A Rule Based System for Speech Language Context Understanding

Imran Sarwar Bajwa[1], Muhammad Abbas Choudhary[2]

1 CISUC –Department of Informatics Engineering, University of Coimbra, 3030 Coimbra, Portugal
2 Balochistan University of Information Technology and Management Sciences, Quetta, Pakistan

imran@dei.uc.pt, abbas@buitms.edu.pk

Abstract

Speech or Natural language contents are major tools of communication. This research paper presents a natural language processing based automated system for understanding speech language text. A new rule based model has been presented for analyzing the natural languages and extracting the relative meanings from the given text. User writes the natural language text in simple English in a few paragraphs and the designed system has a sound ability of analyzing the given script by the user. After composite analysis and extraction of associated information, the designed system gives particular meanings to an assortment of speech language text on the basis of its context. The designed system uses standard speech language rules that are clearly defined for all speech languages as English, Urdu, Chinese, Arabic, French, etc. The designed system provides a quick and reliable way to comprehend speech language context and generate respective meanings.

Keywords: *automatic text understanding*, *speech language processing*, *information extraction*, *language engineering*.

1. Introduction

In daily life, natural languages or speech languages are main source of commutation. Particular and typical set of words and vocabulary collections are used for each field of life and they are quiet exclusive to their use. In computer science, Natural Language Processing (NLP) is the field used for the analysis of speech language text [1]. NLP introduces many new concepts and new terminologies to describe its models, techniques and processes. Some of these terms come directly from linguistics and the science of perception, while others were invented to describe discoveries that did not fit into any previous category [2]. Natural Language understanding is a hefty field to grasp. Understanding and comprehending a speech language means to know that what concepts a word or a particular phrase stands under a particular perspective. To give meanings to a particular sentence a system should know how to link the concepts together in a meaningful way. Speech languages are easy for human beings in terms of learning, understanding and using. On the other hand it is quite complicated to model natural languages for a computer to understand [3]. NLP includes techniques like word stemming (removing suffixes), lemmatization (replacing an inflected word with its base form), multiword phrase grouping, synonym normalization, part-of-speech (POS) tagging (elaborations on noun, verb, preposition etc.), word-sense disambiguation, anaphora resolution, and role determination (subject and object), etc [1].

In modern era, computer machines have attained the ability of solving complex mathematical and statistical problems with speed and grace but still they are inefficient to comprehend the basics of spoken and written languages [4]. The designed automated system can be useful in various business and technical software by only acquiring the requirements from the user. The designed system will extract the required information from the given text and provide to the computer for further automated processing. Applications can be automated generation of UML diagrams, query processing, web mining, web template designing, user interface designing, etc.

2. Related Work

Natural languages have been an area of interest from last one century. In the late nineteen sixties and seventies, so many researchers as Noam Chomsky (1965) [5], Maron, M. E. and Kuhns, J. L (1960) [6], Chow, C., & Liu, C (1968) [7] contributed in the area of information retrieval from natural languages. They contributed for analysis and understanding of the natural languages, but still there was lot of effort required for better understanding and analysis. Some authors concentrated in this area in eighties and nineties as Losee, R. M (1998) [8], Salton, G., & McGill, M (1995) [9], Krovetz, R., & Croft, W. B (1992) [10]. These authors worked for lexical ambiguity and information retrieval [7], probabilistic indexing [9], data bases handling [10] and so many other related areas.

In this research paper, a newly designed rule based framework has been proposed that is able to read the English language text and extract its meanings after analyzing and extracting related information. The

conducted research provides a robust solution to the addressed problem. The functionality of the conducted research was domain specific but it can be enhanced easily in the future according to the requirements. Current designed system incorporate the capability of mapping user requirements after reading the given requirements in plain text and drawing the set of speech language contents.

3. Speech Language Analysis

Natural Language processing is field of interest for the computer science researchers from last few decades. Various statistical and non-statistical methods have been used to design a robust algorithm to understand true semantics of a natural language text. Neural networks can be a great help in this context, especially imply the power of Natural Language processing (NLP) in complex, real world problems. NLP bases systems have abilities: part of Speech (POS) tagging, error detection in annotated corpora and self organization of semantic maps [9]. Human Language Technology (HLT) helps in processing human languages for various applications: Speech Recognition, Machine Translation, Text Generation and Text Mining. Text understanding and mining is one of the major applications of natural language processing (NLP).

In Natural language processing various linguistic techniques i.e. lexical and semantic analysis, have been developed [10]. Text parsing in NLP can be a detailed process or trivial process. Some applications e.g. text summarization and text generation needs detailed text processing. In detailed process every part of each sentence is analyzed. Some applications just need trivial processing e.g. text mining and web mining. In trivial processing of text, only certain passages or phrases within a sentence are processed [11].

