A Novel Approach For Syntactic Similarity Between Two Short Text

Anterpreet Kaur

ABSTRACT: Syntactic similarity is an important activity in the area of high field of text documents, data mining, natural language processing, information retrieval. Natural language processing (NLP) is the intelligent machine where its ability is to translate the text into natural language such as English and other computer language such as c++. Web mining used for task such as document clustering, community mining etc to performed on web. However to find the similarity between the two documents is the difficult task. So with increasing scope in NLP require technique for dealing with many aspects of language, in particular, syntax, semantics and paradigms.

Keywords: Syntactic, similarity, natural language processing, semantic word distance, snippets, stopwords.

1Introduction

Measuring syntactic similarity between words, short text in the area of data mining plays an important role. In the field of data mining syntactic similarity is exploited in application like cleansing data for mining and warehousing, duplicate detection, mining knowledge from text etc. The problem of measuring of similarity between two short segments has become increasingly important for many tasks. Task such as: similarity between two queries, similarity between the user's query and advertiser's keywords, similarity between the given product name and suggested words, similarity between the question paper. Similarity is the complex concept which has been widely discussed in the linguistic, philosophical and information theory communities. Similarity means that to find relevant meaning of the given sentence or the verb and identify the accuracy between them. The main objective to find the similarity is that to identify the repeated questions in the question paper (a.k.a automatic question paper vetting) and try to reduce this problem with the help of the NLP or machine learning technique. Frequently asked question (FAQ) is a question answer retrieval system which finds the question sentence from question- answer collection and then returns its corresponding answer to the users. The task of matching questions to corresponding questions-answer pairs has become a major challenge in a FAQ system. In [6] Zhong Min Juan proposed a method to find matching system in the user query and question in FAQ corpus. Combining semantic and statistical techniques, an effective similarity method is proposed, which firstly build semantic knowledge base, namely, co-occurrence word corpus, then count term frequency of question sentence by using statistic method. In earlier, the work is on a syntactic approach [1] for searching similarities within sentence. This paper proposes a solution based on a purely syntactic approach for searching similarities with sentence, named sub sequence matching.

- *Anterpreet kaur, masters degree program in computer science engineering in Lovely professional university,India, 9779344622. Anterlp@gmail.com.*

Some approaches to find similarity of text is computed as a function of the number of matching tokens or sequence of token they contain. However they fail to identify similarities when the same meaning is conveyed using synonymous terms or phases. Example: "The Dog sat on the mat" and "The Hound sat on the mat." Or when the meanings of the text are similar but not identical. Example: "The Cat sat on the chair" and "The Dog sat on the mat The remaining portion of paper is organized as: Next section is described the background study of papers which study, in next proposed work with formulation ,in other results and discussion and atlast conclusion.

2 BACKGROUND STUDY

The R. Menaha and G. Anupriya [1] present the semantic similarity between words using the semantic word distance and snippets technique. SWD measure the frequency of the word in each document and normalizes it over all document. The page count measure can also be used to find semantic similarity but it does not indicate the number of times a word has occurred in each of this page. A word may appear many times in a document and once in another document, but the page count measure can ignore this type of condition. So the page count measure is not sufficient to measure the semantic relation between two words. SWD considers only the global context of a given words in web pages and it doesn't give importance to the semantic relationship that exit between the word pairs. Therefore snippets are used for finding semantic similarity in local context. In earlier [2002] Federica Mandreoli presents a method which is based on syntactic information for searching similarities within sentences, named estimated sub sequence matching. This method assume a sentence as a series of terms and characterizes the problem of approximate matching between sentences as a problem of searching for similar sequences equivalent to the whole sentences or parts of them. These methods are successful in some fields, but not for others. The disadvantage for using these types of approaches is that they focus only on part features, may be one features, may be two features, but not all particular features.

3 PROPOSED METHOD

Syntactic similarity of sentences is based on to measure the similarity of the given words. If two sentences are similar then structural relations between words will be similar and vice versa. To measure the syntactic similarity between the two documents is not more difficult work, but as the deeply

