Statistical Measure to Compute the Similarity between Answers in Online Question Answering Portals

Shashank
Assistant Professor
Dept. of Computer Science
UIET-KUK
Kurukshetra, India

Shailendra Singh, Ph.D.
Associate Professor
Dept. of Computer Science
PEC University of Technology,
Chandigarh, India

ABSTRACT
Sentence similarity plays an important role in the field of natural language processing where it can be used for information processing, text mining and online question answering portals. Among all these applications sentence similarity is one of the most critical issues which has attracted the attention of many researchers. In this paper, a system which automates the process of subjective answer checking with high accuracy is proposed. Answer assessment system is an arrangement where answers given by users are matched with the stored answers in question answer database to judge their correctness. The main focus here is to improve the similarity of user's answer with the stored answer in question answer database. In addition to this, the intention is to remove the mistakes from answer checking process through automated system. For experimental purpose a set of questions and their answers have been taken and these questions have been answered by a random user. Results show that the method gives higher accuracy as compared to original method. Proposed method makes question answering process fast and efficient.

General Terms
Answer assessment system, Statistical similarity, sentence, and Question answering portal et al.

1. INTRODUCTION
Determining similarity between sentences is one of the decisive tasks which include a wide range of impact in many applications of Natural Language Processing. Sentence similarity refers to quality or state of sentence in which structure or meaning or both are in common with other sentence. Calculation of similarity between two sentences is the basis of measuring the similarity between texts which is the key of documents. Sentence similarity computation has been used in the field of information processing, where similarity calculation is performed to assign a ranking score between query and texts in corpus. Question answering system, where calculation is performed to compute similarity between two questions or between two answers. Online question answering portal is an extension of question answering system. In these systems users are provided with number of questions for which user can store answers which further is compared with the stored answer in question answer database to check the correctness of answers. With the advancement of techniques the need to check the correctness of answers to asked question has become important task in online question answering portal. One can achieve this objective using different similarity measures specifically semantic similarity measures and statistical similarity measures. Semantic similarity measures are very difficult to implement and are cost effective in comparison to the statistical measures. Although there exist a number of methods based on statistical measures, there is a need to improve the accuracy of these methods to achieve the objective. In this paper one such application i.e. online question answering portal is being addressed.

Online question answering portal is about storing the user's response and checking the accuracy of the answers against the stored answers therefore answer assessment system plays an important role in question answering portal. Answer Assessment System is an approach which can be used in online Question answering Portal to check the accuracy of the answers. It includes a question answer database. For each answer, there is only one correct answer. In the front-end of this system, users are asked a question and, in back-end of this system, their answers which may include multiple lines are assessed for correctness against the answer stored in question answer database. It takes the answers provided by the end user as an input and compares it with the corresponding answer. Here answer assessment system is using a statistical measure to calculate the similarity between two Answers. The statistical similarity measures between sentences are based on symbolic characteristics and structural information. It could measure the similarity between answers without any prior knowledge but only on the basis of statistical information of sentences. In fact statistical sentence similarity can be calculated by taking the word count of two sentences. Different statistical methods have been proposed to measure the similarity between two sentences. Earlier authors were focused on word or short segment similarity but here focus is on syntactically well formed sentences. There exit various methods which are based on word sets. Jaccard's method is one such method which uses word sets for sentence similarity but its accuracy is not up to the mark. Here a method is proposed to improve the accuracy of Jaccard's method.

Rest of the paper is organized as follows: The next section explores the related work which has been done in this field. Proposed method and assumptions has been discussed in the subsequent section. After that the evaluation and results section is there. Finally conclusions derived from the study are presented.

2. RELATED WORK
Till date most of the researchers have focused on calculating similarity of words and short segments but very less attention has been paid to sentence similarity. In this section the work carried out in the field of sentence similarity using statistical measures is summarized in brief. The problem of sentence similarity which considers contents of sentence in order to get the similar parts is addressed [1]. Further in a different

attempt term frequency-inverse document frequency (TF-IDF) similarity measure has proposed for detecting topically similar sentences in documents [2]. TF-IDF similarity measure is a numerical statistics, reflects how important a word is to a document in a collection or corpus. It is very often used as a weighting factor in information retrieval and text mining. The proposed solution is based on sum of product of the term frequency and inverse document frequency of the words that appear in both the sentences. The TF-IDF approach which uses confidence and characteristic words to improve the precision and recall values are represented in [3]. There are various ways present to calculate the TF and IDF. In case of term frequency TF (t, d) the simplest choice is to use the raw frequency of a term in a document i.e. number of times that term t is present in document. If one denotes the raw frequency as f(t, d) then the simple TF scheme is TF(t, d) is f(t, d).

