International Journal of Computer Theory and Engineering, Vol. 2, No. 1 February, 2010
1793-8201

Context Based Meaning Extraction by Means of Markov Logic

Imran Sarwar Bajwa

Abstract—**Understanding meanings and semantics of a speech or natural language is a complicated problem. This problem becomes more vital and classy when meanings with respect to context, have to be extracted. This particular research area has been typically point of meditation for the last few decades and many various techniques have been used to address this problem. An automated system is required that may be able to analyze and understand a few paragraphs in English language. In this research, Markov Logic has been incorporated to analyze and understand the natural language script given by the user. The designed system formulates the standard speech language rules with certain weights. These meticulous weights for each rule ultimately support in deciding the particular meaning of a phrase and sentence. The designed system provides an easy and consistent way to figure out speech language context and produce respective meanings of the text.**

Index Terms—**Text Processing, Markov Logic, Meaning Extraction, Language Engineering, Context Awareness.**

I. Introduction

Grammar analysis and meanings extraction are primary steps involved in almost every natural language processing based systems. This scenario becomes more significant and critical when the meanings of a piece of text have to be extracted in a particular **context**. Context based meaning extraction is important for many NLP (Natural Language Processing) based applications i.e. Automated generation of UML(Unified Modeling Language) diagrams, query processing, web mining, web template designing, user interface designing, etc. Modern information systems tend to base on agile and aspect oriented mechanisms. The aspect oriented and component oriented software engineering are becoming quite popular in software designing communities. These newly emerging methodologies can be assisted by providing natural languages' based used interfaces. The research area of improved and effective interfaces for modern software paradigms is one of the major and modern research interests. NLP based context aware user interfaces can be effective in this regard. This research highlights the prospects of assimilation of Markov Logics in designing of various business and technical software.

For analyzing natural language text, many statistical and non statistical models have been presented [3]. Some of the models are knowledge-base model, stochastic logic, probabilistic relational model, and relational Markov model. A probabilistic model can also be represented as a Markov network including Bayesian networks, decision trees, logistic regression, etc [4]. First order logic (FOL) has been used as a tool of predicate logic. But first order logic deals only with linear data. Natural languages tend to be non-linear. The feature of non-linearity in natural languages cannot be knobbed by conventional first order logic. Markov Logics (ML) is simple extension to first-order logic. In Markov Logic, each formula has an additional weight fixed with it [5], in variation of first order logic. In ML, a formula's associated weight reflects the strength of a constraint. The higher weight of a formula represents the greater the difference in log probability and it also satisfies the formula. A first-order KB is a set of formulas in first-order logic, constructed from predicates using logical connectives and quantifiers.

In this article, the section 2 presents review of related work. Section 3 presents the architecture of designed system. Section 4 describes the used methodology based on compositely using rule based approach for extracting the required information from the given text and then employing Markov Logics for further text understanding. The results and analysis has been presented din the last section.

II. Related Work

In this section a short overview over existing approaches for processing and computing the natural language text has been presented.

A. General Text Processing

The major research contributions in the area of computational linguistics have brought into being for last many decades, Major contributors are Maron, M. E. and Kuhns, J. L (1960) [6], Noam Chomsky (1965) [5], Chow, C., & Liu, C (1968) [7]. They presented the methods of information retrieval from natural languages. Other contributions were analysis and understanding of the natural languages, but still there was lot of effort required for better understanding and analysis.

B. Meaning Extraction

Natural language text can be helpful in multiple ways to make computer applications more convenient and easy to use. A CASE tool named REBUILDER UML [1] integrates a module for translation of natural language text into an UML class diagram. This module uses an approach based on

Imran Sarwar Bajwa is with Department of Computer Science and IT, The Islamia University of Bahawalpur, Pakistan, Ph: +92 62 925 5466, imran.sarwar@iub.edu.pk

Case-Based Reasoning and Natural Language Processing. For information extraction and processing at semantic level, Pedro Domingos [11] also presented a technique Markov Logics. Markov logics are simple extension to FOL. There are many applications of Markov logics; link prediction, collective classification, entity resolution, social network analysis, etc. Alchemy system [12] facilitates implement the inference and learning algorithms. Alchemy is based on a declarative programming language that is similar to Prolog. Markov logic has ability to handle uncertainty and learn form the training data. We also have presented a rule based system [13] that is able to extract desired information from the natural language text. The system understands context and then extracts respective information. This model is further enhanced in this research to capture the information from NL text that is further used for semantic understating. The implementation details have been provided in section 4.

