CUSAT TEAM@DPIL-FIRE2016: Detecting Paraphrase in Indian Languages-Malayalam

Manju K
Research Scholar,
Department of Computer science and Engineering,
Cochin University of Science and Technology,India.
manju@mec.ac.in

Sumam Mary Idicula
Head of Department,
Department of Computer science and Engineering,
Cochin University of Science and Technology,India.
sumam@cusat.ac.in

ABSTRACT

This paper describes the work done as part of the shared task on Detecting Paraphrases in Indian Languages(DPIL) in Forum for Information Retrieval and Evaluation(FIRE 2016). Paraphrase identification is the task of deciding whether two given text fragments have the same meaning. Our detection system is for Malayalam language and makes use of the cosine similarity measure, an existing state of the art method for determining the similarity between sentences. The experiments were done on the standard data set and the results showed that the system was able to give performance comparable to methods employing more sophisticated procedures.

CCS Concepts

•**Information Processing** → Similarity Measures; •**Natural Language Processing** → Paraphrase Identification; •**Text Mining** → Text Summarization;

Keywords

Paraphrase; Cosine similarity; text tagging

1. INTRODUCTION

Paraphrases are alternate ways to convey the same information. In natural languages, we can express a single event in different ways which conveys the same information. Paraphrase identification, the ability to determine whether two formally distinct strings are similar or not, have application in various NLP tasks like Information retrieval, Question Answering, Plagiarism detection, Text Mining and Automatic summarization. Paraphrase identification basically uses a simple lexical matching comparison of sentences.

In order to select a sentence pair as paraphrase, they should describe the same event and should contain same information about the event. However there are instances when the concept behind the sentences are difficult to identify, even for humans this is a difficult task.

The rest of the paper is organized as follows: Section 2 discusses related work in the area of Paraphrase detection. Section 3 presents the Task Description. Section 4 tells about the data set provided by the DPIL task[2] organizers. Section 5 explains the methodology used and Section 6 gives the Result and evaluation. Section 7 presents the conclusion and the future improvements that can be made.

2. RELATED WORKS

Paraphrase identification has a lot of significance in different areas of Natural language Processing. Paraphrase identification techniques are mainly classified into statistical and semantic methods. In statistical methods, the similarity between sentences is measured only on the basis of statistical information in the sentences whereas semantic method makes use of word meanings. Work which shows the comparison of statistical and semantic similarity measures[1], which was tested on the same data set stated that the performance of both measures are comparable. One of the most commonly used corpora for paraphrase detection is the MSRP corpus[3], which contains 5,801 English sentence pairs from news articles manually labelled with 67% paraphrases and 33% non-paraphrases. Since there are no annotated corpora or automated semantic interpretation systems available for Indian languages till date, the initiative made as part of the open shared task competition is highly appreciable and is of great help to the research community. The automatic plagiarism detection framework for Malayalam documents[5] uses Jaccard similarity for determining the relation between sentences.

The proposed method implements Paraphrase Identification for Malayalam Language using similarity measures[4].

3. TASK DESCRIPTION

The task is focused on sentence level paraphrase identification for Indian languages-Tamil, Malayalam, Hindi and Punjabi. The proposed method considers only Malayalam language. Malayalam is one among the 22 scheduled languages of India. It is the official language in the state of Kerala and in the Union territories of Lakshadweep and Puduchery. Malayalam belongs to the Dravidian language family and is spoken by approximately 33 million people.The task provided is divided into two sub tasks where sub task 1 is to classify the given pair of sentences to paraphrase or non paraphrase and in sub task 2 the sentences are classified on a 3 point scale, to completely equivalent(P), roughly equivalent(SP) or not equivalent(NP).

4. DATA SET

The shared task challenge provided data for four languages Tamil, Malayalam, Hindi and Punjabi. We were provided with 2500 sentence pairs for sub task 1 and 3500 sentence pairs for sub task 2 as training data and 900 sentence pairs for sub task 1 and 1500 sentence pairs for sub task 2 as test data. The data set available was in XML format taken from prominent Newspapers.

