ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 6, Special Issue 1, January 2017

International Conference on Recent Trends in Engineering and Science (ICRTES 2017)

20th-21st January 2017

Organized by

Research Development Cell, Government College of Engineering, Jalagon (M. S), India

The Study and Review of Paraphrase Detection Techniques in Machine Learning

Darshana S Bhole, Sandip S. Patil

PG Student, Department of Computer Engineering, SSBT's College of Engineering & Technology, Bambhori, Jalgaon[M.S.] India

Associate Professor, Department of Computer Engineering, SSBT's College of Engineering & Technology, Bambhori, Jalgaon[M.S.] India

ABSTARCT: Paraphrase is a process of computing the semantic similarity between sentences, which are not lexicographically similar. Though a number of metrics for English language have been proposed in literature, to quantify textual similarity; it addresses the problem for detection of monolingual text-text lexical similarity. Existing system for Indian Language paraphrase detection uses lexical similarity, based on supervised as it requires large scale training paraphrase corpus. The proposed method employs machine learning metrics, based on lexicography similarity with producing output as +1 for paraphrased, -1 for not paraphrased that means it takes a sentence as input and produces another sentence without changing its semantic. The Results are promising as they are independent to the latent variable and produce output even in missing input condition. This approach may also be used to identify Plagiarism of documents and to eliminate duplicates in a text repository.

KEYWORDS: Semantic similarity, Preprocessing, paraphrase detection.

I. INTRODUCTION

Paraphrase is emphasizing the meaning of something written or spoken using different words, especially to achieve greater clarity. Similarity refers to two or more documents that contain identical words, which may be exactly equal or nearly equal. It process token matching with the help of synonyms of tokens and produces the result [1]. The paraphrase detection is via identifying the synonym of the words in a sentence and match it with another sentence and also match the words and its phrasal words in it.One reason for the interest in such generation systems isthe possibility to automatically learn monolingual text-to-text generation strategies from aligned paraphrase corpora [2]. In fact, text-to-text generation is a particularly promising research direction given that there are naturally occurring examples of comparable texts that convey the same information yet are written in different styles.

II. CONTRIBUTION

In order to achieve the main aim of paraphrase detection of Hindi language, the machine learning algorithms are uses for the paraphrase detecting of the words. The training datasets are applied to the preprocessing and transformation algorithms; after that, a machine learning algorithmis used for learning how to classify sentences i.e. creating a model for input-output mappings, the key point is how to use a variety of strategies to get the semantic similarity of the two sentences. When changing Word-Form orPart-Of-Speech to do transformations on the sentences, and then different forms of a sentence will be turned into a same one, for that purpose use the supervised learning method. First the system is trained with a corpus of labeled sentences pairs. Then, use the learned knowledge to identified whether two sentences are paraphrases.Accordingly gives the output in the form of binary numbers as +1 for paraphrased and -1 for not paraphrased.

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 6, Special Issue 1, January 2017

International Conference on Recent Trends in Engineering and Science (ICRTES 2017)

20th-21st January 2017

Organized by

Research Development Cell, Government College of Engineering, Jalagon (M. S), India

III. OBJECTIVES

1) Checking of the semantic similarity and synonyms of the lexical words with the help of word net or others between two sentences.
2) Forms the result as if two sentences are lexicographically similar then, output as paraphrased otherwise not paraphrased and to simplify the sentences.

