IJDI-ERET

INTERNATIONAL JOURNAL OF DARSHAN INSTITUTE ON ENGINEERING RESEARCH & EMERGING TECHNOLOGIES
Vol. 7, No. 1, 2018
www.ijdieret.in

(Research Article)

Sentence Level Paraphrase Identification System for Tamil Language

Dr. S. V. Kogilavani[1*], Dr. C. S. Kanimozhiselvi[2], Dr. S. Malliga[3]

[1*, 2, 3]*Department of Computer Science and Engineering, Kongu Engineering College, Erode, Tamil Nadu, INDIA*

Abstract

Automatic detection of the paraphrase is a process which has immense applications like plagiarism detection and new event detection. Paraphrase is the representation of a given fact in more than one way by means of different phrases. Identification of a paraphrase is a classical natural language processing task which is of classification type. The aim is to detect sentence level plagiarism through paraphrase identification of sentences in Tamil. The sentences in Tamil language are processed using Tamil shallow parser. Shallow parsing is used to analyze a sentence to identify Part of Speech of sentences such as nouns, verbs, adjectives etc. Sentences are also processed using word2vec tool to identify word order between sentences. From the output of the shallow parsing process and word2vec, the feature file is constructed where the text values are converted into numerical matrix. This feature file is given as input into machine learning algorithms which in turn classify the sentence pair into paraphrase or not-a-paraphrase. If the result is paraphrase means, that sentence will be considered as plagiarized sentence. The accuracy and performance of these methods are measured based on evaluation parameters like accuracy, precision, recall and f-measure. The analysis based on these performance measures shows that Random Forest method classifies the sentence pair into paraphrase or not-a-paraphrase with high accuracy compared to other methods.

Keywords: machine learning algorithm, paraphrase identification, Plagiarism detection, Shallow parser

1. Introduction

Paraphrase is the process of identifying whether the two different text have the same meaning or not. Paraphrase identification plays an important role in information retrieval, information extraction, natural language processing and machine translation. To identify paraphrases the similarity between the pair of the sentences are calculated. Sometimes the two sentences may have the same meaning but that can be expressed by different texts. 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 distances between words. The similarity between sentences is measured based on statistical information of sentences. The statistical similarity between two sentences is calculated based on word vector using Cosine similarity measures. The semantic similarity between two sentences is calculated based on word order.

Plagiarism is defined as considering another person's content as one's own work. Plagiarism is not in itself a crime, but can constitute copyright infringement. In academia and industry, it is a serious ethical offense. Plagiarism and copyright overlap to a considerable extent, but they are not equivalent concepts, and many types of plagiarism do not constitute copyright infringement. Paraphrase identification is the task of determining whether two or more sentences represent the same meaning or not. Plagiarism detection is the task which needs the paraphrase identification technique to detect the sentences which are paraphrases of others. If the two sentences are paraphrases of each other, ultimately those sentences are plagiarized sentences. In this way paraphrase identification leads to detect whether plagiarism is there or not in sentence level.

The similarity between two Tamil sentences is done through shallow parsing where the basic parts of sentences are identified. Word2Vec tool is used for finding cosine similarity measure and the word order is also calculated. Feature file is constructed using all these values and it is classified as paraphrase or not-a-paraphrase using various classifier algorithms.

2. Literature Review

Paolo Rosso [1] analyze the relationship between paraphrasing and plagiarism, paying special attention to which paraphrase phenomena underlie acts of plagiarism and

