SEMANTIC QUANTIFICATION OF TEXT: A SURVEY

[1]Shruthi C. J, [2]Mouneshachari .S

[1,2] Department of CSE, GSSSIETW (VTU), Mysore, (India)

ABSTRACT

Measuring semantic role in the text document is very essential for extracting meaningful information because text document is a source of required information So it leads to the acquisition of valuable knowledge. When the amount of information will be growing tremendously in the real world it makes the necessity of quantifying the text document with respect to semantics then we can get the relevant information in faster way. The quantification of text document will be the most required technology for growing set of data so that we can make the machine to understand the things in meaningful way and it leads to effective interaction or communication between machines and human beings and this semantic role measuring technique has very much attracted great concern in the field of Artificial Intelligence, Psychology, Text mining and Text classification. In this paper we are going to review some of the methods that have been proposed for measuring the semantic role in the text documents.

Keywords: Semantic Role, Quantification, Information Extraction, Machine Learning.

I. INTRODUCTION

The communication between human beings was done very straightly by using natural languages because we have the ability for understanding the languages or thing with respect to different context, situation, but when we communicate with the machine it is not in the position to understand the languages with respect to semantics so, measuring the semantic role with in the text is very essential.

Semantic quantification of text document plays a central role in information extraction or in information management with respect to different context of the text. The text will be quantified based on the relationship between the words that is different feature or properties, synonyms or thesaurus, taxonomy or ontology, terms, frequency, co-occurrence, order of the words.

The amount of information will be growing tremendously in the society and it makes the necessity of quantifying the text for getting most relevant information according to the meaning of the text. The potential application for quantifying the text document includes, knowledge discovery for the decision making systems by constructing the knowledgebase, efficient text summarization, text classification and semantic search engines, semantic similarity.

Yuhua Li have discussed the text can be measured based on the semantic nets and corpus statistics[1], and semantics role is measured in the sentence by using the depth and path length between the words in the semantic nets of the words by constructing the database of the word using synonyms sets and using these two measures best semantic similarity of the words found and along with semantic net the index values are assigned for each word in the sentence for finding the semantics of word and this index value is simply the order number in which the word appears in the sentence and the similarity is calculated by using the cosine similarity measure and finally overall semantic similarity between the two sentence is calculated. Semantic similarity can also be calculated for selected pair of sentences by constructing the semantic vector for two sentences in this vector if two words are similar with respect to meaning then both the word assigned with same value otherwise the value is assigned based on the distance between the words.

Andreia Dal Ponte Novelli and Jose Maria Parente de Oliveira[2] have reported that, the semantic of the text will be calculated by using the vectors of the terms which is present in the text document, this terms are extracted with respect to syntactic structure of the sentence after that they have measure the semantic role of each sentence, finally the total text similarity will be calculated. In this paper, Section 2 presents Related work on this topic, Section 3 discussion regarding most frequently used measures or some important methods, Section 4 Conclude the paper.

II. RELATED WORK

Some of the works reported in the literature that focused on text mining with respect to semantics and semantic role quantification of text documents. However, some of methods available in the literature are reviewed in this Section.

Shaidah Jusoh [3] have discussed Text documents is source of information, this relevant information can be extracted by using the semantics of the text so, the text is processed and segmented into sentences by syntactically for recognizing the part-of-speech which is present in the sentence and that word is considered as entity after that this word is processed with respect to semantics but while processing the text for semantics system would face ambiguity to resolve this problem subject context knowledge should be considered and according to context of word different meaning of the word is stored in the database and value for different context word is calculated by using fuzzy membership function along with subject of the sentence and the most relevant preceding sentences for resolving the ambiguity is preserved for extracting the meaningful information .

Brandon Beamer[4] have reported information regarding the extracting Semantic relation between the words automatically by quantifying the noun features from the wordnet's IS-A backbone and it separates the positive and negative sentences based on the boundary value by using the SemScat learning model after they will find the noun- noun semantics for the better semantic relation extraction of the text .

