An NLP Based Plagiarism Detection Approach for Short Sentences

Shikha Pandey, Arpana Rawal

***Abstract:** The notable issue in the fields of plagiarism detection is, to assess the semantic similarity between obfuscated sentences, and it becomes more completed in case of short sentences (only 4-8 words). An innovative approach, typed dependencies relationship (TDR), based on Natural Language processing is presented for detecting plagiarism on short sentences. In this study proposed approach performed on previous datasets of short sentences and compared results with 3 state-of-art methods. The investigation shows that the proposed calculation has exceptional execution in taking care of sentences with complex linguistic structure.*

***Keywords:** Type Dependencies Relationship, Plagiarism Detection, Sentence Similarity, Syntactic and Semantic Similarity*

I. INTRODUCTION

The present human communication happens as a form of short text scraps of composed content like News features, messages and tweets- although the length of this short content is small but the meaning and utilization is broad, crossing an range of areas and NLP applications. Investigation and analysis of such short-information can uncover data that is essential, in various regions of present day human life. We cannot also ignore it in the field of education and research. In research, the short text or sentence plays an import role and smartly modified by the plagiarist without crediting the original source. The detection of plagiarism in short text is a very complicated task because short text consist only 4-8 words with 50% of syntactical part, which is also import and cannot be ignored by pre-processing task. Detection of the plagiarism between words, sentences, paragraphs and documents is an important component in plagiarism detection process. Finding matching between words is initial process of plagiarism detection system which is then used as a primary stage for sentence, paragraph and document detection. So the objective is to implement a viable technique to compute the similarity between short messages, more often than not around one sentence length. These processed sentence similarity could be helpful for plagiarism detection tools.

This paper is organized as follows: Section 2 presents related work and methods for measuring fundamental similarity on short text. Sections 3 introduce proposed approach for plagiarism detection based on type dependency relationship model with illustrating one example. Section 4 presents the experiment and results from the proposed approach with Li (2006) data sets, and discuss our results with the results obtained from different state-of-the-art baselines. Finally, section 5 draws some conclusions on this work and outline possible future research in this area.

Revised Manuscript Received on 30 September 2018.
***** Correspondence Author

Shikha Pandey*, Department of CSE, Chhattisgarh Swami Vivekanand Technical University, Bhilai, Chhattisgarh, India

Dr. Arpana Rawal, Department of CSE, Chhattisgarh Swami Vivekanand Technical University, Bhilai, Chhattisgarh, India

II. RELATED WORK

A wide literature and increasingly approaches based on pre-processing techniques are available for measuring similarity on text [1]. Text similarity measure can be done from two ways: Lexical similarity and Semantic similarity; Lexical similarity measures uses String-Based algorithms which are further uses character based: Longest Common Substring (LCS) is based on the length of both strings. Damerau-Levenshtein (2,3) , Jaro (4,5) , Winkler (6), Needleman-Wunsch (7), Smith-Waterman (8), N-gram(9). Term based similarity measures are: Block Distance [10], Cosine similarity, Dice's coefficient [11], Euclidean distance Jaccard similarity [12], Matching Coefficient, Overlap coefficient.

Semantic similarity is introduced through Corpus-Based and Knowledge-Based algorithms.

Corpus-Based semantic similarity measure are: Hyperspace Analogue to Language (HAL) [13,14] Latent Semantic Analysis (LSA) [15], Generalized Latent Semantic Analysis (GLSA) [16], Explicit Semantic Analysis (ESA) [17], The cross-language explicit semantic analysis (CL-ESA) [18], Point wise Mutual Information - Information Retrieval (PMI-IR) [19].Knowledge-Based semantic similarity the most popular measures that is based on identifying the degree of relatedness between words using information derived from semantic networks [20]. WordNet [21]] introduced by Miller in 1990, is the most prominent evaluation for plagiarism detection system to detection of semantically similar word. WordNet changed the dimension of research and new approaches were introduced for measuring semantic similarity: Resnik (res) [22], Lin (lin) [23] and Jiang & Conrath (jcn) [24]. The other three measures are based on path length: Leacock & Chodorow (lch) [25], Wu & Palmer (wup) [26] and Path Length (path).