To prevent the bias towards long documents the term frequency is calculated by dividing the raw frequency through maximum raw frequency of any term in document.

$$TF(t,d) = 0.5 + \frac{0.5 \times f(t,d)}{\max\{f(w,d) : w \in d\}}$$

where t and d refers to term and document. *TF(t, d)* is term frequency of any term in document, f(t, d) is raw frequency of the term and f(w, d) is the raw frequency of words in document.

The Inverse Document frequency is a measure of whether the term is common or rare across all the documents. It is obtained by dividing the total number of documents by number of documents where term t is present and then taking logarithm of that quotient.

$$IDF(t,D) = \log \frac{N}{/\{d \in D : t \in D\}/}$$

where IDF (t, D) is the inverse document frequency of term in a document and N is number of documents in corpus whereas $|\{d \in D : t \in D\}|$ represents number of documents where the term t is present.

Finely TF-IDF is calculated as

$$TF - IDF(t,d,D) = TF(t,d) \times IDF(t,D)$$

Some other attempts of using TF-IDF as sentence similarity measure are explored in [4][5]. It chooses to take up the standard vector space approach to estimate the similarity between the vectors of two sentences. Different problems of similarity calculation between short segments of texts are explored and detailed account of similarity measures are presented in [6] that can be used to compute the similarity. These similarity measure can be simple lexical matching, stemming & text representation that are reinforced using web search results with a language modeling framework.

The overlap measure based on Zipfian relationship between length of the phrases and their frequencies in the text collection is discussed in [7]. Their incitement came from the fact that word overlap measure simply treat sentences as a pack of words and does not take into account the differences between single words and multiword phrases.

Another method for text similarity based on vector space model has been proposed in [8]. It puts forward the technique to use key sentences to calculate text similarity. Further Vector space model has been used in Question answering system to calculate the similarity between questions [9]. After that efforts have been made for calculating similarity between questions in frequently asked question (FAQ) system which uses the Word co-occurrence model [10]. Here measurement of similarity of questions is done by number of relative terms as well as the length of question sentence. Further the performance of statistical translation models in identifying related sentences compared to several simplistic approaches such as word overlap fraction is evaluated in [11]. Word overlap fraction is defined as the proportion of words that appear in both sentences normalized by sentence length.

Discussions about Jaccard similarity measure are carried out in [12][13][14] where similarity is computed based on the number of words shared by two sentences. It compares the similarity between two word sets. When applied for sentence similarity, it is defined as the size of intersection of words in two sentences compared to size of the union of words of those sentences. If Sa and Sb are two sentences then in order to compare the two sentences word set need to be prepared first. Word set construction includes tokenization of sentences in which sentences are fragmented into words. Let's consider Sa and Sb are two sentences which are as follows.

Sa= I do not like green vegetables.

Sb= I do not like them.

After tokenization Sa and Sb becomes as follows

Sa= {'I','do','not','like','green','vegitables'}

Sb= {'I','do','not','like','them'}

Total number of words present in Sa i.e W(Sa) = 6

Total number of words present in Sb i.e W(Sb) = 5

Total number of common words present in Sa and Sb i.e $W(Sa) \cap W(Sb)$ = 4

Total number of different words present in Sa and Sb i.e $W(Sa) \cup W(Sb)$ = 7

Therefore Jaccard Similarity JS(W(Sa),W(Sb))= 4/7

Jaccard similarity measure is quite efficient to calculate the similarity but still it does not address the problem of accuracy of answers in online question answering portal completely. This paper discusses certain improvement in the Jaccard similarity measure to get better similarity between two Answers. The Experimental results indicate that the improvements make positive effects and it improves the accuracy of answer checking.

