A Review on WordNet and Vector Space Analysis for Short-text Semantic Similarity

Anjali Kaundal
Department of Computer Science Engineering,
Chandigarh University,Gharuan, Punjab, India.

Arvinder kaur
Department of Computer Science Engineering,
Chandigarh University, Gharuan, Punjab India

Abstract- **Meaningful sentences are the combination of meaning words, if a system wants to process natural language it should have essential knowledge regarding words and their meanings. The assessment of semantic similarity between the words of a short text is one of the challenging task knowledge based tasks and the tasks in NLP like text summarization, information retrieval, search, categorization of text and machine learning etc. which uses the sentence similarity measures for assessing the similarity between the short-text or sentences. In this paper the survey of two techniques is done which are helpful in generating the extractive text summaries WordNet and Vector space analysis. In vector space model words can be represented as numeric vectors based on different semantic similarity measures, the similarity between the word numeric vectors can be calculated with the semantic measures called WordNet. The information regardin the word and teir meaning in earlier days was provided with the help of traditional dictionaries, but these dictionaries were only helpful for human readers not for machine, WordNet provide combination of traditional lexicographic information and modern information. It is a online lexical database designed for use under the program control, it uses the measure like for calculating the semantic similarity between the concepts. Nouns, verbs, adjectives, adverbs are organized into set of synonyms and semantic relationship between the synonym sets called as Synset.**

Keywords –NLP, Vector space Model, WordNet, Short-text Semantic similarity, Synsets.

I. INTRODUCTION

The sentence semantic similarity also known as short text semantic similarity (STSS) [1] and are used to measure the similarity between the text that are typically of length 10-20 words. Different measures for assessing the semantic similarity between the two concepts some of these are SRA [2], WordNet [1], Vector space analysis [3] etc. SRA defines as the approach in which elaboration of RDF graphs with entity role and sense identification is there. It takes as input the sequence of tokens, extracted in the lexical analysis and applies the Semantic Role Annotation to define the roles of each of the entities and to identify their meaning in the sentence. Semantic similarity in this kind of approach is benefitted from using FrameNet [4]. Evaluating the semantic similarity between the long documents is trending now as compare to the short text semantic similarity but these methods are not suitable for the semantic similarity findings between the short text because they do not work well in case of short text. Vector space model is classically used for evaluating the semantic similarity between the two documents For example in short text semantic similarity word co-occurrence plays vital role but in case of long text or documents it is less significant. Vector space model works on three aspects word frequency (IDF), word weighting and similarity calculations. Sentence similarity methods should also be capable of measuring the degree of relatedness between short with partial information, as when one sentence is split into two or more short texts and phrases that contain two or more sentences. [5,6] represents the sentence as bag of word vectors or a tree of syntactic information among [7] or degree of sentence similarity is calculated by three level architecture [2]. It deals with the word specificity and word ordering.

1.1 Vector Space analysis

The ability to judge the similarity between the sentences is a critical part for measuring the similarity between the short-text. If these calculations are not done in efficient way then the output will be highly affected. Hence we consider two factors for calculating the semantic similarity between the short-text first is word specificity and semantic word similarity.

Word specificity refers to the discriminatory power of terms in a given context. It is important for the similarity computations. Vector space model works on three aspects word frequency (TF-IDF), word weighting and similarity calculations. Vector space model that classically uses the cosine similarity for calculating the similarity aspects.

$$cosine\ sim.(d1, d2) = Dot\ product\ \frac{d1,\ d2}{\|d1\| * \|d2\|} \ldots\ldots (1)$$

$$Dot\ product\ (d1,\ d2) = (d1[0] * d2[0] + d1[1] * d2[1] + d1[2] * d2[2] \ldots\ldots\ldots\ldots d1[n] * d2[n]) \ldots\ldots\ldots\ldots (2)$$

$$\|d1\| = squareroot\ of\ (d1[0] * d1[0] + d1[1] * d1[1] + d1[2] * d1[2] \ldots\ldots\ldots\ldots d1[n] * d1[n]) \ldots\ldots\ldots\ldots (3)$$

$$\|d2\| = squareroot\ of\ (d2[0] * d2[0] + d2[1] * d2[1] + d2[2] * d2[2] \ldots\ldots\ldots\ldots d2[n] * d2[n]) \ldots\ldots\ldots\ldots (4)$$

Vector space model (TF-IDF) or term vector model is used for representing the documents as vectors of identifiers such as index terms. IDF between the two sentences is calculated with the help following equation:

$$IDF(s, w) = \log(|Sa \cap Sb|) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3)$$

Sa = first sentence, Sb = second sentence, w = total no. of words present in both the sentences.