5. SYSTEM DESCRIPTION

Data was given in XML format and that file was processed to extract each pair of sentences for paraphrase detection. Cosine similarity measure was used for paraphrase identification and the concerned two sentences in each pair was considered as two documents D_1 and D_2. D_1 and D_2 contain only one sentence each. The overall architecture of the system is shown in Fig 1. D_1 and D_2 are subjected to tokenization and stop word removal. A look up table was used for stop word removal. Due to the agglutinative nature of the language, the same word can appear with different inflections in the sentences. To eliminate these inflections, stemming was performed. Even though literature related to stemming in Malayalam language is available, there is no full fledged tool which can be used in the work. We have custom tailored the Silpa Stemmer[6] by Swathanthra Malayalam Computing group for our purpose. The stemmer removes longest matching suffix from each word with proper replacement to get the base word.

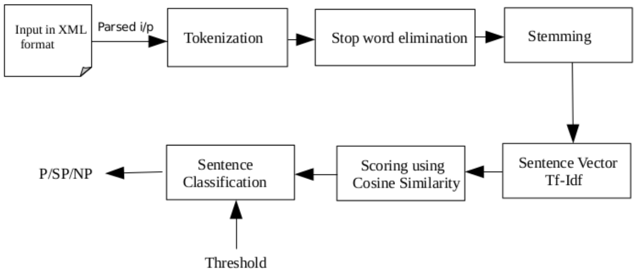


Figure 1: System Architecture

The words in the resulting sentences after preprocessing are the bag of words(vocabulary) for the vector representation of the sentences. The sentence vector is formulated using bag-of-words model to extract frequency information of words in the sentence. The size of the vector will be the size of the vocabulary set and the value at each vector index i represents the count of word i in the sentence. This is the Term Frequency(TF) Vector.For determining the importance of each word with respect to the two documents its Inverse Document Frequency (IDF) is also calculated according to equation(1).

$$Idf_t = log\frac{N}{N_t} \tag{1}$$

where N is the total sentences in a document D, here it is 2 and N_t is the number of sentences in which the term t occurs. The sentence vector is computed according to equation(2).

$$S_i = Tf_{t,i} * Idf_t \tag{2}$$

where $Tf_{t,i}$ is the frequency of term t in Sentence S_i and Idf_t gives the information, how important is the term t. Using equation(3) the similarity between documents are computed where D_1 contains the first sentence and D_2 contains the second sentence in the pair.

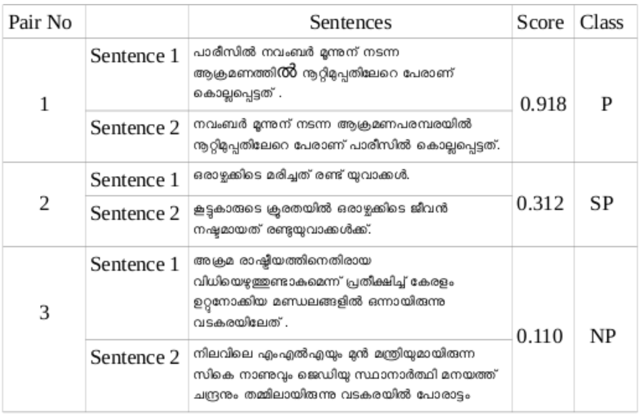
$$Sim(D_1, D_2) = \frac{D_1 * D_2}{\sqrt{D_1^2} * \sqrt{D_2^2}} \tag{3}$$

Similarity score will be a value between 0 and 1.

It was decided to set a threshold for determining the classes Paraphrase, Semi Paraphrase and Non Paraphrase. Through experiment using the training data given for task1 and task2 a threshold of 0.4 was set for Paraphrase, 0.3 for SemiParaphrase and any value less than that as NonParaphrase.

