CUSAT_NLP@DPIL-FIRE2016: Malayalam Paraphrase Detection

Sindhu.L
Department of computer Science
College of Engineering
Poonjar
sindhul.cep@gmail.com

Sumam Mary Idicula
Department of computer Science
CUSAT
Kochi
sumam@cusat.ac.in

ABSTRACT
This paper describes an approach for paraphrase detection in Malayalam sentences developed as part of FIRE 2016 Shared Task on Paraphrase detection in Indian Languages. The task of paraphrasedetection is finding a sentence with the same meaning of another sentence expressed using same or different words. This detection is done by a semantic approach which is language dependent. Individual words, their root forms and synonyms are used in finding similarity between two given sentences. We present an algorithm for paraphrase identification which makes use of word similarity information derived fromCUSAT Malayalam WordNet Padasrinkala.. The approach is evaluated using the Malayalam corpus made available as part of of FIRE 2016 Shared Task on Paraphrase detection in Malayalam.

CCS Concepts
• **Computing methodologies~Natural language processing** • **Computing methodologies~Lexical semantics** • **Computing methodologies~Language resources** • *Computing methodologies~Information extraction*

Keywords
Paraphrase detection; semantic matching;tokenization;POS tagging;lemmatization;corpus.

1. INTRODUCTION
Paraphrase is defined as the reuse of text or its meaning in another sentence using the same or similar words or phrases. Paraphrase detection is used to determine whethertwo texts (sentences) of different lengths have the samemeaning. Such detection is used in various natural language applicationssuch as plagiarism detection, text summarisation, WSD, machine translation etc.Paraphrasing may be due to morphology based changes, lexicon-based changes, syntax-based changes, discourse-based changes, semantics-based changes etc. This approach to paraphrase detection comprises of pure lexical matching and also the similarity between sentences which use synonyms to convey the same meaning.
The outline for the rest of the paper is as follows. Section 2 describes some of the previous approaches to paraphrase identification and their limitations. The approach proposed here is described in Section 3. Section 4 gives a brief description of the Paraphrase Corpus which is used for evaluation. Section 5 presents the results of this evaluation. Conclusions and suggestions for future work are presented in Section 6.

2. PREVIOUS APPROACHES
Purely lexical based matching techniques for paraphrase detection was used by (Clough et al., 2002; Qiu et al., 2006; Zhangand Patrick, 2005).A two-phase process was used by (Qiu et al., 2006) where thecommon semantic units in each sentence are first identified and pairedoff. The significance of the other units are also judged. If there are no unpaired units orif all unpairedunits are insignificant then a positive classificationis given. Comparison is done using a simple lexicalmatching technique.
(Zhang and Patrick, 2005)proposed to create intermediate forms of the sentences so that similartexts are transformed into thesame surface representation.Next, simplelexical matching techniques are used to compare thetransformed text.(Mihalcea etal., 2006) proposed word-to-word similarity measures anda word specificity measure to estimate thesemantic similarity of the sentence pairs.

3. PROPOSED SEMANTIC APPROACH
The proposed task at FIRE 2016 is focused on sentence level paraphrase identification for Indian languages (Malayalam). Sub Task 1: Given a pair of sentences from newspaper domain, the task is to classify them as paraphrases (P) or not paraphrases (NP). Sub Task 2: Given two sentences from newspaper domain, the task is to identify whether they are paraphrases (P), semi-paraphrase (SP) or not paraphrases (NP).
Our proposed semantic approach foridentifying theparaphrases comprisesof three phases – matching identical tokens, matching lemmas and matching with synonyms replaced. Similaritycomparison is performed at the sentence level using the Jaccard, Containment, Overlap and Cosine similarity metrics and if thesimilarity score of a sentence pair is higher than a predetermined threshold, the pair ismarked as plagiarised.The steps are illustrated in Figure 1.

3.1 Tokenization
The two input sentences are broken down into individual words or tokens and compared for similarity. Given two sentences S1 and S2, thetokens produced from S1 will be $\{W_1,W_2. . .W_N\}$, where N is the number of words in the sentence S1.

3.2 Lemmatization and POS tagging
The individual words in the two input sentencesare reduced to their root form or lemmas using a suffix stripping algorithm.Lemmatization is the technique of transforming words into their dictionary base forms.

Suffix stripping algorithm:

The inflected words for similarity analysis are converted to a valid root wordby means of suffix stripping along with some transformational rules. Each rule set consists of suffixes and their corresponding transformations that can generate the root word. This rule set is considers plurals and Vibhakthis in case of nouns and the different tense forms in case of verbs. Suffixes in Malayalam inflected word may range from a single character to a group of characters. So the algorithm starts stripping from the right side of the inflected word character wise. Each time a character which is a valid suffix in the rule set is stripped,

corresponding transformations are done and the resulting word in checked in the dictionary. If it is found the algorithm terminates. Otherwise the procedure continues until a valid word is found.

The root words are checked for correctness with the part of speech tag.These lemmas are then compared for similarity.