3. PRAPOSED METHOD

The method used in Answer Assessment system is based on word set model of sentence similarity. To calculate the Answer similarity, the word set of the sentence must be constructed first. Since the sentences may have different tenses & voices, there are two ways to calculate word based sentence similarity. One is to calculate sentence similarity with the words in sentence other is to calculate sentence similarity with stemmed words in sentences. Here the first approach to calculate the answer similarity is considered. After construction of word sets the similarity of Answers are calculated. Similarity of answers is the ratio of number of

common words between the two answers and minimum of the number of words present in Answer Sa & Sb.

$$\text{Similarity} = \frac{W(Sa) \cap W(Sb)}{\min(W(Sa), W(Sb))}$$

This method is an improvement of Jaccard's algorithm in which similarity is the ratio of common words present in both sentences & total number of different words present in both sentences. Jaccard's similarity measure is very effective method for calculating similarity but in Online Question answering portal it does not address certain problem. Specifically, in question answering portal there exists a few questions for which answer could be given in one word, one line or in paragraph. For example if question asked is

What is the capital of India?

The answer for this question can be given in three ways.

- Capital of India is Delhi.
- Delhi.
- Delhi is the capital of India.

All these answers are correct with respect to the given question. If similarity is calculated by simple Jaccard method the similarity is poles apart therefore an improvement is made in this method to overcome the above said problem. In this process if all words (Tokens) of user's answer are present in Answer stored in Question-Answer database then similarity is considered as one. The steps involved in answer assessment are shown in Fig-1.

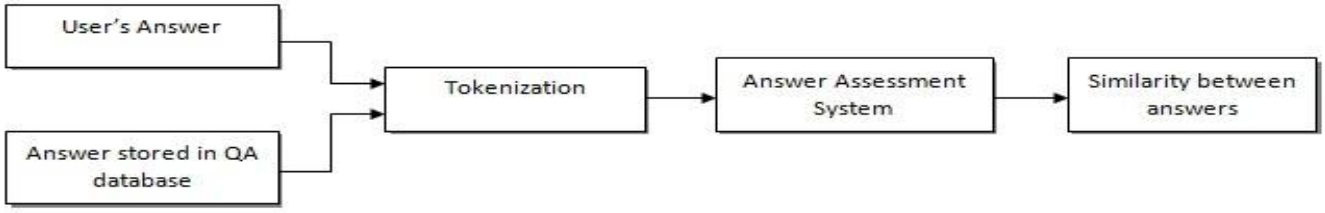

Fig-1. Answer Assessment System

4. EVALUATION CRITERIA

Evaluation of online question answering system is one of the most concerned issues for the researchers. With the increasing result in sentence similarity measure many data sets have developed so it is need of the hour to properly evaluate the method using experimental results. The evaluation criteria used in this process are as follows. ***Accuracy*****:** percentage of sentence correctly compared by the Answer Assessment System.

Precision**:** percentage of relevant sentence correctly retrieved by the system with respect to all sentences retrieved by the system.

$$\textbf{\textit{Precision (P)}} = \frac{User'sAnswer \cap StoredAnswer}{User'sAnswer}$$

Recall**:** percentage of relevant sentence correctly retrieved by the system with respect to all sentences relevant for the human.

$$\textbf{\textit{Recall(R)}}: \frac{User'sAnswer \cap StoredAnswer}{StoredAnswer}$$

F-Measure**:** Combine in a single measure precision (P) and recall (R) giving a global estimate of the performance of a Question-Answer System.

$$\textbf{\textit{F-Measure}}: \frac{2PR}{P+R}$$

Based on sentence similarity measure one can calculate the precision, recall and F-Value. F-Value is used to work out the accuracy of the sentence similarity by using both precision and recall. In this process User's answer is fetched and compared with the Answer stored in the question answer database and then the precision, recall and F-Value could be calculated.

5. DATA SET

For experimental purpose a set of questions have been taken in which correct answer is stored for every question in question answer database. In the data set, length of answers is between 10 words to 20 words. Here answers are chosen with minimum and maximum length size because focus in this research is on subjective answer checking. A user randomly chosen and his answers also are been stored and later it is compared to the answer stored in question answering database. Following table1 describes the data sample which has been used. Data set has been taken in limited domain which includes questions related to computer science. Moreover, answers are not case sensitive i.e. RAM and ram both are considered as same in data sets.

