MANTECH Publications

Recent Trends in Computer Science and Software Technology
Volume 4, Issue 3, September-December 2019

Machine Learning Approaches for Sentence Level Paraphrase Identification

S. V. Kogilavani[1], S.Malliga[2], C. S. Kanimozhiselvi[3]
Assistant Professor[1], Professor[2], Associate Professor[3]
Department of CSE
Kongu Engineering College, Perundurai
Corresponding Authors' Email: *kogilavani.sv@gmail.com[1], mallisenthil@kongu.ac.in[2]*

Abstract

In any language, sentences with different structures may convey the same meaning. Paraphrase identification is defined as the task of deciding whether two given text fragments have the same meaning or not. The proposed work focuses on the detection of paraphrases in Tamil language using the statistical and semantic analysis of sentences. The statistical analysis calculates the similarity between two Tamil sentences based on Jaccard, Dice, Cosine and Word distance. The semantic analysis determines the similarity of two sentences based on Word order. Classification is a data mining technique that assigns categories to a collection of data in order to aide in more accurate predictions and analysis. The proposed approach utilizes five different machine learning algorithms for classification of given sentence pair using statistical and semantic features. The accuracy and performance of different classification methods are measured on the basis of evaluation parameters like accuracy, precision, recall and f-measures. The combination of statistical and semantic similarity features helps to identify whether the pair of sentences is paraphrase or not.

Keywords: *Machine learning, Statistical analysis, Semantic analysis, Paraphrase identification, Classification*

Page 106-119 © MANTECH PUBLICATIONS 2019. All Rights Reserved

INTRODUCTION

Text mining refers to the process of deriving high-quality information from text. Text mining involves the process of structuring the input text, deriving patterns within the structured data and evaluation and interpretation of the output. High quality in text mining usually refers to some combination of relevance, novelty and interestingness. Typical text mining tasks include text categorization, text clustering, concept extraction, production of granular taxonomies, sentiment analysis, document summarization and entity relation modeling. Information can be extracted to derive summaries for the words contained in the documents are to compute summaries for the documents based on the words contained in them.

Natural language is the most powerful tool through which people establish communication and relate to each other. In our daily life we can use different words and phrases to express the same meaning. This is related to our knowledge and cultural habits, that later reflect on our written and spoken skills. The web is the largest text repository, where millions of people share and consult information daily. In the context of Information Retrieval, given a natural language query, the search engine should identify and return documents that have similar or related meanings to the query. However, the relevant information may be present in different forms. For example, ”operating systems” should retrieve document about ”unix”. In order to identify that although neither ”operating” nor ”systems” appear, the document is still relevant as ”unix” is a type of operating system, a paraphrase identification module is needed.

Classification is the task of automatically classifying a set of text documents into different categories from a predefined set. It is a form of data analysis that extracts models describing important data classes. Such models, called classifiers, predict categorical class labels. Many classification methods have been proposed by researchers in machine learning, pattern recognition and statistics. It has numerous application, including fraud detection, target marketing, performance prediction, manufacturing and medical diagnosis. The automatic classification of documents into predefined categories has observed as an active attention, the documents can be classified by three ways, unsupervised, supervised and semi supervised methods. The classification includes Bayesian classifier, Decision Tree, K-nearest neighbor (KNN), Support

Page 106-119 © MANTECH PUBLICATIONS 2019. All Rights Reserved

Vector Machines (SVMs), Neural Networks, Rocchio's.

Machine learning is gaining mainstream presence for data scientists. Machine learning applications are highly automated and self-modifying which continue to improve over time with minimal human intervention as they learn with more data. Its tasks are typically classified into three broad categories, depending on the nature of the learning "signal" or "feedback" available to a learning system. These are supervised learning, unsupervised learning and reinforcement learning. Another categorization of machine learning tasks arises when one considers the desired output of a machine-learned system in classification, regression, clustering, density estimation and dimensionality reduction.

