IJCSE International Journal of Computer Sciences and Engineering Open Access
Survey Paper Vol.-6, Issue-9, Sept. 2018 E-ISSN: 2347-2693

Textual Similarity Detection from Sentence

S.L. Patil[1*], K.P.Adhiya[2]

[1,2]Computer Engineering, SSBT COET Bambhori, Jalgaon, North MaharashtraUniversity, Jalgaon.

Available online at: www.ijcseonline.org

Accepted: 19/Sept./2018, Published: 30/Sept./2018

Abstract— In computer science, textual similarity used for detecting the similarity between words, terms, sentences, paragraph, and document. In natural language processing, sentence similarity performs the tasks such as document summarization, word sense disambiguation, short answer grading , and information retrieval. The lexical overlapping approach evaluates the similarity between the sentence and finds whether a sentence pair is semantically equivalent or not. Existing methods are used for checking the similarity of long text documents. These methods process sentences in high-dimensional space and are not much efficient, requires human input and also not adaptable to some application domains. Semantic textual similarity methods improved in two areas -(a) in the semantic relation between the words and (b) in semantic resources to reduce the dimension. The proposed architecture uses the two methods for directly computing the similarity between very short texts of the sentence and long text sentences. The Weighted Overlap Approach based proposed method provides a nonparametric similarity by comparing the similarity of the rankings for an intersection of the senses in both the sentences. The Cosine similarity based proposed method identify all distinct words from the sentences. In the proposed work the similarity detection methods are focused to check the synonyms similarity between the sentences.

Keywords— Natural Language Processing, Semantic Textual Similarity, Word Similarity, Sentence similarity, Text Similarity.

I. INTRODUCTION

In natural language processing, the recent application is presented a need for an effective method to compute the similarity between sentences. An example of this is machine translation, which needs to find out the closest corpus sentence of user input, according to sentence similarity. In machine translation system needs to compute the similarity between user input and each sentence in the corpus, and then find out the closest corpus sentence of user input according to sentence similarity and return translation of this corpus sentence as a final result. Sentence similarity also has important application in the document digest system, document classification system and information retrieval system. Traditional methods used for detecting similarity between documents, and have centered on analysing shared words. Such methods are usually effective when detecting the similarity of long texts documents because similar long texts documents usually contain a more same words. However, in short texts sentences, word co-occurrence may be rare. This is mainly due to the inherent flexibility of natural language enabling people to express similar meanings using quite different sentences in terms of structure and word contents. Since such information in short texts is very limited, this problem poses a difficult computational challenge.

II. LITERATURE SURVEY

The strategies for finding the similarity between long text or documents have focused on investigating shared words. These procedures are just accessible to manage long text since they contain sufficient co-appearing words that express fundamentally the same as implications. In recent days, to find out the similarity between the text is become the crucial task in the various application, paraphrase detection is a major area of research because of its significance in the various applications of the natural language processing.

2.1 Related Work

Yuhua, et.al, in [1], proposed the novel algorithm for computing similarity between very short texts of sentence length. The author introduced a method that takes account of not only semantic information but also word order information implied in the sentences. This results in a conversational agent knowledge base that is easier to compile, far shorter, more readable and much easier to maintain.

Courtney Corley, in [6], proposed a method that combines word to word similarity metrics into a text to text metric. This method outperforms the previous text similarity metrics based on lexical matching. In natural language processing, the text similarity has been used for a long time in applications and related areas. Text similarity has been also

used for relevance feedback and text classification, word sense disambiguation, and more recently for extractive summarization, and methods for automatic evaluation of machine translation or text summarization.

Liu and Zong, in [2], proposed the example-based machine translation approach for Chinese sentences that are translated into english sentences. The First approach is to find the most similar examples as the input sentence. The second approach is about recombining the translation of the input sentence according to the most similar example and bilingual dictionary. The third approach is the translation of the input sentence.

Jiang, et.al, in [8], proposed the new method for measuring the semantic similarity distance between words and sentence. It combines the statistical methods and lexico-syntactic patterns so that lexical distance between the semantic node is constructed by taxonomy can be better quantified with computational evidence derived from distributional analysis of corpus data.

III. PROPOSED SOLUTION

In natural language processing, the recent application is present a need for an effective method to compute the similarity between sentences. Text similarity detection plays an increasingly important role in text related research and applications in tasks such as information retrieval, text classification, document clustering, machine translation, text summarization and others. Proposed methods are use the WordNet semantic dictionary to capture the semantic similarity between two sentences. In natural language processing (NLP), sentence similarity is one of the core element.