Above mentioned techniques are fundamental and soul of all present approaches which is based on lexical and semantic similarity on sentence.

A strategy for estimating the semantic likeness between short sentences or short messages, based on semantic and word order arrangement was introduced by Lee(2006) and named it Semantic Text Similarity (STS). Li firstly, begins comparison with semantic similarity checking, extract information from knowledge base and corpus statistics. Secondly, proposed a strategy to consider the effect of word order on sentence.

STS strategy accomplished a decent Pearson correlation coefficient for 30 pair of sentences and beat the outcomes [27]. Ercan in 2013 also worked on (Li, 2006) data sets and used methodology which is almost similar with STS (Li, 2006) method [28]. A remarkable work done by (Alzahrani's et al 2015) on detecting highly obfuscated plagiarized texts gained the maximum popularity that was based on fuzzy semantic-based comparability model and compared the result with (Li,2006)[27] and (Lee, 2011)[29]. Alzahrani's model can work on both short and moderate length of sentence. This model outperformed well as compared to five baseline (word-to-word and sentence-based) approaches [30].

III. PROPOSED SYSTEM FRAMEWORK:

A text is thought to be a grouping of words, each of which conveys important data. The word along with their structure, word order and relations with other participating words in sentence, shows specific meaning. The proposed approach detected similarity from syntactic as well as semantic information contained in the compared texts.

Fig.1 shows the proposed system framework which is divided into two similarity computing measures between two sentences- Syntactic similarity and Semantic similarity, which are described as follows:

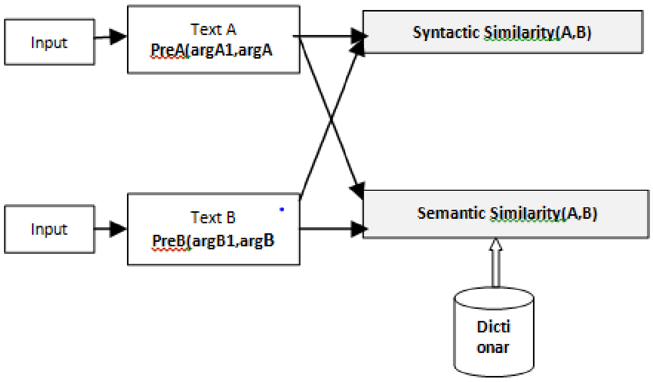

Figure 1: Proposed Sentence Similarity System Framework

3.1 Type Dependency Relationship (TDR):

Previous work (Li 2006[27], Ercan in 2013[28] and Alhzarani in 2015 [30]) worked on semantic meaning of word and word order of sentence for sentence comparison. Most of the previous methods applied text pre-processing techniques like removing stop words, stemming, lemmatization etc, which lost syntactic information of sentence which is also more import in case of short sentence. Proposed approach have not applied any text pre-processing techniques, although, it uses parse dependency relationship from Stanford University typed dependencies representation which was designed to provide a simple description of the grammatical relationships in a sentence that is easily understandable and easily used by people who have no knowledge about word relation in a sentence. So along with word meaning, proposed approach also added relationships between words in a sentence. For example- select any pair of sentence and found the typed dependency parse relationship from Stanford University [31]. Suppose first sentence is A and second sentence is B then the type dependency relationship fetched from Stanford University manual is in form of:

PredicateA (argumentA1, argumentA2)
PredicateB (argumentB1, argumentB2)

Where PredicateA shows a relationship between argumentA1 and argumentA2 in sentence A (an arguments are participating words in any sentence). Also PredicateB defines the relationships between argumentsB1and argumentB2 in a sentence B. The current representation contains approximately 50 grammatical relations defined by Schuster and Manning in 2016 [32]. Now, typed dependency structures of sentential pairs are checked for syntactic and semantic similarity denoted by simrel (A, B), such that

$$Sim_{rel}\ (A,B) = \frac{1*|S(A,n)\cap(B,n)|+0.67*|S(A,n)\cap(B,n)|+0.33|S(A,n)\cap(B,n)|}{\min(countA.countB)} \qquad (1)$$

Where S(A) and A(B) represent the typed dependency relationships of paired sentence. Their intersection represents set of common relationships with n numbers and countA and countB shows the total number of relationships found from Stanford type dependency parser. Similarity measures with different overlapping factors like complete overlapping, partial overlapping and minimum overlapping, observed upon the matched typed-dependency relations extracted out of the input sentence pairs. It should be noted that expression 1 is same for both syntactic and semantic calculation, in semantic calculation synonym of arguments are compared between sentences.