Machine learning and data mining often employ the same methods and overlap significantly, but while machine learning focuses on prediction, based on known properties learned from the training data, data mining focuses on the discovery of unknown properties in the data. Data mining uses many machine learning methods, but with different goals; on the other hand, machine learning also employs data mining methods as "unsupervised learning" or as a preprocessing step to improve learner accuracy. Much of the confusion between these two research communities comes from the basic assumptions they work with: in machine learning, performance is usually evaluated with respect to the ability to reproduce known knowledge, while in Knowledge Discovery and Data Mining (KDD) the key task is the discovery of previously unknown knowledge. Evaluated with respect to known knowledge, an uninformed (unsupervised) method will easily be outperformed by other supervised methods, while in a typical KDD task, supervised methods cannot be used due to the unavailability of training data.

Machine learning techniques present a machine learning approach for paraphrase identification which uses lexical and semantic similarity information. The main objective of machine learning techniques is to increase the final performance of the system. Sentence similarity focuses on computing the order information implied in the sentences. The semantic similarity of two sentences is calculated using information from a structured knowledge and the incorporation of corpus statistics allows our method to be adaptable to different domains.

Page 106-119 © MANTECH PUBLICATIONS 2019. All Rights Reserved

Paraphrase is the task of recognizing whether the text fragments have the same meaning. Paraphrase identification is important for information retrieval, information extraction, natural language processing, and machine translation. If the two sentences are similar, then words in the two sentences may or may not be similar. For example

These two sentences have the same meaning but that can be expressed by different texts. It is mainly focuses on identifying whether the given sentences are paraphrase or not in Tamil language. The similarity between sentences is measured based on statistical information of sentences. The statistical similarity between two sentences are calculated based on word distance using Euclidean measures, word set using Jaccard and Dice measures, word vector using Cosine similarity measures. The semantic similarity between two sentences is calculated based on word order between the sentences.

This paper is organized as follows. Section 2 discusses about related work on paraphrase detection system. Section 3 presents an overview of the proposed approach. The experimental and evaluation measures are discussed in section 4. Finally section 5 concludes this paper.

LITERATURE REVIEW

Ekaterina Pronoza et al (2016) proposed a low level features for paraphrase identification deals with the task of sentential paraphrase identification which is challenging but helpful in many natural language processing (NLP) applications like plagiarism detection, question answering, authorship detection, information retrieval, text summarization. It focuses on the low-level string, lexical and semantic features which unlike complex deep ones do not cause information noise and can serve as a solid

Page 106-119 © MANTECH PUBLICATIONS 2019. All Rights Reserved

basis for the development of an effective paraphrase identification system.

Zornitsa Kozareva et al (2006) stated that the Machine learning techniques presents a machine learning approach for paraphrase identification which uses lexical and semantic similarity information. The main objective of machine learning techniques is to increase the final performance of the system. Sentence similarity focuses on computing the order information implied in the sentences. The semantic similarity of two sentences is calculate edusing information from a structured knowledge and the incorporation of corpus statistics allows our method to be adaptable to different domains.

Steven burrows et al (2012) represented a Paraphrase Acquisition via Crowd sourcing and Machine Learning presents the recognition and generation of paraphrases forms the heart of numerous analysis and synthesis tasks in information retrieval, information extraction, and natural language processing.

To paraphrase means to rewrite content whilst preserving the original meaning. Plagiarism detection is a paraphrase recognition task that has received much recent attention. Plagiarism refers to the use of another author's information, language, or writing, when done without proper acknowledgment of the original source.

L Li et al (2009) stated that measuring sentence similarity from different aspects proposes to determine sentence similarities from different aspects. Based on the information people get from a sentence, Objects-Specified Similarity, Objects-Property Similarity, Objects-Behavior Similarity and Overall Similarity are defined to determine sentence similarities from four aspects.

