Word net based Method for Determining Semantic Sentence Similarity through various Word Senses

Madhuri A. Tayal
Research Scholar, G. H. Raisoni College of Engineering, Nagpur.
Asst. Prof. RCOEM, Nagpur, INDIA.
madhuri.tayal@gmail.com

M. M. Raghuwanshi
Principal, Rajiv Gandhi College of Engineering and Research Nagpur, INDIA.
m_raghuwanshi@rediffmail.com

Latesh Malik
HOD(CSE),
G. H. Raisoni College of Engineering, Nagpur, INDIA.
latesh.malik@raisoni.net

Abstract

Semantic similarity is a confidence score that replicates semantic equivalence between the meanings of two sentences. Determining the similarity among sentences is one of the critical tasks which have a wide-ranging impact in recent NLP applications. This paper presents a method for identifying semantic sentence similarity among sentences using semantic relation of word senses across the different synsets using Wordnet for different part of speech of words. This method firstly detects all the semantic relations (hypernym, hyponym, holonym, meronymy etc.) considering the word as a noun and all the sense relations considering word as a verb from Wordnet respectively. Then it uses common senses between the two sets as Noun and Verb, for two input words for the calculation of semantic word similarity score. As sentence is made up of different words, these word similarity scores have been used for calculation of semantic sentence similarity among the sentences. It is difficult to achieve a high precision score because the exact semantic meanings will not be understood simply. However proposed method outperforms in comparison with existing methods. The evaluation is done for sentences using SemEval-12 Task 6 (Test-Gold-Set) with respect to human ratings.

1. Introduction

Now-a-days web is the largest and utmost useful knowledge base for users with bulky amount of information. The additional sources of information are newspapers, magazines, textbooks etc. Web consists of billions of text documents and daily many different documents are added to it. For understanding these types of documents/ texts many applications have been made such as Text Mining, Storytelling, Machine Translation, Deep Question and answering and Text Summarization etc. To develop and understand this kind of applications, there is a requirement of semantic similarity utility.

Various techniques for semantic similarity have been receiving escalating attention since their introduction by (Miller et al. 1997). Researchers have investigated that finding semantic similarity for the sentences is not an easy task as text document may contain complex sentences. Most of these techniques are based on statistics which indirectly use corpus. Some of them use multiple information sources, lexical chains etc. (Jiang et al., 1997; Li et al., 2009). Some of them are based on wordnet (Simpson et al., 2010; Pederson et al., 2004; Leacock et al., 1998; Budanitsky et al., 2006). WordNet is a lexical catalogue which is accessible online, and provides a large source of English lexical items. The proposed approach incorporates the various senses of the words and there relations from Wordnet.

This paper presents a method for identifying semantic sentence similarity among sentences using Wordnet. This method firstly identifies all the sense relations (hypernym, hyponym, holonym, meronymy etc.) considering the word as a noun and all the sense relations considering word as a verb from Wordnet respectively. Then it uses common senses between the two respective sets as Noun and Verb for two input words for the calculation of semantic word similarity score. The similar senses and their respective counts are useful for the calculation of semantic similarity among the words. Then these Word to Word values will be used for the calculation of semantic sentence similarity.

This paper is organized as follows: Section 2 comprises various methods which are available for

D S Sharma, R Sangal and J D Pawar. Proc. of the 11th Intl. Conference on Natural Language Processing, pages 139–145, Goa, India. December 2014.

determining Semantic sentence Similarity. Section 3 provides the motivation for this work. Section 4 proposes the method for the calculation of Semantic word similarity and sentence similarity. Section 5 contributes experimental results of proposed approach for identifying the semantic sentence Similarity and comparison with other method. The paper ends with parametric analysis, conclusion and work to be carried out in future.

2. Literature Survey

Based on the various ideas of computing similarity, the semantic similarity could use either the path distance between concepts or the information content of a concept as a quantifying measure (B. Plank et al., 2013; Veal et al., 2013; Simpson et al., 2010; Agirre et al., 2009; Turney et al., 2005; Dolan W. et al., 2004; Jiang et al., 1997;). In certain contexts, the combination of both the path distance and information content based methods has been tried out. Following section describes these measures.