Bag-of-words model is built at sentence level with the usual weighted termed-frequency and inverse sentence frequency paradigm [8], in which the sentence frequency is the no. of sentences in the document which contains that term. These sentence vectors are then scored by similarity and highest scoring sentences are picked to be a part of summary in case of text summarization. This is a direct adaptation of information retrieval method for retrieving the similarity. This paradigm can be used for calculating the similarity between the words that are included in the short-text. From each document, drive the vector. The set of documents in a collection then is viewed as a set of vectors in a vector space. Vectors deals only with the numbers. In this example we are dealing with the text documents this was the reason why we used TF and IDF to convert text into numbers so that it can be represented by a vector, d1, d2, d3 are the sentences and q is the query. θ is the similarity between the short-text d1, d2, d3.

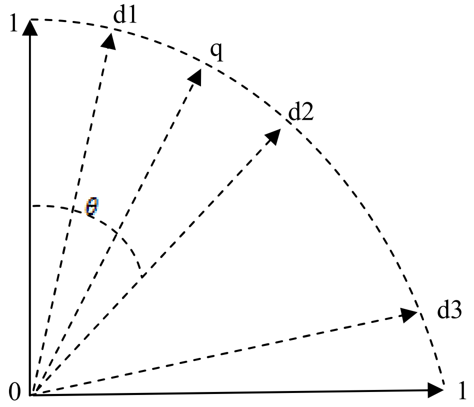
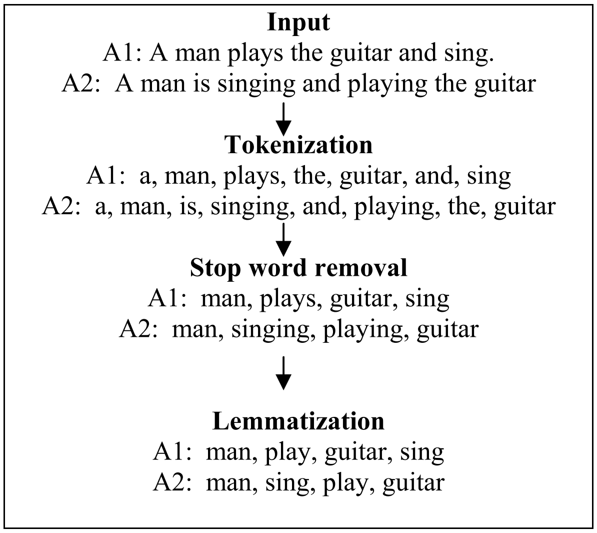


Figure1. Graphical representation of documents

1.2 WordNet semantic similarity

Like many other approaches that are being used for calculating the semantic similarity between two concepts. WordNet semantic similarity method can also be used for computing the semantic similarity between the two concepts. WordNet is defined as the Lexical database which measures the relatedness between the two concepts with the help of following mentioned measures. Eight measures are used to calculate the similarity between the concepts. The performance of the proposed measure is dependent on the basic unit that is similarity computations between the short-text.



Input
A1: A man plays the guitar and sing.
A2: A man is singing and playing the guitar

Tokenization
A1: a, man, plays, the, guitar, and, sing
A2: a, man, is, singing, and, playing, the, guitar

Stop word removal
A1: man, plays, guitar, sing
A2: man, singing, playing, guitar

Lemmatization
A1: man, play, guitar, sing
A2: man, sing, play, guitar

Figure2. Lexical analysis of short-text

Before applying the similarity measures the short-text undergoes the lexical analysis phase the output generated from the lexical analysis phase act as input for similarity computation phase, for example shown in figure 1.

Path measures (PATH) [9]: Stands for the length of the path between the two concepts in the WordNet graph.

Hirst and st. onge measure (HSO) [10]: This takes in to account many other WordNet relations, beyond “is- a” relation (antonyms and synonyms).