NPA Vo et al (2015) represented that measuring semantic similarity is the objective of many works. Many measures perform well in evaluation framework for a specific task like synonymy extraction. Paraphrase Identification and Semantic Similarity are important tasks that can be used as features to improve many other Natural Language Processing (NLP) tasks. For example Information Retrieval, Machine Translation Evaluation, Text Summarization, Question and Answering.

Y Li et al (2006) represented that Sentence Similarity Based on Semantic Nets and Corpus Statistics has the recent

Page 106-119 © MANTECH PUBLICATIONS 2019. All Rights Reserved

applications of natural language processing present a need for an effective method to compute the similarity between very short texts or sentences. Sentence similarity will have Internet-related applications as well. In Web page retrieval, sentence similarity has proven to be one of the best techniques for improving retrieval effectiveness, where titles are used to represent documents in the named page finding task.

Boubacar et al (2014) represented that measuring semantic similarity is the objective of many works. Many measures perform well in evaluation framework for a specific task like synonymy extraction, short text comparison, sentence similarity, and natural language processing etc. The challenge is to find a single measure that performs accurately as long as semantic extraction is needed.

The smallest unit for human communication is "word". Semantic matching certainly cannot work without using words. The combination of words can really express an object or idea and by the same way give to a word a meaning. The association of words can be studied at a statistical level using frequency estimation to define similarity measures.

SM Yimam et al (2016) stated that paraphrases are alternative ways of writing texts while conveying the same information. Paraphrasing consists of mainly two tasks, paraphrase generation and paraphrase identification. Paraphrase generation is the task of obtaining candidate paraphrases for a given target. Paraphrase identification estimates whether a given paraphrase candidate can replace a paraphrase target without changing the meaning in context.

PROPOSED SYSTEM

Overview

In the proposed method, the statistical and the semantic analysis is used to determine whether the given sentence is paraphrase or not. In this system, the statistical analysis is based on word set, word vector, word order, and word distance and semantic analysis is based on word order between sentences. The overall similarity is calculated by combining these two measures. The following Figure 1 represents the proposed system design.

The sentence pair is taken as input which is passed to statistical and semantic similarity. If the two sentences are similar, then words in the two sentences may or may not be similar. Structural relations include relations between words and the

Page 106-119 © MANTECH PUBLICATIONS 2019. All Rights Reserved

distances between words. The similarity between sentences is measured based on statistical information of sentences. The statistical similarity between two sentences are calculated based on word distance using Euclidean measures, word set using Jaccard and Dice measures, word vector using Cosine similarity measures. The semantic similarity between two sentences is calculated based on word order.

By using statistical and semantic similarity, the overall sentence similarity is obtained. After obtaining the sentence similarity, it is passed to machine learning algorithms which contain the classifiers like Support Vector Machines, Maximum Entropy, Boosting, Random Forest and Bagging. Accuracy, sensitivity, specificity, Positive Prediction Value, Negative Prediction Value, Prevalence, detection rate, detection prevalence and balanced accuracy can be calculated for different classifiers.

1 Similarity Based On Statistical Measures

Statistical technique is tokenization and it is mainly needed for preprocessing. Tokenization is the process of breaking a stream of text up into words called tokens and also removing extraneous punctuation. It is useful for both in linguistics, and in computer science, where it forms part of lexical analysis. Statistical methods are used to find the similarity of sentences.

