ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE
International Journal of Advanced Research in Computer and Communication Engineering
ICRITCSA
M S Ramaiah Institute of Technology, Bangalore
Vol. 5, Special Issue 2, October 2016

Techniques of Semantic Analysis for Natural Language Processing – A Detailed Survey

Rajani S[1], M. Hanumanthappa[2]

Computer Science and Applications, Bangalore University, Bangalore, India [1, 2]

Abstract: Semantic analysis is an important part of natural language processing system. It determines the meaning of given sentence and represents that meaning in an appropriate form. Semantics, as a part of linguistics, aims to study the meaning in language. The language demonstrates a meaningful message because of the semantic interaction with the different linguistic levels. In this paper, survey is done on semantic analysis and explores different works that have been done in semantic analysis by different researchers. Few research papers have been considered for the analysis. In the examination, two important research fields are noticed, one of the popular statistical model called as LSA model and another active research area called as ontology which represents a set of primitives of domain of knowledge. In the analysis, it is noted that, LSA is used in automated evaluation against human evaluation and also used for extracting semantic information from textual information. Ontology technique is used to extract structure information from unstructured data, retrieving the information from database and in the semantic web applications.

Keywords: NLP, Semantics, LSA, spring graph, Ontology, NLIDB, SW, SVD

I. INTRODUCTION

The arena of natural language processing (NLP) is computer science, artificial intelligence, and linguistics. NLP mainly focuses on the interactions between computers and human languages or natural languages. Natural Language Processing is used to make computers recognize the statements or words written in human languages [1]. "Natural language processing" make the systems to process sentences in a natural language such as English, rather than in a computer language such as C, C++, and JAVA [2]. Natural language processing systems is a base of linguistic study and used in highly developed semantic representations. NLP consists of following four steps, 1. Morphological processing and Lexical Analysis 2. Syntax analysis (parsing) 3. Semantic analysis 4. Pragmatic analysis. Morphological processing is used to break strings of language into sets of tokens, corresponding to distinct words, sub-words and punctuation forms.

The tokens are classified according to their use (grammatical class). Morphology is recognizing how base words have been changed to form other words with alike meanings. Modification typically occurs, a new word can be formed by adding a prefix or a suffix to a base word.

For example in + active = inactive. Lexical analysis is dividing the whole chunk of text into paragraphs, sentences, and words [3]. The purpose of syntax analysis is to check whether string of words or a sentence is well-formed and to break it up into a structure that shows the syntactic relationships between the different words. A syntactic analyzer or parser uses the lexicon, which contains the syntactic category of each word.

A simple grammar describes, how syntactic groups can be combined to form phrases of different types [3]. Syntactic analysis can be used in punctuation corrector, dialogue systems with a natural language interface, or as a building block in a machine translation system [4]. For example “the large tiger chased the deer”. Following figure shows the simple syntactic tree structure for the above sentence.

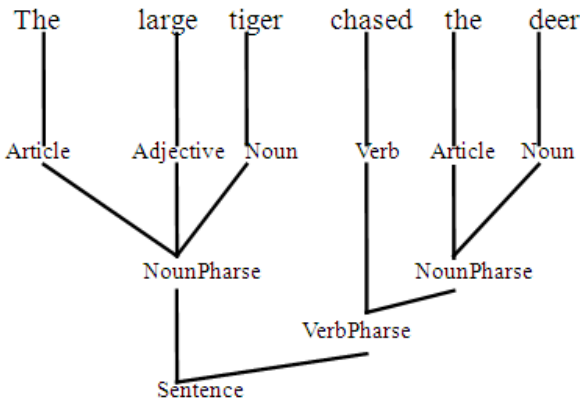

Fig 1: Syntactic tree structure

Semantic analysis gives the exact meaning or the dictionary meaning from structures created by syntactic analysis. Semantics, as a branch of linguistics, aims to study the meaning in language. Language demonstrate a meaningful message because of the semantic interaction with the different linguistic levels, phonology (phones), lexicon and syntax. Semantic analysis deals with the meaning of words and sentences, these words and sentences refers to the elements in the world [5]. The main

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

ICRITCSA

M S Ramaiah Institute of Technology, Bangalore

Vol. 5, Special Issue 2, October 2016

objective of semantic analysis is to minimize the syntactic structures and provide the meaning, finding synonyms, word sense disambiguation, translating from one natural language to another, and populating base of knowledge [6]. For example "colorless red idea". This would be rejected by the analyser as colorless red do not make any sense together. Sentences cannot be fully disambiguated during the syntax and semantic analysis phases but it can be disambiguate during pragmatic analysis. Pragmatic analysis understands the results of semantic analysis from the viewpoint of a specific context [3]. For example "do it fast" should have been interpreted as a request rather than an order.