[*]*Corresponding Author: e-mail: kogilavani.sv@gmail.com*
ISSN 2320-7590

which of them are detected by plagiarism detection systems. Yuhua Li [2] discusses about how to calculate the similarity between short text messages. The method proposed by these authors taken into account semantic as well as word order information in any sentence. R.Thangarajan [3] utilize sixteen different features of parts of speech tagging to represent the similarity between sentences. Using machine learning algorithms such as Support Vector Machine and Maximum Entropy, given sentences have been classified into Paraphrase and Not-a-Paraphrase. S. G. Ajay [4] compares two word embedding models for identifying semantic similarity in Tamil language sentences. These models are used to predict the relationship between words in a corpus. Dhanalakshmi V [5] develops a model in which Part Of Speech tagging and are done using machine learning techniques and the linguistic knowledge is extracted from the annotated corpus. Dhanalakshmi V [6] develops the grammar teaching tools for analyzing and learning character, word and sentence of Tamil Language. They developed other tools like Character Analyzer for character level analysis, Morphological Analyzer and Generator and Verb Conjugator for the word level analysis and Parts of Speech Tagger, Chunker and Dependency parser for the sentence level analysis using machine learning based technology. Timothy P. Jurka [7] represents that Social scientists have long hand-labeled texts to create datasets which is useful for studying topics from congressional policymaking to media reporting. Many social scientists have begun to incorporate machine learning into their toolkits. RTextTools was utilized to make machine learning accessible by providing a start-to-finish product with minimal steps.Tomas Mikolov [8] proposed the skip-gram model is an efficient method for learning high-quality distributed vector representations that capture a large number of precise syntactic and semantic word relationships. S. Rajendran [9] proposed that WordNet which plays an important role both in the development of NLP applications such as a machine translation system and a question answering system as well as for lexical studies of a language. While WordNet have been already compiled for most of the European languages, these resources are under preparation for Indian languages.

3. Proposed System Design

In the proposed method, both statistical and the semantic similarity analysis is used to determine whether the given sentences are paraphrase or not. In this system, the statistical analysis is based on word set, word vector, word order, and word distance. Shallow parsing is done for identifying the basic parts of the sentences and the semantic analysis is based on the Word2Vec tool and the word order is also calculated. The following Figure 1 represents the proposed system design.

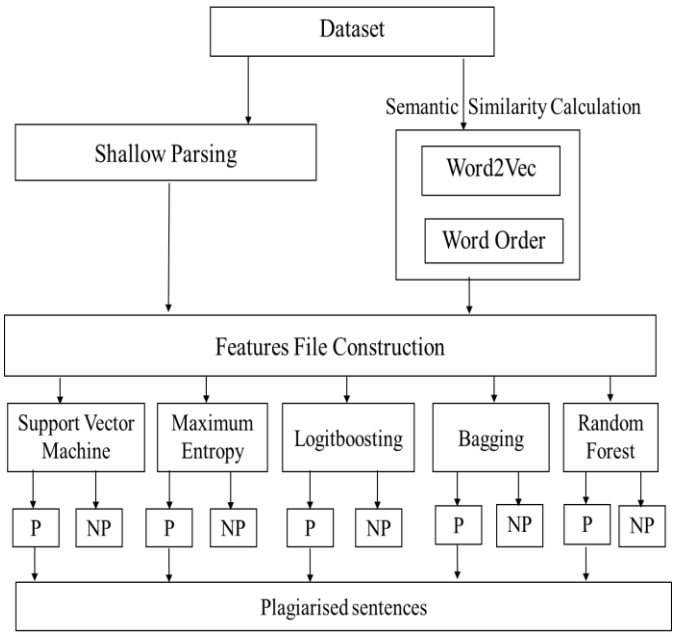

Figure 1. Proposed System Design

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 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 morphological features and Word2Vec tool.

Statistical as well as semantic features are passed to machine learning algorithms which contain the classifiers like Support Vector Machines, Maximum Entropy, Boosting, Random Forest and Bagging. In the proposed method, the sentences are given as input to shallow parserwhich is used to produce sixteen different features for the given sentence pairs. The semantic similarity is calculated using Word2Vec tool. Word2Vec tool is a two layer neural network word embedding algorithm which is used to find the cosine similarity. Word order between sentences are also calculated. Based on these Eighteen features, the feature file is constructed. The feature file is given as input to different machine learning algorithms like Support Vector Machine, Maximum Entropy, Logitboosting, Bagging, Random Forest to classify the given sentence pair. Confusion matrix is computed for different classifiers and the performance measures are analyzed. The sentences which are classified as paraphrases are said to be plagiarized sentences.The table 1 represents the feature file consisting of eighteen features which are derived from shallow parsing, word2vec and word order. It is given as input to various classifiers.

