parsing is the process of converting a sentence into a tree that represents the sentence's syntactic structure. The statement: "The green book is sitting on the desk" consists of the noun phrase: "The green book" and the verb phrase: "is sitting on the desk." The sentence tree would start at the sentence level and break it down into the noun and verb phrase. It would then label the articles, the adjectives and the nouns. Parsing determines whether a sentence is valid in relation to the language's grammar rules.

C. Semantics:

It builds up a representation of the objects and actions that a sentence is describing and includes the details provided by adjectives, adverbs and propositions. This process gathers information vital to the pragmatic analysis in order to determine which meaning was intended by the user.

D. Pragmatics:

Pragmatics is "the analysis of the real meaning of an utterance in a human language, by disambiguating and contextualizing the utterance". This is accomplished by identifying ambiguities encountered by the system and resolving them using one or more types of disambiguation techniques .

1) Ambiguity:

Ambiguity is explained as "the problem that an utterance in a human language can have more than one possible meaning.

Types of Ambiguity:
Syntactic Ambiguity is present when more than one parse of a sentence exists. “He lifted the branch with the red leaf.” The verb phrase may contain “with the red leaf” as part of the imbedded noun phrase describing the branch or “with the red leaf” may be interpreted as a prepositional phrase describing the action instead of the branch, implying that he used the red leaf to lift the branch.

- Semantic Ambiguity is existent when more than one possible meaning exists for a sentence as in “He lifted the branch with the red leaf.” It may mean that the person in question used a red leaf to lift the branch or that he lifted a branch that had a red leaf on it.
- Referential Ambiguity is the result of referring to something without explicitly naming it by using words like “it”, ‘he” and “they.” These words require the target to be looked up and may be impossible to resolve such as in the sentence: “The interface sent the peripheral device data which caused it to break”, it could mean the peripheral device, the data, or the interface.
- Local Ambiguity occurs when a part of a sentence is unclear but is resolved when the sentence as a whole is examined. The sentence: “this hall is colder than the room,” exemplifies local ambiguity as the phrase: “is colder than” is indefinite until “the room” is defined.

IV. METHODS AND APPROACHES

A. Natural Language Processing for Speech Synthesis:

TTS synthesis makes use of NLP techniques extensively since text data is first input into the system and thus it must be processed in the first place. [1] describes the different high-level modules involved in this sequential process: Text Normalization Adapts the input text so as to be synthesized.

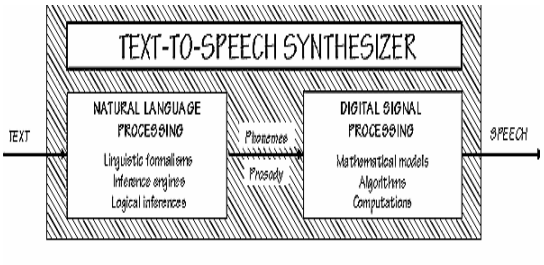

Figure1: TTS System [10]

It contemplates the aspects that are normally taken for granted when reading a text. The sentence segmentation can be achieved though dealing with punctuation marks with a simple decision tree. But more confusing situations require more complex methods. Some examples of these difficulties are the period marking, the disambiguation between the capital letters in proper names and the beginning of sentences, the abbreviations, etc. The tokenization separates the units that build up a piece of text. It normally splits the text of the sentences at white spaces and punctuation marks. This process is successfully accomlished with a parser. Finally, non-standard words such as certain abbreviations (Mr., Dr., etc.), date constructs, phone numbers, acronyms or email and URL addresses need to be expanded into more tokens (units) in order to be synthesized correctly. Rules and dictionaries are of use to deal with non-standard words. Part-of-Speech Tagging assigns a word-class to each token. Thus this process consecutes the Text Normalization. Part-of-Speech taggers have to deal with unknown words (Out-Of Vocabulary problem) and words with ambiguous POS tags (same structure in the sentence) such as nouns, verbs and adjectives. As an example, the use a participle as an adjective for a noun in “broken glass”.
Grapheme-to-Phoneme Conversion Assigns the correct phonetic set to the token stream. It must be stated that this is a continuous language dependent process since the phonetic transcriptions of the token boundaries are influenced by the transcriptions of the neighboring token boundaries. Thus, accounting for the influence of morphology and syllable structure can improve performance of Grapheme-to-Phoneme conversion [5].
Word Stress Assigns the stress to the words, a process tightly bound to the language of study. The phonological, morphological and word class features are essential characteristics in this assignment: the stress is mostly determined by the syllable weight (phonological phenomena which treat certain syllable types as heavier than others [6]). See [1] for a wide set of references for this process.