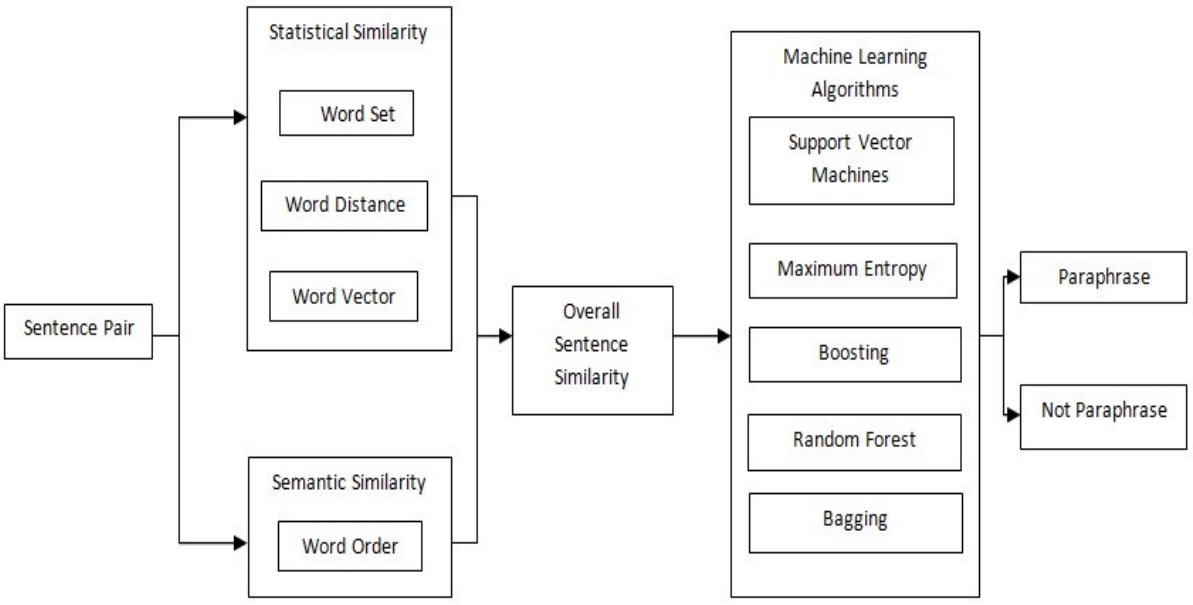


Figure 1 Proposed System Design

Page 106-119 © MANTECH PUBLICATIONS 2019. All Rights Reserved

Jaccard Similarity

Jaccard similarity is a word set based measure in which the word sets of the two sentences are taken into account for similarity calculation. Let w(Sa) be the set of words in first sentences Sa and w(Sb) be the set of words in second sentence Sb. After forming the word set, Jaccard similarity is computed using the equation 1.

$$Jaccard(S_a, S_b) = \frac{w(S_a) \cap w(S_b)}{w(S_a) \cup w(S_b)} \quad (1)$$

Dice Similarity

Dice Similarity is also a word set based measure. Let w(Sa) be the set of words in first sentence Sa and w(Sb) be the set of words in second sentence Sb. After forming the word set, Dice similarity is computed using the equation 2.

$$Dice(S_a, S_b) = \frac{2|w(S_a) \cap w(S_b)|}{|w(S_a)| + |w(S_b)|} \quad (2)$$

when

Word Distance Similarity

The Euclidean distance between points p and q is the length of the line segment connecting them. If the sentence p has words (p1,p2,…..,pn) and sentence q has words (q1,q2,…..,qn) then word distance of p and q are represented as follows.

$$p = (p_1, p_2, \ldots, p_n) \quad (3)$$

$$q = (q_1, q_2, \ldots, q_n) \quad (4)$$

Then the similarity between p and q can be calculated based on the distance of words by using equation 5.

$$d(p,d) = d(q,p) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2} = \sqrt{\sum_{i=1}^{n} q_i - p_i^2} \quad (5)$$

Cosine Similarity

In Cosine Similarity, Word vectors of sentences are constructed and they are assigned with weights. The words in w (Sa) and w(Sb) are assigned with weights, word vectors of Sa and Sb can be represented as follows.

$$v(S_a) = \{(w_1, w_{a1}), (w_2, w_{a2}), \ldots\ldots\ldots\ldots, (w_{i+j}, w_{a(i+j)})\} \quad (6)$$

$$v(S_b) = \{(w_1, w_{b1}), (w_2, w_{b2}), \ldots\ldots\ldots\ldots, (w_{i+j}, w_{b(i+j)})\} \quad (7)$$

word vectors by using equation 8.