2.1 Path Length based Measures

The similarity measurement between 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 major relations that connect different concepts is only is-a type relations. These measures compute similarity in terms of shortest path between the target synsets (group of synonyms) in the taxonomy. The different path length based similarity measures viz. Rada measure (Rada et al., 1989), Hirst Onge measure (Budanitsky and Hirst, 2006), Bulskov measure (Bulskov et al., 2002).

A. Rada Measure (1989):- The semantic distance is calculated by including the number of edges between concepts in the taxonomy. Let C1 and C2 be the two concepts in is-a semantic net. The conceptual Distance among C1 and C2 is specified by

Distance (C1, C2) = Minimum number of edges separating C1 and C2 (1)

B. Budanitsky and Hirst Measure (2006):- Similarity between concepts is described as a path distance between two concepts. The weight of the path linking the concepts C1, and C2 is given by

Weight= c – length (C1, C2) – k * turns (C1, C2) (2)

Where c and k are constants, length (C1, C2) is the distance of the shortest acceptable path connecting the synsets C1 and C2 and turns (C1, C2) is the number of changes in direction in the shortest allowable path.

C. Bulskov Measure (2002):- Similarity is based on the concept inclusion is-a relation for atomic and compound concepts of ontology. The quantification of similarity is based on the direction of concept inclusion.

D. Wu and Palmer Similarity Measure (Wu et al., 1998):-Wu and Palmer suggested a new method on semantic representation of verbs and investigated the influence on lexical selection problems in machine translation. Wu and Palmer describe semantic similarity measure amongst concepts C1 and C2 as

$$Sim_{WP}(C_1, C_2) = 2 \times \frac{N_3}{N_1 + N_2 + 2 \times N_3} \quad (3)$$

Where N1 is the length given as number of nodes in the path from C1 to C3 which is the minimum collective super concept of C1 and C2 and N2 is the length given in number of nodes on a path from C2 to C3. N3 signifies the global depth of the hierarchy and it serves as the scaling factor. Wu and Palmer describe semantic similarity measure between concepts C1 and C2 as

Dist(C1,C2)=1-Sim(C1,C2) (4)

(Leacock and Chodorow, 1998; Budanitsky and Hirst, 2006) advised an approach for measuring semantic similarity as the straight path using is-a hierarchy for nouns in the WordNet.

2.2 Information content based measures (corpus)

In the literature, the information content based approaches are also referred to as corpus based approaches or information theoretic based approaches. Some of them are listed here.

A. Resnik Measure (1995)

Similarity depends on the amount of information of two concepts have in common. This shared information is given by the Most Specific Common Abstraction (MSCA) concept that includes both the concepts.

B. Lin Similarity Measure (Lin et al., 1998)
Lin extended the Resnik(1995) method of the material content (Lin et al., 1998). He has defined three intuitions of similarity and the basic qualitative properties of similarity.
C. Jiang and Conrath measure (Jiang et al., 1997)
Semantic distance is derived from the edge-based view of distance. In order to reimburse for the unpredictability of edge distances, Jiang and Conrath weigh each edge by associating probabilities based on corpus data and also consider the link strength of each concept.

2.3 Hybrid approach

Hybrid approach combines the knowledge derived from different sources of information. The major advantage of these approaches is if the knowledge of an information source is insufficient then it may be derived from the alternate information sources. In this Direction, Li (Li et al. 2003) and Zuber and Faltings (Zuber and Faltings 2007) have been contributed. Li et al. overcomes the weakness of Rada edge counting method. Zuber and Faltings computed the similarity between two concepts using ontology structures.

The proposed method is based on Thanh Ngoc Dao, Troy Simpson's method (2010). This method uses Wordnet in background and Wu and Palmer distance measure for identifying sentence similarity. Given two sentences X and Y, indicate m to be length of X, n to be length of Y. The semantic similarity has been calculated as follows.

Overall score= 2*Match(X, Y) / |X| +|Y| (4)

This method has following disadvantages.