3.2 .Concepts based dictionary construction

Every research article has their own theme and idea and meaning of any words which is used in article may be different in meaning from meaning depicting in WordNet. WordNet is a substantial lexical database of English language, created at Princeton University by a gathering drove by George A. Miller [21]. WordNet dictionary changed the direction of research and it plays a very import role in the area of intelligent and idea plagiarism detection because it contains synonyms, antonyms and other important information of a word, with the help of this information semantic similarity detection is possible. It is freely and publically available for researcher (Simpson T., 2005)[33]. But due to limited vocabulary in WordNet dictionary here we extend the WordNet dictionary according to our input text for that extracted all possible vocabulary words from input texts and try to fetch the synonyms of each word from WordNet, at any situation, synonyms found or not found proposed module will ask to user, "Do you want to add more meaning for this word ", and we can append more synonyms according to theme. Proposed approach used this dictionary for semantic calculation.

3.3 Illustrative Example- Detailed execution of proposed method

To illustrate how proposed method works for a pair of sentence, below we explain detail description of our method with taking a sentence pair from our study. In this study we optimized the majority of the computations in Python 3.6 (32 bit) framework,

so as to make the usages as simple as possible for the end client. For Example:

Sentence1-I like that bachelor.

Sentence 2 - I like that unmarried man.

Above sentences looks like exact same from the view of human estimation. Proposed method produced two output similarity measures: syntactic similarity (without using dictionary) and semantic similarity (with using dictionary) which is close to human estimation. Now perform following steps for checking sentence similarity.

Step 1: Fetch the Type dependency relationship from Stanford University [32]. Total no of relationship fetched for sentence A and B is 4 and 5 as shown below:

TDR of sentence A	**TDR of sentence B**
nsubj(like-2,I-1)	*nsubj(like-2,I-1)*
root(ROOT-0,like-2)	*root(ROOT-0,like-2)*
det(bachelor-4,that-3)	*det(man-5,that-3)*
dobj(like-2,bachelor-4)	*amod(man-5,unmarried-4)*
	dobj(like-2,man-5)

Step2. Type dependency relationship of sentence A should be matched with Type dependency relationship of sentence B in the following manner:

PredicateA == PredicateB

ArgumentA1 == ArgumentB1

ArgumentA1== ArgumentB2

Step3. Now Search overlapping of relationships for 100%: In 100% overlapping, there should be one-to-one matching between TDR relationships of both sentences. In 67% overlapping predicateA should be matched with predicateB with one of the its similar argument or if predicates are not similar but both argumentA1, argumentA2 is similar with argumentB1,argumentB2 then it also comes in the categories of 67% overlapping. similarly, In 33% overlapping any one of the argument of sentence A is matched with argument of sentence B with same word position.

Step3: Now the syntactic similarity and semantic similarity can be formalised from the expression (1) and categories them according to their overlapping. The syntactic overlapping similarity in sentence A and B can be formularized as follows:

Syntactic Similarity		
100 %Match	67%Match	33%Match

nsubj(like, I)
<>
nsubj(like, I)
root(ROOT,like)
<>
root(ROOT,like)
	det(bachelor,that)
	<>
	det(man,that)
	dobj(like,bachelor)
	<>
	dobj(like, man)
		nsubj(like, I)
		<>
		dobj(like, man)

Now the semantic overlapping similarity in sentence A and B by using self made concept based dictionary can be formularized as follows:

Semantic Similarity		
100 %Match	67%Match	33%Match

nsubj(like, I)
<>
nsubj(like, I)
root(ROOT, like)
<>
root(ROOT, like)
	det(bachelor, that)
	<>
	det(man, that)
	dobj(like, bachelor)
	<>
	dobj(like, man)
		nsubj(like, I)
		<>
		root(ROOT, like)
		nsubj(like, I)
		<>
		dobj(like, man)

Execution: From the table of syntactic similarity, number of relationships different matching categories are 2, 2,1(100% match,67% match and 33%match) respectively. Our method produced syntactic similarity is .9175% (2*1+2*(.67)+1*(.33)/4) which is better than previous work(Li2006,Ercan2013,Alhzarani2015), again if we refine our method and uses our concept based dictionary for calculating semantic similarity then result is 100% (2*1+2*(.67)+2*(.33)/4), Here we put, the unmarried is synonym of bachelor. The proposed technique gives a generally high closeness. This case exhibits that the proposed technique can catch the significance of the sentence despite the co-event of words.

