A Survey on Paraphrase Detection Techniques for Indian Regional Languages

Shruti Srivastava
PG Student
Department of Computer Engineering-PCE,
Mumbai University, India

Sharvari Govilkar
Associate Professor
Department of Computer Engineering-PCE,
Mumbai University, India

ABSTRACT

Whenever the text contains multiple ways of saying "the same thing," but the application requires the same treatment of those various alternatives, an automated paraphrase recognition mechanism would be useful. One reason why paraphrase recognition systems have been difficult to build is because paraphrases are hard to define. Although the strict interpretation of the term "paraphrase" is quite narrow because it requires exactly identical meaning, in linguistics literature paraphrases are most often characterized by an approximate equivalence of meaning across sentences or phrases. This paper presents a survey of paraphrase detection techniques for Indian regional languages.

General Terms

Your general terms must be any term which can be used for general classification of the submitted material such as Pattern Recognition, Security, Algorithms et. al.

Keywords

Paraphrase detection, Textual Similarity metrics, String similarity, Semantic relatedness, Statistical and semantic analysis, Bi-CNN-MI.

1. INTRODUCTION

Natural Language Processing (NLP) focuses on developing computer systems that can analyze, understand and generate natural human-languages. One of the major difficulties faced in natural language processing is ambiguity where the same text has several possible interpretations. Another equally challenging aspect is that the same content can be conveyed in different ways. This is termed as Paraphrasing. Paraphrases can occur at the word level, phrase level, sentence level or discourse level.

This paper is organized into 6 sections. The section 1 presents the introduction, section 2 describes about paraphrase detection. Related work is presented in section 3 which describes various paraphrase detection techniques applied on Indian regional languages. Section 4 gives brief idea about the general architecture for paraphrase detection. Various paraphrase detection techniques are discussed in section 5. Section 6 offer comparison and observations of various paraphrase detection techniques applied on Indian regional Languages. Conclusion is made in section 7.

2. PARAPHRASE DETECTION

The Paraphrase detection is a classic NLP task which is a classification problem. Given a pair of sentences, the system is required to assess if the two sentences carry the same meaning, to classify them paraphrase, or not paraphrase otherwise. In order to obtain high accuracy on this task, thorough syntactic and semantic analysis of the two text entities is required.

In a paraphrasing system the need to estimate the semantic equivalence between two sentences is crucial. Therefore, it is important to come up with a process which will be able to recognize, extract or generate paraphrases, yielding as a result sentences (or phrases or patterns) that will maintain the semantic notion of the original sentence.

Semantic notion considers the meaning of the sentence which relies on the relation between words or phrases. Paraphrasing methods recognize, generate or extract phrases, sentences or longer natural language expressions that convey almost the same information.

In syntactic approach the paraphrase is decided on the number of syntactic rules used to transform a sentence into the other; the rules can be derived from grammatical analysis or be based on a statistical approach.

Such detection is widely used to remove the tremendous amount of duplicate information on the Internet. It is also used to handle the overlap of semantic components in texts. Such components are used in various natural language applications such as word sense discrimination, summarization, automatic thesaurus extraction, question-and answer generation, machine translation and plagiarist or analogical relation identification.

3. RELATED WORK

This section cites the related work that uses various techniques for paraphrase identification. Paraphrase detection is important for applications such as summarization, information retrieval, information extraction and question answering, etc. Various approaches have been developed for automatic paraphrase identification which was summarized here.

Jun Choi Lee & Yu-N Cheah [1] presented a semantic relatedness measures that based on Synset shortest path in WordNet for paraphrase detection. This method used distance per word in the sentence to measure semantic relatedness between two sentences. The proposed method is then evaluated using a paraphrase detection evaluation based on the Microsoft Research Paraphrase Corpus and showed reasonable results compare to other similarity and semantic relatedness. Author suggested that this approach may be used for question answering or content ranking and in text understanding also.