Natural Language Processing require a series of actions to process the text and retrieve required information. In this information, first of all knowledge about phonetics and phonology is required. These tasks further demand recognition of variations in words that is called morphology. Afterwards, syntax of the text is understood which requires the knowledge needed to order and group words together. After syntax, semantical analysis is performed, in which the meanings of the sentences are understood, mainly due to compositional semantics [12]. To have a more compound understanding of the sentence, pragmatics are required, where the sentence is analyzed according to the context. Discourse analysis is the last part of NLP, where the semantic analysis of the linguistic structure is performed beyond the sentence level [13].

4. Designed System Architecture

The designed speech language contents system has ability to draw speech language contents after reading the text scenario provided by the user. This system draws in five modules: Text input acquisition, text understanding, knowledge extraction, generation of speech language contents and finally multi-lingual code generation as shown in following fig.

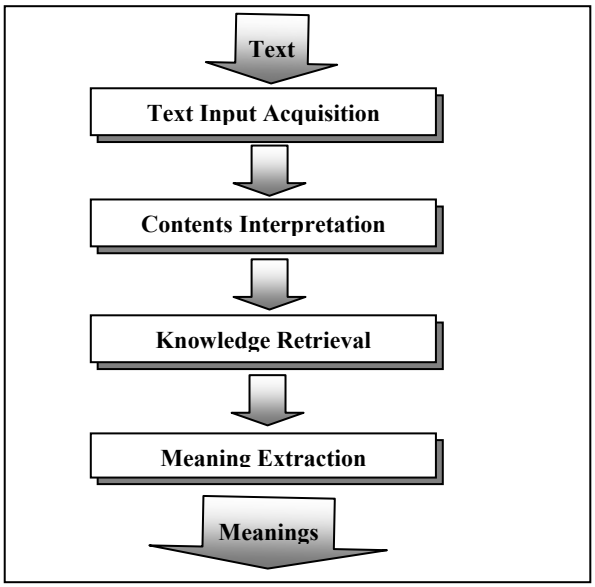

Figure 1- *Architecture of the designed Speech Language Context Understanding Model*

1. Text input acquisition

This module helps to acquire input text scenario. User provides the business scenario in from of paragraphs of the text. This module reads the input text in the form characters and generates the words by concatenating the input characters. This module is the implementation of the lexical phase. Lexicons and tokens are generated in this module.

2. Parts of Speech Tagging

Part of speech tagging is the process of assigning a part-of-speech or other lexical speech marker to each word in a corpus. *Rule-based taggers* are the earliest taggers and they used hand-written disambiguation rules to assign a single part-of-speech to each word [12]. This module categorizes tokens into various classes as verbs, nouns, pronouns, adjectives, prepositions, conjunctions, etc.

3. Information Repossession

In information repossession module, distinct objects and classes and their respective attributes are extracted on the basses of the input provided by the prior module. Nouns are symbolized as classes and objects and their associated characteristics are termed as attributes. In pattern analysis phase, irrelative rules and patterns are eliminated for efficient pattern discovery process [13].

4. Meaning extraction

This module finally uses speech language contents symbols to constitute the various speech language contents by combining available symbols according to the information extracted of the previous module. Distinct representations are created and assigned to various linguistic inputs. The process of verifying the meanings of a sentence is performed by matching the input representations with the representation in the knowledge base [14].

5. Used Methodology

Conventional methods for natural language processing use rule based techniques besides Hidden Markov Method (HMM), Neural Networks (NN), probabilistic theory and statistical methods. Agents are another way to address this problem [8]. In the research, a rule-based algorithm has been used which has robust ability to read, understand and extract the desired information. First of all basic elements of the language grammar are extracted as verbs, nouns, adjectives, etc then on the basis of this extracted information further processing is performed. In linguistic terms, verbs often specify actions, and noun phrases the objects that participate in the action [14]. Each noun phrase specifies that how an object participates in the action.

A robust method that has robust ability to understand a natural language sentence must discover the actor. Actor performs the action in a sentence. To manifest the meanings of a sentence some categorization has been employed. In a sentence different words represent different course of actions. Some are doing work and for some one a work is being done. Some tool can be used to perform a certain task on a certain place on a specific time. This type of labeling can make the process of meaning representation easy and simple. Following are major labels that are applied to the different parts of the sentences: actor, co-actor, recipient, conveyance, route, site, time, duration. These labels can be identified by the help of prepositions.

a. Actor object

The actor represents the person that causes an action in a sentence. For example in sentence "Ahmed hits the ball," Ahmed is agent who performs the task. But in this example a passive sentence, the agent also may appear after a preposition for example: "The ball was hit by Ahmed." The "*subject*" part in a sentence; that are Nouns and pronouns appearing before the verb phrase in a sentence are nominated as actors.

b. Co-Actor object

Co-actor is the joining phrase that serves as a partner of the principal actor. Actor and co-actor carry out the action together for example "Ahmed played tennis with Ali." In this example Ahmed is actor and Ali is co-actor as Ahmed is playing with Ali. Nouns and pronouns coming after the verb phrase are normally co-actors.