Table1. Data sample for analysis

Sample 1	**Question**	What is Computer?
	User's Answer	Computer is an electronic device which is used for computations.
	Stored Answer	Computer is an electronic device which is used to perform logical & arithmetic computations.
Sample 2	**Question**	What is RAM?
	User's Answer	RAM is Random Access memory which is used in cpu to provide memory to store instructions and data.
	Stored Answer	RAM is random access memory which is a volatile memory used to store the data and instructions.
Sample 3	**Question**	What is ROM?
	User's Answer	rom is read only memory which can be used in boot of the computer.
	Stored Answer	ROM is read only memory which is a non volatile memory and used to boot the computer.

Sample 4	**Question**	4 What is CPU?
	User's Answer	CPU is central processing unit which is used to perform different logical and arithmetic calculations.
	Stored Answer	CPU is central processing unit which is the brain of the computer where most calculations take place.
Sample 5	**Question**	5 What is ALU?
	User's Answer	ALU is arithmetic logic unit used in performing logical operations in computer system.
	Stored Answer	ALU is arithmetic logic unit which is used to perform logical operations.

6. RESULT AND DISCUSSION

In order to check the accuracy and simplicity and to evaluate the performance of proposed system five questions answer sets are used which are presented in table1. Table2 shows the results of proposed method which shows the precision rate, recall rate, F-value and accuracy of different measures of sentence similarity. The evaluation results shows that the sentences similarity based on proposed method has the better performance than the Jaccard algorithm and other previous existing algorithms like TF-IDF, edit distance etc. For the comparative analysis the same data set is used on Jaccard algorithm before and after modifications and a table is created which is shown in table3 and further graph is been plotted which shows the considerable amount of improvements in accuracy. One problem with this method is that it only considers the word count but not the syntactic structure and meaning so even if syntactically Answer is wrong but if words are present which is there in QA Database then it will count those words too in order to calculate the similarity.

Table2. Results for the proposed system

Sample	Proposed Method			
	Precision	**Recall**	**F-Value**	**Accuracy**
S1	1	.78	.87	1
S2	1	.94	.96	1
S3	.85	.70	.76	.85
S4	.56	.52	.53	.60
S5	.64	.66	.64	.66

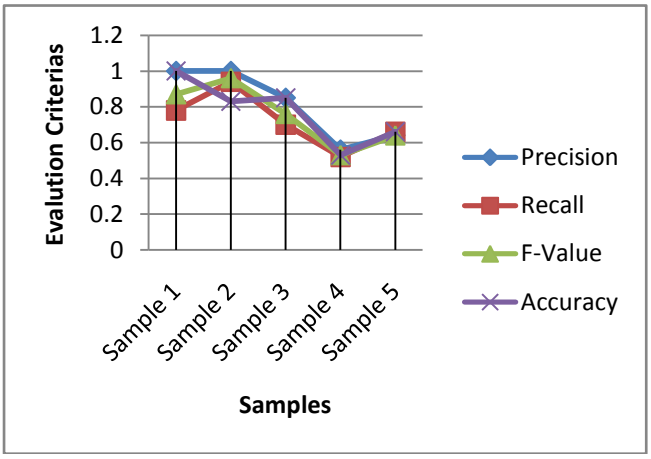

Fig2. Results Analysis of the proposed system

Table3. Comparative analysis between Jaccard and Modified Jaccard method

Sample	Jaccard Method	Modified Jaccard Method
S1	0.66	1
S2	0.65	1
S3	0.5	0.85
S4	0.33	0.6
S5	0.44	0.66

