International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 04 Issue: 05 | May -2017 www.irjet.net p-ISSN: 2395-0072

Detecting Paraphrases in Tamil Language Sentences

Dr.S.V.Kogilavani[1], Dr.R.Thangarajan[2], Dr.C.S.Kanimozhiselvi[3] Dr.S.Malliga[4]

[1]*Assistant Professor (SRG), Department of CSE, Kongu Engineering College, Tamil Nadu, India*

[2,4]*Professor, Department of CSE, Kongu Engineering College, Tamil Nadu, India*

[3]*Associate Professor, Department of CSE, Kongu Engineering College, Tamil Nadu, India*

Abstract - *In text mining, sentence similarity is used as a criterion to discover unseen knowledge from textual database. 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. This paper 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. The proposed approach utilizes machine learning algorithms like Support Vector Machine and Maximum Entropy for classification of given sentence pair using statistical and semantic features. The accuracy and performance of these 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.*

Key Words: *Paraphrase Identification, Machine Learning Approach, Support Vector Machine, Maximum Entropy, Statistical Analysis, Semantic analysis.*

1. INTRODUCTION

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, machine translation [25].It can be identified by calculating the similarity between the pair of the sentences. Paraphrase detection system improves the performance of a paraphrase generation by choosing the best sentences from the list of paraphrase sentences. This paper is mainly focuses on identifying whether the given sentences are paraphrase or not in Tamil language. To illustrate the concept of paraphrase consider the following sentence pair,

S_1:நான்என்றசொல்லைஒழித்துநாம்என்றுஉருவாக்குவோம்
S_2:நான்என்றசொல்லைஒழிப்பதன்மூலம்நாம்என்றஒற்றுமை யைஉருவாக்கிடுவோம்

These two sentences 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[1]. Structural relations include relations between words and the distances between words. The similarity between sentences is measured based on statistical information of sentences[4]. The statistical similarity between two sentences are calculated based on word distance using Euclidean measures, word set using Jacquard and Dice measures, word vector using Cosine similarity measures. The semantic similarity between two sentences is calculated based on word order. The paper is organized as follows. Section 2 describes about the related work. Section 3 represents the proposed approach and methodology. Section 4 presents experimental results and evaluation. Section 5 concludes the work.

2. RELATED WORK

Sentences with different structures may convey the same meaning[5]. Paraphrase identification mainly focuses on the statistical measures and semantic analysis of Tamil sentences to detect the paraphrases. The semantic representation of Universal Networking Language(UNL), represents only the inherent meaning in the sentence without any syntactic details. Combination of statistical similarity and semantic similarity score results the overall similarity score [4].

The low-level features for paraphrase identification deals with the task of sentential paraphrase identification[2]. 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 basis for the development of an effective paraphrase identification system. This experiment show the improvement of the paraphrase identification model based on the standard low-level features.

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[3]. 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.

Paraphrase Acquisition 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 [7].

Measuring sentence similarity from different aspects proposes to determine sentence similarities from different aspects. It shows that the proposed method makes the sentence similarity comparison more exactly and gives out a more reasonable result[19].

Measuring semantic similarity is the objective of many works. Many measures perform well in evaluation framework for a specific task like synonymy extraction [18].

3. RELATED WORK

In the proposed method, the statistical and the semantic analysis is used to determine the paraphrases. 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.

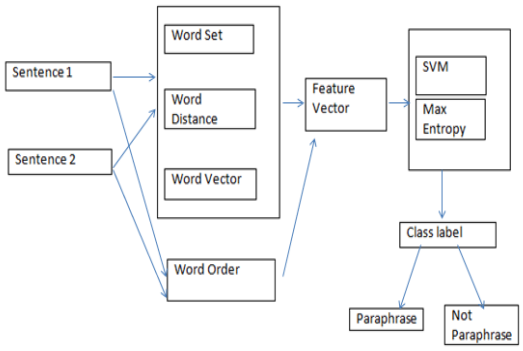

Fig-1 Proposed System Design

3.1. Jaccard Similarity Measure

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 $w(S_a)$ be the set of words in first sentences S_a and $w(S_b)$ be the set of words in second sentence S_b. After forming the word set, Jaccard similarity is computed using the equation 1.

