Vandana Publications
IJEMR

Volume-7, Issue-3, May-June 2017
International Journal of Engineering and Management Research
Page Number: 789-791

Natural Language Processing with Semantic Measurement

Komalpreet Singh Gill[1], Dr. Ankit Jagga[2]
[1]Student, Asian Educational Institute, Patiala, Punjab, INDIA
[2]Assistant Professor, Asian Educational Institute, Patiala, Punjab, INDIA

ABSTRACT

Natural language is one of the most complicated structures a man has met with. It plays a fundamental role not only in human communication but even in human way of thinking and regarding the world. Therefore, it is extremely important to study it in all its respects. Much has been done in understanding its structure, especially the phonetic and syntactic aspects. Less, however, is understood its semantics. There are many linguistic systems, often based on set theory and logic, attempting to grasp (at least some phenomena) of the natural language. However, none them is fully acceptted and satisfactory in all respects.

***Keywords*--** accepted, language, relationships, concepts

I. INTRODUCTION

There are different levels of tasks in NLP, from speech processing to semantic interpretation and discourse processing. The goal of NLP is to be able to design algorithms to allow computers to "understand natural language in order to perform some task. Example tasks come in varying level of difficulty.

A number of studies have demonstrated links between linguistic knowledge and performance in math. Studies examining these links in first language speakers of English have traditionally relied on core relational analyses between linguistic knowledge tests and standardized math tests. For second language (L2) speakers, the majority of studies have compared math performance between proficient and non-proficient speakers of English. In this study, we take a novel approach and examine the linguistic features of student language while they are engaged in collaborative problem solving within an on-line math tutoring system.

In Natural Language Processing (NLP), semantic similarity plays an important role and one of the fundamental tasks for many NLP applications and its related areas. Semantic Textual Similarity (STS) can be defined by a metric over a set of documents with the idea is to finding the semantic similarity between them. Similarity between the documents is based on the direct and indirect relationships among them.

II. SEMANTIC SIMILARITY

The information processing systems require knowledge sources to computationally assess the similarity between concepts. Generally human information processing involves the process of concept matching. Category membership and similarity are two important aspects of concept matching. In the context of ontologies, the vital use of similarity measurement is to determine how one concept of ontology is similar to concept of ontology. Based on the various notions of quantifying similarity, the semantic similarity could use either the path distance between concepts or the information content of a concept as a quantifying measure. In certain contexts, the combination of both the path distance and information content based approaches has been tried out. In this direction, whatever may be the quantifying factor, the knowledge sources are mandatory to compute similarity. These semantic similarity measures are useful mechanisms in Information retrieval systems, natural language processing systems and ontology mapping systems. The task of developing a semantic similarity measure which completely agrees with human assessment of similarity is extremely difficult.

2.1 Path Length based Measures

The similarity measurement among concepts is based on the path distance separating the concepts. In this method, the quantification of similarity is based on the taxonomy or ontology structure. In these taxonomical or ontology structure, it is assumed that the predominant relations that connect different concepts is only *is-a* type relations. These measures compute similarity in terms of the shortest path between the target synsets (group of synonyms) in the taxonomy. The degree of similarity is determined on the basis of the path, and generally will correspond inversely with the path length.

2.2 Information Content Approach

Information Content (IC) approaches [] are used to determine the semantic similarity between concepts. In this method, each of the concept or node poses IS-A taxonomy are kept in one set called C and all of these nodes carry unique concepts. Intuitively, one key to the similarity of two concepts is that to which they share information in common. In taxonomy direct relation between two concepts can be found by an edge counting method. In this method, if the minimal path between two nodes is long, that means it is necessary to go high in the hierarchy to find a least upper bound.

2.3 Hybrid Approach

Node and edge based methods discussed in previous sections have many differences in between them. The edge-based methods, looks true without any concise reasoning and on the other hand, node-based approach looks more accurate than distance-based. The distance measure was relayed on the subjective knowledge of the network while the WordNet was used not for measuring the similarity, but for the construction of the network layers

III. LITERATURE REVIEW

Literature survey plays an imperative role in our research work. It is the documentation of a comprehensive review of particular theme, which holds the information of past and present development of the topic. Thus it motivates to develop innovative techniques and models. This work describes the work of eminent researchers and highlights the challenges, which still require to be addressed.