Table 1. Feature File

Para Id	ben	nom	RB	Present	Past	acc	future	loc	gen	dat	JJ	NNP	NN	soc	VM	count	cosine	Word order	Class
TAM 0001	0	10	2	0	2	0	2	3	0	0	2	2	11	0	6	6	0.935	0.104	P
TAM 0002	0	8	1	0	2	0	0	1	0	0	1	1	9	0	2	4	0.886	0.24	P
TAM 0003	0	12	1	2	3	0	0	0	0	5	0	0	19	0	5	6	0.946	0.138	P
TAM 0004	0	15	2	0	1	0	3	2	0	3	2	0	20	0	3	4	0.912	0.03	P
TAM 0005	0	11	0	1	0	0	0	0	0	0	1	3	16	0	1	7	0.962	0.096	P

4. Classification Using Machine Learning Algorithms

4.1 Support Vetor Machine Classification Method: Support Vector Machine (SVM) is based on the structural risk minimization principle from computational learning theory. This method analyzes data and defines decision boundaries by having hyper-planes. In binary classification problem, the hyper-plane separates the given vector in one class from other class, where the separation between hyper-planes is desired to be kept as large as possible. One property of SVM is that their ability to learn can be independent of the dimensionality of the feature space. SVM measure the complexity of hypotheses based on the margin with which they separate the data, not the number of features. Since SVM requires input in the form of a vector of numbers, the constructed feature file is given as input to SVM.

*4.2 Maximum Entropy Classification Method:*Maximum Entropy(ME) is a general technique for estimating probability distributions from data. The over-riding principle in maximum entropy is that when nothing is known, the distribution should be as uniform as possible, that is, have maximal entropy. Labeled training data is used to derive a set of constraints for the model that characterize the class-specific expectations for the distribution. Constraints are represented as expected values of “features” any real-valued function of an example. The improved iterative scaling algorithm finds the maximum entropy distribution that is consistent with the given constraints.

Due to the minimum assumptions that the maximum entropy classifier makes, we regularly use it when we don’t know anything about the prior distributions and when it is unsafe to make any such assumptions. Moreover maximum entropy classifier is used when we can’t assume the conditional independence of the features. This is particularly true in text classification problems where our features are usually words which obviously are not independent. This method requires more time to train comparing to Naive Bayes, primarily due to the optimization problem that needs to be solved in order to estimate the parameters of the model. Nevertheless, after computing these parameters, the method provides robust results and it is competitive in terms of CPU and memory consumption.In our text classification scenario, maximum entropy estimates the conditional distribution of the class label given pair of sentences. Entire document is represented by a feature file. The labeled training data is used to estimate the expected value on a class-by-class basis.

4.3 Logitboosting Classification Method: Logiboosting(LB) is a machine learning ensemble meta-algorithm for primarily reducing bias, and also variance in supervised learning, and a family of machine learning algorithms which convert weak learners to strong ones. A weak learner is defined to be a classifier which is only slightly correlated with the true classification,it can label examples better than random guessing. In contrast, a strong learner is a classifier that is arbitrarily well-correlated with the true classification. While boosting is not algorithmically constrained, most boosting algorithms consist of iteratively learning weak classifiers with respect to a distribution and adding them to a final strong classifier. When they are added, they are typically weighted in some way that is usually related to the weak learners' accuracy. After a weak learner is added, the data are reweighted: examples that are misclassified gain weight and examples that are classified correctly lose weight.

Only algorithms that are provable boosting algorithms in the probably approximately correct learning formulation can accurately be called boosting algorithms. Other algorithms that are similar in spirit to boosting algorithms are sometimes called "leveraging algorithms", although they are also sometimes incorrectly called boosting algorithms. The main variation between many boosting algorithms is their method of weighting training data points and hypotheses.

4.4 Bagging Classification Method: Bagging means Bootstrap Aggregation. Bootstrapping is a process of selecting samples from original sample, or population, and using these samples for estimating various statistics or model accuracy. Bagging, Bootstrap aggregating, was proposed for improving classification accuracy. It is a process of creating

random samples with replacement for estimating sample statistics. One of the way to select samples or bootstrap samples is to select n items with replacement from an original sample, N.

A bootstrap sample may have a few duplicate observations or records, as the sampling is done with replacement.Bagging is a process, where a model is trained on each of the bootstrap samples and the final model is an aggregated models of the all sample models. For a numeric target variable regression problems, the predicted outcome is an average of all the models and in the classification problems, the predicted class is defined based on plurality.