1. It is likely for two synsets from the same part of speech to have no common sub-sumer. Since all the different top nodes of each part of the speech taxonomy did not joined, a path cannot continuously be found between the two synsets and that provides the incorrect results.
2. Multiple inheritances are allowed in Word Net, some synsets belongs to more than one taxonomy. So, if there is more than one way between two synsets, the direct such path is selected and that can give wrong value of semantic similarity.
3. Even it does not check the semantic similarity amongst positive and negative sentences.

These weaknesses tried to be removed in proposed approach, described in the next section.

3 Motivation for Proposed Work

Finding semantic sentence similarity is a very complex task in literature because, understanding will be done through interpreting the information from the sentence by a human brain. The task of replacing human brain with computer program is a challenging task. Following assumptions were taken into account while finding the similarity between the sentences.

1. Subject-Subject contains Noun/ Pronoun in the sentence. Nouns plays important role in the sentence, According to (Wren and Martin) noun entity is responsible for doing the actions. So Noun of sentence-1 and sentence-2 is important to check, as well as the similarity in between them.
2. Verb-Verb plays important role in the sentence; According to (Wren and Martin) this entity is responsible which actions are taking place in the sentence. So verb of sentence-1 and sentence-2 is important to check, as well as the similarity in between them.
3. Object-Object plays important role in the sentence, According to (Wren and Martin) object entity is responsible on whom the actions will be taken. So object of sentence-1 and sentence-2 is important to check, as well as the similarity in between them.
4. Similarity-To verify weather particular pairs of words are semantically related or not, human brain verifies the fact that for how much ways (Senses) they are identical.

For processing these observations, we have the whole process is structured for two types of sentences, simple and complex. Simple sentence comprises single Verb and complex sentence comprises of more than one verb. Example for simple sentence is "Stack uses arrays". Example for complex sentence is "Data structures arrange the required data properly for any application". Further two pairs (Noun-Noun, Verb-Verb pair) have been taken to find the identical senses wherever they are same using WORDNET. After this, average of mentioned pairs will be considered as a result of sentence similarity. Similarity for Simple and complex sentences is calculated as:

Similarity % (Simple sentence) = Average % [(Word-Similarity-Subject pair of sentence 1 and 2) + (Word-Similarity- Verb pair of sentence 1 and 2) + (Word-Similarity-Object pair of sentence 1 and 2)

Similarity % (Complex sentence) = Average % [(Word-Similarity-Subject pair of sentence 1 and 2) + (Word-Similarity- verb pair of sentence 1 and 2) + (Word-Similarity-Object pair of sentence 1 and 2) + (Word-Similarity-Additional pair of Verb Verb2)….n terms.

Above Presented Logic has been used throughout the procedure mentioned in the next section.

4. Proposed Method for Semantic Similarity between the Sentences

Semantic score is the key value to be used for numerous applications in the arena of Natural Language Processing. Following method is identified for the calculation of semantic similarity for the words and sentences.

4.1 Semantic Similarity for Words

Firstly the semantic similarity between words will be identified because sentence consists of words. The bottommost unit in a WordNet is synset, which indicates a sure meaning of a word. It contains the word, its description, and its synonyms. The exact meaning of one word under one type of POS is called a sense. At this time, all senses of the word (synonyms, hypernym, hyponym, holonym etc.) are utilized for the calculation of sentence similarity. In the wide sense the steps for word similarity are as follows.

Step-1 Find the Noun and Verb senses for the entered word from wordnet.
Step-2 For checking similarity between two words, Counts the noun and Verb senses, using

$$\textbf{Score} = A \cap B / \text{Min}(A, B) \quad (5)$$

Consider word1 and word2 for which similarity is to be checked. Word1 having n1 as a count of all senses collected from different synsets through wordnet, considering as a Noun. And v1 as a count of all senses collected from different synsets through wordnet, considering as a Verb. Similarly, n2 and v2 for Word2. The common matches for word-1's and word-2's Noun and Verb senses are 'n' and 'v' respectively. C is the overall count for the calculation of semantic similarity. The complete steps are as follows.