3.1 Proposed Approach

The proposed architecture focuses on the two methods for computing the semantic similarity between the sentences. The first method is weighted overlap approach based method and second method is cosine similarity based method. The proposed method established a computational method that able to measure the similarity between very short texts sentences as well as long text sentences. In the proposed system is first the preprocessing of data is done. Then the preprocessed data are given as an input to the alignment base disambiguation algorithm. Then proposed weighted overlap based method and cosine similarity based method are used for detecting similarity between sentences.

A. Proposed Arichitecture

The architecture of the proposed system is shown in Figure 3.1 Inputs to the proposed system are training datasets. The data sets contain the collection of sentences. The data preprocessing consists of the stemming, removal of stop words, tokenization. After preprocessing the file is to be given as input to the proposed methods. The goal is to identify an overall strategy to capture the semantic similarity between two sentences. The proposed architecture is given below:

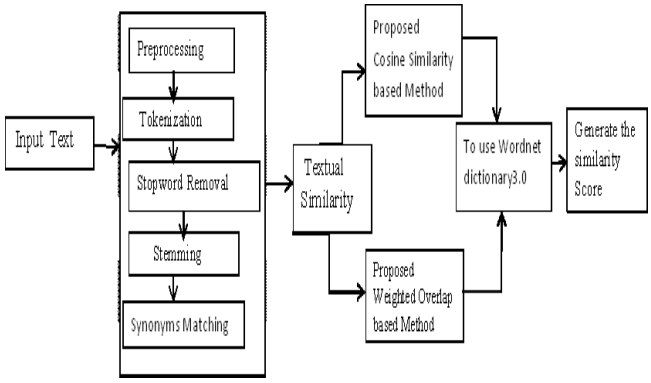

Fig: 1 Structure Of Textual Similarity Detection

Figure 3.1 shows the proposed architecture for textual similarity detection, in which preprocessing steps are performed. In preprocessing, the sentences are tokenized in terms of tokens. In the tokenization, sentences are separated by dot and the words are separated by space also the special symbols are removed. In the next stage of preprocessing the stop word removal algorithm is used. The stemming process performs in the next stage. The proposed method have been used to check sentence similarity and also check the synonyms similarity between the sentences. The synonyms word are checked from wordnet dictionary 3.0. The proposed architecture uses the two methods for checking the synonyms similarity between the sentences. The first is based on the weighted overlap method and the second is based on the cosine similarity method.

B. Semantic Signature Similarity

The proposed methods uses two techniques for calculating textual similarity between the sentences. The first is proposed weighted Overlap based method and second is cosine similarity based method use to compute the textual similarity between the sentence.

a. Cosine Similarity

The Cosine similarity based proposed method identify all distinct words from the sentences and detecting similarity between two vectors of an inner product space that measures the cosine of the angle between them. The cosine of 0 is 1, and it is less than 1 for any other angle.

- Identify all distinct words in both texts.
- Identify the frequency of occurrences of these words in both text and treat it as vector.
- Apply cosine similarity function.

$$\text{Similarity } \cos(S1,S2) = \frac{S1 * S2}{||S1|| * ||S2||}$$

- if a = a1,a2,...,an and b = b1,b2,...,bn
- a.b = Sum(a1*b1 + a2*b2 + ... + an*bn)
- ||a||= sqrt(a12 + a22 + ::: + an2)
- ||b||= sqrt(b12 + b22 + ::: + bn2)

b. **Weighted Overlap Approach**

The measure computes the similarity between a pair of ranked lists by comparing the relative rankings of the sentences. Let H denote the intersection of all non-zero dimensions in the two signatures and rh(S) be a function returning the rank of the dimension h in the sorted signature S.

where the denominator is a normalization factor that guarantees a maximum value of one. The proposed weighted overlap based method first sorts the two sentences according to their values and then harmonically weights the overlaps between them. The minimum value is zero and occurs when there is no overlap between the two sentences, i.e., H =0. The measure is symmetric and satisfies the top weightedness property, i.e., it differences in the higher rankings more than it does for the lower ones. Note that rh(S)is the rank of the dimension h in the original vector Sand not that in the corresponding vector truncated to the overlapping dimensions H. In our setting, we experiment with the untruncated semantic signatures and all our signatures are equally-sized (the size is equal to the number of nodes in the network.