$$Cosine(S_a, S_b) = \frac{\sum_{k=1}^{i+j} W_{ak} w_{bk}}{\sqrt{\sum_{k=1}^{i+j} w_{ak}^2} \sqrt{\sum_{k=1}^{i+j} w_{bk}^2}} \quad (8)$$

Page 106-119 © MANTECH PUBLICATIONS 2019. All Rights Reserved

similarity for each sentence, a raw semantic vector is derived with sustenance of lexical database and also a word order vector is formed for each of the sentences using the same information form lexical database. For each word in a sentence contributes differently to the complete meaning of the whole sentence, the importance of a word is weighted by using information content resultant from corpus by combining the raw semantic with information from the corpus, a semantic vector is found for each of the two sentences. Semantic similarity is computed based on the two semantic vectors. An order similarity is computed using two order vectors. Finally, the overall similarity is derived by combining order similarity and semantic similarity.

Word Order Similarity

Sentence similarity based on the word order requires constructing the order vectors of the two sentences. If the sentence Sa has words (wa1,wa2,.....,wai) and sentence Sb has words (wb1,wb2,.....,wbi) then word order vectors for Sa and Sb are represented as follows.

$$L(S_a)=\{(w_{a1},w_{a2}),(w_{a1},w_{a3}),.....,(w_{a(i-1)},w_{ai})\} \quad (9)$$

$$L(S_b)=\{(w_{b1},w_{b2}),(w_{b1},w_{b3}),.....,(w_{b(i-1)},w_{bi})\} \quad (10)$$

where (wx,wy) ∈ L(Sa) U L(Sb) means wx is before wy. Then the similarity between Sa and Sb can be calculated based on the orders of words by equation 11.

$$\text{Word Order } S_a, S_b = \frac{(|L(S_a)\cap L(S_b)|)}{(|L(S_a)|\cup|L(S_b)|)} \quad (11)$$

Classification Using Machine Learning Algorithms

Machine learning is a method of data analysis that automates analytical model building. Using algorithms that iteratively learn from data, machine learning allows computers to find hidden insights without being explicitly programmed where to look. The process of machine learning is similar to that of data mining. Machine learning the concept of combining classifiers is proposed as a new direction for the improvement of the performance of individual classifiers. A classifier is a supervised function is a machine learning tool where the learned target attribute is categorical. It is used after the learning process to classify new records by giving them the best target attribute. The proposed method utilizes machine learning algorithms like support vector machine, maximum entropy, boosting, random forest and bagging.

Page 106-119 © MANTECH PUBLICATIONS 2019. All Rights Reserved

PERFORMANCE EVALUATION

One of the most commonly used corpora for paraphrase detection in Tamil language consists of 2500 sentence pairs as training dataset and 900 sentence pairs as test dataset. In that Conference, under the track Detection of Paraphrases on Indian Languages especially Tamil, corpus is morphologically extracted using Shallow Parser and classified through Support Vector Machine and Maximum Entropy. The following evaluation measures are used in the proposed system.

Accuracy

Accuracy is the traditional way to measure the performance of the system [9]. It is the most common measure of classification process. It can be calculated as the ratio of correctly classified sentences to total number of sentences. It can be calculated using equation 12.