Word1's senses as (Noun sense/Verb sense) = n1 / v1, Word2's senses as (Noun sense/Verb sense) = n2 / v2, Common Matching word senses (Noun and Verb) among Word1 and Word2= as 'n' (for Noun) and 'v' (for Verb) calculated as n = n1∩n2 and v = v1∩v2. C = Final Similarity count, C1, C2, temporary Similarity counts.
1. If ((n1 || n2 || n) == 0) then C = C1 If (v1 < v2) C1 = v / v1 Else C1 = v / v2
2. If ((v1 || v2 || v) == 0) then C = C1 If (n1 < n2) C1 = n / n1 Else C1 = n / n2
3. Else if (n1<n2) C1 = n / n1 Else C1 = n / n2 If (v1 < v2) C2 = v / v1 else C2 = v / v2
Word Similarity = C = C1+C2

If word1 and word 2 has no noun senses then word similarity will be based on verb senses and vice versa, as indicated in step 1 and 2. Otherwise, it will be a summation of noun and verb senses. All senses of the word (synonym, hypernym, hyponym, holonym etc.) are taken into account while identifying different Verb and Noun senses for the words. If for both words, the noun senses and verb senses identified as nonzero then the summation of matching percentage for noun and Verb is taken into account for the calculation of word similarity.

Example-word similarity: For two words, "stack", "queue", n1=39, n2=77, v1=6, v2=18(counted from all the senses holonym, synonym, hypernym etc. using wordnet.) n=19 (common word senses of n1 and n2), v = 0, since n1 < n2:. C1=n/n1, 19/39=0.48= 0.5(rounding off)

4.2 Semantic similarity for Sentences

Sentence is a collection of words. Once the word to word similarity is known, this count has been used to calculate semantic similarity between the sentences, following steps to be followed for the calculation of sentence similarity.

Step-1 Tag entered two sentences using POS tagger.
Step-2 Identify the word similarity for all combinations of Noun-Noun pairs and Verb-Verb pairs.
Step-3 Calculate the final value of semantic similarity between the sentences by

averaging of all these pair's word similarity score from step-2.

Final Score =Average ($\sum_{i=1}^{n}$ NounScore +$\sum_{i=1}^{n}$ VerbScore).

Semantic similarity= $\sum_{i=1}^{n}$(Subject, Verb, Object)

Consider, Sentence 1 and 2 for which the similarity is to be checked. Both the sentences get tagged using POS tagger. The Noun-Noun and Verb-Verb pairs will be identified between the sentences. Then there in between word similarity scores has been calculated. The total average of word similarities of all these pairs will be considered as the sentence similarity for two sentences.
Example-sentence similarity: - Consider, two sentences, 1. "Database keeps data.", 2. "Data is important.". N-N pairs are: Database-Data, data-Data. Their respective similarities are 0.3 and 1. Verb-Verb pair is "keeps" and "is" and its similarity is 0.2. So, the average of these three pair=0.5 and it is the sentence similarity for these two sentences. (50% matching). The experimental results for word and sentence similarity are presented in the next section.

5 Performance Evaluation and Comparison with other Methods

The accuracy of word similarity is tested through; Miller & Charles test set is used (Miller et al., 1997). This test set contains 353 word pairs with semantic similarities and their respective human ratings for pairs of words. Table 1 shows the word similarity results for some of the random words given in MC set. For this set, human ratings have been calculated from five language experts, and mean value is taken. Various scores for these pairs with Jiang and Conrath's method (Jiang et al.,1997), Wordnet based method based on Internet and knowledge by (Liu et al.), Wordnet based method by (Simpson et al., 2010), and also human ratings have been compared and presented. It is observed that proposed method's results closely matches with respect to human ratings compared to other methods. The minimum correlation value indicates the words are semantically closely related to human ratings.

MC-set	Jiang and Conrath's Method (1997)	WordNet-based internet knowledge (Liu et al.)	Wordnet-Wu/Palmer	**Proposed method**	Human ratings
car – automobile	0.341	0.347	1	1	0.894
gem – jewel	0.340	0.349	1	1	0.896
journey – voyage	0.306	0.330	0.5	1	0.929
boy – lad	0.289	0.325	0.5	0.6	0.883
asylum – madhouse	0.323	0.318	0.5	0.9	0.887
magician – wizard	0.343	0.344	1	0.8	0.902

Table 1. Semantic Similarity distance for various methods

From Table 1 it is found that the proposed method's results are closer to human ratings. The average semantic difference between proposed method and Wordnet based Simpsons's (Wu and Palmer Distance) method with respect to human ratings for hundred pairs of words of MC set is calculated and shown in Table 2 below.

