Scientific Journal Impact Factor (SJIF): 4.364

International Journal of Modern Trends in Engineering and Research

ISSN (ONLINE) : 2349 - 9745

ISSN (PRINT) : 2393 - 8161

MEASURES TO CALCULATE SEMANTIC SIMILARITY: A SURVEY

Tanuja Lohnari[1]

[1] *Assistant Professor, Department of Computer Engineering, DYPIEMR, Pune, India*

Abstract- Semantic similarity is a metric used to measure extent of similarity of meaning between two concepts. The concepts can be two words, sentences, or paragraphs. It tries to find the distance between two concepts in the semantic space, lesser the distance greater the similarity. Semantic similarity techniques compute similarity between two concepts which are lexicographically different. The methods used to find semantic similarity between two words can be extended to find similarity between sentences, phrases, or paragraphs. Finding semantic similarity between two concepts has many practical applications. It plays increasingly important role in the fields of Natural Language Processing (NLP), Information Retrieval (IR), Text Mining, etc. Text similarity techniques can be effectively employed for tasks such as text summarization [14], text classification [15], redundancy removal, document retrieval [16], question generation, question answering [17], etc. Effectiveness of these tasks can be immensely improved if semantic similarity measures are used to determine text similarity.

Keywords-Semantic similarity, Text similarity, Corpus-based Methods, Knowledge-based Methods, Semantic Relatedness

I. INTRODUCTION

Semantic similarity is a metric used to measure extent of similarity of meaning between two concepts. The concepts can be two words, sentences, or paragraphs. It tries to find the distance between two concepts in the semantic space, lesser the distance greater the similarity. Two concepts can be similar in two ways: lexically or semantically. Lexical structure of a word means the sequence of characters that form the word. Two words are lexically similar if they have the same sequence of characters. While semantics deal with meaning of terms. Two terms or concepts are said to be semantically similar if their meaning is similar even if their lexical structure is different (e.g. Synonyms). Semantic similarity techniques compute similarity between two concepts which are lexicographically different. The methods used to find semantic similarity between two words can be extended to find similarity between sentences, phrases, or paragraphs.

Finding semantic similarity between two concepts has many practical applications. It plays increasingly important role in the fields of Natural Language Processing (NLP), Information Retrieval (IR), Text Mining, etc. Text similarity techniques can be effectively employed for tasks such as text summarization [14], text classification [15], redundancy removal, document retrieval [16], question generation, question answering [17], etc. Most common application of semantic similarity measuring techniques can be found in search engines. Basic job of search engines it to search a huge repository of documents based on the search query entered by the user and get the documents relevant to the search term. The relevance models used in search engine such as Boolean models, vector space model, and probabilistic models [1] are majorly based on lexicographical similarity rather than semantic similarity. This greatly affects the relevance factor of retrieved information. The effectiveness of retrieval can be improved using semantic similarity methods during query expansion [2]. Authors Courtney Corley and Rada Mihalcea [8] have evaluated several techniques of semantic similarity calculation and have

DOI:10.21884/IJMTER.2017.4015.E381W

observed that accuracy of tasks such as paraphrase recognition increases significantly if semantic information is incorporated in text similarity measures.

The approaches being used to calculate semantic similarity can be classified as corpus-based models and knowledge-based models [3]. Corpus-based models try to determine degree of similarity between two terms based on the information obtained by analyzing a large collection of documents (corpus) where as knowledge-based models try to find how similar two terms are based on some sort of semantic network like an ontology [4].

II. CORPUS-BASED METHODS

Corpus is a large collection of written material, writings, conversations, and speeches that people use to study and describe a language [5]. There is a huge collection documents available on the Web that can be used to gather information which will help in determining the degree of semantic relatedness. A large number of corpus-based methods are available, few of which are discussed here.

Latent Semantic Analysis: It is one of the earliest methods of semantic similarity calculation proposed by Landauer [6]. It is based on the assumption that words having similar meaning will occur in similar texts. This model analyzes a large piece of text, calculates term frequencies, and creates a matrix where columns represent test samples, rows represent unique terms found in the text, and cells represent term frequencies per test sample. This matrix is subjected to a technique called singular value decomposition (SVD). SVD is a dimensionality reduction technique which reduces number of columns while preserving similarity information in rows. Similarity between two terms is then measured using cosine between vectors thus generated. This model is successful in removing some of the shortcomings such as sparseness and high dimensionality of traditional vector space model.

