Volume 6, Issue 5, May 2016 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Online Assessment of Similarity between Sentences in Question Analogous System: A Review Paper

Neha Kumari, Sukhbir Kaur

Computer Science and Engineering, Lovely Professional University,
Punjab, India

Abstract: Sentence similarity play a vital role in the area of Natural Language processing where it has different kinds of benefits like it is used for processing the information, retrieving useful information out of large amounts of data and is also used for question answer portals. Given two sentences, an effective similarity measure should be able to determine whether the sentences are semantically equivalent or not, taking into account the variability of natural language expression. That is, the similarity scheme should be able to calculate similarity even if the sentences do not share similar surface form.There are many methods suggested that are used to calculate the similarity of sentences and they mainly focus on the one or more feature for example word, structure and semantic information etc. In this paper we surveyed different research papers in order to find various techniques used for calculating similarity between two sentences.

Keywords: Similarity, Syntactic similarity, Sentence similarity, Stop Word, NLP

I. INTRODUCTION

Similarity is a quality or state having something common. Similarity mean to check that how two word are similar. Similarity in between the two word, two query, question to question or question answer system. There are many techniques to in which we find the similarity between the two words. These techniques are ordering method, distance method or primitive method.Similarity is concept which has been defined in linguistic, philosophical and information theory communities. Similarity means that to find the relevant meaning of given sentence or the verb and find the accuracy between them. The main objective to find the similarity between the repeated questions in the question paper. To solve this problem is a major task. And to solve with the help of a natural language process NLP or machine learning and find the accuracy between the question similarities. Whenever people talk about words, usually they think about the semantic similarity. Semantic similarity means that the synonyms of the given words and Syntactic similarity is the concept in which the similarity will be measured by word to word. But the main issue to find the similarity is that the some common words which are mostly used in the text such as: “THE, WHAT, A, WHY, IS, ARE” if we can’t ignored these types of words then the similarity percentage will be high. So to ignore these types of words we can use the “STOPWORDS”. Basically in computer stop words are the words which are filtered out or removes the common words. Some tools specifically avoid removing these stop words to support phrase search.For a good performance to measure the similarity we can use the stop words.NLP is an intelligent machine to ability to translate a text into a natural language such as English and others so the increasing scope with the help of NLP.Text Similarity to be used a string based similarity, corpus based similarity or knowledge based similarity. In this paper Section II describes types of similarity, Section III describes review and section IVdescribes conclusion.

II. TYPES OF SENTENCE SIMILARITY

There are three types of the sentence similarity measure:-

1. Statistical measure
2. Semantic measure
3. Syntactic measure

1. Statistical Measure

The statistical similarity measures between sentences are based on symbolic characteristics and structural information. It could measure the similarity between sentences without any prior knowledge but only on the statistical information of sentences. The statistical similarity can be calculated by taking the word counts of the two sentences.

For Example:
Text 1: What is Computer?
Text 2: what is Robotics?

2. Semantic Measure

Two sentences with different symbolic and structure information could convey the same or similar meaning. Semantic similarity of sentences is based on the meanings of the words and the syntax of sentence. If two sentences are similar,

structural relations between words may or may not be similar. Structural relations include relationsbetween words and the distances between words. If the structures of two sentences are similar, they are more possible to convey similar meanings.
For Example:
Text 1: Servant cleans the house
Text 2: Maid cleans the house

3. Syntactic Measure

Syntactical similarity is one part of text analysis, it might be misunderstand what is actually means. The text can have many information hidden into itself. Syntactical structure is one of them. Syntactical means structure of the words and phrases. A common analysis type (much more complex though) is lexical analysis which is analyzing meaning of the text.
Example: -
Text 1: LPU is the good university to study in.
Text 1: Studying in LPU is really good.
Example 2:
Text 1: USA is the great city to live in.
Text 2. USA is great city to live.
These two another phrases are syntactically similar (many words are the same, not all though) and also lexically similar the meaning is almost same

III. REVIEW

Yuhua Li et.al. [1],Introduced about the semantic similarity. It is used in the field of a text mining, information extraction, and dialogue system. In previous similarity is measured in the long text but here similarity is measure in a short text. Sentence similarity, semantic nets, corpus, natural language processing, and word similarity. Firstly semantic similarity is to be derived from lexical data base and corpus Lexical knowledge base is based on the human knowledge about the word in natural language. The corpus reflect the actual use of language and word. We focus not only the common human knowledge but using corpus to application also. Secondly impact of word order on sentence meaning. Different word and number of word pair in a different pair.