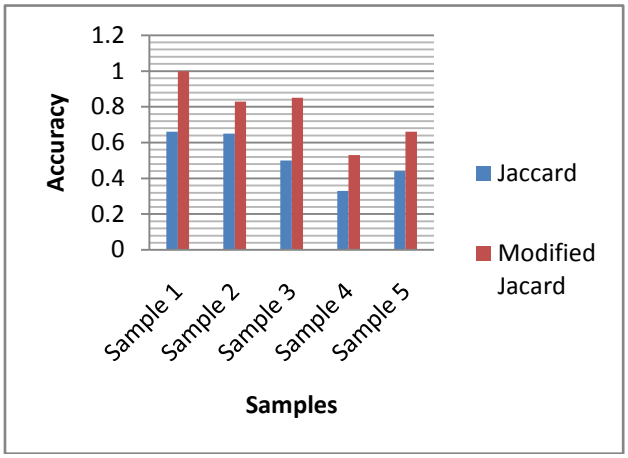

Fig3. Accuracy Comparison

7. CONCLUSION AND FUTURE WORK

In this paper a critical issue related to online question answering portal is addressed. Statistical similarity measure is used to calculate the similarity between two answers. Statistical similarity can be calculated without any prior knowledge of sentence and it only depends on word counts. Evaluation is carried out on a set of question answers. Experimental results show that proposed method performs better than some of the existing statistical methods. The advantage of using Jaccard method is that its efficiency and accuracy are better as compared to the other methods. Future research work mainly focuses on the syntactic and semantic structure of answers which could help in improving answer similarity.

8. REFERENCES

[1] F. Mandreoli, R. Emilia, R. Martoglia, and P. Tiberio, "A Syntactic Approach for Searching Similarities within Sentences," in *Proceedings of eleventh international conference on information and knowledge management*, 2002, pp. 635–637.

[2] J. Allan, C. Wade, and A. Bolivar, "Retrieval and novelty detection at the sentence level," in *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval - SIGIR '03*, 2003, pp. 314–321.

[3] Y. Zhang, L. Gong, and Y. Wang, "An improved TF-IDF approach for text classification," *J. Zhejiang Univ. Sci.*, vol. 6, no. 1, pp. 49–55, Jan. 2005.

[4] Y. Li, D. McLean, Z. a. Bandar, J. D. O'Shea, and K. Crockett, "Sentence similarity based on semantic nets and corpus statistics," *IEEE Trans. Knowl. Data Eng.*, vol. 18, no. 8, pp. 1138–1150, Aug. 2006.

[5] P. Achananuparp, X. Hu, and S. Xiajiong, "The Evaluation of Sentence Similarity Measures," in *Proceedings of the 10th international conference on Data Warehousing and Knowledge Discovery (DaWaK 08)*, 2008, vol. 5182, pp. 305–316.

[6] D. Metzler, S. Dumais, and C. Meek, "Similarity Measures for Short Segments of Text," in *Proceedings of the 29th European conference on IR research ECIR 07*, 2007, pp. 16–27.

[7] S. Banerjee and T. Pedersen, "Extended Gloss Overlaps as a Measure of Semantic Relatedness," in *Proceedings of the 18th international joint conference on Artificial intelligence IJCAI03*, 2003, pp. 805–810.

[8] H. Dong, J. Wu, X. Zhao, and Y. Li, "Study on the Calculation of Text Similarity Based on Key-sentence," in *2010 International Conference on E-Business and E-Government*, 2010, pp. 1952–1955.

[9] D. Wang, "Improved sentence similarity algorithm based on VSM and its applications in question answering system," in *Intelligent Computing and Intelligent Systems (ICIS), 2010 IEEE International Conference*, 2010, pp. 368–371.

[10] Z. M. Juan, "An Effective Similarity Measurement for FAQ Question Answering System," in *2010 International Conference on Electrical and Control Engineering*, 2010, pp. 4638–4641.

[11] D. Metzler, Y. Bernstein, W. B. Croft, A. Moffat, and J. Zobel, "Similarity measures for tracking information flow," in *Proceedings of the 14th ACM international conference on Information and knowledge management - CIKM '05*, 2005, pp. 517–524.

[12] A. Islam and D. Inkpen, "Semantic text similarity using corpus-based word similarity and string similarity," *ACM Trans. Knowl. Discov. Data*, vol. 2, no. 2, pp. 1–25, Jul. 2008.

[13] J. Zhang, Y. Sun, H. Wang, and Y. He, "Calculating Statistical Similarity between Sentences," *J. Converg. Inf. Technol.*, vol. 6, no. 2, pp. 22–34, 2011.

[14] S. Chauhan, P. Arora, and P. Bhadana, "Algorithm for Semantic Based Similarity Measure," *Int. J. Eng. Sci. Invent.*, vol. 2, no. 6, pp. 75–78, 2013.