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

3.2 Dice Similarity

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

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

3.3 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 $(p_1,p_2,.....,p_n)$ and sentence q has words $(q_1,q_2,.....,q_n)$ then word distance of p and q are represented as follows.

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

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

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

$$d(\mathbf{p},\mathbf{q}) = d(\mathbf{q},\mathbf{p}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2}$$

$$= \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}. \quad (5)$$

3.4 Cosine Similarity

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

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

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

Then the cosine similarity between sentences can be calculated based on the word vectors by using equation 8.

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

3.5 Word Order Similarity

Sentence similarity based on the word order requires constructing the order vectors of the two sentences.

If the sentence S_a has words (w_{a1},w_{a2},.....,w_{ai}) and sentence S_b has words (w_{b1},w_{b2},.....,w_{bi}) then word order vectors for S_a and S_b 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 $(w_x,w_y) \in L(S_a) \cup L(S_b)$ means w_x is before w_y. Then the similarity between S_a and S_b can be calculated based on the orders of words by equation 11.

$$WordOrder(S_a,S_b) = \frac{\left|L(S_a) \cap L(S_b)\right|}{\left|L(S_a)\right| \cup \left|L(S_b)\right|} \quad (11)$$

3.6 Sample Sentences

T1.Test_Tam0001
S1=
சங்கராபுரம்தொகுதியில்போட்டியிடும்ஸ்டாலின்நடைபயணமாக சென்றுபிரசாரம்செய்தார்.
S2= தி.மு.க.,
வேட்பாளர்ஸ்டாலின்போட்டியிடும்சங்கராபுரம்தொகுதியில்சின்ன சேலம்பகுதியில்நடைபயணமாகசென்றுஓட்டுசேகரித்தார்.
T1.Test_Tam0002
S1=கேரளமாநிலம்திருச்சூரில்கூடல்மாணிக்கம்கோயில்திருவிழாது வங்கியது.
S2=கூடல்மாணிக்கம்கோயில்திருவிழாகோலாகலமாகதுவங்கியது.
T1.Test_Tam0003
S1=விஜய்மல்லையாவின்பீர்நிறுவனத்துக்குகுறைந்தவிலைக்குநில ம்விற்பனைக்குசர்ச்சைஏற்பட்டுள்ளது.
S2=விஜய்மல்லையாவின்தொழில்குழுமத்துக்குகேரளஅரசுகுறைந்த விலைக்குநிலம்விற்பனைசெய்ததாகசர்ச்சைகிளம்பியுள்ளது.
T1.Test_Tam0004
S1=சர்க்கரைநோயாளிகளுக்குஉதவும்நோக்கில்ரத்தம்மற்றும்சிறுநீ ர்பரிசோதனைக்குகுறைந்தவிலைகருவிகண்டுபிடிப்பு.
S2=சர்க்கரைநோயாளிகள்ரத்தம்மற்றும்சிறுநீர்பரிசோதனைசெய் யபயன்படுத்தப்படும்குளுக்கோமீட்டர்கருவிகள்ரூ.5-
க்கும்குறைவானசெலவில்தயாரிப்பதற்கானதொழில்நுட்பம்.

The above sentences are represented the syntactic analysis are word set, word distance and word vector and semantic analysis are word order are calculated values in below table 1.

Table – 1: Syntatic and Semantic Measures

Sentences Id	Syntactic Measures			Semantic Measures	Total value
	Word Set Value	Word Distance Value	Word Vector Value	Word Order Value	
T1.Test_Tam0001	0.4857	1	0.5714	0.1667	0.5559
T1.Test_Tam0002	0.4663	1	0.5601	0.1609	0.5468
T1.Test_Tam0003	0.6477	1	0.5477	0.1562	0.5879
T1.Test_Tam0004	0.2102	0	0.2767	0.0273	0.1285

4. PERFOMANCE 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. The following evaluation measures are used in the proposed system.

4.1 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.

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

4.2 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[8]. It can be calculated using equation 13.

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

4.3 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.