Sergio Ceron-Figueroa et.al [1] describes the model for ontology matching over two educative content repositories. The basic idea behind this work is to automatically improve the efficiency of homogeneity resources for e-learning.

Rana and Singh et.al [2] have proposed a Semantic Web Mining interface, which is competent to handle heterogeneity issue and provide meaningful information in non-redundant way. Their work focused on finding the most ambiguous words and finding the relatedness measure with other important keywords in the query. It also considered removing redundancy among keywords being matched, but performs little effort to enhance the social web and e-commerce activities

Tim Berners-Lee et.al [3] in their visionary article laid down the foundation of Semantic Web. They gave a new direction for the information oriented WWW to be knowledge oriented in future. Their work enlightened the powerful role of agents and ontologies in semantic web

Mathieu et.al [4] have proposed a semantic web search engine called Watson, which provides various functionalities to discover and locate ontologies and semantic data online. It provides new possibilities in terms of enlarging semantic applications used by the content of the semantic web by exploiting a with set of API tool consisting of high level elements for searching, exploring and retrieving semantic data from web.

Heidari et.al [5] describes the need for combining the semantic web and e-commerce and evaluated the benefit of this combination. Their work describes the importance of semantic web technique and emphasized that the technology has the ability to extremely persuade the prospect improvement of the internet nation. They proposed innovative approach that support conventional web development and e-commerce to join the semantic web resulting.

IV. PROBLEM DEFINITION

A major issue on the Semantic Web Mining is inability to predict the user web surfing behavior appropriately on web and gain new knowledge through this interaction. Recently the web mining communities have focused on classifying standards that evaluating user browsing behaviors on e-commerce websites but failed to enhance user's satisfaction towards ambiguous queries from different perspectives. Ambiguous queries arise whenever two queries share the similar kind of information and numerous intended meanings are related with the same word leading to semantic ambiguity, which leads by the semantic similarity problem of ontology context. The semantic similarity problem arises when those contexts are concepts are parallel in meaning but have dissimilar names (Truck and Lorry). If we can automatically identify the semantic similarity between ontologies contexts, it is probable to apply appropriate tools to handle particular kind of queries, instead of for all.

V. PROPOSED MODEL: OPINION BASED ANNOTATION (OBA)

Our system comprises of four modules namely query phase (QA), analysis phase (AP), domain analysis (DA)

- **Query Phase (QA):** a set of key words is given as an input to QA, which describes the user information needs. In contrast to existing search engines that retrieves only the results on the basis of the probable search while ignoring the semantics of the user requirements, QA uniquely contributes a different and novel algorithm that focus on finding the relevant meaning that describes the user's desires. The algorithm tries to find the user behavior and prevents from the repetition of accessed data. Thus QA is an intelligent module as it improves the probability of success by finding the appropriate results. It performs several tasks to achieve the preprocessing phase: Tokenization and part of speech.

Tokenization is the process of splitting up a query string into a set of tokens or words. It usually splits words by blank, punctuation and quotation marks at both sides of a sentence. The tokens not only considered as words but also numbers, punctuation marks, parentheses and quotation marks. Parsing is the process of analyzing a string of symbols, either in natural language or in computer languages, conforming to the rules of a formal grammar.

- **Analysis phase (AP):** the second module of (OBA) is AP, which contains the Log Files, cookies and session identification. A log file is a file that records either events that occur in an operating system or other software runs, or messages between different users of communication software. Logging is the act of keeping a log. For Web searching, a transaction log is an electronic record of interactions that have occurred during a searching episode between a Web search engine and users searching for information on that Web search engine. More recent entries are typically appended to the end of the file. Information about the request, including client IP address, request date/time, page requested, HTTP code, bytes served, user agent, and referrer are typically added. This data can be combined into a single file, or separated into distinct logs, such as an access log, error log, or referrer log. However, server logs typically do not collect user-specific information.