Chen Liang, Praveen Paritosh, Vinodh Rajendran, Kenneth D. Forbus [2] proposed a new alignment based approach to learn semantic similarity. They used a hybrid representation of attributed relational graphs to encode lexical, syntactic and semantic information. Three important components of the method were graph extractor, structural aligner and similarity estimator. Two components structural alignment and similarity estimation integrated through alignment as feature extraction and as latent variable. This alignment improved through two structural constraints and achieved results competitive with other state-of-the-art models on the MSRP corpus. It can be used to test on other semantic NLP task like textual entailment and question answering.

Hoang-Quoac Nguyen-son, Yusute Miyao and Isao Echizen [3] proposed approach for paraphone detection using identical phrase and similar word matching. Author explains Similarity Metric (SimMat) which was calculated by matching identical phrases with maximum length, removing minor words, matching of similar words by Kuhn and Munkre algorithm, calculate related matching metric (RelMat) and finally calculate the SimMat by combining penalty metric and RelMat. The system achieves 77.6% accuracy which was higher than the previous method.

Jun Choi Lee & Yu-N Cheah [4] presented a novel approach in using synonyms to enhance the similarity metrics for paraphrase detection. Instead of using different text features such as hypernyms and word similarity, it expands the lexical in the text using synonyms. This method receives two texts and tokenizes them into two arrays, then compares the arrays for finding match list. If match was found then calculate similarity otherwise lemmatize the remaining terms and evaluates the synonyms for that term and then calculate similarity. The synonyms are extracted using WordNet lexical database.

Wenpeng Yin and Hinrich Schiitze [5] presented a new deep learning architecture Bi-CNN-MI for paraphrase identification. Most paraphrase identification was focused only on one level of granularity but they explained that paraphrase identification requires multiple levels of granularity. In Bi-CNN-MI, the sentence analysis network CNN-SM, the sentence interaction model CNN-IN and a logistic regression explained which performs paraphrase identification. The architecture has seven layers and also consists of creation of unigram feature matrix, short n-gram feature matrix, long n-gram feature matrix and sentence level feature matrix. These matrices are then put to logistic classifier for paraphrase identification. Adding MT metric as input to the Bi-CNN-MI logistic regression improves the performance by the accuracy of 78.4% and F1 of 84.6%. Author concluded that Bi-CNN-MI can also be applied to sentence matching, question answering and other tasks in future.

Wenpeng Yin & Hinrich Schutze [6] presented an extension to Ji and Eisenstein’s proposed TF-KLD to weight features and used non-negative factorization to learn latent sentence representation. An extensive TF-KLD-KNN computes its weights as the average of the weights of its k nearest neighbours which can be determined by cosine measure over embedding space. MSPR corpus was used for evaluation. In this approach paraphrase identification could be improved by the utilization of continuous and discontinuous phrase embeddings.

Majid Mohebbi and Alireza Talebpour [7] presented a EMM (Extended Maximum Matching) algorithm for paraphrase identification for measuring semantic similarity of text using graph theory. EMM applied separately to each pair of nouns, verbs, adjectives, adverbs, cardinals and no edges between different classes. This graph based algorithm used word similarity information extracted from WordNet. This algorithm did not find the max similarity for each word but it selected only certain weights. The approach affected by the order of appearance of the words and by choosing special edge.

Zia.Ul-Qayyam and Wasif Altaf [8] proposed paraphrase identification approach based on improved pre-processing and semantic heuristics based enhanced features set. The system produces better result than the state-of-art system. Cosine similarity measures have used to analyze pre processing applied by base-line system. Five different systems were developed and applied for base-line pre-processing analysis. After which paradetect misclassification resulted in highlighting advantages and disadvantages of semantic heuristic based features.

Nitin Madani, Joel Tereault and Martin Enedorow [9] employed 8 different MT Metrics for identifying paraphrases across 2 different data sets –the Microsoft Research Paraphrase Corpus (MSRP) and Plagiarism Detection Corpus (PAN). They described the use of meta-classifier and try to find evidence of each metrics strength in each data set. Results for PAN dataset are much better than for MSRP corpus and also discovered that TERP metric provide good performance and outperforms various previous approaches. The method explained the specific examples to prove the strength of new metric over simple n-gram overlap (BLUE, NIST) and edit distance based metrics (TER,WER,PER). They released error analysis data of 100 pairs for MSRP corpus and 100 pairs for PAN corpus as they will be useful to other researchers.