IV. RELATED WORK

- Vigneshvaran et al.in[1], presented the approach for detecting paraphrase by counting the tokens and produces the result from the SVM classifier using Plagiarism detection corpus(PAN). With the help of synonyms of tokens, it processes the token matching and finally produces the result by SVM classifier. SVM is binary classifier that processes threshold values and binary values and produces output as +1 or -1. This SVM classifier is not suitable for optimal multiclass design.
- Cordeiro et al.in[2], uses a new metric, the Sumo-Metric,for finding paraphrases,the performance of 5 metrics over 4 corpora, it is for monolingual text to text generation. The conclusion is that the Sumo- Metric performed betterin terms of F-Measure and Accuracy. It forms the lexical links to the sentences, various corpora sets are compared determinant or between verbs or nouns/names.
- Sazianti et al. in[3], introduced the metaphor for the semantic similarity of two sentences with the synonyms matching from WorldNet while applying the vector to the words. The result get after concerning the output as 3 similarity measures comparisons as better in measuring the semantic similarity between sentences. However more testing and evaluation work to be conducted for real test data and human experts.
- Bing-Quan et al.in[4], presented the a combination of rule based and supervised learning method to recognize paraphrases. Sentence pairs are collected form QA system gives the exceed precision as compared to existing methods. After comparing gets these method is not suitable for single corpus or parallel corpus.
- El-Sayed al.in[5], evaluated the five lexical similarity metrics by using the three machine learning paradigms to detect paraphrase. Applying statistical test result get the multilayer perceptron perform significantly better than Naïve Bayes in terms of accuracy and precision. However MLP wants more training datasets.
- Diego Uribe al in[6], uses the paraphrasing pairs to making the relationship between the sentences. It uses classification technique based on logistic regression. The experimental result shows the acceptable trade off between precision and recall values. This can detect paraphrase pairs but erroneously sometimes it can classified positive as negative instances.
- Sethi et al.in[7], addresses the problem of statistical analysis of the plagiarized text up to certain extent. Reframing of the Hindi sentences and produce the output. This is helpful in making the robot understand different forms of sentences. This system extend to the decision support system that is work in similar manner.

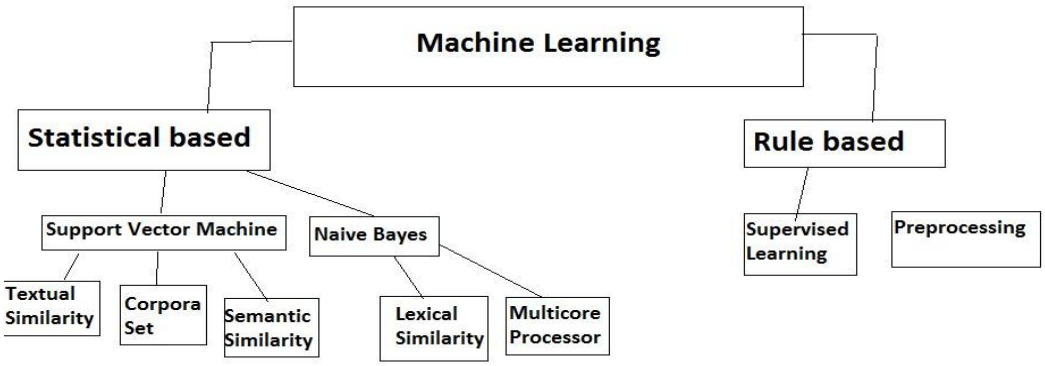

Figure 1: Structure of Literature Review

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 6, Special Issue 1, January 2017

International Conference on Recent Trends in Engineering and Science (ICRTES 2017)

20th-21st January 2017

Organized by

Research Development Cell, Government College of Engineering, Jalagon (M. S), India

The Machine Learning deals with powerful way to categorized the rule based and statistical based approach to paraphrases the pairs.

V. PROPOSED SOLUTION

The proposed solution focuses on paraphrase detection of Hindi language in order to facilitate and improve in a statistical based model of classifier that is uses in the architecture, to detect the lexical similarity of words.The integration of the statistical model of the proposed architecture omitting the best result as in binary values, accordingly forms the output as +1 for paraphrased the both sentences and -1 for not paraphrase. Supervised learning is provided while in preprocessing, during the words are recorded. Therefore, adding up the number of times word occurs in one sentence is calculated and accordingly find out the paraphrasing sentences by applying it to the classifier.The approach used in proposed system is first the preprocessing of data is done. Then the preprocessed data is given as an input to the statistical based support vector machine algorithm. Finally the paraphrase detection is done and the results of the model is tested by the testing datasets that are again provides to the classifier.The model uses WorldNet as their central resource. Most of the words are same as the once in other sentence, so only substitute the synonyms of the word.Simple sentence similarity for paraphrase:

$$\text{sim}(\text{sen1}, \text{sen2}) = \frac{\sum(\alpha i(\text{MaxS}(\text{wi}, \text{wj})) + \sum \alpha j(\text{MaxS}(\text{wi}, \text{wj}))}{\sum(\beta(\alpha i + \alpha j))}$$