4.5 Random Forest Classification Method: Random Forest (RF) is an ensemble learning based classification and regression technique. It is one of the commonly used predictive modeling and machine learning technique.Breiman in 2001 proposed random forests, which add an additional layer of randomness to bagging. In addition to constructing each tree using a different bootstrap sample of the data, random forests change how the classification or regression trees are constructed. In standard trees, each node is split using the best split among all variables.

In a random forest, each node is split using the best among a subset of predictors randomly chosen at that node. This somewhat counterintuitive strategy turns out to perform very well compared to many other classifiers, including discriminant analysis, support vector machines and neural networks, and is robust against over fitting. In addition, it is very user-friendly in the sense that it has only two parameters, the number of variables in the random subset at each node and the number of trees in the forest, and is usually not very sensitive to their values. The entire forest is contained in the forest component of the RandomForest object. It can take up quite a bit of memory for a large data set or large number of trees. If prediction of test data is not needed, set the argument forest as false when running RandomForest. This way, only one tree is kept in memory at any time, and thus lots of memory, potentially execution time, can be saved.

5. Performance Evaluation

Corpora for paraphrase detection in Tamil language is taken from the Eighth International Forum for Information Retrieval and Evaluation which consists of 2500 sentence pairs as training dataset and 900 sentence pairs as test dataset. The following evaluation measures are used in the proposed system.

The table 2 represents the precision, recall and F-Measures values obtained using different classifier techniques. The result shows that Random Forest classifier's precision, recall and F-Measure values are high compared to other classifiers in detecting paraphrases.

Table 2. Precision, Recall and F-Measure Values

Class Method	Class	Precision	Recall	F-Measure
SVM	Paraphrase	0.76	0.78	0.77
	Not Paraphrase	0.68	0.66	0.67
ME	Paraphrase	0.79	0.84	0.81
	Not Paraphrase	0.75	0.68	0.71
LB	Paraphrase	0.75	0.90	0.82
	Not Paraphrase	0.81	0.59	0.68
BAGGING	Paraphrase	0.80	0.79	0.79
	Not Paraphrase	0.71	0.72	0.71
RF	Paraphrase	0.81	0.82	0.81
	Not Paraphrase	0.74	0.72	0.73

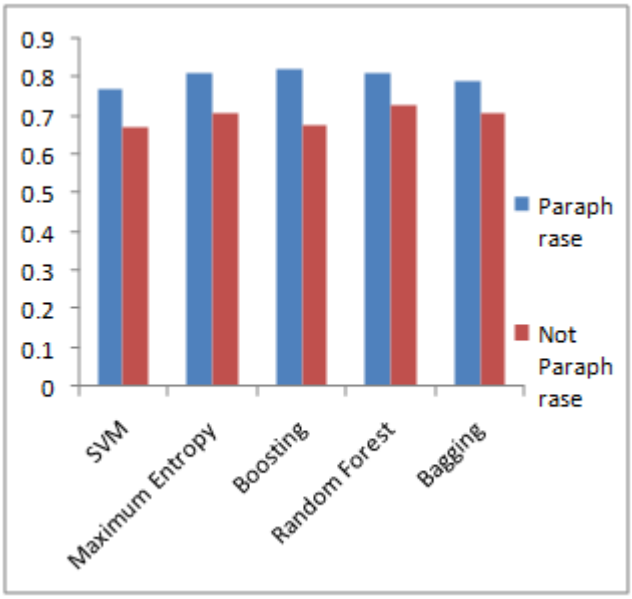

Figure 2. Comparison of Confusion Matrix

The figure 2 represents confusion matrix generated by the various classifiers. The results show that the Random Forest classifier gives more True Positive and True Negative values compared to other classifiers. The following Figure 3 shows that Random Forest algorithm gives more accuracy as compared to other algorithms. Previously, Maximum Entropy gives the maximum accuracy of 0.741 and got the second position in the International Forum for Information Retrieval and Evaluation Conference, which is slightly lesser than first place accuracy of 0.776. Now the accuracy is increased to 0.7789 by additional processing of sentences through semantic analysis such as word ordering and cosine similarity.