II. LITERATURE SURVEY

1 Peter W. Folitz proposed a paper for text comprehension of the semantic similarity between pieces of textual information using LSA. In text-comprehension research, subjects is read from textual information and provide some form of summary. To study text comprehension cognitive model has been used. In this model, semantic information from original text and the reader's summary represent a sets of semantic components called propositions. Performing propositional analysis on text and subject's provides information contained in the text and representation of the subject's memory of the text. This summary documents the researcher to identify what information the subject has extracted from the text. For analysis, researchers examine each sentence in the subject's summary and match the information contained in the original sentence in the text. It's not easy to matching information from summary to original text.

To compare a text, LSA used information-retrieval methods. To analyze a text, LSA first generate a matrix of occurrence of each words in sentence or paragraph. Latter LSA uses the SVD (Singular Value Decomposition) technique. SVD decomposes the word–by-document matrix into a set of k. LSA representing documents and terms in k orthogonal indexing dimensions. Here experimenter made the match on the basis of the semantic content of the text. Author sum up results from three experiments to demonstrate applications of LSA for text comprehension. In first experiment authors have taken example of text containing 6,097 words and written summary from original text. The purpose of this experiment is to match each individual sentences from the subject's summary against original texts. Summary highly semantically similar to those in the original texts. Second experiment describe how semantically similar and for rating how much relevant information is cited in the essay on the basis of cosine between the vectors of the two texts. The last experiment describes, how to measure the coherence and comprehensibility of texts. LSA is well-matched for researchers in education who is interest to find out semantic similarity between textual resources. LSA is automatic and fast quick measurements to find out semantic similarity in textual information [7].

2 Joao Carlos Alves dos Santos and Eloi Luiz Favero present a paper on application of latent semantic analysis (LSA) for automatic evaluation of short answers to open ended questions. The author explained, how automatic evaluation produces more accuracy rate by using LSA than human evaluation. For evaluation authors considered entrance examination from Federal University of Para. Automatic evaluation system is nothing but a computational technology. This paper define how computers measure students learning degree and rate written answers. Automatic evaluation uses n-grams approaches. The n-grams typically are collected from a text or speech corpus. An n-gram of size 1 is referred to as a "unigram", size 2 is a "bigram" (or, less commonly, a "digram"), and size 3 is a "trigram". In order to compare accuracy between human evaluator scores and LSA scores, following six steps are considered: (1) preprocessing, (2) weighing, (3) singular value decomposition (SVD), (4) rating, (5) adjustments, and (6) accuracy. 1. Preprocessing: Making of the initial matrix: counts the unigrams and bigrams in each answer, 2. Weighing of the entries: a weight function expresses the importance of words in each answer, 3. SVD: (a) SVD calculation: the initial matrix is broken down into a product of three other matrices. (b) Reduction to semantic space: we empirically choose the dimension of semantic space, 4. Rating: each answer is compared to the reference answer, 5. Adjustments: (a) Penalty factor: based on the mean value and standard deviation of number of words per answer. (b) Re-rating: after applying the penalty factor, each answer is again compared to the reference. 6. Accuracy: (a) Error calculation: calculates the arithmetic mean of errors in each comparison. (b)

$$\text{Accuracy} = \frac{6 - \text{Error mean}}{6} * 100$$

Above steps will be repeated several times by changing parameters and keep the best result. Authors found 79% of similar answer is present in LSA model against human evaluators. From the experiment 84.94% of accuracy index found from LSA and 84.93% accuracy index from human evaluators [8].

3 Large volume of information spread across the web becomes useless if we are unable to locate, and without extracting correct piece of information from it. Many challenges are going, how to extract useful information from unstructured data and build semantic structured data. Harish Jadhao, Dr. Jagannath Aghav, Anil Vegiraju were used semantic tools for extracting information form unstructured data and they present it to the user through spring graph, is a visualization mechanism. Unstructured

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

ICRITCSA

M S Ramaiah Institute of Technology, Bangalore

Vol. 5, Special Issue 2, October 2016

data are from different resources and in different formats. Information are gained from varied sources like news and magazine articles, audio and video content, and blog entries. It's very difficult for analyst to extract useful information from different resources. Structured data provide a set of entities within a domain and the associations between those entities. The term entity refers to any item, company, person, location, and organization. Based on domain ontology they extract relevant entities and represent this structure into a RDF (Resource Description Framework) graphs. Further data stored in RDF knowledge base and queried using SPARQL (Simple Protocol and RDF Query Language) query language. For extracting structured data from unstructured data, authors have used Ontology learning method. Ontology learning primarily focused on defining the concepts and associations between them. It extracting domain terms, concepts, individuals, concept attributes and relations from textual data. To represent ontology extraction in a graphical format, Graph visualization has been used, called as spring graph [9]. Another use of spring graph is used in information extraction from a digital library after semantic analysis [11]. Structuring analysis is performed by removing scripting code and HTML comments. This model present noise free articles. This paper provide the extraction of significant and relevant results from a well-structured ontology file and also demonstrated a solution in spring graph [9].