research there is not more work on the syntactic similarity. So I have decided to make an improvement in the syntactic similarity between two papers. There are various algorithms which are help to find the similarity between words, Algorithms such as Edit distance, longest common substring, bi-gram algorithm and Soundex algorithm. But in these algorithms there is some problem to find syntactic similarity between words. Those approaches don't work on the some conditions. The Soundex Algorithm is a similarity algorithm, which simply defined that given two strings are similar or not. However, it would not describe any similarity between 'FRENCH' and 'REPUBLIC OF FRANCH', because they don't start with the same letters they started with different letters. On the other hand the Edit Distance algorithm would distinguish some better result than the Soundex algorithm between the two strings, but would rate 'FRANCE' and 'FRENCH' (with a distance of 6) to be more similar than 'FRENCH' and 'REPUBLIC OF FRENCH'. And at last The Longest Common Substring would give 'FRENCH' and 'REPUBLIC OF FRENCH' having a good rating of similarity(a common substring of length 6). However, it is undesirable that according to new approach, the string 'FRENCH REPUBLIC' is equally similar to the two strings 'REPUBLIC OF FRANCE' and 'REPUBLIC OF CUBA'. Having to seen the drawbacks of the existing algorithms, the proposed approach is new string similarity metric that doesn't matter on the ordering method. In addition, its decided to present a new approach which not only considers the single longest common substring, but also other common substrings too. If the two strings are pronounced same then the similarity of that string are usually high, but there is difference in both of that strings, so it doesn't mean that there is not similarity between that words. So Firstly its decided to check that how many adjacent characters are contained in both the strings. Let clear this statement by taking the algorithm:

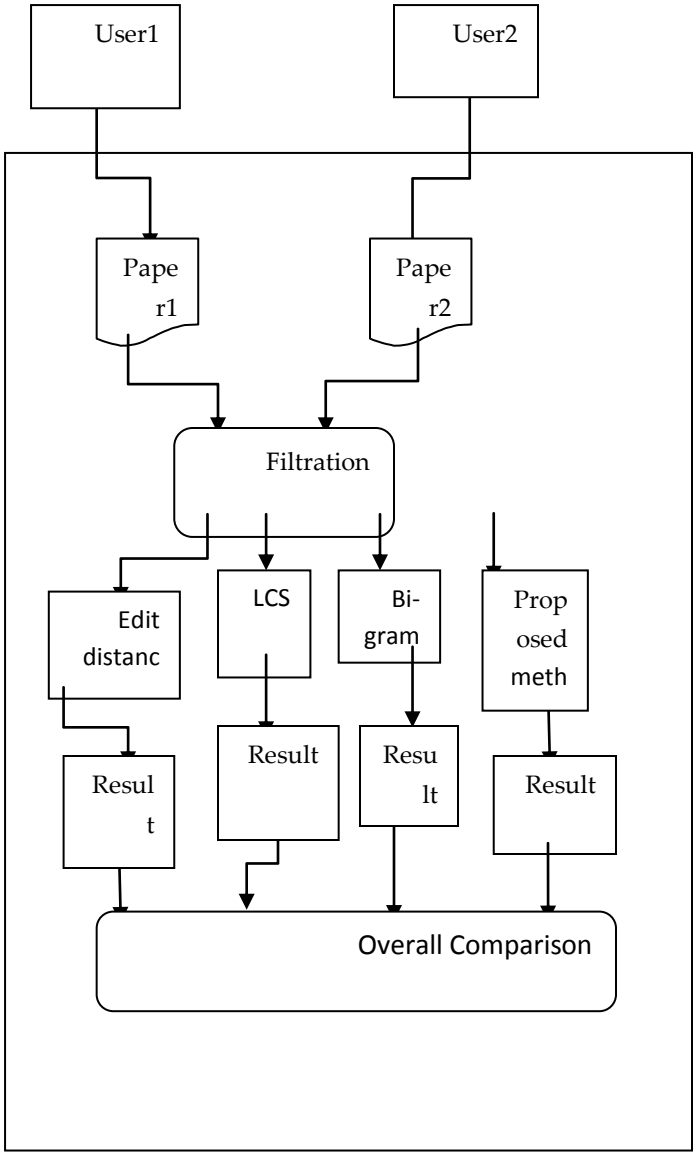
1) Firstly take the two strings in which its decided to find the similarity between them.
   Example:
   SYNTACTIC
   SEMANTIC
2) Then map them both to their upper case characters and then decided to split them up into their character pairs. Example of such statement is that:
   SYNTACTIC: {SY, YN, NT, TA, AC, CT, TI, IC}
   SEMENTIC: {SE, EM, MA, EN, NT, TI, IC}
1) Then I check out which character pairs are in both strings. So in the given example, the intersection is {TI, IC}.
2) At last, I would like to explain the way of finding the similarity as a mathematically which reflects the size of the intersection comparative to the sizes of the given strings.

$$\text{Similarity (S1, S2)} = \frac{2 \times |character\ (C1) \cap character(c2)|}{|character\ (c1)| + |character\ (c2)|}$$

This new algorithm is also work in the following on the following requirements:

A true indication of lexical similarity:
This means that two string or the words which have the small differences should be accepted as similar. It means that a considerable two string which have common characteristics should point to a gave a high level of similarity between the strings.

1. **Its not possible to changes of word order**: The given two strings which contain the same words in the given documents, but they are in a different order, should be re-nowned as being similar. On the other hand, the given two documents should be renowned as dissimilar, if one string is just a same anagram of the characters contained in the other document.
2. **Language Independence** - This algorithm should also work on many different languages not easily only in English, and gave a better result to find the similarity between two documents.



