International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 3 Issue 11, November-2014

A Novel Approach for Similarity Between Two Short Text

Anterpreet Kaur
department of computer science
m.tech
guide: Sukhbir Kaur
lovely professional university
jalandhar, india

***Abstract*—Syntactic similarity plays a significant role in the area of data mining, information retrieval, text mining and natural language processing. In the technology of the computer's environment, it's difficult to find the similarity between two short texts. Natural language processing (NLP) is the intelligent machine, where its ability is to translate the text into the natural language such as English and other computer language such as c++. In text processing, analysis may followed an appropriate translation or a summary of original text. So with increasing scope NLP require technique for dealing with many aspects of language, in particular, syntax, semantics and paradigms. Although related work has been done in this field such as measuring semantic similarity between words using page counts and snippets, using semantic word distance and snippets. This paper presents advancement in syntactic similarity between two questions with the help of data mining technique.**

Keywords- Similarity, natural language processing, semantic word distance, snippets.

I. INTRODUCTION

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 keyword, 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.

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, and then next section is present literature of papers. At last the conclusion of the papers.

II. 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.

III. LITRATURE RIVEW

In [1] r. menaha and g. anupriya present a approach which is proposed to measure the similarity between words are: SWD and snippets. To recover the disadvantages of [2] they proposed a method to measure the similarity between words. Semantic word distance(SWD) helps to find the accuracy of similar word in each document and normalizes it over all documents. Snippets are a programming term for a small region of re-usable source code, machine code and text. It helps to provide information regarding the local context of query term. In fig 1show the outline of a proposed method.

Methodology used:

A. Pattern extraction
B. Pattern clustering
C. Training SVM (Support vector machine)

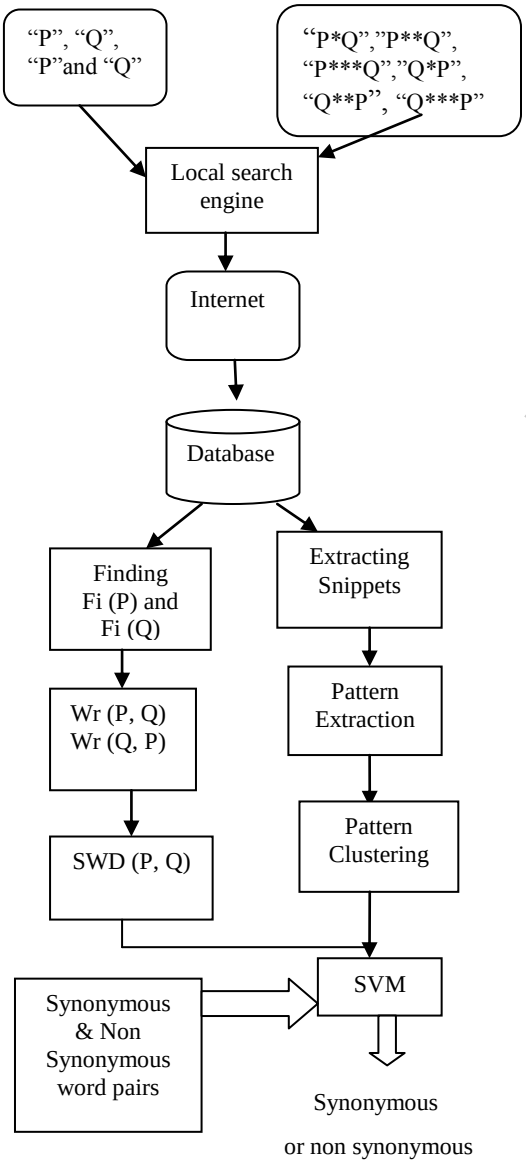

Fig 1. Outline work of developed method

Result of this paper by using google as a search engine to extract a web pages for a given word pair. The clusters score of the word pairs are measured and the SVM is trained to classify either the given word pair as synonyms or non synonyms word pairs. Now in the given table 1, they compare the accuracy of the developed system and the previous system which is page count measure &snippets.

TABLE I.

Performance evaluation	
Method	***Accuracy***
SWD & Snippets	95%
page count measure & snippets	92%

These given methods integrates the SWD and snippets for measuring the similarity and uses SVM as a classifier to classify the given word pairs. But for more effective result the developed system can be applied for query expansion application.

In [2] similarity between the words is also identified by using the lexical dictionary, lexical dictionary such as word net. But the main problem for using the lexical dictionary is that they are not having the recent information of words in various contexts. For instance the word "Apple", in the field of computer science this word have another meaning. It is the name of the company in the hardware as well as software technology. However this word is ignored in the lexical dictionaries, they consider it as a fruit. Many new words are created which have their different meaning and relationships with other words, which are not in the lexical dictionaries.

To overcome this disadvantage a new method is present that automatically finds the semantic similarity between words based on the page count and text snippets from web search engine like Google.

Methodology used:

A. Page count based co-occurrence measures
B. Lexical pattern extraction
C. Lexical pattern clustering

In the case of Page count based co-occurrence, the user can send their input of the two words A and B to the search engine and these words are given to page count by the search engine. The four major word co-occurrence measure jaccard, overlap, dice and point wise mutual information (PMI) are used in proposed work to find the similarity between words.