3.3 Synonym replacement

For the remaining lemmas that are not matched, substitute synonyms from the CUSAT Malayalam wordnet-PADASRINKALA. An example is given below

WORD : സമുദ്രം

SYNONYMS : സമുദ്രം , കടല് , ആഴി , അകൂപാരം , അപാംപതി , അപ്പതി , അബ്ധി , അര്ണ്ണവം , ഉദധി , ജലനിധി , പാരാവാരം , സാഗരം

POS : Noun

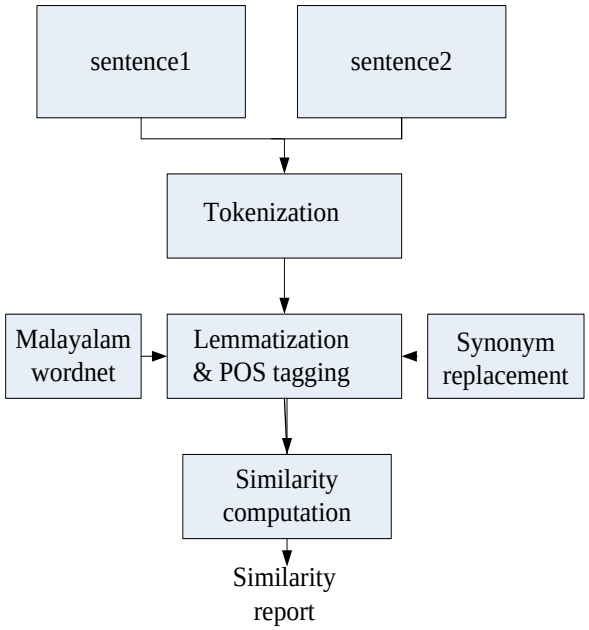

Figure 1. Paraphrase detection method

3.4 Similarity computation

The combined similarity obtained from direct word matches, lemma matches and synonym match produces a score between 0 and 1 that indicates the similarity between sentences S1 and S2.

a) Jaccard Similarity

$$S_{jaccard}(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

b) Containment measure

The similaritybetween two sentences is calculated using the containment similarity measure proposed by Clough and Stevenson (2010) given in equation.

$$S_{containment}(A,B) = \frac{|A \cap B|}{|A|}$$

A and B represent the sets of n-grams in the sentencesS1 and S2 respectively. The containmentmeasure calculates the intersecting n-grams but normalises them only with respectto the count of n-grams in the first sentence S1.

c) Overlap coefficient

The overlap coefficient is also proposed by Clough and Stevenson (2010) .

$$S_{overlap}(A,B) = \frac{|A \cap B|}{\min(|A|,|B|)}$$

A and B are the unique n-grams contained in the sentence S1 and sentence S2 respectively. The intersecting n-grams of both sentences is dividedby the sentence with the smaller word count.

d) Cosine Similarity

The similaritybetween two sentences is calculated using the cosine similarity given in equation.

$$S_{cosine}(A,B) = \frac{A \cdot B}{|A||B|}$$

Sentences S1 and S2 are represented as vectors A and B respectively.

Consider the example sentence pairs

S1: മകളെ പീഡിപ്പിച്ച പ്രതിയുടെ കൈരണ്ടും പിതാവ്മുറിച്ചുമാറ്റി.

S2:എട്ടുമാസം പ്രായമുള്ള പെണ്കുഞ്ഞിനെ പീഡിപ്പിച്ച പ്രതിയുടെ ഇരുകൈകളും കുട്ടിയുടെ അച്ഛന് മുറിച്ചുമാറ്റി.

From S1 and S2 we get

Direct matches: 3 (പീഡിപ്പിച്ച , പ്രതിയുടെ , മുറിച്ചുമാറ്റി)

Lemma match: 0

Synonym match: 2 (മകളെ↔പെണ്കുഞ്ഞിനെ
പിതാവ്↔ അച്ഛന്)

So the similarity or intersecting word count will be

Direct match + lemma match + synonym match

which is 3 + 0+ 2 = 5

If we find the overlap coefficient

Overlap-similarity = 5/6 = 0.8

Similarly all other measures are calculated.

Jaccard similarity = 0.5

Containment similarity= 0.8

Cosine similarity = 0.7

4. PARAPHRASE CORPUS

There are no annotated corpora or benchmark data for paraphrases available for Indian languages till date. The data provided for this shared task have been splitinto two training sets containing 2500 and 3500 examples respectively and two test sets containing 900 pairs of sentences for task1 and 1400 pairs of sentences for task2. The training data-set -1 contains 1000 sentencepairs that have been marked by human judges as paraphrases and1500 sentencepairs that have been marked as not paraphrases.
The training data-set -2 contains 1000 sentencepairs that have been marked as paraphrases , 1000 sentencepairs that have been marked as semi-paraphrases and 1500 sentencepairs that have been marked as not paraphrases.This train/test partitionhas been observed by all the approaches evaluatedhere.