Where , Sim(sen1,sen2) means the semantic similarity of the two sentences. $\text{MaxS}(\text{wi}, \text{wj})$shows themax similarity of the word. The parameter α is decided by the (wi, wj)

The architecture of the proposed system is shown in Figure 2. Inputs to the proposed system is training datasets. The data sets contains the collection of sentences. The data preprocessing consists of the stemming, removal of stop words, tokenization as shown . After preprocessing the data is to be given as input to the classifier. The goal isto identify overall strategy to capture paraphrase between two sentences.The Proposed architecture is given as:

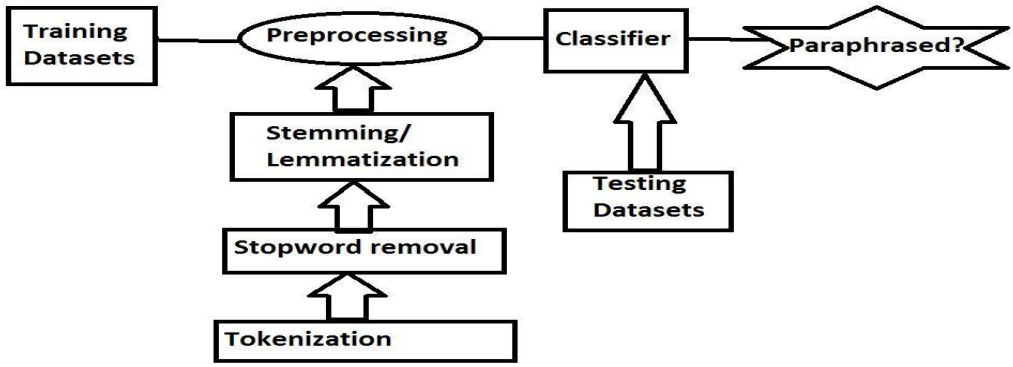

Fig 2: Architecture of the proposed system

This system is applied to the Indian language for paraphrase detection, in this; Hindi dataset are giving as a training datasets. In preprocessing the tokenization of the words are form then stop word removal stemming/ lemmatization of the words for this uses the porter algorithm with the suffix, prefix removal. It uses the Hindi word datasets applying to the classifier gets the output that the given sentences are paraphrased or not paraphrased.

VI. EXPERIMENTAL RESULTS

Figures shows the results after preprocessing of the sentences, in fig 3 the output of the stop word removal after comparing the both sentences. In fig 4 the output after removing the suffixes from the sentences. It uses the concept

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 6, Special Issue 1, January 2017

International Conference on Recent Trends in Engineering and Science (ICRTES 2017)

20th-21st January 2017

Organized by

Research Development Cell, Government College of Engineering, Jalagon (M. S), India

from the Porter Stemmer algorithm. The DPIL dataset by Amrita University is used for the detection of paraphrasing the sentences.

```
Enter Pid:=HIN
0001
Enter pid=HIN0001
Sentence1:=भारतीय मुस्लिमों की वजह से नहीं पनप सकता आईएस|
After removing Stopwords Sentence1:=[भारतीय, मुस्लिमों, वजह, पनप, आईएस]
Sentence2:=भारत में कभी वर्चस्व कायम नहीं कर सकता आईएस|
After removing Stopwords Sentence2:=[भारत, कभी, वर्चस्व, कायम, आईएस]
```