Semantic Similarity using Web Search Engines: This approach proposed by Danushka Bollegala, Yutaka Matsuo, and Mitsuru Ishizuka [7] tries to find co-occurrence value between two words using results of web search. The method uses information such as page count and snippets provided by web search engines for each search. To find similarity between two words **P** and **Q**, the method determines page counts for queries **P** and **Q** searched separately and also page count for the combined query **P AND Q**. The method then goes on to find numerous lexico-syntactic patterns of the two query terms and find the frequency of occurrence of these patterns in the snippets returned by the web search engine. The patterns are then ranked according to their ability to express similarity between two words. Similarity scores based on page counts are integrated with lexico-syntactic patterns using support vector machine (SVM) to calculate semantic similarity between given words.

Similarity Based on Corpus Statistics and Lexical Taxonomy: Authors Jay J. Jiang and David W. Conrath [12] have proposed a novel approach in which they have combined lexical taxonomy structure with statistical information derived from a corpus. The method is derived from edge-based approach of similarity calculation in which the information gathered from the corpus is used as a deciding factor. The proposed method thus enhances the normal edge-counting method by adding the information content of the node-based methods. The authors have claimed that their proposed model outperforms other computational models.

Explicit Semantic Analysis: This is an approach that tries calculates semantic relatedness between two arbitrary texts by augmenting word knowledge with text representation [9]. Meaning of a word is explicitly represented in the form of a weighted vector of concepts derived from Wikipedia which in itself is a huge repository of online articles. The two texts that are to be compared are converted into weighted vectors of concepts by employing machine learning. The vectors thus generated are then compared using some method such as cosine similarity to calculate semantic relatedness between texts.

III. KNOWLEDGE-BASED METHODS

Taxonomies, ontology, semantic nets are forms of knowledge representation used in Information Retrieval (IR). These represent entities, ideas, their properties and relationship between entities often in a hierarchical manner. These knowledge representations are used by many methods to find semantic similarity between two terms. Knowledge-based approached can be further classified as node-based approached and edge-based approaches. Node-based approaches use information content to evaluate similarity while edge-based approaches use conceptual distance.

One of the pioneering works done in this field was a model proposed by Wu and Palmer [10]. A conceptual domain can be represented by a taxonomy where concepts are arranged in a hierarchical manner. The similarity between two concepts represented within the same domain can be calculated as a measure of distance between two concepts as represented in the hierarchal structure. If *C1* and *C2* are two concepts whose similarity is to be calculated and *C3* is their least common subsumer concept then similarity between two concepts is expressed as

$$\boldsymbol{Sim_{WP}(C1, C2) = \frac{2 * N3}{(N1 + N2)}}$$

where *N1* is edge count on the path between *C1* and root, *N2* is edge count on the path between *C2* and root and *N3* is the edge count on the path between *C3* and root node. This is an edge counting based technique which assumes that all edges have equal weights.

In his research Resnik [11] observed that the problem with edge-counting based measures of semantic similarity is that all the edges in the taxonomy do not represent equal semantic distance between the terms that they are connecting. For example, in the WordNet taxonomy there is only one link between ***coin and dime*** and one link between ***treasure and gold*** but meaning of *dime* is more closely related with *coin* than meaning of *gold* with *treasure*. Resnik has observed in his paper that finding semantic similarity between two words is essentially finding the extent to which they share information in common. This can be done by finding the least subsumer. As we move up the hierarchy in the taxonomy we move from more specific concepts to more abstract concepts. So, we can safely assume that the lesser the distance between two concepts and their common subsumer, greater the semantic similarity between them. Thus, Resnik has proposed to quantify the information content shared by two concepts in order to measure similarity between them. More the shared information content, more similar the concepts are. Information content shared by two concepts is the information content of the concept that subsumes them. Taking the basis of information theory, the model calculates semantic similarity between concepts c1, and c2 using following formula

$$\boldsymbol{Sim(C1, C2) = MAX_{c \in S(C1,C2)}[-\log p(C)]}$$

where *S(C1, C2)* is the set of concepts that subsume both *C1* and *C2* and *–log p(C)* is quantified information content of concept *c*. This approach has proven to give better results as compared to earlier edge counting methods which are very close to results obtained by human judgment.