- **Domain Analysis Module (DAP):** The data after preprocessing is transferred to the log files where the IP address of the user is collected and the data accessed by the user is determined from the database. The match between the users input query and the accessed query by the semantic similarity using the Information Content(IC) factor and the W-path algorithms. If the similarity found between the input and the cookies is similar then the user receives the result directly from the previous search. There is no need of accessing again and again the same inputs. The data of the cookies is stored in the database of the log files. So fetching the data from the database if the probability of the similar data is high then the desired result is given to the user, taking the data from the logfiles database. If the similarity between two contexts is found then move to ontology matching module.

VI. CONCLUSION

The most basic concept in NLP is that of a "bag-of-words", also called a frequency distribution. It's a way to turn a piece of free text (a tweet, a Word document, or whatever else) into a numerical vector that we can plug into a machine learning algorithm. The first part of any NLP process is simply breaking a piece of text into its constituent parts (usually words). This process is called "tokenization", and it is complicated by issues such as punctuation markers, contractions, and a host of other things. The chapter is about extensions and refinements of the basic idea of bag-of-words, and discusses some of the more intricate (and error-prone) sentence parsing toward the end.

REFERENCES

[1] Sergio Ceron-Figueroa (2017) Instance-based ontology matching for e-learning material using an associative pattern classifier, International Journal of Computational Analysis, Vol 45, N0.5, Elesvier, pp (218-255).

[2] Vijay Rana, Singh G (2014) “Analysis of Web Mining Technology and Their Impact on Semantic Web”, DOI: 10.1109/CIPECH.2014.7019035, IEEE Xplore, pp (5-11).

[3] Berners T, Hendler J and Lassila O (2001), “The Semantic Web”, Scientific American: Feature Article, Vol.284, No5. pp (1-4).

[4] Mathieu Aquin and Enrico Motta (2011), Watson. More than a Semantic Web search engine, INRIA, 2011-ACM Digital Library, pp (55-63).

[5] Karim Heidari (2009), the Impact of Semantic Web on E-Commerce, World Academy of Science, Engineering and Technology”, Vol 25, pp (303-306).

[6] Franzoni (2017), Just an Update on PMING Distance for Web-based Semantic Similarity in Artificial Intelligence and Data Mining, vol [1],pp (39-44).

[7] Samar Fathy, Nahla El-Haggar and Mohamed H. Haggag (2017), A Hybrid Model for Emotion Detection from Text, International Journal of Information Retrieval Research, Vol 7, N0.1, pp (31-37).

[8] Cooley, r., tan, p-n., and Srivastava (1999), WebSIFT: The web site information filters system. In Proceedings of the Web Usage Analysis and User Profiling Workshop (WEBKDD’99), Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp (1-18).

[9] Lee et.al (1993) Joon Ho Lee, Myoung Ho Kim, and Yoon Joon Lee. Information retrieval based on conceptual distance in IS-A hierarchies. Journal of Documentation, vol 49(2), pp (188–207).

[10] R. Cooley, B. Mobasher, and J. Srivastava (1997). Grouping web page references into transactions for mining World Wide Web browsing patterns. IBM , pp (115,127).

[11] Wu, K., Yu, P. S., & Ballman (1998), Speed- tracer: A web usage mining and analysis tool. IBM Systems Journal, Vol.37, No.1, pp (19-27).

[12] Wenpu Xing and Ali Ghorbani (2004) Weighted PageRank Algorithm: Proceedings of the Second Annual Conference on Communication Networks and Services Research CNSR’04© 2004 IEEE .

[13] Vijay Rana, Dhir V, (2016), WIRS: Wisdom Information Retrieval System, International Journal of Computing Academic Research (IJCAR), Vol .5, No. 2, pp 122-132.

[14] Vijay Rana, Vijay Dhir, (2016), ICS: A Knowledge Based Technique to Identifying Ambiguous Queries, International Journal of Technology and Engineering (IJTE), Vol 1, No.3, pp 124-128.

[15] Vijay Rana, Singh G, (2015) “MBSOM: Meaning Based Semantic Ontology Matching Technique”, INSUSH ERA-2015, DOI: 978-1-4799-8432-9, IEEE Xplore, pp 14-19.