B. Natural Language Processing for Speech Recognition:

Automatic Speech Recognition systems make use of NLP techniques in a fairly restricted way: they are based on grammars. This paper refers to a grammar as a set of rules that determine the structure of texts written in a given language by defining its morphology and syntax. ASR takes for granted that the incoming speech utterances must be produced according to this predetermined set of rules established by the grammar of a language, as it happens for a formal language. In that case, Context-Free Grammars (CFG) play an important role since they are well capable of representing the syntax of that language while being efficient at the analysis (parsing) of the sentences. For this reason/restriction, such language cannot be considered natural. ASR systems assume though that a large enough grammar rule set enable any (strictly formal) language to be taken for natural. NLP techniques are of use in ASR when modeling the language or domain of interaction in question.
Through the production of an accurate set of rules for the grammar, the structures for the language are defined. These rules can either be 1) hand-crafted or 2) derived from the statistical analyses performed on a labelled corpus of data. The former implies a great deal of hard-work since this process is

neither simple nor brief because it has to represent the whole set of grammatical rules for the application. The latter is generally the chosen one because of its programming flexibility at the expense of a tradeoff between the complexity of the process, the accuracy of the models and the volume of training and test data available (notice that the corpus has to be labelled, which implies a considerably hard workload). Since hand-crafted grammars depend solely on linguistics for a particular language and application they have little interest in machine learning research in general. Thus, the literature is extensive on the datadriven approaches (N-gram statistics, word lattices,
etc.) bearing in mind that by definition a grammarbased representation of a language is a subset of a natural language. Aiming at a flexible enough grammar to generalize the most typical sentences for an application, [2] and [3] end up building N-gram language models.
N-grams model a language through the estimates of sequences of N consecutive words. While the former tackles the problem with a binary decision tree, the latter chooses to use more conventional Language Modeling theory (smoothing, cutoffs, context cues and vocabulary types) also makes use of N gram structures but it pursues a unified model integrating CFGs. Refer to the cited articles for further information. Lastly, [4] presents a means of dealing with spontaneous-speech through the spotlighting addition of automatic summarization including indexing, which extracts the gist of the speech transcriptions in order to deal with Information Retrieval (IR) and dialogue system issues.

V. 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. The state-of-the-art Natural Language Processing techniques applied to speech technologies, specifically to Text-To-Speech synthesis and Automatic Speech Recognition. In 3TTS. The importance of NLP in processing the input text to be synthesized is reflected. The naturalness of the speech utterances produced by the signal-processing modules are tightly bound to the performance of the previous text-processing modules. In ASR the use of NLP particularly is complementary [7].
It simplifies the recognition task by assuming that the input speech utterances must be produced according to a predefined set of grammatical rules. Its capabilities can though be enhanced through the usage of NLP aiming at more natural interfaces with a certain degree of knowledge. Reviews the major approaches proposed in language model adaptation in order to profit from this specific knowledge.

VI. FUTURE WORK

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 [9].
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).

REFERENCES

[1] L. R. Bahl, P. F. Brown, P. V. de Souza, and R. L. Mercer, "A tree based statistical language model for natural language speech recognition," in Acoustics, Speech and Signal Processing, IEEE Transactions on, vol. 37, Issue 7, (Yorktown Heights, NY,USA), pp. 1001–1008, July 1989.

[2] P . Clarkson and R. Rosenfeld, "Statistical language modeling using the cmu-cambridge toolkit," in Proceedings EUROSPEECH (N. F.G. Kokkinakis and E. Dermatas, eds.), vol. 1, (Rhodes, Greece), pp. 2707–2710, September 1997.

[3]] J. Tejedor, R. Garca, M. Fernndez, F. J. LpezColino, F. Perdrix, J. A. Macas, R. M. Gil, M. Oliva, D. Moya, J. Cols, , and P. Castells, "Ontology-based retrieval of human speech," in Database and Expert Systems Applications, 2007. DEXA '07. 18th International Conference on, (Regensburg, Germany), pp. 485–489, September 2007.

[4] J. R. Bellegarda, "Statistical language model adaptation: Review and perspectives," vol. 42, no. 1, pp. 93–108, 2004.

[5] Y.-Y. Wang, M. Mahajan, and X. Huang, "A unified context-free grammar and n-gram model for spoken language processing," in IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. III, (Istanbul, Turkey), pp. 1639–1642, Institute of Electrical and Electronics Engineers, Inc., 2000

[6] L. Zhou and D. Zhang, "NLPIR: a theoretical framework for applying natural language processing to information retrieval," J. Am. Soc. Inf. Sci. Technol., vol. 54, no. 2, pp. 115–123, 2003

[7] L. Zhou and D. Zhang, "NLPIR: a theoretical framework for applying natural language processing to information retrieval," J. Am. Soc. Inf. Sci. Technol., vol. 54, no. 2, pp. 115–123, 2003.

[8] Wohleb, R. "Natural Language Processing: Understanding Its Future," PC/AI, November/December, 2001.

[9] Guerra, A. "T. Rowe Price to hone in on voice systems," Wall Street and Technology, Vol. 19, No. 3, 2000.

[10] "TTS SYSTEM" Internet http://tcts.fpms.ac.be/synthesis/introtts_old.html [jan1 2013].