Socher, Eric Huang, Pennington, Andrew, Christopher [10] proposed an approach that incorporates the similarities between both single word features and multi-word phrases extracted from the nodes of parse trees. The recursive auto-encoder is a recursive neural network that learns feature representations for each node in the tree, such that the word vectors underneath each node can be recursively reconstructed. These feature representations are used to compute a similarity matrix that compares both the single words as well as all nonterminal node features in both sentences. The dynamic pooling layer is used to keep as much of the resulting global information of this comparison as possible, and to deal with the arbitrary length of the two sentences. A softmax classifier, relying on features based on the similarity matrix together with simple features such as the difference in sentence length, or the percentage of words and phrases in one sentence that are in the other sentence and vice-versa, is finally used to classify whether the two sentences are paraphrases or not.

Anupriya Rajkumar and Dr. A. Chitra [11] presented the work carried out on paraphrase recognition using Neural Network classifier. Neural network architecture explained as a foremost machine learning techniques. Automatic paraphrase recognition can be implemented by machine learning algorithm which undergoes in 3 steps. In the first step, sentence have been collected from 4076 training set and 1726 test set on MSRP corpus. In second step, feature extraction involves combination of purely lexical, syntactic, lexical-semantic and lexical-syntactic features. In third step, architecture of back propagation network consisting 3 layers has been utilized. The neural network based recognizer system can be used in Question answering systems and for plagiarism detection in document collection.

Samuel Fernando and Mark Stevenson [12] presented a matrix similarity approach for paraphrase identification. This method represents word-to-word similarities rather than maximal similarities between sentences. They experimented with six word-net similarity metrics viz Jcn, Lch, Lesk, Lin, Res, Wup to populate the similarity matrix. The WordNet-based lexical similarity measure attempts to find the highest similarity score for a word pair. The approach described was evaluated against the MSRP corpus to find classification threshold for similarity score which maximized accuracy. The matrix similarity approach outperforms both random and vector-based baselines for all six of the similarity measures.

Mihai Lintean and Vaile Rus [13] proposed an approach which considers both similarity and dissimilarity between two sentences. The approach used word semantics and weighted dependencies to compute degree of similarity at word and syntactic level. The similarity and dissimilarity scores computed by mapping the input sentences into set of dependencies and then identify common and non-common dependencies between them. The paraphrase score was calculated by the ratio of similarity and dissimilarity score. If the score is above threshold then the sentence pair considered as paraphrase otherwise considered as non-paraphrase.

Dipanjan Das and Noah A. Smith [14] used generative model that creates paraphrases of sentences and probabilistic inferencing to reason about whether or not two sentences have paraphrase relationship. Model applied used quasi-synchronous dependency grammars effectively incorporating syntax and lexical semantics and hidden alignment between trees of two sentences. They also experimented with combination of their model with a complementary logistic regression model using product of experts. Highest performance accuracy of 83.42%, with 1.0000 precision and 95.29% recall was achieved using oracle ensemble..

Rus, Philip, Mihai, Danielle, Arthur [15] addressed the task of paraphrase identification by computing the degree of subsumption at lexical and syntactic level between two sentences in a bidirectional manner: from Text A to Text B and from Text B to Text A. The approach relied on a unidirectional approach that was initially developed to recognize the sentence-to-sentence relation of entailment. They used similarity to decide paraphrasing, simply discarding dissimilarities without carefully analyzing their importance to the final decision. The similarity was computed as a weighted sum of lexical matching, i.e. direct matching of words enhanced with synonymy information from WordNet, and syntactic matching, i.e., dependency overlap. Dependencies were derived from a phrase-based parser which outputs the major phrases in a sentence and organizes them hierarchically into a parse tree. The approach offers competitive results with other approaches on a standard MS Paraphrase corpus.