Rada et. al. have proposed a method based on the conceptual distances between terms measured using a taxonomy defining these terms [13]. Based on the assumption that shorter the distance between two nodes in taxonomy, greater will be semantic similarity between those nodes, the approach tries to find the shortest path between the two terms in the given taxonomy. Similarity between two terms is then calculated using following formula:

$$\boldsymbol{Sim(C1, C2) = 2MAX - L}$$

where *MAX* is the maximum path length between nodes *C1* and *C2* in the given hierarchy and *L* is the length of the shortest path between these nodes. This method enjoys the advantage of having low complexity while calculating the distance.

IV. CONCLUSION

Assessing semantic similarity has been a crucial component in many applications of Natural Language Processing, Information Retrieval, and Artificial Intelligence. This paper has reviewed methods of semantic similarity from both corpus-based as well as knowledge-based classes. Incorporating semantic similarity measures in retrieval methods give exceedingly better results as compared to results based on plain lexicographic matching. Still, a similarity matching technique which goes closer to human perception of similarity is yet to be invented and that remains as a challenge in this field.

REFERENCES

[1] Christopher C Yang, "Search engines information retrieval in practice", Wiley Online Library, 2010.

[2] Wessel Kraaij, Jian-Yun Nie, Michel Simard, "Embedding web-based statistical translation models in cross-language information retrieval", Computational Linguistics, MIT Press, vol 29, pp. 381—419, 2003.

[3] Wael H Gomaa, Aly A Fahmy, "A survey of text similarity approaches", International Journal of Computer Applications, Foundation of Computer Science, vol 68, 2013.

[4] Rada Mihalcea , Courtney Corley, Carlo Strapparava , "Corpus-based and knowledge-based measures of text semantic similarity", American Association for Artificial Intelligence, vol 6, pp. 775—780, 2006.

[5] https://www.merriam-webster.com/dictionary/corpus

[6] Thomas K Landauer, Susan T Dumais, "A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge", Psychological review, American Psychological Association, vol 104, 1997.

[7] Danushka Bollegala, Yutaka Matsuo, Mitsuru Ishizuka, "Measuring semantic similarity between words using web search engines", WWW, vol 7, pp. 757—766, 2007.

[8] Courtney Corley, Rada Mihalcea, "Measuring the semantic similarity of texts", Proceedings of the ACL workshop on empirical modeling of semantic equivalence and entailment, Association for Computational Linguistics, pp. 13—18, 2005.

[9] Evgeniy Gabrilovich, Shaul Markovitch, "Computing semantic relatedness using Wikipedia-based explicit semantic analysis", IJCAI, vol 7, pp. 1606—1611, 2007.

[10] Zhibiao Wu, Martha Palmer, "Verbs semantics and lexical selection", Proceedings of the 32nd annual meeting on Association for Computational Linguistics, Association for Computational Linguistics, 133—138, 1994.

[11] Philip Resnik, "Using information content to evaluate semantic similarity in a taxonomy", arXiv preprint cmp-lg/9511007, 1995.

[12] Jay J Jiang, David W Conrath, "Semantic similarity based on corpus statistics and lexical taxonomy", arXiv preprint cmp-lg/9709008, 1997.

[13] Roy Rada, Hafedh Mili, Ellen Bicknell, Maria Blettner, "Development and application of a metric on semantic nets", IEEE transactions on systems, man, and cybernetics, Vol 19, 17—30, 1989.

[14] Ramiz M Aliguliyev, "A new sentence similarity measure and sentence based extractive technique for automatic text summarization", Expert Systems with Applications, Elsevier, vol 36, pp. 7764—7772, 2009.

[15] Yung-Shen Lin, Jung-Yi Jiang, Shie-Jue Lee, "A similarity measure for text classification and clustering", IEEE transactions on knowledge and data engineering, vol 26, pp. 1575—1590, 2014.

[16] Wenhai Sun, Bing Wang, Ning Cao, Ming Li, Wenjing Lou, Y. T. Hou, Hui Li, "Verifiable privacy-preserving multi-keyword text search in the cloud supporting similarity-based ranking", IEEE Transactions on Parallel and Distributed Systems, vol 25, pp. 3025—3035, 2014.

[17] Dekang Lin, Patrick Pantel, "Discovery of inference rules for question-answering", Natural Language Engineering, Cambridge Univ Press, vol 7, pp. 343—360, 2001.