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad (12)$$

Precision

Precision is the faction of retrieved instances that are relevant. Precision is also used in recall. The usage of "precision" in the field of information retrieved differs from the definition of accuracy and precision. It can be calculated using equation 13.

$$Precision = \frac{TP}{(TP+FP)} \quad (13)$$

Recall

Recall in information retrieval is the fraction of the documents that are relevant and the recall is also referred to as the true positive rate or sensitivity. It can be calculated using equation 14.

$$Recall = \frac{TP}{(TP+FN)} \quad (14)$$

F-Measure

F-Measure is a measure of test the accuracy and also defines as the harmonic mean of precision and recall of the test. It is required to optimize the system towards either precision or recall, which have more influence on final result. It can be calculated using equation 15.

$$F - Measure = 2 * \frac{(Precision * Recall)}{(Precision + Recall)} \quad [15]$$

Results and Discussion

Table 1 represents the precision, recall and F-M easure values obtained using SVM, Maximum Entropy, Boosting, Random

Page 106-119 © MANTECH PUBLICATIONS 2019. All Rights Reserved

Forest and Bagging classifiers. The result shows that SVM classifier's precision, recall and F-Measure values are high compared to other classifiers.

Page 106-119 © MANTECH PUBLICATIONS 2019. All Rights Reserved

Table 1 Performance Evaluation Measures

Class Method	Class	Precision	Recall	F-Measure
SVM	Paraphrase	0.80	0.81	0.80
	Not Paraphrase	0.73	0.72	0.72
Max Entropy	Paraphrase	0.74	0.66	0.70
	Not Paraphrase	0.58	0.68	0.63
Boosting	Paraphrase	0.78	0.86	0.82
	Not Paraphrase	0.77	0.66	0.71
RF	Paraphrase	0.79	0.79	0.79
	Not Paraphrase	0.70	0.70	0.70
Bagging	Paraphrase	0.78	0.78	0.78
	Not Paraphrase	0.69	0.69	0.69

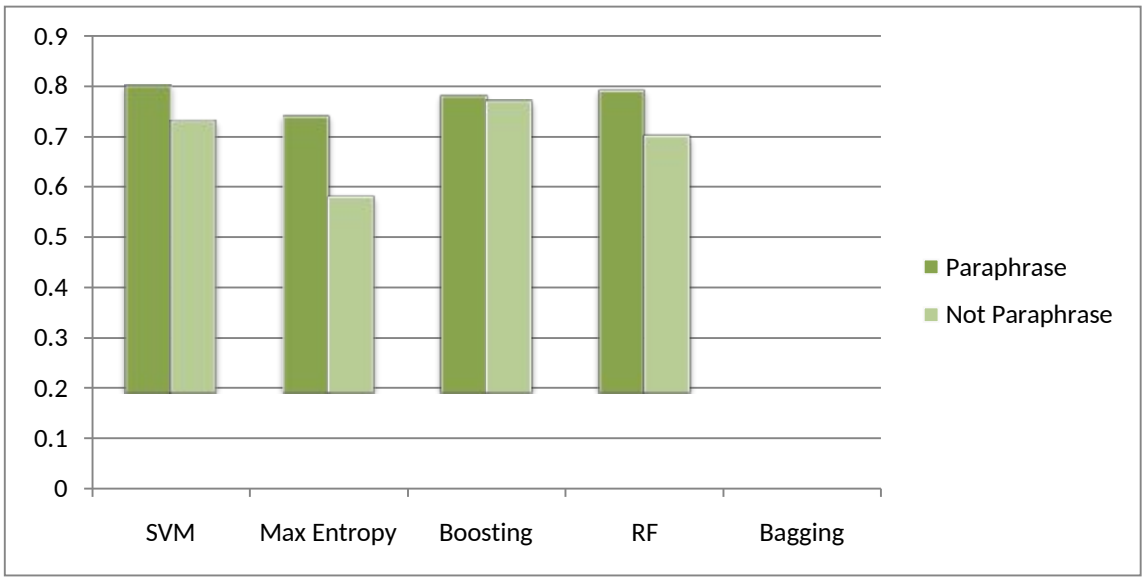


Figure 2 Comparison of Precision values

Figure 2 represents the precision value of all five classification algorithms. The result shows that SVM has high precision for paraphrase sentence and boosting technique has high precision for Not a Paraphrase sentence.