Joao Cordeiro, Gael Dais and Pravel Brazil [16] proposed a Sumo-metric for paraphrase detection which solves the limitations of previous ones.Author tested the performance of 5 metrics over 3 corpora. Out of 5 metric, Lerenshtein Distance and BLUE measure performs poor in all cases. Whereas Word Simple N-graph overlap stood second and Exclusive LCP (Longest Common Prefix) stood third in experimental result ranking. The Sumo-metric outperforms by an average of 98.53% of all 9974 sentence pairs.

Kozareva and Montoyo [17] created feature vectors around lexical and semantic similarity attributes to train SVM, k-Nearest Neighbour technique and Maximum Entropy classifiers. The features they used skip grams, longest common subsequence, and cardinal similarity information based on WordNet similarity package, where all features used were bidirectional. The experiments revealed that simple features relying on common consecutive or in sequence matches can resolve a large number of paraphrases correctly, and that combining multiple classifiers can also be beneficial.SVM was found to be outperforms all classifiers in all experimental setting.

N. Sethi, P. Agrawal, V. Madaan, S.K. Singh [18] proposed a technique for paraphrasing or re-framing Hindi sentences using NLP. The main steps involved dividing the paragraph into sentences, tokenizing the sentences into words, applying reframing rules and then combining the results to form new paragraphs. Using this algorithm Hindi sentence as an input could produce another semantically equivalent sentence by applying synonyms and antonyms replacement. This replacement could be possible by sentence mapping with database. With the help of reframing, complex Hindi sentences can be changed into its simplified form. This method may be further useful in making a robot as well as use as Hindi tutor which understood different forms of sentences.

Ashwini Gadaag, Dr. B.M. Sagar, Mr. Rajshekar Murthy [19] presented an approach to paraphrase a Kannada sentence using synonym substitution, statistical model and semantic feature method. The paper described different paraphrase recognition technique for different languages like Chinese and Spanish and discussed about merit and demerits of each of them. They concluded that it was better to use the combination of different techniques instead of using only one.

Author Ditty Mathew & Dr. Suman Mary Idicula [20] proposed that paraphrase can be detected using the sentence similarly measures such as symbolic similarly, structure similarity and semantic similarity for Malayalam sentences. The statistical methods like Jaccard Similarity, Dice similarity, cosine similarity, word order similarity, word distance similarity are used to find similarity of sentences. For finding semantic similarity, UNL graph based matching score is used. Universal Natural Language (UNL) gives the semantic representation of sentences in a graphical form which can be easily understandable by computer. Overall similarity of two sentences was calculated by combining these two measures. Future scope suggested by the paper is the method can be used for complex sentence too.

4. GENERAL ARCHITECTURE FOR PARAPHRASE DETECTION

Paraphrase detection determines whether two phrases of arbitrary length and form capture the same meaning. Identifying paraphrases is an important task that is used in information retrieval, question answering, text summarization, plagiarism detection and evaluation of machine translation and other NLP tasks.

Paraphrases can occur at the word level, phrase level, sentence level or discourse level. A typical example of sentence level paraphrasing is the following pair of statements “Tata acquires Jaguar” and “Jaguar sold to Tata”. Hence, Paraphrase detection or paraphrase identification is the task to identify sentences with similar meaning by evaluating the similarity between two texts based on lexical, semantic and structural similarity.

Lexical similarity is a measure of the degree to which the word sets of two given languages are similar. A lexical similarity of 1 (or 100%) would mean a total overlap between vocabularies, whereas 0 means there are no common words. Semantic similarity is a confidence score that reflects the semantic relation between the meanings of two sentences. Structural relations include relations between words and the distances between words. If the structures of two sentences are similar, then there is a possibility that they convey similar meanings. Thus for paraphrase identification the similarity score in two sentences must be calculated.