C. *Information Retrieval*

The second category of prior studies concentrates on contexts consisting of a single word only, typically modeling the combination of a predicate p and an argument a. Kintsch (2001) uses vector representations of p and a to identify the set of words that are similar to both p and a. In the nineties, major contributions were turned out by Krovetz, R., & Croft, W. B (1992) [10], Salton, G., & McGill, M (1995) [9], Losee, R. M (1998) [8], These authors worked for lexical ambiguity and information retrieval [8], probabilistic indexing [9], data bases handling [10] and so many other related areas.

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]. Scientists are used to formerly employ rule-based/ statistical algorithm with a text data base i.e. WordNet to identify the data type of a text piece and then understand and extract the desired information from the given piece of text. Parts of speech tagging is a typical phase in this procedure in which all basic elements of the language grammar are extracted as verbs, nouns, adjectives, etc. Detailed description has been provided in experiments section.

III. Designed System Architecture

In this research paper, a newly designed Markov Logic based system has been presented that is able to read the English language text and extract its meanings after analyzing and extracting related information. Various linguistic phases are common for processing natural language i.e. lexical and semantic analysis, pragmatic analysis Text parsing in NLP can be a detailed 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].

An architecture based on three component modules is presented in the Figure 1.0. First component provides support for phonetics and phonology related issues, if required. This component further provides recognition of variations in words and this step is formally called morphology. Afterwards, syntax of the text is understood which requires the knowledge needed to order and group words together.

The second component provides support for semantical analysis of POS (parts of speech) tagged text. At this level, the Markov logic has been used to understand meanings of the sentences [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].

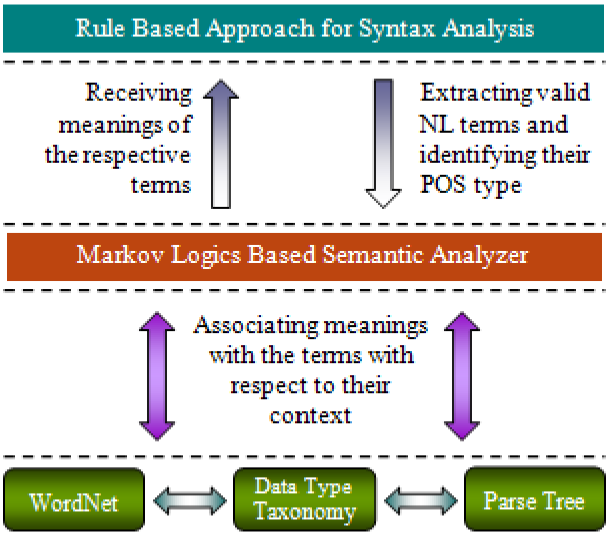
Rule Based Approach for Syntax Analysis
Receiving meanings of the respective terms
Extracting valid NL terms and identifying their POS type
Markov Logics Based Semantic Analyzer
Associating meanings with the terms with respect to their context
WordNet
Data Type Taxonomy
Parse Tree

Fig.1 Employing Markov Logic for Meaning Extraction

The third and last component is the knowledge base of the system that consists of English word libraries; data type taxonomies and parse trees. The designed system has ability to understand English language contents after reading the text scenario provided by the user. This system is based on a layered design and common layers are text input acquisition layer, parts of speech tagging layer, Markov Logic based Network layer, pattern analysis layer of speech language contents and finally meaning extraction layer. The complete design is shown in Figure 2.0.

A. *Getting Text Input*

Typically, the natural language text has no formal format and accordance. To make the take more appropriate for processing, input text is acquired. User can provide the input scenario in paragraphs of the text. Major issues in this phase are to read the input text in the form characters and generate the words by concatenating them. Another major issue is to remove the text noise and abnormalities from text. In general, this unit is the implementation of the lexical phase of text processing.