IV. EXPERIMENTS AND RESULTS

Currently, there are no appropriate benchmark data sets available for the assessment of proposed sentence (contains 4-8 words) similarity method. In spite of the fact, a couple of close studies have been published timely by researcher. So this research performed on Li (2006) data sets which are borrowed from different papers and books on natural language understanding. Table 1 show eight sentence pairs chosen from Alhzarani (2015) experimental data set, here we set threshold value 1.67 for syntactic similarity and .25 for semantic similarity, if it is greater than this then acceptable otherwise it should be 0. Results shown in Table 1 stated that proposed computed syntactic similarity values were found to be fairly consistent with previous values and semantic similarity values were found is more close to the human intuition and more better than previous methods.

V.CONCLUSION

This paper presented a two practical sentence similarity evaluation approach. Firstly, a syntactic similarity approach is only based on type dependency relationships without using the concept based corpus (self made dictionary). Secondly, semantic similarities approach with using both TDR and dictionary. Our approach tackled the issue of text pre processing, as we accept sentence without changing their structure or loss of any information.

It is observed that both proposed approach worked well for short length sentence with good time complexity. As a future work we will try this extended approach for moderate length text.

Table 1: Experimental Result on Raw Sentences of Short Lengths

	Sentence 1	Sentence 2	Li (2006)	Canvas 2013	Alhzarani (2015)	Syntactic Similarity (TDR)	Semantic Similarity (TDR) with explanation
Pair A	I like that bachelor	I like that unmarried man	0.561	0.558	0.649	**.9175%**	**1.00**
Pair B	I have a pen	Where do you leave	0.000	0.277	0.000	**0.000**	**0.000** Because we are considering greater than the threshold.
Pair C	John is very nice	Is john very nice	0.977	0.599	1.00	**.833**	**.833** here the strings are exact similar but the type of sentence is totally different so exact match should not be 100%
Pair D	It is a dog	It is a log	0.623	0.182	0.737	**.670**	**.670** Here dog & log cannot be matched anyhow. So matching should be less.
Pair E	It is a red alcoholic drink	It is an dictionary	0.000	0.000	0.074	**0.000**	**0.000** Because we are considering greater than the threshold.
Pair F	Canis familiaris are animals	They are common pets	0.362	0.806	.391	**.335**	**.832** Considering pets are animals & put it as a synonym of animal in our dictionary.
Pair G	It is a glass of cider	It is a full cup of apple juice	0.678	0.253	0.652	**.558**	**.723** By putting synonym (Cider) is juice and synonym (glass) is cup.
Pair H	Dogs are animals	They are common pets	0.738	0.756	0.494	**.446**	**1.00** By putting synonym (Animal) is pet.