Alignment-Based Disambiguation Algorithm

Commonly, semantic comparisons are between word pairs or sentence pairs that do not have their lexical content sense-annotated, despite the potential utility of sense annotation in making semantic comparisons.

Algorithm 3 Alignment-based Sense Disambiguation

```
Input:T1and T2, the sets of word types being compared
Output:P, the set of disambiguated senses for T1.
Step 1:P←— 0
Step 2:for each token ti ϵ T1
Step 3:max-similarity←— 0
Step 4:best-sense(i) ←— null
Step 5:for each token tj ϵ T1
Step 6:for each sense iϵ Senses(ti), sense jϵ Senses(tj)
Step 7:similarity←— R(sense i,sense j)
step 8:if similarity > max-similarity then
step 9:max-similarity←— similarity
step 10:best-sense i←— sense i
step 11:if best-sense i <> null then
step 12:P←— P⋃ best-sense i
step 13:return P
```

Fig. 2

To find this maximum we use an alignment procedure which, for each word type wi in item T1, assigns wi to the sense that has the maximal similarity to any sense of the word types in the compared text.The Algorithm 1 Shows the alignment-base sense disambiguation for checking synonyms similarity.

IV. RESULTS AND DISCUSSION

The resulting point of view, the proposed methods are more effective as compared to the other machine learning approach. The first method is based on weighted overlap approach and second is based on cosine similarity method both are used for the textual similarity, semantic matching, textual entailment. The MSRpair dataset of English language is used for carrying out the experiment. Results are carried out by using the English dataset on the proposed system. Before evaluating performance, preprocessing is done on various sentences and extracted features are stored in a new file.

A. Performance Matrics

Performance Metrics used for finding out the similarity between sentences. The proposed algorithms are designed in such a way that to obtained the positive integer value of word with the help of sentence scoring method. In preprocessing, sentences are separated through stop words and words are separated through space. Each words have score which is calculated through following Equation.

The precision, recall and f-measure values are uses in paraphrase detection.

Precision is defined as the ratio of the number of relevant records retrieved to the addition of the number of relevant records retrieved and number of irrelevant records retrieved.

Recall is defined as the ratio of the number of relevant records retrieved to the addition of the number of relevant records retrieved and number of relevant records not retrieved.

F-Measure is defined as a measure that combines precision and recall is the harmonic mean of precision and recall.

- Precision =TP(TP+TN)
- Recall =TP(TP+FP)
- F-Measure =2 (Precision * Recall)/(Precision + Recall)

The F-Measure calculation for semantic similarity detection, the values of precision and their respective recall values are considered. By applying the formula given in Equation 3, calculates the F-Measure values. The graphs shows is plotted by considering the average F-score values of the both methods given in Table 1. The figure 1 shows the F-measure of the proposed methods are weighted overlap based method and cosine similarity based method.

Table 1: Comparison of Simplified Weighted Overlap Approach with Cosine Similarity Method Based on F-score

No.of Sentences	F-score of Weighted Overlap Based Method	F-score of Cosine Similarity
10	0.62	0.82
20	0.77	0.73
30	0.71	0.80
40	0.57	0.60
50	0.60	0.67
60	0.67	0.67
Average	0.65	0.71

output of both the algorithm. Average F-score of the simplified weighted overlap approach is 0.65 and cosine similarity is 0.71. The proposed system using cosine similarity gives the better F-score value than simplified weighted overlap approach method.

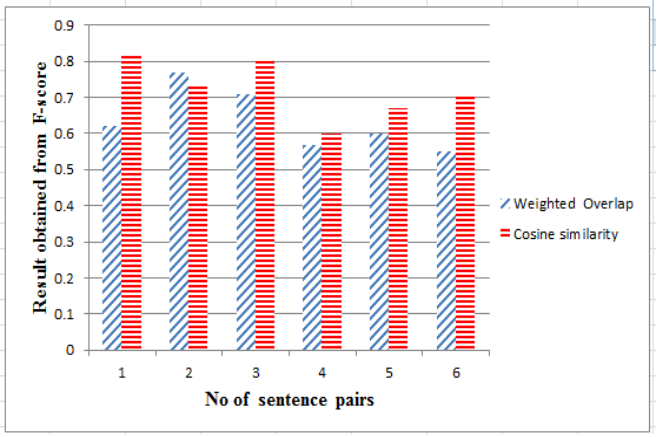

Fig:3 score

V. Discussion

Results of the experiments show the effectiveness of the textual similarity detection from sentences. The textual similarity detection uses the two methods. First is weighted overlap based method and second is cosine similarity based method. The cosine similarity based method provide the better similarity results than weighted overlap based method because it considers synonyms similarity along with cosine similarity.