*[**Fig-1:** Schematic diagram of semantic similarity measurement of two Answers]*

The similarity between two given strings s1 and s2 is twice the number of character pairs that are common to both strings is divided by the sum of the number of character pairs in the two strings. Note that the formula rates completely dissimilar strings with a similarity value of 0, since the size of the letter-pair intersection in the numerator of the fraction will be zero.

On the other hand, if you compare a (non-empty) string to itself , then the similarity is 1. For our comparison of 'SYNTACTIC' and 'SEMANTIC', the metric is computed as follows:

Similarity

$$Syntactic, Semantic) = \frac{2\times|\{TI,IC\}|}{|\{SY,YN,NT,TA,AC,CT,TI,IC\}|+|\{SE,EM,MA,AN,NT,TI,IC\}|}$$

$$= \frac{2\times2}{8+7}$$

$$=0.27$$

Given that the values of the metric always lie between 0 and 1, it is also very natural to express these values as percentages. For example, the similarity between 'SYNTACTIC' and 'SEMENTIC' is 27%. From now on, we will express similarity values as percentages, rounded to the nearest whole number. Suppose we don't want to know how similar two strings are? But want to know which of the string is more similar to the given string. Suppose the given string is "SEALED" and check that which of the strings is most similar to given string?

TABLE 1
RESULT RANK

Word	Similarity
Dealed	80%
Healthy	36%
Heard	22%
Herald	20%
Hold	0%

Find the Most Similar Word to 'Healed.

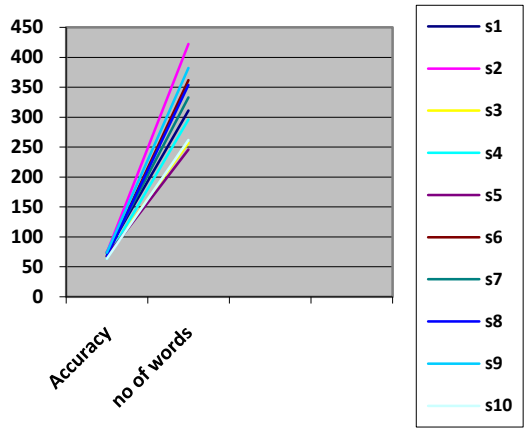
5 DATA SET

For experimental result, collect a set of questions from different resources. In the data set, length of questions is from 10 to 15, which helps to measure the similarity between the questions. Here questions are chosen with minimum and maximum length size because focus in this research is on to measure the similarity between questions. A user has randomly chosen the questions and the accuracy of similarity is stored on database. Following table1 describes the data sample which has been used. Data set has been taken in limited field which includes different kinds of questions related to computer science.

TABLE-2
ACCURACY OF PAPER

	Sample	No of words	no of questions	Accracy
S1	311	15		73.86%
S2	422	15		73.36%
S3	257	15		68.22%
S4	296	15		63.2%
S5	246	15		67.5%
S6	362	15		66.91%
S7	333	15		71.86%
S8	354	15		67.62%
S9	382	15		73.74%
S10	262	15		63.36%

The overall accuracy of my project is 70%. In the given data-set we take 10 samples of the question paper and in one sample, two set of questions has been taken. Then with our proposed method we show the accuracy between the two set of question paper.



It represent the graphically presentation of the ten sample of the question paper.

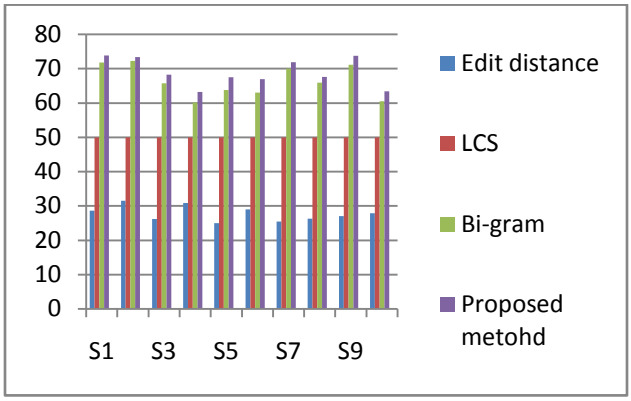
6 RESUTS AND DISCUSSIONS

In order to check the accuracy and simplicity and to evaluate the performance of proposed system ten samples of questions sets are used which are presented in table1. Following Table3 shows the results of proposed method comparative to the different algorithms LCS, Edit distance and Bi-gram algorithms.. The evaluation results shows that the similarity based on proposed method has the better performance than the existing algorithms. For the comparative analysis the same data set is used on proposed algorithm and a table is created which is shown in table-3 and further graph is been plotted which shows the considerable amount of improvements in accuracy.