REFERENCES

1. Gomaa Wael H. & Fahmy A. (2013) A Survey of Text Similarity Approaches, International Journal of Computer Applications (0975 – 8887)
2. Hall, P. A. V. & Dowling, G. R. (1980) Approximate string matching, Comput. Surveys, 12:381-402.
3. Peterson, J. L. (1980). Computer programs for detecting and correcting spelling errors, Comm. Assoc. Comput. Mach., 23:676-687.
4. Jaro, M. A. (1989). Advances in record linkage methodology as applied to the 1985 census of Tampa Florida, Journal of the American Statistical Society, vol. 84, 406, pp 414-420.
5. Jaro, M. A. (1995). Probabilistic linkage of large public health data file, Statistics in Medicine 14 (5-7), 491-8.
6. Winkler W. E. (1990). String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage, Proceedings of the Section on Survey Research Methods, American Statistical Association, 354–359.
7. Needleman, B. S. & Wunsch, D. C.(1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins", Journal of Molecular Biology 48(3): 443–53.
8. Smith, F. T. & Waterman, S. M. (1981). Identification of Common Molecular Sub sequences, Journal of Molecular Biology 147: 195–197.
9. Alberto, B., Paolo, R., Eneko A. & Gorka L. (2010). Plagiarism Detection across Distant Language Pairs, In Proceedings of the 23rd International Conference on Computational Linguistics, pages 37–45.
10. Eugene F. K. (1987). Taxicab Geometry, Dover. ISBN 0-486-25202-7.
11. Dice, L. (1945). Measures of the amount of ecologic association between species. Ecology, 26(3).
12. Jaccard, P. (1901). Étude comparative de la distribution florale dans une portion des Alpes et des Jura. Bulletin de la Société Vaudoise des Sciences Naturelles 37, 547-579.
13. Lund, K., Burgess, C. & Atchley, R. A. (1995). Semantic and associative priming in a high-dimensional semantic space. Cognitive Science Proceedings (LEA), 660-665.
14. Lund, K. & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. Behavior Research Methods, Instruments & Computers, 28(2),203-208.
15. Landauer, T.K. & Dumais, S.T. (1997). A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge", Psychological Review, 104.
16. Matveeva, I., Levow, G., Farahat, A. & Royer, C. (2005). Generalized latent semantic analysis for term representation. In Proc. of RANLP.
17. Gabrilovich E. & Markovitch, S. (2007). Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis, Proceedings of the 20th International Joint Conference on Artificial Intelligence, pages 6–12.
18. Martin, P., Benno, S. & Maik, A.(2008). A Wikipedia-based multilingual retrieval model. Proceedings of the 30th European Conference on IR Research (ECIR), pp. 522-530.
19. Turney, P. (2001). Mining the web for synonyms: PMI-IR versus LSA on TOEFL. In Proceedings of the Twelfth European Conference on Machine Learning (ECML).
20. Mihalcea, R., Corley, C. & Strapparava, C. (2006). Corpus based and knowledge-based measures of text semantic similarity. In Proceedings of the American Association for Artificial Intelligence.(Boston, MA).
21. Miller, G.A., Beckwith, R., Fellbaum, C.D., Gross, D. & Miller, K. (1990). WordNet: An online lexical database. Int. J. Lexicograph. 3, 4, pp. 235–244.
22. Resnik, R. (1995). Using information content to evaluate semantic similarity. In Proceedings of the 14th International Joint Conference on Artificial Intelligence, Montreal, Canada.
23. Lin, D. (1998b). Extracting Collocations from Text Corpora. In Workshop on Computational Terminology , Montreal, Kanada, 57–63.
24. Jiang, J. & Conrath, D. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. In Proceedings of the International Conference on Research in Computational Linguistics, Taiwan.

25. Leacock, C. & Chodorow, M. (1998). Combining local context and WordNet sense similarity for word sense identification. In WordNet, An Electronic Lexical Database. The MIT Press.
26. Wu, Z.& Palmer, M. (1994). Verb semantics and lexical selection. In Proceedings of the 32nd Annual Meeting of International Journal of Computer Applications (0975 – 8887) Volume 68– No.13, April 2013 18 the Association for Computational Linguistics, Las Cruces, New Mexico.
27. Li, Y., McLean, D., Bandar, Z.A., O'Shea, J.D., Crockett, K., 2006. Sentence similarity based on semantic nets and corpus statistics. IEEE Trans. Knowledge Data Eng. 18, 1138–1150.
28. Ercan Canhasi 2013, measuring the sentence level similarity, ISCIM 2013, pp. 35-42 © 2013 Authors
29. Lee, M.C., 2011. A novel sentence similarity measure for semanticbased expert systems. Expert Syst. Appl. 38, 6392–6399.
30. Alzahrani, S. M., Naomie Salim, Vasile Palade(2015) ,:Uncovering highly obfuscated plagiarism cases using fuzzy semantic-based similarity model, Journal of King Saud University – Computer and Information Sciences 27, 248–268(2015).
31. Marie-Catherine de, Marneffe, Bill MacCartney, Christopher D. Manning . In LREC (2006).
32. Sebastian Schuster and Christopher D. Manning2016. In LREC 2016.