4 To organize information, the fields of artificial intelligence, the Semantic Web, systems engineering, software engineering, biomedical informatics, library science, enterprise bookmarking, and information architecture all create ontologies. Ontology is a types, properties, and interrelationships of the entities. Authors Avinash J. Agrawal and Dr. O. G. Kakde used domain ontology for semantic analysis of natural language queries for Natural Language Interface to Database (NLIDB). Following applications like railway inquiry, airway inquiry, resort inquiry, bank, corporate or government call centers needs implementation of NLIDB. In this paper authors have taken railway inquiry application for semantic analysis. Mainly NLIDB is used for taking structured information from database and information hunter uses natural language for submitting his or her query to database.

Applications like railway inquiry, airway inquiry, bank, corporate etc. require a detail analysis of input query for database. Detail analysis input is represented in intermediate form, in order to generate database query for target database. This intermediate form is created by the method called as domain ontology. In order to improve computerized text processing, ontological semantic provides the language independent, meaning based representations of concepts. Unlike words in a language, each ontological concept is unambiguous. Three types of concepts are Object, Events, and properties. Authors had consider railway domain like Train, Station, Seats, Fare and Concession. These concepts are described with the associated property set. Property links from one concept to another and defines relations between concepts. For example concept train is related to concept station by from. In database query there are two important constituents i.e. what is expected and what are constraints. For example query: "list superfast trains from Mysore to Mangalore". Expected entity is: superfast train, Constraints: 1) Train [from] = Mysore 2) Train [to] = Mangalore 3) Train [type] =Superfast {additional constraint}. These constraints mapped into the standard concepts and properties in the domain ontology. In natural language interface to database, intension is to determine the meaning of input question from the viewpoint of database concepts. Finally a database query generated from the question which will essentially bring the preferred information from the database [12].

5 Authors Kouji Kozaki, Yusuke Hayashi, Munehiko Sasajima, Shinya Tarumi, and Riichiro Mizoguchi had explained review about semantic web application published in the semantic web conferences (ISWC, ESWC, and ASWC). Semantic web (SW) was proposed by Tim Berners-Lee about ten years ago. In W3C's vision web of linked data refers to semantic web. Semantic web technologies allow people to stores data on the web, build dictionaries, and write guidelines for handling data. Web of linked data are empowered by technologies such as RDF, SPARQL, OWL (Web Ontology Language), and SKOS (Simple Knowledge Organization System). RDF and OWL are the basic technologies of SW. Important features of SW include allowing computers to process semantics on various resources of WWW. This can be defined by ontology [13]. In this paper authors focus on what type of ontologies used and usage of ontology in SW. Following are the types of usage of ontology used in analysis: 1) Common Vocabulary, 2) Search, 3) Index, 4) Data Schema, 5) Media for Knowledge Sharing, 6) Semantic Analysis, 7) Information Extraction, 8) Rule Set for Knowledge Models, 9) Systematizing Knowledge. Types of Ontology: 1) Simple Schema, 2) Hierarchies of is-a relationships among Concepts, 3) Relationships other than "is-a" is Included, 4) Axioms on semantics are Included, 5) Strong Axioms with Rule Descriptions are Included. Authors have depicted both positive and negative points of current SW applications from the viewpoint of usages and types of ontologies [13].

III. CONCLUSION

It is evident that in order to process any natural language semantics is necessary. Without the syntax and semantic analysis the machine translation result may be ambiguous. Semantics is a sub part of linguistics which focuses on the study of meaning. In this paper, survey done on semantic

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

ICRITCSA

M S Ramaiah Institute of Technology, Bangalore

Vol. 5, Special Issue 2, October 2016

analysis and its research area in different fields. From all the above discussion, survey had noticed that LSA is a very good approach for finding more accuracy rate from LSA model than human evaluation. Same approach was implemented in written examination and find out the better accuracy rate than human evaluation. Another application of LSA is to find out semantic similarities between pieces of textual information. Most analysis are currently performed on UNIX workstations. LSA analysis should be able to be performed on desktop machines as well.

