ISSN:0975-9646

A Novel Approach of Syntactic Similarity of Question Analogous System

Neha kumari [1] Sukhbir Kaur [2]
[1,2] *Lovely professional university*

***Abstract*-----Similarity is the one of the major problem in the various areas of data mining, text mining, NLP, Geo informatics and biomedical informatics. Now a days Similarity is a big time consuming process because data is repeated again and again. Syntactic similarity is one of the similarity type which means structure of the words and phrases and similarity is measured by word to word. In this paper we reviewed previous papers and techniques and proposed approach and algorithm that can be used for finding similarity and accuracy. The aim of this paper is to be present an approach which can be used for find out the similarity between the questions and to be remove the duplicacy with the help of syntactic similarity.**

Keywords- Syntactic similarity, Naïve Based, Rabin Karp, boyer Moore, multiple pattern search, fast indexing.

I. INTRODUCTION

Determining the similarity between the sentences is a major task which impact in the application of NLP, detecting and tracking the topics, question generating and questions answers system. Sentence similarity is a state in which structure or meaning or both are same in two or more sentences, words are compared and similar patterns are found. There are three types of sentence similarity those are statistical measure, semantic measure and syntactic measure. We will work on the syntactic similarity with the help of three algorithms namely naïve based, Rabin karp and boyer Moore. Our work is to improve the duplicacy of the data, there may happen many times that there are repeated question in question paper so main objective is to find the same questions in the question system.

For example:-

1. What do you mean by artificial intelligent?
2. What is artificial intelligent?
1. Where I can put on
2. Where I can put it

These questions are different but meaning are same so our approaches will be remove these types of questions in the system.

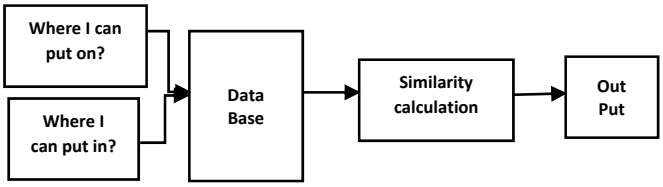

Fig. System Calculating Similarity and Accuracy

The diagram shows that firstly we have two questions or strings in the text box and they are stored in the database and after that similarity is calculated for the two question and results are compared with each other.

II. LITERATURE 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], proposes 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], proposes 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], proposes 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], proposes 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], proposes 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 Measures Sentence Semantic Similarity Measures word order similarity**,** The Combined Semantic and Syntactic Measures

Prathvi Kumari et.al. [7], proposes 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], suggests 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], suggests 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], proposes 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] proposes 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 introduced 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

In this paper to use a small sub set because of a computational cost. 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 Word Net 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 macine.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].

Megha Mishra et.al. [28], Proposed an approaches in which combined the three features that is semantic, syntactic and lexical and used a SVM classifiers with the help of this classifier to improve the accuracy of a system.

Shashank et.al. [29], Proposed a method for calculating the similarity and jaccord method is used with the help of this to improve the accuracy with compare of previous method.

III. COMPARISON TABLE

S.No	Ref. Id	Author Name	Technique used	Result	Efficiency
1	[4]	Zhong Min Juan	Semantic and Statistical methods	Proposed a method that builds semantic knowledge base namely co-occurrence words corpus, then count term frequency of question sentence by using statistic method.	Better performance with new methods.
2	[10]	Kai Wang, Zhaoyan Ming, Tat-Seng Chua.	Syntactic Tree Matching.	The proposed method uses Syntactic Tree Technique to improve the accuracy rate as compared to the previously used methods for finding the accuracy.	8.3% accurate from previous methods 50% accurate if semantic features are used.
3	[12]	Anterpreet Kaur,	LCS, Edit Distance and Bi-gram algorithms	The proposed method finds similar questions and removes the possibility of relevant question in future time.	70% accuracy rate.
4	[16]	U.L.D.N Gunasinghe, W.A.M De-silva, N.H.N De-silva, A.S Perera, W.A.D Sashikha, W.D.T.P Premasiri	Semantic and Syntactic methods	The proposed system can be used for variable length strings means the size of question is not fixed for calculating similarity.	More accurate than previous system.
5	[21]	Jehad Q. Odeh	FLFC and RSMA algorithms.	In proposed system both FLFC and RSMA algorithms are compared with Brute Force and Boyer-Moore, and FLFC is found to be best If it merges with recursive matching technique.	50% improvement in accuracy rate from previous work.
6	[28]	Megha Mishra, Vishnu Kumar Mishra and Dr. H.R. Sharma	Linear SVM(Support Vector Machine)	The proposed method combines three features namely Semantic, Syntactic and Lexical with SVM classifier to improve the accuracy.	91.1 % for fine grain and 96.2 % accuracy rate for coarse grain.
7	[29]	Shashank, Shailendra Singh	Jaccard Method	The proposed system shows that the Jaccard method is more efficient than any other statistical methods.	More efficient and accurate than other methods.