Resnik measure (RES) [11]: The idea is that it measure the how much content is similar to each other, two contents are semantically similar on the information they share with each other. The information content is based on the lowest common subsumer (LCS) of two contents.

$$Sim.\,RES(Ca, Cb) = IC\big(LCS(Ca, Cb)\big) \ldots (6)$$

$\mathrm{IC} = \mathrm{Information\ content},\ \mathrm{LCS} = \mathrm{Least\ common\ Subsumer},\ \mathrm{Ca} = \mathrm{Concept1}, \mathrm{Cb} = \mathrm{Concept2}$

Jiang and Conrath measure (JCN) [12]: In this approach or measure the sum of the individual information contents is similar to that of their LCS, than the concepts are closed together.

$$sim.\,JCN(Ca, Cb) = \frac{1}{\mathrm{IC(Ca)} + \mathrm{IC(Ca)} - 2 * \mathrm{IC(LCS\ (Ca, Cb))}} \ldots . (7)$$

$\mathrm{IC} = \mathrm{Information\ content},\ \mathrm{LCS} = \mathrm{Least\ common\ Subsumer}\ ,\ \mathrm{Ca} = \mathrm{Concept1}\ ,\ \mathrm{Cb} = \mathrm{Concept2}$

Lin measure (LIN) [13]: Ratio of the information content of LCS to the information content of the each of the concept.

$$sim.\,LIN(Ca, Cb) = \frac{22 * IC\big(\mathrm{LCS\ (Ca, Cb)}\big)}{\mathrm{IC(Ca)} + \mathrm{IC(Cb)}} \ldots . (8)$$

$IC = \text{Information content}, LCS = \text{Least common Subsumer}, Ca = \text{Concept1}, Cb = \text{Concept2}$

Extended Gloss Overlap measure (LESK-E) [14]: Attempts to calculate the similarity of each concept from the overlapping of the glosses associated with each concept and with their related concepts in the WordNet.

Wu and Palmer measure (WP): Compares the global depth value of two concepts using the WordNet taxonomy.

$$Sim.\,WP = GDV\,(Ca, Cb) \ldots (9)$$

$GDV = \text{Global Depth Value of two concepts}$

Leacock and Chodrow measure (LC): Quantifies the length of the shortest path and the maximum depth of the taxonomy of two concepts to measure the similarity between them.

$$LC = Ls \text{ and } Md\,(Ca, Cb) \ldots .\,(10)$$

$Ls = \text{Length of the shortest path}, Md = \text{Md max. depth}$

The major reason behind the Google success is its pageRank algorithm. PageRank algorithm includes the following steps:

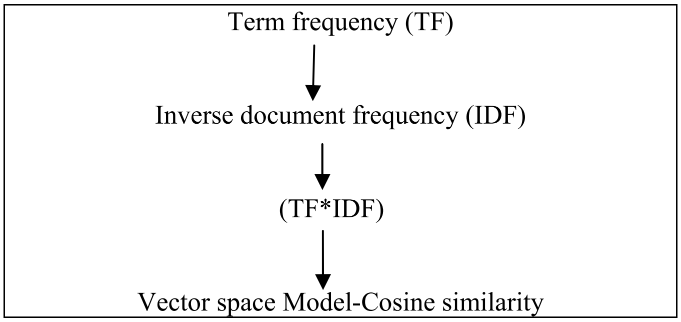

Term frequency (TF)
Inverse document frequency (IDF)
(TF*IDF)
Vector space Model-Cosine similarity

Figure3. PageRank algorithm

There are many applications of both approaches in Natural language processing defined as follows:

Table1. Applications of WordNet and VSM

Approach	Application
Vector space model (VSM)	Information retrieval Computational lexical semantics Word sense disambiguation Named entity disambiguation Text document clustering Text document categorization Collaborative recommendations
WordNet	Conceptual identification Image retrieval Machine translation Query expansion Information retrieval Document classification

II. LITERATURE SURVEY

This section represents the survey of techniques that can be used for semantic similarity. Method that can be proposed, WordNet and Vector space analysis.

2.1 Vector space analysis and WordNet

Jovita et al. (2015) approached a technique that is used to represent knowledge and retrieve the answer for a given question by utilizing Vector space Model. Question answering is a technique in information retrieval system that gives directly answer to the queries generated by the user. Many different approaches like N-gram, template-driven response, reversible transformation are attempt to use vector space model.

L.-C. Yu et al. (2016) introduced near-synonym substitution using discriminative vector space model, there are two components: a vector space model and discriminative training. Near synonyms are fundamental and essential knowledge resources for language learning tasks. Near synonym substitution is a difficult task in online learner system because it is it is not easily grasped in practical use, especially in second level language learners (L_2). Vector space model act as baseline classifier to classify the test example into near-synonym in a given near-synonym set.