Figure.3.1 depicts the general work-flow of Paraphrase Detection System. The four main modules and the sub modules are described in the following subsections:

1. Pre-processing
2. String similarity metrics
3. Classifier
4. Performance evaluation

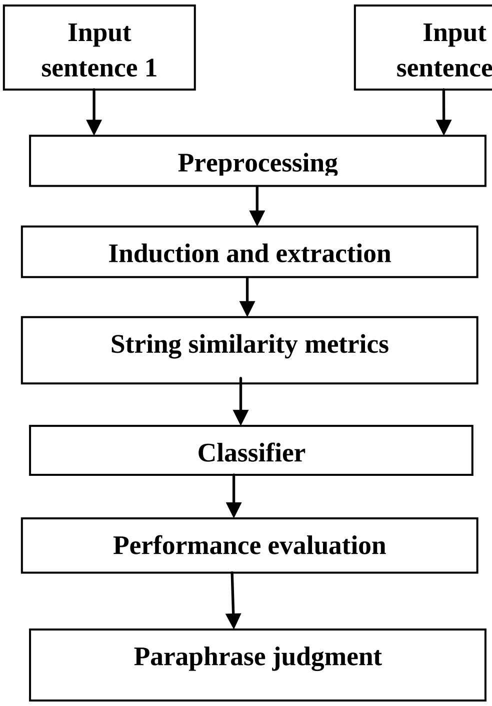

Figure 1 Modules of the paraphrase detection system

5. TECHNIQUES FOR PARAPHRASE DETECTION

There are different approaches, both supervised and unsupervised, have been proposed for paraphrase detection , ranging from simple level like word/n-gram overlapping, string matching, to more complicated ones like semantic word similarity, word alignment, syntactic structure, etc. In order to obtain high accuracy on this task, thorough syntactic and semantic analysis of the two text entities is required.

5.1 Textual Similarity metrics

Various measures [16] have been proposed for detecting similarity between pairs of English text fragments (sentences or phrases) for. The Levenshtein Distance, also known as the Edit Distance, sim_{o} metric, sim_{exo} i.e. exclusive longest common prefix (LCP) n-gram metric ,the Longest Common Prefix (LCP), BLUE metric , SumoMetric. SumoMetric performed better than any other measure over all corpora either in terms of F-Measure and Accuracy.

5.2 Using weighted dependencies and word semantics

This approach [12] used word semantics and weighted dependencies to compute degrees of similarity and dissimilarity at word/concept level and at syntactic level between two English sentences being judged as being paraphrases or not. The paraphrase score was calculated by taking the ratio of similarity score and dissimilarity score. The proposed approach offers state of-the-art performance due to the use of syntactic information.

5.3 Identical Phrase and Similar Word Matching

SimMat metric [3] developed in this approach quantifies the similarity between two English language sentences. It was calculated using the matching of identical phrases and similar words. Phrase-by-phrase matching is done using a heuristic algorithm that determines the longest duplicate phrase in each iteration. Word matching is done using the Kuhn-Munkres algorithm. The metric achieved the highest paraphrase detection accuracy (77.6%) when it was combined with eight standard machine translation metrics.

5.4 String similarity with synonyms

This is most common method [4] of paraphrase detection as it can be used for Hindi, English and Spanish languages. Instead of using different text features such as hypernyms and word similarity this method expands the lexical in the text using synonym. It evaluates the synonyms only for the term that not match in the comparing text to avoid unnecessary comparisons. It provides a simpler way in considering synonyms in text similarity calculation. The cosine similarity with synonyms shows better results in Precision and Recall as compared to the original cosine similarity in detecting paraphrases.

5.5 Using Semantic relatedness

The method [1] could detect paraphrases from English as well as Hindi language sentences. The semantic relatedness measure uses the shortest path value between synsets in WordNet to compute sentence-to-sentence semantic relatedness measurement. Semantic relatedness identifies text with exact match as well as measure the relatedness among the compared text which outperforms other similarity and semantic relatedness methods.