IV. PROPOSED APPROACH

This approach can be used for a questions analogues system and to find out the similarity. Many repeated questions presented in a question system so this approaches will remove the duplicacy of the two strings. It can be search for multiple patterns at once. In our algorithm along with multiple pattern search and also used fast indexing of boyar Moore algorithm. If string length of L and pattern length of P then the complexity will be O(LP) which is very less when compared to other algorithms used in implementation. This approach will find out the similarity and accuracy.

Basic Steps to used are as Follows:-

1. **Two Strings**
2. **Storing in a data base**
3. **Compared them**
4. **Find similarity and accuracy**
5. **Display them**
6. **Compared with others**

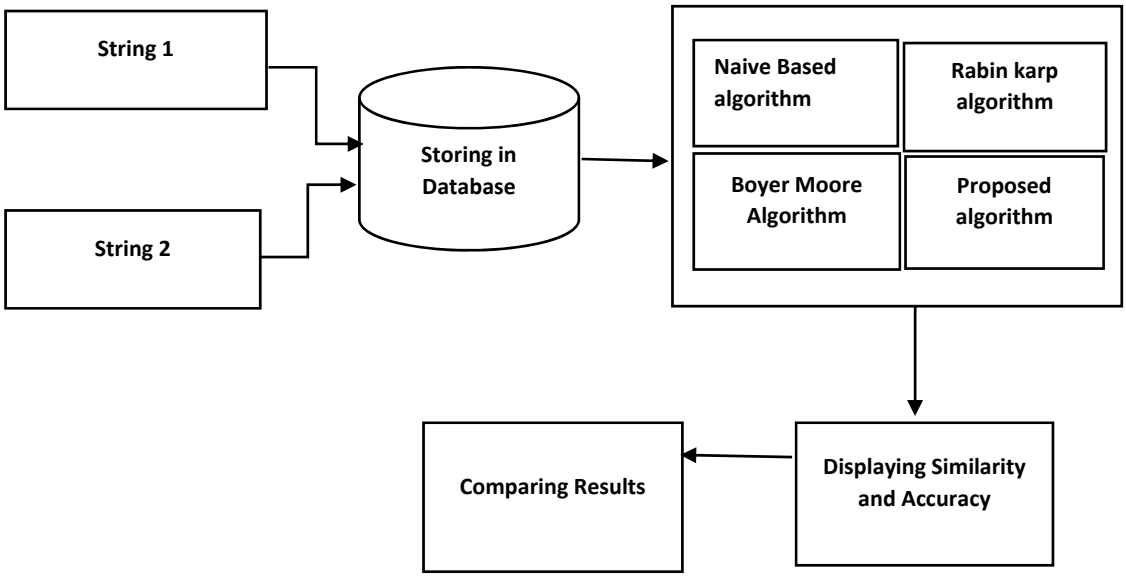

Fig. Proposed System

The basic steps to follow to get the similarity and accuracy:-

First of all we take two strings namely string 1 and string 2 that user enters for comparison in the form of questions and then the strings are stored in the data base. After that the strings are compared with each other and lengths of the two strings that the user had entered are calculated. The calculated length of strings are used for calculating similarity and accuracy. This is done by comparing all the characters of the two strings as and also we use three algorithms for this purpose naive based, Rabin-karp, boyer-moore algorithms. These three algorithms are used for comparing the strings and are used for calculating similarity and accuracy of strings. These algorithms calculate similarity and accuracy by comparing characters of the one string with characters of another string. The results that are generated by the three algorithms are compared with our proposed algorithm and is found that results from our algorithm are better than those three algorithms. In our proposed algorithm for calculating similarity and accuracy we divide one of the string into two parts and then we compare the two parts of the string with another string one at a time this will help in reducing the time complexity of the algorithm means the proposed algorithm will take less time to generate results and will be efficient than all the other algorithms. At last step we compare the results that are generated from the three algorithms and proposed algorithm and it is found that the results from our algorithm are better than those three algorithms. This is done by comparing the values of similarity we get from three algorithms with the value we get from proposed algorithm.

V. SOFTWARE AND DATABASE

We are using Matlab for programming and development of the GUI and other functions. We use different library functions in Matlab for the purpose of calculating similarity and accuracy in the algorithms mentioned in the proposed system. For storing strings we are using MySQL database and we connect the database with Matlab using jdbc driver.