Heng Zhang, Guoqiang Zhong (2016) presented the framework for short text classification by learning vector representation of both words and hidden topics together. Topic of words and text is viewed as new words for enriching the data, on the enrichment of data learning of vector representation of both words and topic is done. In this way the representation of short text in form of vectors is done.

W. Lu et al. (2015) proposed semantic similarity assessment using differential evolution algorithm in continuous vector space, which assess the semantic similarity between the short-text which is a quite challenging task in knowledge-based applications. In this approach the combining of the corpus based method with the WordNet is there. Corpus based approaches map a given corpus into vector space. The words act as vectors in which the semantic similarity is calculated with the help of different measures of WordNet.

Oliva et al. (2011) represented the SyMSS, syntax based measure for short-text similarity. The meaning of the sentence is not only dependent on the words that are included in the sentence but also on the structural way of the words that are combined to form a sentence. SyMSS captures and combines the semantic information obtained from the lexical database called WordNet. With this information the proposed measure captures the semantic similarity between the short-text that plays the same syntactic roles.

R. Ferreira et al. (2016) proposed a three level architecture for assessing the sentence similarity through lexical syntactic and semantic similarity methods. The degree of similarity can be assessed with the help so many different similarity measures that are available recently. Two major problems are not addressed by such approaches word order and meaning of the sentence as whole this method largely improves the problem that are being faced by the previous approaches.

Islam and Inkpen et al. (2008) presented a method for measuring the semantic similarity of text using corpus-based measures and modified version of Longest Common Subsequence (LCS) string matching algorithm. In this approach the main concern is about finding the similarity between the two sentences and short paragraphs because in earlier approaches the similarity between the large documents was calculated.

Mihalcea et al. (2006) introduced a measure which represents the sentence as bag of vectors and perform similarity measure that works according to the sequence of the sentence first sentence is act as reference for finding the similarity between the second sentence or we can say that the words that are similar to the first sentence present in the second sentence are grasped. In the same fashion the similarity between the second and the third sentence is calculated. At the end the average of the similarity computation is calculated to find the semantic similarity between the sentences.

Li et al. (2006) approached a word order vector that converts each sentence in to semantic vector by using lexical database. A new word vector is proposed for each sentence using information from lexical database that calculate the weights of the semantic similarity between the texts by obtaining similarity from the corpus-based methods. The semantic similarity is calculated by grasping the semantic vectors at last the sentence similarity is combined by using semantic similarity and word order vector.

Jimenez et al. (2012) used a recursive model to compare sentences by dividing them into words, and in turn, words and divided into q-grams of characters. The idea of soft cardinality is proposed instead of classical string cardinality. Based on this cardinality, they proposed a similarity model to assess the similarity using seven parameters. It make use of bag of words vector in a regression algorithm, along with reduced-error pruning tree (REP tree) using a set of 17 feature obtained from the combination of soft cardinality with different similarity functions for comparing pairs of words.

Table2. Comparison of Literature survey

Author	Problem identified	Proposed solution
Jovita et al. (2015)	Question answering (QA)	Vector space model (VSM)
L.-C. Yu et al. (2016)	Near-synonym substitution	Discriminative Vector space model (VSM)
Heng Zhang et al. (2016)	Short-text classification	Learning vector representation of both words and hidden topics tog-ether
W. Lu et al. (2015)	Semantic similarity	Differential evolution algorithm in continuous vector space
Oliva et al. (2011)	Word ordering	WordNet
R. Ferreira et al. (2016)	Semantic similarity of text	Three level architecture
Islam and Inkpen et al. (2008)	Semantic similarity	Corpus-based measures and LCS
Mihalcea et al. (2006)	Semantic similarity	Bag of vectors
Li et al. (2006)	Semantic similarity	Word order vector
Jimenez et al. (2012)	Semantic similarity	Soft cardinality

III. CONCLUSION AND FUTURE SCOPE

The advances in technology of computers and electronics, the increasing popularity of the internet and WWW lead to vast amounts of increase in electronic text information. In order to find the similarity between the text documents or short text or we can say that in the field of informational retrieval both the approaches are quite beneficial. Vector space analysis can be done for finding the short-text similarity instead of finding the similarity between the larger documents.