B. *POS Tagging of Text*

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 has also used a rule based algorithm to categorize tokens into various classes as verbs, nouns, pronouns, adjectives, prepositions, conjunctions, etc.

C. Markov Logic based Network

In Markov Logics [], every FOL formula has an associated weight that represents the strength of the constraint. Markov logic allows contradiction between various formulas and thee contradictions are resolved by comparing the evidence weights of multiple constraints. A Markov Logic Network [] (MLN) can be viewed as a template for constructing a Markov network. Markov logic is used in this layer to define inference rules similar to the used in the Markov networks and Bayesian networks. These inference rules are used to find most probable meanings of the words given some evidence. MLN weights another ability that is learning. MLN weights can learn through the maximizing of the relational database.

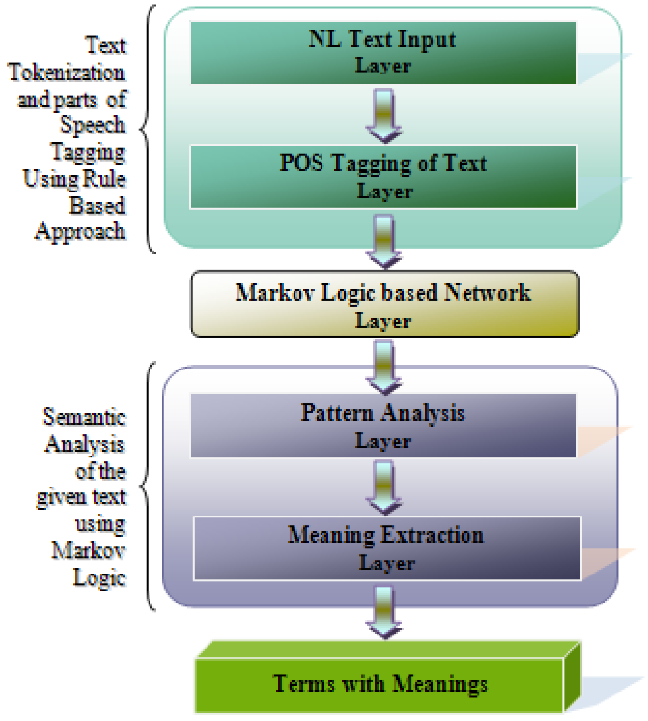
Text Tokenization and parts of Speech Tagging Using Rule Based Approach
NL Text Input Layer
POS Tagging of Text Layer
Markov Logic based Network Layer
Semantic Analysis of the given text using Markov Logic
Pattern Analysis Layer
Meaning Extraction Layer
Terms with Meanings

Fig.2 Markov Logic based Meaning Extraction

D. Pattern Analysis

In linguistic terms, verbs often specify actions, and noun phrases the objects those participate in a particular action [14]. Each noun phrase specifies that how an object participates in the action. This module typically provides support to identify distinct objects and their respective attributes. Nouns are symbolized as objects and their associated characteristics are termed as attributes. In pattern analysis phase, irrelative rules and patterns are also eliminated for efficient pattern discovery process [13].

E. Meaning Extraction

Prepositions can help out in developing relations among objects identified in the previous module. This module finally uses MLN to identify the relations and conclude appropriate meanings on the basis of the extractions. Distinct representations are ultimately 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].

IV. Results and Discussion

To validate the presented model, four classes have defined: subject class, verb class, object class, adverb class. These classes are defined on the basis of classical division of a typical English sentence.

$$\underbrace{\text{Student}}_{\textbf{Subject Part}} \quad \underbrace{\text{is reading}}_{\textbf{Verb Part}} \quad \underbrace{\text{a book}}_{\textbf{Object Part}} \quad \underbrace{\text{in library}}_{\textbf{Adverb Part}}$$

First of all it was observed that at which percentage the words are accurately classified. For this purpose, two types of data sets have been used; first data set of text for training of the designed system and other data set with text for testing. The rules of Markov Logic have been trained with the training data set. The weights of the rules were randomly selected and during training these weights were adjusted to get the optimal output.