VI. CONCLUSION AND FUTURE WORK

In this paper the proposed algorithm will calculate accuracy and similarity based on syntactic methods and we are comparing the results with already defined algorithms Naive based algorithm, Rabin karp algorithm and Boyer Moore algorithm. Our proposed algorithm will be able to remove duplicacy and will improve accuracy rate and similarity. In future we are going to implement the proposed work and show results experimentally.

ACKNOWLEDGEMENT

I expressed my heartiest gratitude to all the peoples who helped me in completing this work.

REFERENCE

[1] Yuhua Li, David McLean, Zuhair A. Bandar, James D. O'Shea, and Keeley Crockett," Sentence Similarity Based Semantic Nets and Corpus Statistics," *IEEE international conference of semantics similarity*, 2014 vol.4 pp.45-50.

[2] Wanpeng Song, Min Feng2 Naijie Gu1and Liu Wenyin," Question Similarity Calculation for FAQ Answering," *Third International Conference on Semantics, Knowledge and Grid IEEE*,2007, vol.3, pp. 1-9.

[3] MuthukrishananUmamehaswari, MuthukrishnanRamprasath, and Shanmugasundaram Hariharan, "Improved Question Answering System by semantic reformulation," *IEEE- Fourth International Conference on Advanced Computing, ICoAC 2012 MIT*, Anna University, Chennai. December 13-15, 2012.

[4] Zhong Min Juan, "An Effective Similarity Measurement for FAQ Question Answering System, "*International Conference on Electrical and Control Engineering IEEE*, 2010.

[5] Jun sheng Zhang, Yunchuan Sun, Huilin Wang and Yanqing He, "Calculating Statistical Similarity between Sentences, "*Journal of Convergence Information Technology*, February 2011, Vol.6, no.2.

[6] Palakorn Achananuparp, Xiaohua Hu, and Shen Xiajiong," The Evaluation of Sentence Similarity Measures," *International Conference of Computational Linguistics*, Vol. 6, pp.25-32.

[7] Prathvi Kumari, and Ravi Shankar K," Measuring Semantic Similarity between Words using Page-Count and Pattern Clustering Methods, "*International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, July 2013, Vol.3, pp.2278-3075.

[8] Partha Pakray, Sivaji Bandyopadhyay and Alexander Gelbukh," Textual entailment using lexical and syntactic similarity," *International Journal of Artificial Intelligence & Applications (IJAIA)*, January 2011,Vol.2,no.1.

[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," Department of Computer Science The University of York ,2014 vol.7, pp. 1-9.

[10] Kai Wang, Zhaoyan Ming and Tat-Seng Chua, "A Syntactic Tree Matching Approach to Finding Similar Questions in Community-based QA Services," School of Computing National University of Singapore2009.