The goal of the information retrieval is to find the relevant information regarding to the query fired by the user. In which the similarity methods plays the vital role.

Future scope for vector space analysis includes that documents may be represented by some other methods. Dimensionality of the term-document matrix can be reduced by the new techniques. New clustering algorithm may be introduced. Inputting the number of clusters to be may be automated. The system may be modified to have the provision of refining the input query. The system may be extended to handle any type of datasets including images. Evaluation may be conducted on the techniques for documents in various languages and on the study of effects of language in the performance of retrieval process.

There various enhancements that can be done on WordNet too. One of them is multilingual WordNet. One of the most relevant task in the development of multilingual WordNet is EuroWordNet, a project based on WordNet , whose purpose is to build WordNet for European languages or another enhancement of WordNet is Hindi WordNet, it is a system for bringing together different lexical and semantics relations between the Hindi words. The design of the Hindi WordNet is inspired from the famous English WordNet. Punjabi WordNet is an online lexical resource, it is more than a conventional Punjabi dictionary and a combination of conventional Punjabi dictionary and thesaurus. It provides information about the word from different perspectives and relationship between the words too.

WordNet is most frequently used concept in finding the relatedness between the two different concepts. Main aim of researchers is to make WordNet more efficient and effective.

REFERENCES

[1] Oliva et al., SyMSS: a syntax based measure for Short-text Semantic similarity, Data and Knowledge engineering, Elsevier, 2011, 390-405.

[2] R. Ferreira et al., Assessing sentence similarity through lexical syntactic and semantic similarity, Computer Speech and Language, Elsevier, 2016.

[3] Jovita et al., Using Vector space model in Question answering system, Procedia computer science, Elsevier, 2015, pp. 305-311.

[4] Fillmore et al., Background to Framenet. Int. J. Lexicogr. 2016, pp. 235-250.

[5] Mihalcea et al., Corpus-based and knowledge based measure for text semantic similarity, Proceedings of the 21st National Conference on Artificial intelligence, 2006, pp. 75-780.

[6] Qiu et al., Paraphrase recognition via dissimilarity significance classification, Proceedings of the Conference on Empirical Methods in Natural Language Processing, (2006), Association for Computation Linguistics , USA, pp. 18-26.

[7] Islam and Inkpen et al., Semantic text similarity using corpus-based word similarity and string similarity, ACM transaction Knowledge Discovery Data, pp. 1-10.

[8] Vishal Gupta, Gurpreet singh lehal, Survey of text summarization extractive techniques, Journal of Emerging Technologies in Web Intelligence, August 2010, pp. 258-268.

[9] R. Rada et al., Development and application of a metric on semantic nets, IEEE Transaction on Systems, Man and cybernetics (1987), pp. 1-30.

[10] G.Hirst, D. St-Onge, Lexical chains and representation of context for the detection and correction of Malapropisms, in: C. FellBaum (ED.), WordNet: An electronic lexical database, MIT press, 1998, pp. 805-810.

[11] P. Resnik, Using information Content to Evaluate Semantic Similarity in a Taxanomy, Proceedings of the 14th International Joint Conference on Artificial Intelligence, 1995 , pp. 448-453.

[12] J. Jiang, D. Conarth, Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy, Proceedings on International Conference on Research in Computational Linguistics, 1997, pp. 19-33.

[13] S. Banerjee, T. Pedersen, Extended Gloss Overlaps as a Measure of Semantic Relatedness, Proceedings of the 18th International Joint Conference on Artificial Intelligence, 2003, pp. 805-810.

[14] H. Calvo et al., Integerated Concept Blending with vector space models, Computer Speech and Language, Elsevier, 2016.

[15] L.-C. Yu et al., Near-Synonym substitution using a Discriminative Vector space model, Knowledge Based Systems, Elsevier, (2016), pp. 74-84.

[16] Heng Zhang, Guoqiang Zhong, Improving Short Text Classification by Learning Vector Representation of Both Words and Hidden Topics, Knowledge-Based Systems, Elsevier, 2016.

[17] W. Lu et al., Semantic Similarity Assessment using differential evolution algorithm in continuous vector space, Journal of Visual language and Computing, Elsevier, 2015, pp. 246-251.

[18] Li et al., Sentence Similarity based on Semantic nets and corpus statistics. IEEE Trans. Knowledge Data Engineering, 2006, pp. 113-1150.