Wanpeng Song et.al. [2],proposed a method for calculating the similarity which is calculated using statistic similarity & semantic similarity.Experimental results shows that the similarity calculated by the proposed methods is better than existing methods

MuthukrishananUmamehaswari et.al. [3], proposed a method for calculating the similarity between sentences using semantic based reformulation between two sentences. The experimental work shows that using semantic based reformulation helps to improve the Performance of QA system.

Zhong Min Juan [4],proposed a method in which Word co-occurrence corpus is used to improve its ability to match question and answer. Firstly semantic knowledge base, is built namely, co-occurrence words corpus, then count term frequency of question Sentence by using statistic & semantic methods.

Jun sheng Zhang et.al. [5],proposed two methods first is that the statistical similarity measure between sentences is based on symbolic characteristic and structural information. The second one is that the Sentence similarity based on word set & sentence similarity based on word order capture more local information of sentence pair.

Palakorn Achananuparp et.al. [6], proposed a method that calculates the similarity between the sentences. There are widely applications such as text mining, question answering, and text summarization. Sentence similarity is to be measured using Word overlap measures, simple word and IDF overlap, jaccord method, Phrasal Overlap measures , TF-IDF Measures TF-IDF Vector Similarity and Linguistic MeasuresSentence Semantic SimilarityMeasures word order similarity**,** The Combined Semantic and Syntactic Measures

Prathvi Kumari et.al. [7], proposed a method to find the semantic similarity between the two words. Information available on the web and to use the methods that make use of page counts and snippets to measure the semantic similarity. Various word co -occurrence are defined using the page count and integrated the lexical pattern extracted from the text snippets. Pattern extraction and clustering method are used for a numerous semantic relation between the two words.

Partha Pakray1 et.al. [8], suggested a method of Textual entailment recognize system that are use lexical and syntactic features. TE is a rule based. Textual Entailment is relationship of pairs and text expressions. Entailing “Text” (T) and the entailed “Hypothesis” (H). T entails H if the meaning of H can be inferred from the meaning of T.

Enrique Alfonseca et.al. [9], suggested that the previous system had present time constraint and in complete prototype. So we present the system using the syntactic and semantic similarity to Verify the syntactic analysis for QA and experiment with different semantic distance metrics in view of more complete and integrated future system.

Kai Wang et.al. [10], suggested a method to define the simple question it is based on the syntactic tree structure and solve the problem of similar matching questions. Yahoo answer, question matching, syntactic structure, QA keywords are used for this.

Wael H. Gomaa et.al. [11], proposed a method for text similarity that partitions text similarity into three approaches 1. String based 2. Corpus based 3. Knowledge based similarity. Text similarity is an important as a text related research and applications, such as an information retrieval, document clustering, topic detection, topic tracking etc.

Anterpreet Kaur et.al. [12], proposed that Syntactic similarity is an important area of text document, data mining, and natural language process. Proposed method are to be introduced in which it is not possible to change the word order and languages are independent. To measure the similarity between the questions in two questions paper. But it may be happen that questions relate to each other. So ignore this type of problem we proposed a system in which our system may know the similar question in the paper and find that question.so the possibility of relevant question are decreased in a future.

Ercan Canhasi [13]proposed a method used to calculate the similarity between short English texts, specifically of sentence length. The algorithm calculates semantic and word order similarities of two sentences. In order to do so, it uses a structured lexical knowledge base and statistical information from a corpus. The described method works well in determining sentence similarity for most sentence pairs, consequently the implemented method can be used in computer automated sentence similarity measurements and other text based mining problems.