[11] Wael H. Gomaa and Aly A. Fahmy," A Survey of Text Similarity Approaches," *International Journal of Computer Applications*, April 2013, Vol.68, no.13, pp.0975 – 8887.

[12] Anterpreet Kaur," A Novel Approach for Syntactic Similarity between Two Short Texts, "*INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH*, June 2015,Vol.4,issue 06, pp. 2277-8616.

[13] Rafael Ferreira," A New Sentence Similarity Method based on a Three-Layer Sentence Representation," *IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, 2014.

[14] Zhao Jingling, Zhang Huiyun and Cui Baojiang," Sentence Similarity Based on Semantic Vector Model, "*Ninth International Conference on P2P, Parallel, Grid, Cloud and Internet Computing IEEE, 2014.*

[15] Xiao-Ying Liu and Chuan-Lun Ren," Similarity measure based on sentence semantic structure for recognizing paraphrase and entailment, "Hindawi Publishing Corporation Mathematical Problems in Engineering, Vol. 2015, Article ID 203475, 8 pages July 2013.

[16] U.L.D.N Gunasinghe, W.a.m de silva, N.H.N.D de silva, A.S Parera and W.A.D Sashika," Sentence Similarity Measuring by Vector Space Mode, "*International conference on Advances in ICT for emerging regions*,2014,pp. 185-189.

[17] Chi Zhang, Lei Zhang, Chong-Jun Wang, and Jun-Yuan Xie," Text Summarization Based on Sentence Selection with Semantic Representation," *IEEE 26th International Conference on Tools with Artificial Intelligence*,2014,pp.1082-3409.

[18] Jiwoon Jeon, W. Bruce Croft and Joon Ho Lee," Finding Similar Questions in Large Question and Answer Archives," Centre for Intelligent Information Retrieval, Computer Science Department University of Massachusetts, Amherst, MA 010032005.

[19] Ercan Canhasi," Measuring the sentence level similarity," Faculty of Computer Science University of Prizren, Kosovo ISCIM 2013, pp. 35-42.

[20] Jehad Q. Odeh," New and Efficient Recursive-based String Matching Algorithm (RSMA-FLFC)," *International Journal of Computer Applications*, Vol 86,no.15,January 2014,pp.0975 – 8887.

[21] Asli Celikyilmaz, Dilek Hakkani-Tur and Gokhan Tur, "LDA Based Similarity Modelling for Question Answering, "*Proceedings of the NAACL HLT 2010 Workshop on Semantic Search*, Los Angeles, California, June 2010,pp.1-9.

[22] D. Das, Schneider, D.Chen, and N.A.Smith," Probalistics frame-semantic parsing, "in Human Language Technologies," *The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*,2010, pp.948-956.

[23] A. Berger, R. Caruana, D. Cohn, D. Freitag, and V. Mittal, "Bridging the lexical chasm: statistical Approaches to answer-finding," *international conference of artificial intelligence*, 2010, Vol.1, pp. 192–199.

[24] P. F. Brown, V. J. D. Pietra, S. A. D. Pietra, and. L. Mercer," The mathematics of statistical machine Translation: parameter estimation," Compute. Linguist.1993, Vol.2, pp.263–311.

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

[26] E. Sneiders, "Automated question answering using question templates that cover the conceptual model of the database, "*In Proceedings of the 6th International Conference on Applications of Natural Language to Information Systems-Revised Papers*,2002, pp. 235–239.

[27] R. Nigel Horspool, "Practice Fast Searching in String," *Journal of Software Practice and Experience*, vol.10, pp. 501-506.

[28] Megha Mishra, Vishnu Kumar Mishra and Dr. H.R. Sharma," Question Classification using Semantic, Syntactic and Lexical features," *International Journal of Web & Semantic Technology (IJWesT)*,Vol.4, no.3, July 2013.

[29] Shashank and Shailendra Singh, "Statistical Measure to Compute the Similarity between Answers in Online Question Answering Portals," *International Journal of Computer Applications*, Vol.103, no.15, October 2014, pp.0975 – 8887.