Page 106-119 © MANTECH PUBLICATIONS 2019. All Rights Reserved

Figure 3 represents the recall value of all five classification algorithms. The result shows that boosting technique has high precision for paraphrase sentence and SVM has high precision for Not a Paraphrase sentence.

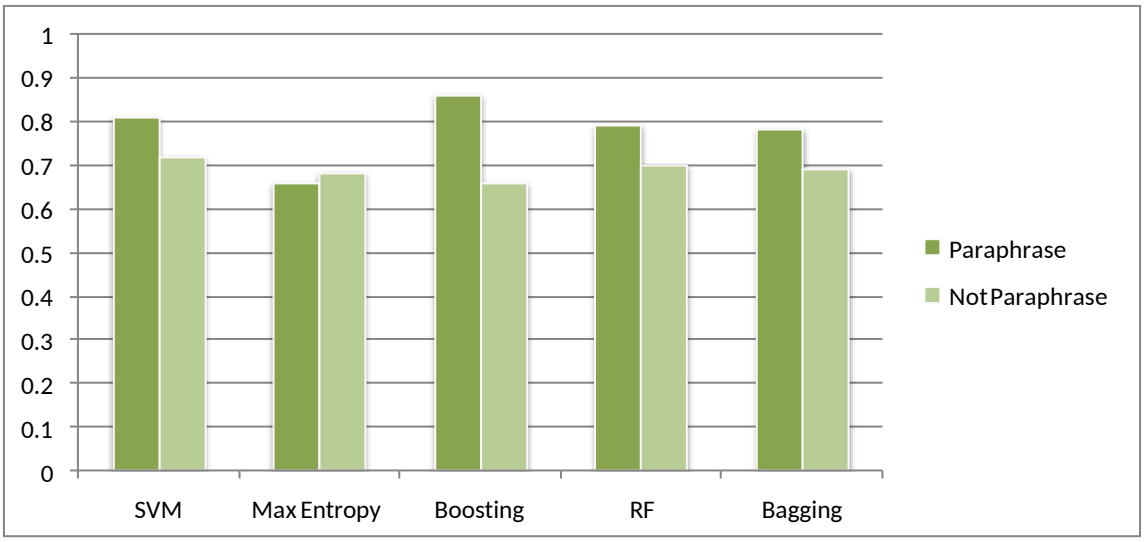


Figure 3 Comparison of Recall values

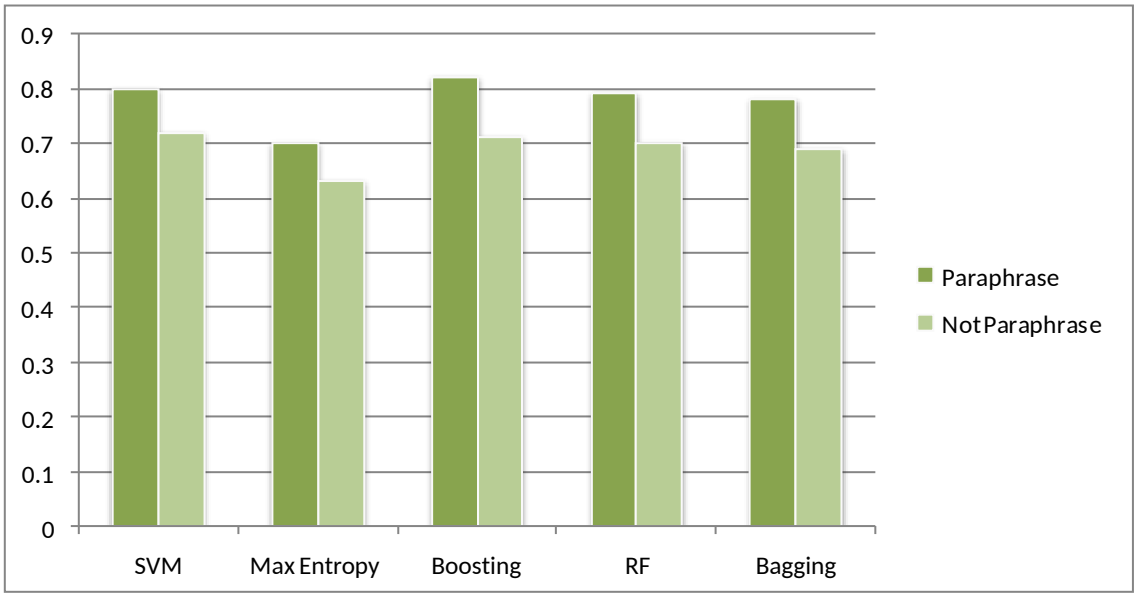


Figure 4 Comparison of F-Measure values

Page 106-119 © MANTECH PUBLICATIONS 2019. All Rights Reserved

Figure 4 represents the measure value of all five classification algorithm. The result shows that Boosting technique has high precision for paraphrase sentence and SVM has high precision for Not a Paraphrase sentence.

CONCLUSION AND FUTURE WORK

Paraphrase identification is important for text classification and retrieval. It represents the methods for measuring the similarity between sentences based on syntactic and semantic word level information. The sentences are classified using five supervised machine learning algorithms, such as Support Vector Machine, Maximum Entropy, boosting, Random forest, bagging. It shows that sixteen different syntactic and semantic features to represent the similarity between sentences.

The five different Machine learning algorithms have been considered for classification of given sentence pair into Paraphrase and Not-a-Paraphrase. The accuracy and performance of these methods are measured on the basis of parameters such as accuracy, precision, recall, F-Measure. In future work, more number of the statistical and semantic measures can be incorporated to improve the accuracy of the classifier models.

REFERENCES

I. Christopher D. Manning and Hinrich Sch¨utze. (1999), “Foundations of Statistical Natural Language Processing”. The MIT Press, Cambridge, Massachusetts.