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

4.4 F-Measure

F-Measure is a measure of test the accuracy and also define as the harmonic mean of precision and recall of the test[15]. 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.

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

4.5 Sensitivity

Sensitivity (also called the true positive rate, the recall, or probability of detection in some fields) measures the proportion of positives that are correctly identified as such.

$$TPR=TP/P=TP/(TP+FN) \quad (16)$$

4.6 Specificity

Specificity (also called the true negative rate) measures the proportion of negatives that are correctly identified as such.

$$SPC=TN/N=TN/(TN+FP) \quad (17)$$

The following table 2 represents the precision, recall and F-Measures values obtained using SVM and Max Entropy classifiers. The result shows that SVM classifier's precision, recall and F-Measure values are high compared to Max Entropy classifier.

Table -2: Precision, Recall and F-Measure Values

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

The table 3 represents confusion matrix generated by SVM classifier for the given dataset. Out of 900 sentences, 427 paraphrase sentences are identified as Paraphrase sentences. 104 paraphrase sentences are identified as not a paraphrase sentences. 99 not a paraphrase sentences are identified wrongly as paraphrase sentences. 270 not a paraphrase sentences are identified as not a paraphrase sentences.

Table-3: Confusion Matrix

	Paraphrase	Not Paraphrase
Paraphrase	427	104
Not Paraphrase	99	270

The table 4 represents the SVM classification Performance such as Sensitivity, Specification, Accuracy, Positive and Negative Predictive value.

Table-4: SVM Classification Performance

SVM Sensitivity	0.8118
SVM Specificity	0.7219
Accuracy	0.7744
Positive Predictive Value	0.8041
Negative Predictive Value	0.7317

5. CONCLUSION

Paraphrase identification is important for text classification and retrieval. This paper represents methods for measuring the similarity between sentences based on syntactic and semantic word level information. After that the sentences are classified using two supervised machine learning algorithms, such as SVM and Max Entropy. In this paper we utilize sixteen different syntactic and semantic features to best represent the similarity between sentences. Two machine learning algorithms such as Support Vector Machine and Maximum Entropy 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 results show that SVM method outperforms than Max Entropy to identify paraphrases.

REFERENCES

[1]https://www.researchgate.net/publication/277013951_Paraphrase_identification_of_malayalam_sentences_-_an_experience
[2]https://www.researchgate.net/profile/Ekaterina_Pronoza/publication/300138252_Low-Level_Features_for_Paraphrase_Identification/links/581e55fd08ae12715af5da63.pdf
[3] http://www.kozareva.com/papers/fintalKozareva.pdf
[4]http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=1644735
[5] http://ceur-ws.org/Vol-1737/T6-11.pdf
[6] Zhang, Sun, Wang and He, "Calculating Statistical Similarity between Sentences," *Journal of Convergence of Information Technology*, February 2011.
[7]http://webcache.googleusercontent.com/search?q=cache:http://www.uni-weimar.de/medien/webis/publications/papers/stein_2013c.pdf
[8]https://cocoxu.github.io/publications/tacl2014-extracting-paraphrases-from-twitter.pdf
[9]https://fenix.tecnico.ulisboa.pt/downloadFile/395145918749/resumo.pdf
[10]http://ijcsit.com/docs/Volume%207/vol7issue4/ijcsit20160704100.pdf
[11]http://www.ijcaonline.org/volume17/number2/pxc3872778.pdf
[12]https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/I05-50015B15D.pdf
[13]http://www.wseas.us/e-library/conferences/2010/Cambridge/ICNVS/ICNVS-41.pdf
[14]http://alt.qcri.org/semeval2015/cdrom/pdf/SemEval004.pdf

[15]https://hal.inria.fr/hal-01426749/document
[16]https://www.ijcai.org/Proceedings/16/Papers/406.pdf
[17]http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.421.489&rep=rep1&type=pdf
[18]http://www.cc.gatech.edu/~jeisenst/papers/ji-emnlp-2013.pdf
[19]http://www.aclweb.org/anthology/S15-2011
[20]https://ijcai.org/Proceedings/13/Papers/441.pdf
[21]http://www.aaai.org/Papers/JAIR/Vol38/JAIR-3804.pdf