Methods	Average Correlation for hundred pairs of words from MC set
For Proposed Method with respect to human ratings.	0.29
for Wordnet (Wu-Palmer)[Simpson et al., 2010] based Method with respect to human ratings.	0.34

Table 2. Average Correlation for hundred pairs of Words with respect to human ratings

Table 2 indicates that average difference for hundred pairs of word from MC set for proposed method with respect to human rating is found to be less as compared to Wordnet (Wu-Palmer) method. The motivation to take Wordnet-based (Wu and palmer) method for comparison, because it uses path length as a criteria for counting similarity from Wordnet. And in some cases the results are not precise. This method (Simpson et al., 2010), which uses Wordnet

based (Wu-Palmer) distance is found to be challenging in comparison with proposed method and proposed method removes its disadvantages carefully. Comprehensive comparison clearly indicates that proposed method results are accurately similar to the human ratings.

Sentence-Evaluation is done for two hundred sentences from SemEval-12 Task 6 using Microsoft Research Paraphrase corpus i.e. Test-Gold-Set (Eneko Agirre et al., 2012) shown in Table 3. For Evaluation of these samples, two hundred sample sentences were given to five language experts. The mean of similarity score is calculated for the sentence similarities given by these experts. The results are compared and shown in Figure 1.

Methods	Average Correlation for 200 pairs for sentences from set
Proposed Method for SemEval-12 Task 6 Test-Gold-Set, with respect to human ratings	0.164
for Wordnet (Wu-Palmer)(Simpson et al., 2010)based Method with respect to human ratings	0.31

Table 3. Average Correlation for two hundred pairs of Sentences

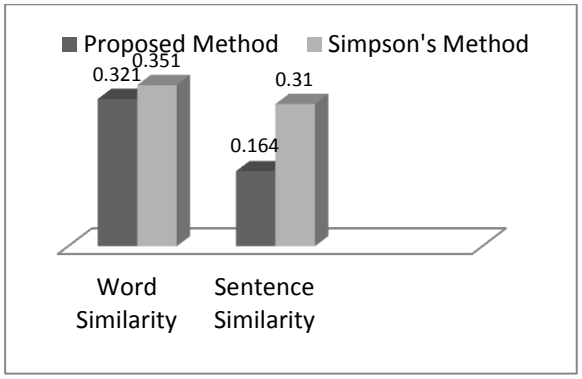

Figure 1. Comparative Chart for Average Correlation for two hundred pairs of words and sentences with respect to Human ratings.

Average correlation for two hundred pairs of sentences from SemEval-12 Task 6 for proposed method with respect to human rating is found to be less as compared to Wordnet based (Simpson et al. 2010) method. It indicates results are closer to human ratings.

Conclusion and Future Work

This paper presents semantic sentence similarity approach for language processing using various word senses of Wordnet, which removes the drawbacks of Simpson's Wordnet based method. This method does not rely on the path distances; between the synsets instead it depends upon all the semantic relation of word senses across the different synsets using Wordnet for different part of speech of words. In addition to this it correctly contributes the semantic similarity amongst positive and negative sentences too. After parametric analysis of Noun-Noun, Verb-Verb and averaging of both pairs, following conclusions have been drawn.

1. Sentence Similarity using averaging of Noun-Noun pair does not give appropriate results and it is less as compared to Verb-Verb pairs with respect to human ratings.
2. Sentence Similarity using averaging of Verb-Verb pair also does not give appropriate results and in comparison the average difference using Verb-Verb pairs is more than Noun-Noun pair with respect to human ratings. It indicates that Noun-Noun pair gives better results and are nearer to human ratings as compared to Verb-Verb pair results.
3. The combined score gives far better results than individual performance of Noun-Noun and Verb-Verb pair. It shows that the combined contribution of Noun and Verb in a sentence while finding the semantic similarity has major role. Because the combination of both, contributes for the meaning of a sentence and therefore it gives better results with respect to human ratings.