II. Ditty Mathew and sumam mary Idicula. (2013), “Paraphrase identification of Malayalam sentences - an experience”, Fifth International Conference on advanced computing(ICoAC)

III. Ekaterina Pronoza and Elena Yagunova. (2015), “Low-level Features for Paraphrase Identification”, Advances in Artificial Intelligence and Soft Computing.

IV. Kozareva Z and Montoyo A. (2006), “Paraphrase Identification on the Basis of Supervised Machine Learning Techniques.” Advances in Natural Language Processing.

V. Li , D. McLean and Z.A. Bandar. (2006), “ Sentence similarity based on semantic nets and corpus statistics”, IEEE Transactions on Knowledge and Data Engineering , Volume: 18, Issue: 8.

VI. LinLi, Xia Hu, Bi-Yun Hu, Jun Wang, Yi-Ming Zhou. (2009), “Measuring Sentence Similarity

Page 106-119 © MANTECH PUBLICATIONS 2019. All Rights Reserved

From Different Aspects", Proceedings of the Eighth International Conference on Machine Learning and Cybernetics, Baoding.

VII. Long Qiu, Min-Yen Kan, and Tat-Seng Chua. (2006), "Paraphrase recognition via dissimilarity significance classification", In Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, pages 18–26, Sydney, Australia, July. Association for Computational Linguistics.

VIII. Palakorn, Xiaohua, Shen Xiajiong. (2008), "The Evaluation of Sentence Similarity Measures", Proceedings of the 10th international conference on Data Warehousing and Knowledge Discovery, Pages 305-316.

IX. Xiaoying Liu, Yiming Zhou, Ruoshizheng. (2008), "Measuring Semantic Similarity within Sentences", Proceedings of the Seventh International Conference on Machine Learning and Cybernetics, Kunming.

X. Zhang, Sun, Wang and He. (2011), "Calculating Statistical Similarity between Sentences," Journal of Convergence of Information Technology.

Page 106-119 © MANTECH PUBLICATIONS 2019. All Rights Reserved