Zhao jingling ET. al. [14], proposed a new method to compute the sentence similarity which is divided into a three part firstly obtain word semantic similarity second obtain semantic similarity between sentences that is based upon word semantic similarity and structure of sentence finally calculate the word order similarity between sentences and combined the semantic similarity and word order similarity as the final similarity between sentences. To use word similarity methods which is divided into two group corpus based method and dictionary based method.

XIAO-YING-LIU et.al. [15], proposed a method which is used to compare the two application with existing one. Sentence semantic structure to overcome the problem from variability language expressions. Verb arguments pairs represents a sentence instead of frames which are smaller structure of frame.so combined the verb - argument pair and word similarity measure based on Word Net from total sentence similarity eliminating the effect of semantic gap. These two approaches are superior to existing one. In future will carried other applications such as text summarization and question answering.

U.L.D.N Gunasinghe et.al. [16], proposed an algorithm for measuring the sentence similarity. This algorithm is based upon semantic and syntactic measures of sentence similarity. This algorithm takes into account a vector space model for measuring the sentence similarity, the vector space model is generated at the word nodes in the sentence. This algorithm has two phases in first phase we consider relationship between verbs in the sentence and in the other we take relationship between nouns in the sentence

Chi Zhang et.al. [17], proposed a method called sentence selection with semantic representation (SSSR). SSSR uses well developed selection strategy to select summary sentences. The selection strategy used in SSSR is to select sentences that can reconstruct the original document with very less distortion with linear combinations. This model uses two selection strategies weighted mean of word embedding's and deep coding.

Asli Celikyilmaz et.al. [18], proposed two method Latent Dirichlet Allocation (LDA) and Hierarchical LDA (HLDA)Discover the hidden concept and to introduce set of method based upon LDA to find the similarity between question and candidate passage those are used for ranking scores. Result of this paper show that extracting information from hidden concepts improves the results of a classifier – based QA model. Increasing the number of training sample then find the more accurate result and accuracy. In future instead of IBM model 1 plan to study advanced techniques that increase the knowledge and accuracy of the system and also plan to use the translation probabilities learned from the QA Archive for document retrieval experiments.

Rafael Ferreira ET. al. [19], proposed a new sentence similarity measures that solve the problem by taking into lexical, syntactic and semantic analysis of sentences. In previous works WordNet was used to evaluate the semantic word which gives the poor result.so in this paper Semantic Role Annotation (SRA)[20]is used to extract the semantic word and two traditional measure Pearson's correlation coefficient (PCC) and Spearman's rank correlation coefficient (SRCC) is used and gives the better results.

Jehad Q. Odeh et.al. [21],Proposed two algorithms first least frequency character algorithm (FLFC) and recursive based string matching algorithm (RSMA).FLFC is an advanced version of scan for lowest frequency character proposed by horspool [22] SFLFC Proposed algorithms were implemented tested, compared and analyzed with naïve brute force and boyer-moore using a different data set and size. The different algorithms were tasted using the same machine. The result were averaged.RSMA-FLFC algorithm enhance the executions time as compare with brute force and boyar moor. Testing to measure the effectiveness of proposed recursive string matching compared to FLFC without deploying the recursive techniques that applying FLFC is more beneficial if it is merged with recursive matching techniques.

Jiwoon jeon et.al. [23], Proposed a method to used automatic way of building collections of semantically similar question pairs from existing QA collections. After then consider the collections of bilingual and run the IBM machine translation model 1 [24] to learn word translation probabilities.To give a new question, a translation based information retrieval model exploited the word relationship to retrieve similar question from QA archives.Different type of approaches are used to solve the mismatch problems between the questions they are knowledge database [25] which is machine readable dictionary. Employee manual rule and template [26],statistical technique to developed the information retrieval and natural language processing [27].

IV. CONCLUSION

It may happen many times, in a question paper similar question can be occurred that the questions are related to each other. So to ignore this type of problem, we are going to proposed a method in which the developed system may try to find those questions which are similar to question paper, so that the possibility of relevant questions are decreased in the future time. From all the literature review it is clear that there is not much work on the syntactic similarity between two short segments, so improvement is done to measures the similarity between questions in two question papers. From this literature review we are able to discover that syntactic similarity an efficient method to find similarity between two sentences, questions and phrases and in future we are going to develop a system based on syntactic similarity that will calculate similarity between two statements in an efficient way and will also improve the accuracy.