Some of the restrictions for this algorithm are as follows. In very few cases, it provides much underestimated results compared to human ratings. Due to which the results for sentence similarity becomes incorrect. This inadequacy is emerged from Wordnet utility which is capable to process available vocabulary. Proposed utility will be used in many NLP applications like question answering, text summarization, etc. and can be applied on Hindi, Marathi etc. languages. The performance can be further improved with more deep analysis of words with diverse mathematical parameters, domain corpora etc.

References

A.Budanitsky and G. Hirst. 2006. *Evaluating WordNet-based measures of semantic distance*, Computation Linguistics. 32(1): pp.13-47.

Barbara Plank, Alessandro Moschitti. 2013. *Embedding Semantic Similarity in Tree Kernels for Domain Adaptation of Relation Extraction*, Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: 1498–1507.

Claudia Leacock and Martin Chodorow. 1998. *Combining local context and WordNet similarity for word sense identification*. In Fellbaum, C. (ed.), WordNet: An Electronic Lexical Database: 265–283.

D.Lin. 1998. *An information-theoretic definition of similarity*, In Proc. of the Intl Conf. on Mach. Learn: pp: 296–304.

Dolan W, Quirk C. and Brockett C. 2004. *Unsupervised construction of large paraphrase corpora: Exploiting massively parallel new sources*. In Proceedings of the 20th International Conference on Computational Linguistics.

Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Pasca and Aitor Soroa. 2009. *Study on Similarity and Relatedness Using Distributional and WordNet-based Approaches*. In Proceedings of NAACL '09, 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics: pp. 19-27.

Eneko Agirre, Daniel Cer, Mona Diab and Gonzalez-Agirre Aitore, 2012. SemEval-2012 task6: A pilot on Semantic textual Similarity. In Proc. 6th International Workshop on Semantic Evaluation (SemEval2012), First joint conference on Lexical and Computational Semantics, Montreal, Canada.

Gang Liu, Ruili Wang, Jeremy buckley, Helen Zhou. *A WordNet-based Semantic Similarity Measure Enhanced by Internet-based Knowledge.*

George A. Miller and Walter. G. Charles.1997. *Contextual correlates of semantic similarity*, Language and Cognitive Processes, 6(1):1-28.

H.Bulskov, R. Knappe, and T. Andreasen. 2002. o*n Measuring Similarity for Conceptual Querying*. in T. Andreasen, A. Motro, H. Christiansen, H.L. Larsen (Eds.): Flexible Query Answering Systems, Lecture Notes in Artificial Intelligence 2522: pp. 100-111.

J.Jiang and D.W. Conrath.1997. *Semantic similarity based on corpus, statistics and lexical taxonomy*, In Proc. of the Intl Conf. on Research in Comput. Linguist: pp. 19–33.

P.Resnik. 1995. *Using information content to evaluate semantic similarity*, In Proc. of the 14th Intl Joint Conf. on Artificial Intelligence: pp. 448–453.

Peter Turney. 2005. *Measuring semantic similarity by latent relational analysis*. Proceedings of the 19th International Joint Conference on Artificial Intelligence: 1136-1141.

R.Rada, H. Mili, E. Bicknell, and M. Bletner. 1989. *Development and Application of a Metric on Semantic Nets*, IEEE Trans. Syst. Sci. Cybern, 19(1): pp. 17-30.

Troy Simpson, Thanh Dao. 2010. *Capturing the semantic similarity between two short sentences based on the WordNet dictionary.*

Ted Pederson, Siddarth Patwardhan and Jason Michelizzi. 2004. *WordNet-Similarity- measuring the relatedness of concepts*. In Proceedings of HLT-NAACL'04 Annual Conference of the North American Chapter of the Association for Computational Linguistics: pp. 38-41.

Vincent Schickel-Zuber, Boi Faltings. 2007. A Semantic Similarity Function based on hierarchical Ontologies, IJCAI.

Wren and martin, English grammar and composition. (Book) s. Chand & company ltd.

Yuhua LI, Zuhair A. andar, David McLean. 2003. *An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources*. IEEE Transactions on Knowledge and Data Engineering, Volume 15, Issue 4: 871-882.