LSA is suitable for researchers in education and psychology who must evaluate from textual material. Another active research area noticed in the survey, called as Ontology. It provide methods and models for extracting pertinent information from unstructured data. In future implementation, ontology can be used to extract structured data from unstructured data like satellite images, scientific data, and social media data. Using natural language, information seeker submit their query in various database like inquiry in airway, bank, government organization etc. In future, to write query for mobiles phones, domain ontology can port in hand held systems, which is useful and general usage. It examined from second [8] and fourth [12] papers, authors have taken the practical examples and provided the accurate result.

REFERENCES

[1] Abhimanyu Chopra, Abhinav Prashar, Chndresh sain, "Natural Language Processing," in INTERNATIONAL JOURNAL OF TECHNOLOGY ENHANCEMENTS AND EMERGING ENGINEERING RESEARCH, vol 1, issue 4 ISSN 2347-4289.

[2] http://www.mind.ilstu.edu/curriculum/protothinker/natural_language_processing.php

[3] https://www.scm.tees.ac.uk/isg/aia/nlp/NLP-overview.pdf

[4] Rashmi S, M Hanumanthappa, Regina L Suganthi, "Processing of natural Language Semantically- A Detailed Survey", IJERT, ISSN: 2278-0181, Vol. 2 Issue 12, December – 2013.

[5] Mallamma V. Redd, M Hanumanthappa, "Semantical and Syntactical Analysis of NLP", IJCSIT, Vol. 5 (3), 2014, 3236 – 3238.

[6] https://www.cs.helsinki.fi/u/myllymak/Teaching/2004/Fall/Seminar/poroshin.pdf.

[7] Peter W. Foltz, "Latent semantic analysis for text-based research", Behaviour Research Methods, Instruments, & computers 1996, 28 (2), 197-202.

[8] Joao Carlos Alves dos Santos and Eloi Luiz Favero, "Pratical use of a latent semantic analysis(LSA) model for automatic evaluation of written answers, Santos and Favero Journal of the Brazilian Computer Society (2015), 21:21, DOI 101.1186/s13173-015-0039-7, a SpringerOpen Journal.

[9] Harish Jadhao, Jagannath Aghav, Anil Vegiraju,"semantic tool for unstructured data", International Journal of Scientific & Engineering Research, Volume 3, Issue 8, August-2012, ISSN 2229-5518.

[10] Ingo Brub and Arne Frick, Universitat Karlsruhe, Fakultat fur Informatik, D-76128 Karlsruhe, Germany, Fast Interactive 3-D Graph Visualization.

[11] Junliang Zhang,University of North Carolina, Chapel Hill,Javed Mostafa,Himansu Tripathy,Laboratory for Applied Informatics Research Indiana University,Information Retrieval by Semantic Analysis and Visualization of the Concept Space of D-Lib Magazine,2010.

[12] Avinash J. Agrawal, O. G. Kakde, "Semantic Analysis of Natural Language Queries Using Domain Ontology for Information Access from Database", I.J. Intelligent Systems and Applications, 2013, 12.

[13] Kouji Kozaki, Yusuke Hayashi, Munehiko Sasajima, Shinya Tarumi, and Riichiro Mizoguchi, "Understanding Semantic Web Applications", J. Domingue and C. Anutariya (Eds.): ASWC 2008, LNCS 5367, pp. 524–539, 2008. © Springer-Verlag Berlin Heidelberg 2008.

[14] Alani, H., Kalfoglou, Y., O'Hara, K., Shadbolt, N.R.: Towards a Killer App for the Semantic Web. In: Gil, Y., Motta, E., Benjamins, V.R., Musen, M.A. (eds.) ISWC 2005. LNCS, vol. 3729, pp. 829–843. Springer, Heidelberg.

[15] Mäkelä, E., Hyvönen, E., Saarela, S.: Ontogator — a semantic view-based search engine service for web applications. In: Cruz, I., Decker, S., Allemang, D., Preist, C., Schwabe, D., Mika, P., Uschold, M., Aroyo, L.M. (eds.) ISWC 2006. LNCS, vol. 4273, pp. 847–860.

[16] Springer, Heidelberg (2006). Chen, H., et al.: Towards a semantic web of relational databases: A practical semantic toolkit and an in-use case from traditional chinese medicine. In: Cruz, I., Decker, S., Allemang, D., Preist, C., Schwabe, D., Mika, P., Uschold, M., Aroyo, L.M. (eds.) ISWC 2006.

[17] Samir Prado Daud, Prof. Dr. Carlos Henrique Costa Ribeiro," NLP –LEXICAL ANALYSISAPPLIED TO REQUIREMENTS", 9th Brazilian conference on Dynamics, control and their applications June 7-11, 2010.