To test the accuracy of the classified text, the testing data set of three types, on the basis if difficulty level, are selected: plain, average, and compound. Plain data set is very simple without any phrasal and idiomatic constraints. Average data set is containing simple sentences but with conjunctions, interjections, etc. Compound data sets are reasonably complex than other two categories due to inclusion of phrases and even idioms. 10 sentences of each category were tested and following results were received.

	Simple	Average	Compound	Average
Subject	98%	96%	89%	94 %
Verb	97%	94%	86%	92 %
Object	95%	93%	82%	90 %
Adverb	92%	89%	78%	86 %

TABLEI. STATISTICS SHOWING RESULTS

We interpret these results as encouraging evidence for the usefulness of Markov logic for judging substitutability in context. Overall accuracy for a complete sentence becomes 90.5%. Following graph shows the results.

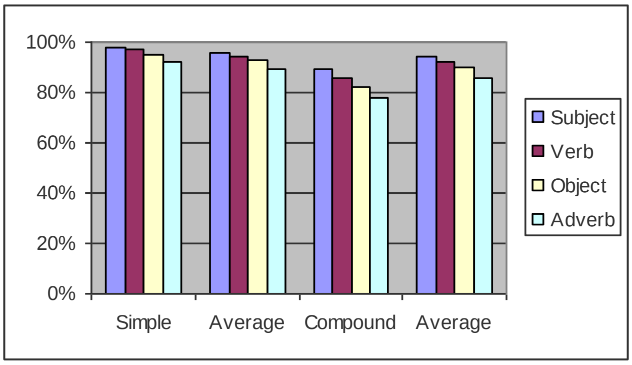

Fig.3 Graph showing results

V. 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.

ACKNOWLEDGMENT

This research is part of the research work funded by HEC, Pakistan and was conducted in the department of Computer Science & IT, The Islamia University of Bahawalpur, Pakistan. I want to thank Prof. Dr. Irfan Hyder (PAF-KIET) and Prof. Dr. Abass Choudhary (CEME-NUST) for their support and guidance during this research.

REFERENCES

[1] Katrin Erk, Sebastian Pad'o (2008) "A Structured Vector Space Model for Word Meaning in Context", In Proceedings of EMNLP (2008)

[2] Imran Sarwar Bajwa, M. Abbas Choudhary, (2006) "A Rule Based System for Speech Language Context Understanding" Journal of Donghua University, (English Edition) 23 (6), pp. 39-42.

[3] J. Henderson, (2003) "Neural network probability estimation for broad coverage parsing," in Proc. of 10th conference of the European Chapter of the Association for Computational Linguistics (EACL 2003), pp. 131–138.

[4] S. Kok and P. Domingos, "Learning the structure of Markov Logic Networks" In Proc. ICML-05, Bonn, Germany, 2005. ACM Press, pp. 441–448.

[5] S. Kok, P. Singla, M. Richardson, and P. Domingos "The Alchemy system for statistical relational AI", Technical report, Department of Computer Science and Engineering, WA, 2005. http://www.-cs.washington.edu/ai/alchemy .

[6] Pedro Domingos, Stanley Kok, Hoifung Poon, Matthew Richardson, Parag Singla, "Unifying Logical and Statistical AI", Proceedings of the Twenty-First National Conference on Artificial Intelligence (2006), pp. 2-7

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

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

[9] Chow, C., & Liu, C. (1968) "Approximating Discrete Probability Distributions with Dependence Trees". IEEE Transactions on Information Theory, 1968, IT-14(3), pp. 462–467.

[10] 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.

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

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

[13] 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, pp. 362-373

[14] Nerbonne, John (2003): "Natural language processing in computer-assisted language learning". In: Mitkov, R. (ed.): The Oxford Handbook of Computational Linguistics. Oxford, pp. 670-698.

[15] D. McCarthy, R. Navigli. (2007), "SemEval-2007 Task 10: English Lexical Substitution Task" In Proceedings of SemEval, pp. 48–53.

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

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

[18] WangBin, LiuZhijing, (2003), "Web Mining Research", Proceedings of Fifth International Conference on Computational Intelligence and Multimedia Applications, 2003. CHINA. pp. 84 - 89