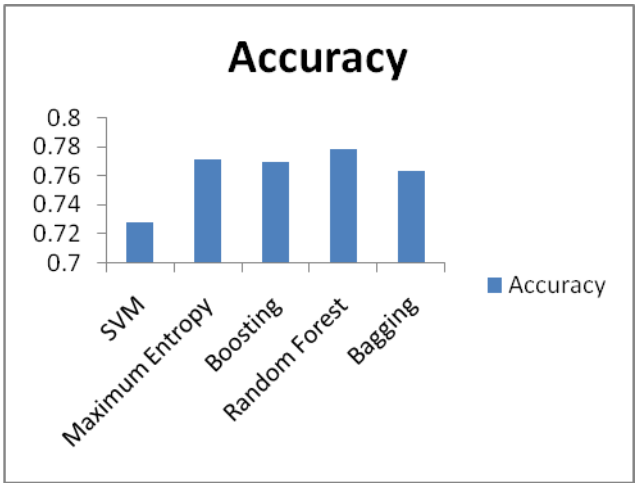

Figure 3. Comparison of Accuracy

6. Conclusion and Future Work

Paraphrase identification is important for text classification and retrieval. This project work represents methods for measuring the similarity between sentences based on semantic word level information. After that the sentences are classified using different supervised machine learning algorithms. The eighteen different semantic features are used to represent the similarity between sentences. 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. The result shows that Random Forest method outperforms than other algorithms to identify paraphrases. The present work identifies plagiarism in sentence level.

The future work is to use Doc2Vec instead of Word2Vec where whole document can be analyzed using the proposed features in order to detect plagiarism through paraphrase identification.

References

1. Alberto Barron-Cedeno, Marta Vila, M. AntoniaMarti and Paolo Rosso, 'Plagiarism Meets Paraphrasing: Insights for the Next Generation in Automatic Plagiarism Detection' - *Association for Computational Linguistics* Vol. 39, No. 4, pp.917-947, 2013.
2. Yuhua Li, David McLean, Zuhair A. Bandar, James D. OS-hea, and Keeley Crockett, 'Sentence Similarity Based on Semantic Nets and Corpus Statistics' - *IEEE Transactions on knowledge and data engineering* Vol. 18, No. 8, 2006.
3. R.Thangarajan, S.V.Kogilavani, A.Karthic and S.Jawahar, 'Detection of Paraphrases on Indian Languages' – *CEUR Workshop Proceedings* Vol.1737, pp.282-288, 2016.
4. S. G. Ajay, M. Srikanth, M. Anand Kumar and K. P. Soman, 'Word Embedding Models for Finding SemanticRelationship between Words in Tamil Language' - *Indian Journal of Science and Technology*, Vol 9(45), 2016.
5. Dhanalakshmi V, Anandkumar M, Rajendran S, Soman K P, 'POS Tagger and Chunker for Tamil Language' - *International Forum for Information Technology in Tamil*, 2009.
6. Dhanalakshmi V, Anand Kumar M, Soman K.P and Rajendran S, 'Natural Language Processing Tools for Tamil GrammarLearning and Teaching' - *International Journal of Computer Applications* Volume 8– No.14, 2010.
7. Timothy P. Jurka, Loren Collingwood, Amber E. Boydstun, Emiliano Grossman, and Wouter van Atteveldt, 'RTextTools: A Supervised LearningPackage for Text Classification' - *The R Journal* Vol. 5/1, 2013.
8. Tomas Mikolov, IlyaSutskever, Kai Chen, Greg Corrado, Jeffrey Dean, 'Distributed Representations of Words and Phrases and their Compositionality' - *Neural Information Processing Systems Proceedings*, pp.3111-3119, 2013.
9. Rajendran, S., S. Arulmozi, B. Kumara Shanmugam, S. Baskaran, and S. Thiagarajan (2002) 'Tamil WordNet' - *Proceedings of the First International Global WordNet Conference. CIIL, Mysore*, pp.271-274

Biographical notes

Dr.S.V. Kogilavani is born in Coimbatore, TN, in the year 1978. She completed her B.E (Computer Science and Engineering) in the year 1999 from Madras University, Chennai and obtained M.E (Computer Science and Engineering) degree in the year 2007 from Anna University, Chennai. She got her Ph.D from Anna University, Chennai in the year 2013. Her research area is information retrieval, summarization and opinion mining. She is associated with the Department of Computer Science and Engineering as Senior Assistant Professor at Kongu Engineering College, Tamil Nadu, India .She has presented 25 papers in national and international conferences and published 20 papers in national and international journals. She conducted many courses for the benefits of students. She conducted various workshops and seminars.