John A. Bullinaria[5] have discussed about selecting the best method for extracting semantic representation from simple word co-occurrence statistics in large text corpora, this can be achieved by three factors that is functional word stop-lists, word stemming, and dimensionality reduction using Singular Value Decomposition (SVD) by using this factors significant semantic vector will be formed for the better semantic representation of text, after that this semantic vector is tested with different tasks i.e, TOEFL, Distance Comparison, Semantic Categorization, Clustering Purity for finding the similarity of work done in the previous methods based on the performance baselines.

Peter D. Turney[6] have reported Computer systems can't understand the meaning of human language, this limitation was addressed by the survey of vector space model of term–document, word–context, and pair–pattern matrices so using this matrices better semantic measures of the text will be achieved for making the systems to understand the natural languages like human beings so, we can effectively communicate with the machines.

Dingding Wang, Tao Li[7] have presented report on Multi-document summarization that can be done by sentence-level semantic analysis and symmetric non-negative matrix factorization methods of text documents, in first method sentences are extracted by using machine learning methods and the semantic similarity matrix is constructed after that semantic role and pair wise semantic similarity is calculated but some times it was very difficult to find the similarity by using rectangular matrix so, this can be resolved by using second method that is symmetric non-negative matrix factorization for sentence-sentence similarity measure.

Mehmet Ali Salahli[8] have given some idea about Semantic relatedness that can be achieved by measuring the semantic relatedness between the words via related terms, in this approach pair of sets of words is considered for which words we want to find the similarity and relatedness is calculated by computing the normalized values for that words and this relation are not calculated directly instead that can be computed by using the synonyms of that words.

S. Anitha Elavarsi[9] have reported survey on semantic similarity measures for text processing in different way. The documents can be classified based on the single ontology and cross ontology similarity measure along with the basic methods used for the semantic measure will be discussed so, it will be useful for the finding or considering the best method for the semantic similarity measure.

Eugene Santos Jr[10] have discussed quantifying the semantics in uncertainty is very difficult and also for constructing the knowledge base so, this can be resolved by using the bayesian knowledge-base because it can handle the uncertain data very effectively and this model assign the numerical values for each conditional probability rule implicitly corresponding to conditional probabilities in target probability distribution construction without considering the explicit semantics assumption.

Giannis Varelas[11] have discussed regarding the semantic similarity can be calculated by mapping the terms into ontology and also relationship between the term and ontology measured. Different semantic similarity methods also reviewed by proposing new method that is semantic similarity retrieval model this model analyze the documents and construct the terms vector and in this term vector each term is assigned with its weights based on the frequency of occurrence of terms in the documents and the similarity between two documents is calculated by using cosine similarity measure and it will be applied for the semantic information extraction from the web documents.

III. METHODS

This section reviews the most relevant methods for measuring the semantics of the text documents.

3.1 BAGS OF WORDS or WORD CO-OCCURRENCE

The text documents contains number of frequently occurring words, these words are grouped or clustered based on the frequency count of the words and each cluster value is different from other cluster so, calculating this clustering value we can get most semantically similar information or meaningful information from text documents so, this method is used commonly in information retrieval systems and in this method all meaningful words are collected so, the count of the words are in hundreds or thousands.

The semantic similarity between the word or sentence can be quantified based on the co-occurrence of the words by using pattern matching methods so, the meaning of the text is conveyed with limited set of patterns and also it requires complete set of meaningful words for avoiding ambiguity and also singular value decomposition(SVD) method used for the reduction of dimensionality of the count of the words.

Table 1. DATABASE OF BAGS OF WORDS

Word	Occurrences
Variation	1
Hard	1
Cluster	2
Observation	2
⋮	⋮
Centre	1

3.2 Word net semantic similarity

In wordnet semantic similarity method, the words in text documents are extracted and synonyms of the words are collected and that will be constructed in the form of hierarchical structure and the semantic similarity will be calculated by computing the path length and depth of the synonyms of the word and also it is more important in determining the semantic distance between the words and also it is one of the form of knowledge representation.