Conclusion and Future Scope

In natural language processing, sentence similarity performs the tasks such as document summarization, word sense disambiguation, short answer grading and information retrieval. Existing methods used for checking the similarity of long text documents. These methods process sentences in high-dimensional space and are not much efficient, requires human input and also not adaptable to some application domains. The traditional methods are not considered the synonyms similarity. Semantic textual similarity methods improved in two areas, first is the semantic relation between the words and second is semantic resources to reduce dimension and overcome disadvantages of existing methods. Using this proposed methods improved the results of text similarity and also checks the synonyms similarity between the sentences.

As future work, try the similarity detection methods on semantic networks obtained from other collaboratively-constructed resources, such as Wikipedia. The Wikipedia graph is expected to be particularly suitable for measuring the similarity between texts as it provides a remarkable coverage of proper nouns and domain-specific terms.

ACKNOWLEDGMENT

First and foremost, I would especially like to thank my adviser and guide Prof. Dr.K.P.Adhiya, for being a wonderful instructor and always being in bright spirits. I am incredibly grateful to Principal K. S. Wani , Head of the department Prof. Dr. Girish K. Patnaik and the Professors in the M.E. Computer Engineering Program at SSBT's COET,Jalgaon. Without their time and support, I would not have been able to complete this work. Last, but not least, I like to thank my family and friends who also supported me throughout my studies.

References

[1]. Y. Li, D. McLean, Z. A. Bandar, J. D. OShea, and K. Crockett. 2006. "Sentence similarity based on semantic nets and corpus statistics". In the procedding of Transactions on Knowledge and Data Engineering IEEE, Vol.18(8), pp. 1138-1150.

[2]. Y. Liu and C.Q. Zong, "Example-Based Chinese-English Machine Translation", In the proceding of.2004 IEEE Intternational Conf Systems, Man, and Cybernetics, Vol.1-7, 2004, pp.6093-6096.

[3]. LiHong, D. Wang, and M. Huang, "Improved Sentence Similarity Algorithm based on VSM and its application in Question Answering System". Intelligent Computing and Intelligent Systems (ICIS), 2010, In the procedding of International Conference on IEEE, 2010, pp. 368 - 371.

[4]. Eneko Agirre, Daniel Cer, Mona Diab and Gonzalez-Agirre Aitore, 2012. SemEval-2012 task6: "A pilot on Semantic textual Similarity". procedding of First Joint Conference on Lexical and Computational Semanticcs,June 7-8, 2012. pages 385-393.

[5]. Z. Wu and M. Palmer, "Verb semantics and lexical selection", In the Proceedings of 32nd annual Meeting of the Association for Computational Linguistics, (1994) June. IEEE, 2005, pp. 27-30.

[6]. R. M. Courtney Corley, "Measuring the semantic similarity of texts," proceding of in ACL workshop on Empirical Modeling of semantic Equivalence and Entailment (EMSEE),2013. IEEE, 005, pp. 13-18.

[7]. R. RL. Xu, D. Wang, and M. Huang, "Improved Sentence Similarity Algorithm based on VSM and its application in Question Answering System." Intelligent Computing and

Intelligent Systems (ICIS), 2010, procedding of International Conference on IEEE, 2010, pp. 368 - 371.

[8]. Jiang, Jay J., and David W. Conrath. "Semantic similarity based on corpus statistics and lexical taxonomy", In the Proceedings of ROCLING X, Taiwan, 1997, https://arxiv.org/abs/cmp-lg/9709008 accessed on November 7, 2017.

[9]. Z.Jingling , Z. Huiyun , Cui .Baojiang ."Sentence Similarity Based on Semantic Vector Model" In the procedding of Ninth International Conference on P2P, Parallel, Grid, Cloud and Internet Computing 2014, IEEE, pp. 499-503.

[10]. Emiliano Giovannetti, Simone Marchi, and Simonetta Montemagni, "Combining Statistical Techniques and Lexico - syntactic Patterns for Semantic Relations Extraction from Text", In the procedding of sixth international conference on Stastical Technique 2008,IEEE, pp. 399-402.