5.6 Statistical and semantic analysis

This method [20] focus on the statistical measures and semantic analysis of Malayalam sentences to detect the paraphrases. The statistical techniques selected in this work are based on word set, word vector, word order and word distance. The UNL graph matching method is used for the Semantic similarity task. The overall similarity of two sentences calculated by combining these two measures provides a better result.

5.7 Bi-CNN-MI

It is a deep learning architecture [5] used for English language based on the fact for paraphrase identification requires comparing two sentences on multiple levels of granularity. It was designed to produce units at fixed levels and only units at the same level are compared with each other due to which Bi-CNN-MI outperforms all other system.

6. COMPARATIVE ANALYSIS TABLE

In this section we compare various paraphrase detection techniques for different Languages:

Table 1.Comparision of various Paraphrase Detection Techniques

Sr No.	Name of the technique	Language Implemented on/Observations
1	**Textual Similarity metrics** (i) Levenshtein Edit Distance (ii)Word Simple N-gram Overlap (iii) Exclusive LCP N-gram Overlap (iv) BLEU measure (v) Sumo-Metric	**English**
		Sumo-Metric performed better than any other metrics over all corpora as it considers all pairs having a high degree of lexical reordering, and different syntactic structure.
2	**Using weighted dependencies and word semantics**	**English**
		Considers both similarity and dissimilarity between two sentences. The similarity between sentences is computed using word-to-word similarity metrics instead of simple word matching or synonymy information in a thesaurus.
3	**Identical Phrase and Similar Word Matching**	**English, Hindi**
		Word matching is done using the Kuhn-Munkres algorithm and phrase matching using heuristic algorithm to determine longest duplicate phrase.
4	**String similarity with synonyms**	**English, Hindi, Spanish**
		This method only evaluates the synonyms for the term that not exist in the matching list to avoid the unnecessary computational task.
5	**Semantic relatedness**	**English, Hindi**
		Semantic relatedness not only able to identify text with an exact match in terms of meaning, bus also measures the relationship degree or relatedness among the compared text.
6	**Statistical similarity and semantic similarity**	**Malayalam,Hindi**
		Combination of statistical similarity and semantic similarity score results the overall similarity score.
7	**Bi-CNN-MI**	**Englis**h
		Bi-CNN-MI for paraphrase identification based on comparing two sentences on multiple levels of granularity.

7. CONCLUSION

Paraphrase identification/detection is an important task as that can be used as a feature to improve many other NLP tasks as Information Retrieval, Machine Translation Evaluation, Text Summarization, Question and Answering, and others. Besides this, analyzing social data like tweets of social network Twitter is a field of growing interest for different purposes.

Due to its vast importance several approaches were proposed for automatic paraphrase detection in English language like logic-based, vector-based model, string similarity, syntactic similarity, machine learning, and machine translation. Among these approaches, Vector based approach (Bi-CNN-MI), Machine translation approach (SimMat Metric, Sumo Metric) have been recognized as most effective methods. Whereas string similarity approaches (Synonyms matching, and semantic relatedness) using WordNet were easy to implement and can be used for the task of paraphrase detection in Hindi languages.

The report hereby presented gives a brief idea of all the techniques used and implemented on most of the English language.

The literature survey and the study from past decades are considered and the conclusion can be drawn that for mostly research is continuing for Indian regional languages like Hindi, Malayalam, and Kannada. Moreover less work has been done in Hindi Language and no work has been done yet for Marathi language. So considering these facts, my future scope is to implement paraphrase detection technique for Hindi or Marathi language.

8. ACKNOWLEDGEMENT

I am using this opportunity to express my gratitude to thank all the people who contributed in some way to the work described in this paper. My deepest thanks to my project guide for giving timely inputs and giving me intellectual freedom of work. I express my thanks head of computer department and to the principal of Pillai College of Engg., New Panvel for extending his support.

9. REFERENCES

[1] Lee, Jun Choi and Cheah, Yu-N (2016) *Paraphrase Detection using Semantic Relatedness based on Synset Shortest Path in WordNet.* In: International Conference on Advanced Informatics: Concepts, Theory and Applications, 16-17 August 2016, Parkroyal Penang Resort.