6. RESULTS AND EVALUATION

The proposed system was experimented with the data set provided by the open shared task. Fig 2 shows the similarity score obtained for the 3 classes of sentence pairs.



Pair No		Sentences	Score	Class
1	Sentence 1	പാരീസിൽ നവംബർ മൂന്നന് നടന്ന ആക്രമണത്തിൽ നൂറ്റിമുപ്പതിലേറെ പേരാണ് കൊല്ലപ്പെട്ടത് .	0.918	P
	Sentence 2	നവംബർ മൂന്നന് നടന്ന ആക്രമണപരമ്പരയിൽ നൂറ്റിമുപ്പതിലേറെ പേരാണ് പാരീസിൽ കൊല്ലപ്പെട്ടത്.		
2	Sentence 1	ഒരാഴ്ചക്കിടെ മരിച്ചത് രണ്ട് യുവാക്കൾ.	0.312	SP
	Sentence 2	കൂട്ടുകാരുടെ ക്രൂരതയിൽ ഒരാഴ്ചക്കിടെ ജീവൻ നഷ്ടമായത് രണ്ടുയുവാക്കൾക്ക്.		
3	Sentence 1	അക്രമ രാഷ്ട്രീയത്തിനെതിരായ വിധിയെഴുത്തുണ്ടാകുമെന്ന് പ്രതീക്ഷിച്ച് കേരളം ഉറ്റുനോക്കിയ മണ്ഡലങ്ങളിൽ ഒന്നായിരുന്നു വടകരയിലേത് .	0.110	NP
	Sentence 2	നിലവിലെ എംഎൽഎയും മുൻ മന്ത്രിയുമായിരുന്ന സികെ നാണുവും ജെഡിയു സ്ഥാനാർത്ഥി മനയത്ത് ചന്ദ്രനും തമ്മിലായിരുന്നു വടകരയിൽ പോരാട്ടം		

Figure 2: Similarity Score Obtained

The accuracy and F-score for this methodology of paraphrase identification is tabulated in Table 1 for subtask 1 and subtask 2

Table 1: Results

Language	SubTask1		SubTask2	
	Accuracy	F1 Score	Accuracy	F1 Score
Malayalam	0.80444	0.76	0.50857	0.46576

7. CONCLUSION

This paper discussed on how cosine similarity can be used for Paraphrase identification. The morphological richness and agglutinative nature of the language demands for stemming of the sentence pairs before paraphrase scoring. The accuracy of the preprocessing phase has got a significant role in the paraphrase identification system. Performance of the system can be improved by considering semantic similarity using word net in addition to statistical measures. An ensemble of different similarity scores may improve the accuracy of the system. The vague demarcation between semi paraphrase and non paraphrase is a challenge in this type of work.

8. REFERENCES

[1] S. S. Abraham and S. M. Idicula. Comparison of statistical and semantic similarity techniques for

paraphrase identification. In *2012 International Conference on Data Science & Engineering (ICDSE)*.

[2] M. Anand Kumar, S. Shivkaran, B. Kavirajan, and K. P. Soman. DPIL@FIRE2016: Overview of shared task on detecting paraphrases in indian languages. In *Working notes of FIRE 2016 - Forum for Information Retrieval Evaluation, Kolkata, India, December 7-10, 2016*, CEUR Workshop Proceedings. CEUR-WS.org, 2016.

[3] W. B. Dolan and C. Brockett. Automatically constructing a corpus of sentential paraphrases. In *Proc. of IWP*, 2005.

[4] S. Fernando and M. Stevenson. 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, 2008.

[5] L. Sindhu, B. B. Thomas, and S. M. Idicula. Automated plagiarism detection system for malayalam text documents. *International Journal of Computer Applications*, 106(15), 2014.

[6] S. Thottungal.
Silpastemmer: http://libindic.org/stemmer.