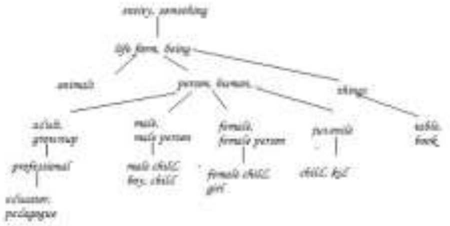

Fig.1 Word Net

Some of the function used for the computation of semantic similarity is,
The path length calculation,

$$\mathbf{S(W_1, W_2) = f(l, h)} \qquad (1)$$

And
Path length and Depth can be calculated,

$$\mathbf{S\ (W_1, W_2) = f_1(l).\ f_2(h)} \qquad (2)$$

The path length and the depth of the words can be calculated by using the figure like, educator- professional-adult-person length is 4 and the path length of male child- male- person is 3 so, based on this value we can compute the similarity.

3.3 Feature based method

In feature based method, some of the predefined text semantic features will be collected which is present in the sentence and these features are differentiated as primary and composite feature. The primary feature compare single units which is present in the text and composite feature compares the pairs of units in the text documents so, the text will be represented in the form of vector value of these feature then the similarity between texts can be calculated through this value.

In another way the terms, words, and characters will be considered as features of the text document and this features will be extracted from the text and based on the position or order of this feature the semantic role within the text is measured.

IV. Conclusion

The main objective of this paper is to highlights the basic methods of quantifying the semantic role in text documents as well as to provide review report carried out in this area. According to this methods we can get the better semantic relatedness of the text documents and also it will give the useful information about the strong methods used for the semantic quantification of text.

REFERENCES

[1] Yuhua Li, “Sentence Similarity Based On Semantic Nets and Corpus Statistics”, IEEE Trans. Knowledge and data engineering, vol.18, NO.8,AUGUST 2006.

[2] Andreia Dal Ponte Novelli, “A Method For Measuring Semantic Similarity Of Documents”, International Journal of Computer Applications(0975- 8887), vol 60- No.7, DECEMBER 2012.

[3] Shaidah Jusoh, “Semantic Extraction From Texts”, International Conference on Computer Engineering and Applications, IPCSIT vol.2(2011).

[4] Brandon Beamer, “Automatic Semantic Relation Extraction with Multiple Boundary Generation”, Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence (2008).

[5] John A. Bullinaria, “Extracting Semantic Representations from Word Co-occurrence Statistics: Stop-lists, Stemming and SVD”, To appear in Behavior Research Methods, 2012.

[6] Peter D. Turney, “From Frequency to Meaning: Vector Space Models of Semantics”, Journal of Artificial Intelligence Research 37 (2010) 141-188, Submitted 10/09; published 02/10.

[7] Dingding Wang, Tao Li, “Multi-Document Summarization via Sentence-Level Semantic Analysis and Symmetric Matrix Factorization”, SIGIR’08,July20–24,2008.

[8] Mehmet Ali Salahli, “AN APPROACH FOR MEASURING SEMANTIC RELATEDNESS BETWEEN WORDS VIA RELATED TERMS ”, Mathematical and Computational Applications, Vol. 14, No. 1, pp. 55-63, 2009 © Association for Scientific Research.

[9] S.Anitha Elavarasi, Dr. J. Akilandeswari, K. Menaga, “A Survey On Semantic Similarity Measure”, International Journal of Research in Advent Technology, vol.2, No.3, March 2014.

[10] Eugene Santos Jr, Eugene S. Santos, Solomon Eyal Shimony, “Semantic and Knowledge Acquisition in Bayesian Knowledge-Bases”, FLAIRS 2002.

[11] Giannis Varelas, “Semantic Similarity Methods in WordNet and their Application to Information Retrieval on the Web”, WIDM’05, November 5, 2005.