c. Recipient object

The recipient is the person for whom an action has been performed. Recipient is that particular part of a sentence, about which that sentence is. "Ahmed brought the balls for Ali." In this sentence, Ali is the recipient, where Ahmed is actor that has brought the ball for Ali.

d. Thematic object

The thematic object is an entity in the sentence, on which the action is being performed. Generally, '*object'* part in a sentence is represented as the thematic object. Often the thematic object is the same as the syntactic direct object, as "Ahmed hit the ball." Here the ball is thematic object.

e. Conveyance object

The conveyance is something in which or on which somebody travels. For example in the sentence "Aslam always goes by train", train is conveyance object. Conveyance object normally appears after by or via prepositions.

f. Route object

Motion from source to destination takes place over at route or trajectory. Several prepositions can serve to introduce trajectory noun phrases: through, from, to "Ahmed and Aslam went in through the front door."

g. Site object

The location is where an action occurs. As in the trajectory role, several prepositions are possible, which conveys meanings in addition to serve as a signal that a location noun phrase is "Ahmed and Ali studied in the library, at a desk, by the wall, near the door."

h. Time object

Time specifies the occasion of the occurrence of an action. Prepositions like at, before, and after introduce noun phrases serving as time. For example, "Ahmed and Ali left before noon", before is the time object.

i. Duration object

Duration specifies how long an action takes. Preposition such as for, since indicate duration. "Ali and Zia jogged for an hour."

6. Conclusion

The designed system for speech language context understanding using a rule based algorithm is a robust

framework for inferring the appropriate meanings from a given text. The accomplished research is related to the understanding of the human languages. Human being needs specific linguistic knowledge generating and understanding speech language contents. It is difficult for computers to perform this task. The speech language context understanding using a rule based framework has ability to read user provided text, extract related information and ultimately give meanings to the extracted contents. The designed system very effective and have high accuracy up to 90 %. The designed system depicts the meanings of the given sentence or paragraph efficiently. An elegant graphical user interface has also been provided to the user for entering the Input scenario in a proper way and generating speech language contents.

7. Future Work

The designed system analyzes and extracts the contents of a speech language as English. Current designed system only works for the active-vice sentences. Passive-vice sentences are still to work for to make the system more efficient and effective for various business applications. There is also periphery of improvements in the algorithms for the better extraction of language parts and understanding the meanings of the speech language context. Current accuracy of generating meanings from a text is about 85% to 90%. It can be enhanced up to 95% or even more by improving the algorithms and inducing the ability of learning in the system.

8. References

[1] Anne Kao & Steve Poteet, (2005) "Text Mining and Natural Language Processing – Introduction for the Special Issue", ACM Press New York, NY, USA

[2] Blaschke C, Andrade M. A, Ouzounis C. and Valencia,A. (1999) "Automatic extraction of biological information from scientific text: protein–protein interactions". In the proceedings of ISMB, 1999, Heidelberg, Germany, pp. 60–67.

[3] C. A. Thompson, R. J. Mooney and L. R. Tang, (1997) "Learning to parse natural language database queries into logical form", in: Workshop on Automata Induction, Grammatical Inference and Language Acquisition, Nashville, Tennessee,1997.

[4] Pustejovsky J, Castaño J., Zhang J, Kotecki M, Cochran B. (2002) "Robust relational parsing over biomedical literature: Extracting inhibit relations". In proc. of Pacific Symposium of Bio-computing, 362-373, 2002.

[5] Maron, M. E. & Kuhns, J. L. (1997) "On relevance, probabilistic indexing, and information retrieval" Journal of the ACM, 1997, 7, 216–244.

[6] Chomsky, N. (1965) "Aspects of the Theory of Syntax. MIT Press, Cambridge, Mass, 1965.

[7] Chow, C., & Liu, C. (1968) "Approximating discrete probability distributions with dependence trees". IEEE Transactions on Information Theory, 1968, IT-14(3), 462–467.

[8] Losee, R. M. (1988) "Parameter estimation for probabilistic document retrieval models". Journal of the American Society for Information Science, 39(1), 1988, pp. 8–16.

[9] Salton, G., & McGill, M. (1995) "Introduction to Modern Information Retrieval" McGraw-Hill, New York., 1995

[10] Krovetz, R., & Croft, W. B. (1992) "Lexical ambiguity and information retrieval", ACM Transactions on Information Systems, 10, 1992, pp. 115–141

[11] Partee, B. H., Meulen, A. t., &Wall, R. E. (1999) "Mathematical Methods in Linguistics". Kluwer, Dordrecht, The Netherlands. 1999.

[12] Voutilainen, A. (1995), "Constraint Grammar: A language Independent system for parsing Unrestricted Text", Morphological disambiguation, pp. 165-284

[13] Jurafsky, Daniel, and James H. Martin. (2000). "Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics". Prentice-Hall

[14] WangBin, LiuZhijing, (2003), "Web Mining Research", Proceedings of Fifth International Conference on Computational Intelligence and Multimedia Applications, 2003. ICCIMA 2003, Xi'an, CHINA. Page(s):84 - 89