Table 1. Training data

Sets	Number of Documents		
	Paraphrase	Semi paraphrase	Not paraphrase
Set-1	1000	0	1500
Set-2	1000	1000	1500

Table 2. Test data

Sets	Number of Documents
Task-1	900
Task-2	1400

Table 3. Examples of sentences from Train dataset

id	Sentence pair	Tag
1	റോയൽ ചലഞ്ചേഴ്സിനെ ആറു വിക്കറ്റിന് തകർത്ത് മുംബൈ വീണ്ടും വിജയവഴിയിൽ.	P
	ബാംഗ്ലൂർ റോയൽ. ചലഞ്ചേഴ്സിനെ മുംബൈ ആറു വിക്കറ്റിന് തോൽപ്പിച്ചു.	
2	സമുദ്രത്തിന്റെ അടിത്തട്ടിലുള്ള തെരച്ചില് വീണ്ടും ആരംഭിക്കും. ഒരു വർഷമെടുക്കും തെരച്ചില് പൂർത്തിയാകാന്.	NP
3	രണ്ടു വർഷമായി ഈസ്റ്റ് വെള്ളിമാടുകുന്ന് ഭാഗത്ത് കനാലില് വെള്ളമെത്തിയിട്ട്. ഈസ്റ്റ് വെള്ളിമാടുകുന്നില് കുടി വെള്ളം വറ്റി.	SP

5. EXPERIMENTS

The approach described in Section 3 was evaluatedagainst the Paraphrase Corpus.All synonyms of Malayalam WordNet were considered when finding the similaritybetween words.
The training data was used to find the classificationthreshold (paraphrase/semi-paraphrase/not-paraphrase) for the two tasks. Considering the four similarity measures, the following observations are made.
Containment measure is useful in cases where thesuspicious text is shorter than the source text. Overlap measure is useful in cases where the size of suspicious and source text varies. Jaccard similarity values are less compared to the Cosine value. Hence only the Cosine value is considered for setting the threshold.
Accuracy, precision, recall and F measurewere evaluated for the test corpus:These are defined as follows:

$$accuraccy = \frac{TP + TN}{TP + TN + FP + FN}$$

where TP are true positives, TN are true negatives,FN are false negatives and FP are false positives.

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

$$F = \frac{2\text{x } precision \text{ x } recall}{precision + recall}$$

Results for the semantic similarity approach on the test data areshown in Table3.

Table3. Results on test data

Task	No. of sentences	Accuracy	F-measure
Task-1	900	0.76	0.75
Task-2	1400	0.52	0.51

6. CONCLUSION AND FUTURE WORK

This paper presented an approach to the problemof paraphrase detection in Malayalam language. Paraphrase has been identifiedbased on the tokens and its synonyms that are common thathas been taken as attribute for checking paraphrase. Thewords are checked against Malayalam Wordnet. Bycalculating the token matching ,lemma match and synonymtoken matching andfixing an appropriate threshold value, the given sentence can be classified as paraphrase, semi-paraphrase sentence or not paraphrase.

From the obtained values of Accuracy and F-measure, we consider combining the similarity approaches in future to improve the efficiency of the system. Also, the accuracy of this method can be further enhanced by including a spell-checker and correcting misspelled words before similarity checking.

7. REFERENCES

[1] AnandKumar, M., Singh, S., Kavirajan, B., and Soman, K .P.2016.DPIL@FIRE2016: Overview of shared task on Detecting Paraphrases in Indian Languages. Working notes of FIRE 2016 - Forum for Information Retrieval Evaluation, Kolkata,India, December7-10,CEUR Workshop Proceedings,CEUR-WS.org

[2] Clough Paul, Robert Gaizauskas, Scott Piao, and YorickWilks.,2002,METER: MEasuring TExt Reuse.InProceedings of the 40th Anniversary Meeting for theAssociation for Computational Linguistics (ACL-02),, Pennsylvania, PA,pages 152–159.

[3] Qiu Long, Min-Yen Kan, and Tat-Seng Chua.,2006,Paraphrase recognition via dissimilarity significanceclassification., In Proceedings of the 2006 Conferenceon Empirical Methods in Natural Language Processing, , Sydney, Australia, July.Association for computational Linguistics,pages 18-26.

[4] Zhang. Y and Jon Patrick.,2005, Paraphrase identification by text canonicalization, In Proceedings of Australasian Language Technology Workshop 2005, Sydney, Australia,pages 160-166.

[5] Mihalcea.R, Courtney Corley, and Carlo Strapparava., 2006,Corpus-based and Knowledge-based Measures of Text Semantic Similarity, In Proceedings of the American Association for Artificial Intelligence (AAAI).

[6] Sundaram, Mahalakshmi Shanmuga, Anand Kumar M, and Soman Kotti Padannayil,"AMRITA CEN@ SemEval-2015: Paraphrase Detection for Twitter using Unsupervised Feature Learning with Recursive Autoencoders." SemEval-2015.

[7] Mahalakshmi, S., Anand Kumar, M., Soman, K.P, 2015, Paraphrase detection for Tamil language using deep learning algorithm, International Journal of Applied Engineering Research, 10 (17), pp. 13929-13934.