Figure 3: Removing the Stop words of two sentences.

```
Enter Pid:=HIN
0250
Enter pid=HIN0250
Sentence1:=उसने उन्नीस -बीस अक्टूबर दो हज़ार छह की रात अनिल मंडल की
                        हत्या करने के बाद उसकी सिरकटी लाश बोरी में डालकर तिहाड़ जेल के गेट
                        नंबर तीन के सामने फेंक दिया था।
After removing Stopwords Sentence1:=[उसने, उन्नीस, -बीस, अक्टूबर, हज़ार, छह, रात, अनिल, मंडल, हत्या, उसकी, सिरकटी, लाश, बोरी, डालकर, तिहाड़, जेल, गेट, नंबर, तीन, सामने, फेंक]
After removing Suffix Sentence1:=[उसने, उन्नीस, -बीस, अक्टूबर, हज़ार, छह, रात, अनिल, मंडल, हत्या, उसकी, सिरक, लाश, बोरी, डालकर, तिहाड़, जेल, गेट, नंबर, तीन, सामने, फेंक]
Sentence2:=झा ने दो हजार छह में अनिल मंडल का सिर और शरीर के अंग काटने
                        के बाद हत्या कर दी थी और उसका शव जेल के बाहर फेंक दिया।
After removing Stopwords Sentence2:=[झा, हजार, छह, अनिल, मंडल, सिर, शरीर, अंग, काटने, हत्या, दी, उसका, शव, जेल, बाहर, फेंक, दिया ]
After removing Suffix Sentence2:=[झा, हजार, छह, अनिल, मंडल, सिर, शरीर, अंग, काट, हत्या, दी, उसका, शव, जेल, बाहर, फेंक, दिया ]
```

Figure 4: Suffix removal of two sentences

Algorithm 1: Porter Stemmer Algorithm

Step 1: Gets rid of plurals and calculate the length

Step 2: If length is greater then 4 then remove ियां, ियो

Step 3: If length is greater then 3 then remove ाए,ाओ, ुआ,ुओ,ये,ेन, ेण, ीय,टी,ार,ाई

Step 4: If length is greater then 2 then remove ा, े, ी,ो,ि,अ,

Step 5: Display the new sentences

This algorithm conflates derived words to a stem(root). Stem is not the morphological root, other stemmers might produce different stems. Porter algorithms works according to the length of the words.

VI. CONCLUSION

This paper examines the approach for measuring the semantic similarity of the two sentences. Stemmer algorithm is uses for removing the suffixes. The two sentences are paraphrase or not is calculating by token count, token matching and synonyms token matching by using the classifier. The experimental result shows the group of root words that are ready for detection of paraphrases. In future uses the classifier algorithm to detect the paraphrases of sentences.

REFERENCES

[1] Vigneshvaran, Jayabalan, VijayaKathiravan,"An Eccentric Approach for Paraphrase Detection Using Semantic Matching and Support Vector Machine",2014 International Conference on Intelligent Computing Applications.
[2] JaonCordeiro, Gael Dias, Pavel Brazdil, "A Metric for Paraphrase Detection",Proceedings of the International Multi-Conferenceon Computing in the Global Information Technology (ICCGI'07) 2007 IEEE.
[3] SazianiMohdSaad,SitiSakiraKamarudin,"Comparative Analysis of Similarity Measures forSentence Level Semantic Measurement of Text",2013 IEEE International Conference on Control System, Computing and Engineering, 29 Nov. - 1 Dec. 2013, Penang, Malaysia
[4] BING-Quan LIU, Shuai XU, BAO-XUN Wang,"A combination of rule and supervised learning approach to recognize paraphrases",Proceedings of the Eighth International Conference on Machine Learning and Cybernetics, Baoding, 12-15 July 2009
[5] El-Sayed M. El-Alfy,"Statistical Analysis of ML-Based ParaphraseDetectors with Lexical Similarity Metrics" IEEE 2014
[6] Diego Uribe,"Recognition of Paraphrasing Pairs",Electronics, Robotics and Automotive Mechanics Conference 2008
[7] NandiniSethi, Prateek Agrawal, VishuMadaan and Sanjay Kumar Singh," A Novel Approach to paraphrase Hindi Sentences using Natural Language Processing", Indian Journal of Science and Technology, vol 9(28), july 2016.