ACKNOWLEDGMENT

I sincerely thank all those people who helped me in completing this comparative analysis.

REFRENCES

[1] Yuhua Li, David McLean, Zuhair A. Bandar, James D. O'Shea, and Keeley Crockett"Sentence Similarity Based Semantic Nets and Corpus Statistics"

[2] Wanpeng Song, Min Feng2 Naijie Gu1, Liu Wenyin "Question Similarity Calculation for FAQ Answering"

[3] MuthukrishananUmamehaswari, MuthukrishnanRamprasath, Shanmugasundaram Hariharan "Improved Question Answering System by semantic refomulation"

[4] Junsheng Zhang, Yunchuan Sun, Huilin Wang, Yanqing He "Calculating Statistical Similarity between Sentences"

[6] Palakorn Achananuparp, Xiaohua Hu, and Shen Xiajiong"The Evaluation of Sentence Similarity Measures"

[7] Prathvi Kumari, Ravi Shankar K"Measuring Semantic Similarity between Words using Page-Count and Pattern Clustering Methods"

[8] Partha Pakray, Sivaji Bandyopadhyay and Alexander Gelbukh" textual entailment using lexical and syntactic similarity"

[9] Enrique Alfonseca, Marco De Boni, José-Luis Jara-Valencia, Suresh Manandhar "A prototype Question Answering system using syntactic and semantic information for answer retrieval"

[10] Kai Wang, Zhaoyan Ming, Tat-Seng Chua "A Syntactic Tree Matching Approach to Finding Similar Questions in Community-based QA Services"

[11] Wael H. Gomaa Aly A. Fahmy"A Survey of Text Similarity Approaches"

[12] Anterpreet Kaur"A Novel Approach For Syntactic Similarity between Two Short Text"

[13] Ercan Canhasi"Measuring the sentence level similarity"

[14] Zhao jingling ET. al, Zhang Huiyun, Cui Baojiang"Sentence Similarity Based on Semantic Vector Model"

[15] xiao-ying liu, chuan-lun ren" Similarity measure based on sentence semantic structure for recognizing paraphrase and entailment"

[16] U.L.D.N Gunasinghe, W.a.m de silva, N.H.N.D de silva, A.S Parera, W.A.D Sashika" Sentence similarity measuring by vector space mode"

[17] Chi Zhang, Lei Zhang, Chong-Jun Wang, Jun-Yuan Xie"Text Summarization Based on Sentence Selection with Semantic Representation"

[18] Asli Celikyilmaz, Dilek Hakkani-Tur, Gokhan Tur "LDA Based Similarity Modeling for Question Answering"

[19] Rafael Ferreira ET. Al" A New Sentence Similarity Method based on a Three-Layer Sentence Representation"

[20] D. Das, N. Schneider, D. Chen, and N. A. Smith, "Probabilistic frame-semantic parsing,"

[21] Jehad Q. Odeh"New and Efficient Recursive-based String Matching Algorithm (RSMA-FLFC)"

[22] R. Nigel Horspool, 1980. Practical Fast Searching in Strings. Journal of Software Practice and Experience

[23] Jiwoon Jeon, W. Bruce Croft and Joon Ho Lee"Finding Similar Questions in Large Question and Answer Archives"

[24. P. F. Brown, V. J. D. Pietra, S. A. D. Pietra, and R. L. Mercer. The mathematics of statistical machine translation:.

[25] A. Berger, R. Caruana, D. Cohn, D. Freitag, and V. Mittal. Bridging the lexical chasm: statistical approaches to answer-finding

[26] E. Sneiders. Automated question answering using question templates that cover the conceptual model of the database.

[27] R. D. Burke, K. J. Hammond, V. A. Kulyukin, S. L. Lytinen, N. Tomuro, and S. Schoenberg. Question answering from frequently asked question files

© 2016, IJARCSSE All Rights Reserved