TABLE 3
ACCURACY OF PROPOSED METHOD

Sample	Edit distance	LCS	Bi-gram	proposed method
S1	28.6%	50%	71.84%	73.86%
S2	31.53%	50%	72.22%	73.36%
S3	26.24%	50%	65.72%	68.22%
S4	30.83%	50%	60.13%	63.2%
S5	24.96%	50%	63.76%	67.5%
S6	28.97%	50%	63.02%	66.91%
S7	25.45%	50%	70.01%	71.86%
S8	26.26%	50%	65.93%	67.62%
S9	27.08%	50%	71.12%	73.74%
S10	27.86%	50%	60.48%	63.36%

In the above table comparison has been done for the same ten samples with apply the different algorithms but the table values indicate the considerable improvement in proposed method.



From the given graph its simply show that accuracy of the proposed method is high comparative to the other's algorithm.

7. CONCLUSION

From all the review it is clear that there is no more work on the syntactic similarity between two short segments, so it is decided to work on to measures the similarity between questions in two question papers (aka automated question vetting).It may happen many times that in two sections a similar question can be occurred or it may also be happened that the questions are related to each other. So to ignore this type of problem we proposed a method in which our system may know the similar questions in two papers and find that questions so that the possibility of relevant questions are decreased in the future time. The future work is on to improve the approaches to measure the syntactic similarity between two short texts. In the data mining field the more work is based on the semantic similarity between short texts. But the result is not more satisfied. The average accuracy of repeated words in the two questions paper is 70%. So in future work, with the help of the NLP it should be increased by using the different methods.

ACKNOWLEDGMENT

I would like to take this opportunity to express my deep sense of gratitude to all who helped me directly or indirectly during thesis work. Firstly, I would like to thank my supervisor Ms Sukhbir kaurfor being great mentor best adviser I could ever have. Her advice, encouragement and critics are so innovative ideas, inspiration and cause behind the successful completion of this dissertation. I am highly obliged to all faculty members of computer science and engineering department for their support and encouragement. I would like to express my sincere appreciation and gratitude towards my friends for their encouragement, consistent support and invaluable suggestions at the time I needed the most. I am grateful to my family for their love, support and prayer.

REFERENCE

[1] R. Menaha and G. Anupriya, "Semantic similarity between words using SWD and snippets," International conference on current trends in advanced computing, 2013.

[2] Manasa.Ch and V. Ramana, "Measuring semantic similarity between words using page counts and snippets," Manasa ch et al, International journal of computer science & communication network, vol 2(4), 553-558

[3] Yi Liu and Qiang Liu, "Sentence similarity computation based on feature set," 13th International conferences on computer support cooperative work in design.

[4] Wenpeng Lu, Jinyong Cheng and Qingbo Yang, " Question answering system based on web," 5th Internatonal conference on intelligent computational technology and automation, 2012.

[5] Vasileios Hatzivassiloglou, Judith L. Klavans and Eleazar Eskin, "Detecting text similarity over short passages: Exploring linguistic feature combinations via Machine learning," unpublished.

[6] Zhong Min Juan, " An effective similarity measurement for FAQ Question Answering system," International conference on electrical and control Engineering,2010.

[7] Federica Mandreoli, Riccardo Martoglia and Paolo Tiberio, "A Syntactic Approach for Searching Similarities within Sentences," unpublished.

[8] Liu, X., Zhou, Y. and Zheng, R., "Measuring Semantic Similarity in Wordnet", Proceeding of ICMLC2007 Conference, Hongkong, 2007.

[9] Linli Li, Xia Hu, Bi-Yun Hu, Jun Wang and Yi-Ming Zhou, "Meassuring Sentence Similarity from diffferent aspects," Eighth International Conference on Machine Learning and Cybernetics,Baoding, 12-15 July 2009.

[10] Ko, Y., Park, J., and Seo, J., "Improving text categorization using the importance of sentences", Information Processingand Management Vol. 40, No.1, pp. 65–79, 2004.

[11] Hongni Dong, Jiang Wu and Xiaohui Zhao, "Study on the calculation of text similarity based on key-sentence," 2010 International Conference on E-Business and E-Government.

[12] Xinxin Zhao, Tiedan Zhu and Yushu Liu, "Document Classification in Different Granularity," Computer Engineering.Vol.32 No.20, p.183-184, Oct 2006(In Chinese)

[13] Abolfazl Keighobadi Lamjiri, Leila Kosseim and Thiruvengadam Radhakrishnan, "Comparing the Contribution of Syntactic and Semantic Features in Closed versus Open Domain Question Answering," International Conference on Semantic Computing.

[14] Takale, S.A. and Nandgaonkar, S.A (2010) 'Measuring semantic similarity between words using web documents', IJASCA-International Journal of Advanced Computer Science and Applications, Vol 1, No.4,pp.78-82.