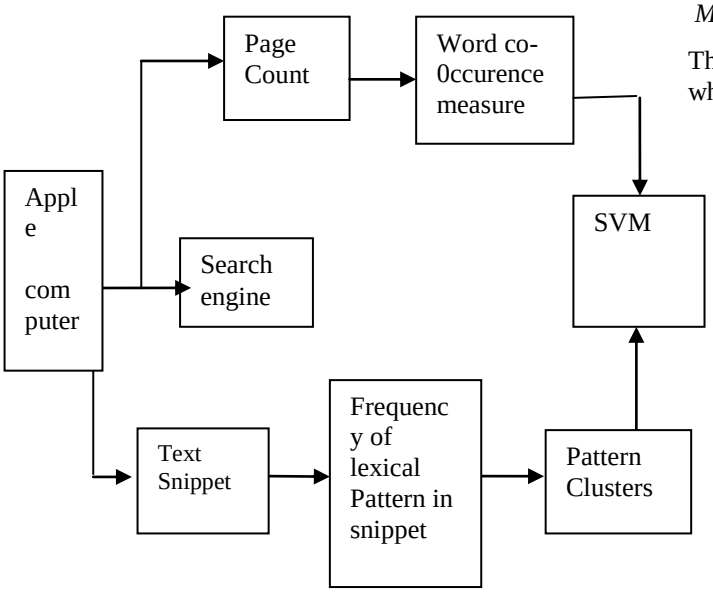

Fig 2 presents the outline of the developed method.

Result:

Using the algorithm like pattern clustering and pattern extraction helps to finding various relationships between words. The experiments are made with synonyms and non synonyms word pair that are collected from the word net synsets. Table 2 show the accuracy of the proposed system and the previous one also which is lexical dictionary.

TABLE II.

Performance evaluation	
Method	***Accuracy***
Page count & snippets	92%
Lexical dictionary	87%

Limitation:

- Using of page count method to measure the similarity between words are not an appropriate solution.
- Because it does not indicate the number of times a word has occurred in each of page.
- A word may appear many times in a document and same word in another document but the page count measure can ignore this type of problem.

In [5] Vasileios Hatzivassiloglou, Judith L. Klavans and Eleazar Eskin, focus on problem that is generated in day today life to detecting whether two small paragraphs contain common information or not. When the large number of text are compared to detect the similarity then the overlap may be sufficient to find similarity; but when the unit of texts are small then simple surface matching of words are used. Their main motive is to recover sets of small textual units from a collection of documents so that each text phrases within a given set describes the same action.

Methodology used:

They present a feature vector over a pair of textual units, where a feature is either primitive or the composite feature.

A. *Primitive feature:* Primitive feature are those that are based on both single words and simplex noun phrases. This feature compares a single word from each text document. It also consisting of one characteristic. So in the primitive feature following methods are presented which matches between text units.

- Word co-occurrence: In this method it is used for sharing of a single word between text documents.
- Matching noun phrases**:** In this method they use a LINKIT tool to identify simplex noun phrases and match those that share same head.
- Word Net synonyms: word net helps to provide common information, placing words in set of synonyms. We match the words which have the same meaning.

B. *Composite features:* In addition to the primitive features, they present a new feature which is called composite feature. Composite features are the combination of primitive features. In this features various methods are used.

- Ordering: In the ordering technique suppose there are two elements A & B. So these two elements have the same order in both textual units. Fig 3 shows the ordering technique. In this example the word "two" in both of text have same order. In both texts it occurs in first order. And the word "contact" in both of text is in the second order.

a) An OH-58 helicopter, carrying a crew of Two, was on a routine training orientation when contact was lost at about 11:30 a.m. Saturday

b) "There were two people on board" said bacon. " we lost radar contact with helicopter about 9:15 EST

- Distance: In the distance method here the distance of both texts will be check. Fig 4 shows the distance technique. In a given example in first text the word "contact and lost" has a distance one. In the second text the word "lost and contact" has a distance one. The distance of both the text has same.

a) An OH-58 helicopter, carrying a crew of two, was on a routine training orientation when contact was lost at about 11:30 a.m. Saturday

b) "There were two people on board," said Bacon. "we lost radar contact with the helicopter about 9:15 EST. "

- Primitive**:** In the primitive feature here we check the words in both the text have the relative match to each other.
  Fig 5 shows the example of primitive.

a) An OH-58 helicopter, carrying a crew of two, was on a routine training orientation when contact was lost at about 11:30 a.m. Saturday.

b) "There were two people on board," said Bacon. "we lost radar contact with the helicopter about 9:15 EST. "

IV. CONCLUSION

From all the literature 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).

TABLE III

Performance evaluation

Year	*Paper tittle*	*Method*	*Accuracy*
2013	Semantic similarity between words using SWD and snippets.	• SWD • Snippets	95%
2010	Semantic similarity between words using page counts and snippets.	• Pattern clustering • Pateern extraction	92%
2012	Detecting text similarity over short passages: Exploring linguistic feature combinations via Machine learning	• Primitive features • Composite features	90%

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 accuracy of repeated words in the document is 9%. So in future work, with the help of the NLP it should be decreased by using the methods.

REFERENCES:

[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.