[2] Chen Liang, Praveen Paritosh, Vinodh Rajendran, Kenneth D. Forbus, Learning Paraphrase Identification with Structural Alignment Conference: IJCAI 2016, At New York

[3] Hoang-Quoc Nguyen-Son, Yusuke Miyao, Isao Echizen, Paraphrase Detection Based on Identical Phrase and Similar Word Matching, 29th Pacific Asia Conference on Language, 2015.

[4] J.C. Lee, and Y. Cheah. “Paraphrase Detection using String Similarity with Synonyms.” The Fourth Asian Conference on Information Systems, ACIS 2015.

[5] Wenpeng Yin, Hinrich Schütze, Convolutional Neural Network for Paraphrase Identification, ,Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL, pages 901–911, Denver, Colorado, May 31 – June 5, 2015,Association for Computational Linguistics.

[6] Wenpeng Yin & Hinrich Schutze, Discriminative Phrase Embedding for Paraphrase Identification, Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL, pages 1368–1373, Denver, Colorado, May 31 – June 5, 2015 Association for Computational Linguistics.

[7] Majid Mohebbi and Alireza Talebpour, Texts Semantic Similarity Detection Based Graph Approach, The International Arab Journal of Information Technology VOL. 13, NO. 2, March 2016.

[8] Zia Ul-Qayyum and wasif Altaf, Paraphrase Identification using Semantic Heuristic Features, , Research Journal of Applied Sciences, Engineering and Technology 4(22): 4894-4904, 2012 ISSN: 2040-7467 © Maxwell Scientific Organization, 2012.

[9] Nitin Madnani, Joel Tetreault, Martin Chodorow, Re-examining Machine Translation Metrics for Paraphrase Identification, Conference of the North American Chapter of the ACL: 2012 Association for Computational Linguistics.

[10] Socher, Eric Huang, Pennington, Andrew, Christopher(2011), Dynamic pooling and unfolding recursive Autoencoder for Paraphrase Detection., "Advances in Neural Information Processing Systems 24".

[11] Anupriya Rajkumar, Dr. A. Chitra, Paraphrase recognition using neural network classification, International Journal of Computer Applications 1(29) · February 2010.

[12] Mihai Lintean and Vaile Rus, Paraphrase Identification Using Weighted Dependencies and word semantics, Proceedings of the Twenty-Second International FLAIRS Conference (2009).

[13] Dipanjan Das and Noah A. Smith, Paraphrase Identification as Probabilistic Quasi-Synchronous Recognition, Proceedings of **ACL-IJCNLP 2009.**

[14] Fernando and Stevenson, 2008.A Semantic Similarity Approach to paraphrase Detection, In Proceedings of the 11th Annual Research Colloquium of the UK Special Interest Group for Computational Linguistics, pages 45–52. Citeseer.

[15] Rus, V. and McCarthy, P.M. and Lintean, M.C. and McNamara, D.S. and Graesser, A.C. (2008). Paraphrase identification with lexico-syntactic graph subsumption, FLAIRS 2008, pp. 201-206.

[16] Cordeiro, Dias, Brazdil A Metric for Paraphrase Detection Proceedings of the International Multi-Conference on Computing in the Global Information Technology (ICCGI'07) ,IEEE

[17] Kozareva and Montoyo, Paraphrase identification on the basis of supervised machine learning techniques, Advances in Natural Language Processing: 5th International Conference on NLP (FinTAL 2006), Turku, Finland, 524-533.

[18] Nandini Sethi, Prateek Agrawal, Vishu Madaan 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.

[19] Survey of paraphrase Extraction Techniques for Kannnada, Ashwini Gadaag,Dr. B.M. Sagar,Mr. Rajshekar Murthy International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 6, June 2014.

[20] Paraphrase Identification using Malayalam sentences,Ditty Mathew, .rD Sumam Mary Idicula, IEEE,repap Advanced Computing (ICoAC), 2013 Fifth International Conference on 18-20 Dec. 2013.