DPIL@FIRE 2016: Overview of Shared Task on Detecting Paraphrases in Indian Languages (DPIL)

Anand Kumar M, Shivkaran Singh
Center for Computational Engineering and Networking (CEN)
Amrita School of Engineering, Coimbatore
Amrita Vishwa Vidyapeetham
Amrita University
m_anandkumar@cb.amrita.edu

Kavirajan B, Soman K P
Center for Computational Engineering and Networking (CEN)
Amrita School of Engineering, Coimbatore
Amrita Vishwa Vidyapeetham
Amrita University
kp_soman@amrita.edu

ABSTRACT
This paper explains the overview of the shared task "Detecting Paraphrases in Indian Languages" (DPIL) conducted at FIRE 2016. Given a pair of sentences in the same language, participants are asked to detect the semantic equivalence between the sentences. The shared task is proposed for four Indian languages namely Tamil, Malayalam, Hindi and Punjabi. The dataset created for the shared task has been made available online and it is the first open-source paraphrase detection corpora for Indian languages.

CCS Concepts
Computing methodologies → Artificial intelligence → Natural language processing → Language resources

Computing methodologies → Artificial intelligence → Natural language processing → Lexical semantics

Keywords
Paraphrase detection; Semantic analysis, Indian languages; DPIL Corpora

1. INTRODUCTION
A Paraphrase can be defined as "the same meaning of a sentence is expressed in another sentence using different words". Paraphrases can be identified, generated or extracted. The proposed task is focused on sentence-level paraphrase identification for Indian languages (Tamil, Malayalam, Hindi and Punjabi). Identifying paraphrases in Indian languages is a difficult task because evaluating the semantic similarity of the underlying content and the understanding the morphological variations of the language are more critical. Paraphrase identification is strongly connected with generation and extraction of paraphrases. The paraphrase identification systems improve the performance of a paraphrase generation in terms of choosing the best paraphrase candidate from the list of candidates generated by paraphrase generation system. Paraphrase Identification is also used in validating the paraphrase extraction system and the machine translation system. In question answering system, Paraphrase identification plays a vital role in matching the questions asked by the user to the original questions for choosing the best answer. Automatic short answers grading is another interesting application which needs semantic similarity for providing grades to the short answers. Plagiarism detection is another task which needs the paraphrase identification technique to detect the sentences which are paraphrases of other sentences.

One of the most commonly used corpora for paraphrase detection is the MSRP corpus[1], which contains 5,801 English sentence pairs from news articles manually labeled with 67% paraphrases and 33% non-paraphrases. Since there are no annotated corpora or automated semantic interpretation systems available for Indian languages till date, creating benchmark data for paraphrases and utilizing that data in open shared task competitions will motivate the research community for further research in Indian languages.

Details about the task and dataset can be found on the website[1] of the shared task. The descriptions of the subtasks and evaluation metrics are discussed in Section 2, Paraphrase corpus creation and statistics are explored in Section 3, System descriptions of participants and result analyses are done in Section 4. We discuss the findings from the results Section 5.

2. RELATED TASKS AND CORPORA
In SemEval-2015[2], shared task on Paraphrase and Semantic Similarity In Twitter (PIT) [2] was conducted with the English Twitter Paraphrase Corpus [3]. The task has two sentence-level sub-tasks: a paraphrase identification task and a semantic textual similarity task. The same dataset was used for both sub-tasks but it differs in annotation and evaluation. A freely available manually annotated corpus of Russian sentence pairs is ParaPhraser [4], which is used in the recently organized shared task on Paraphrase detection for the Russian language [whit pap]. There were two subtasks, one was three-class classification: given a pair of sentences, to predict whether they are precise paraphrases, near paraphrases or non-paraphrases and another was binary classification: given a pair of sentences to predict whether they are paraphrases or non-paraphrases. Microsoft Research Paraphrase (MSRP) corpus is a well-known corpus which is manually annotated and it consists of 5,801 paraphrase pairs in the English language. The PAN plagiarism corpus 2010 (Paraphrase for Plagiarism -P4P) is used for the evaluation of automatic plagiarism detection algorithms. The corpus [5] is manually annotated with the paraphrase phenomena they contain. It is composed of 847 source-plagiarism pairs in English. The complete summary of existing paraphrase corpora and Linguistic phenomenon for paraphrases are discussed in [6]. In [7], issue of text plagiarism for Hindi language using English documents is addressed. For Tamil languages, paraphrase detection using deep learning techniques is applied in [8]. For Malayalam, paraphrase identification using fingerprinting [9] and statistical similarity [10] has been performed.

[1] http://nlp.amrita.edu/dpil_cen/

[2] http://alt.qcri.org/semeval2015/

Table 1. Examples for Hindi and Tamil language

Hindi	मृतका निशा तीन भाई-बहनों में सबसे बड़ी थी। *[The deceased Nisha was eldest of three siblings]* तीन भाई-बहनों में सबसे बड़ी थी मृतका निशा। *[Out of three siblings, deceased Nisha was the eldest]*	P
	उपमंत्री की बेसिक सैलरी 10 हजार से बढ़कर 35 हजार हो गई है। *[The basic salary of deputy minister is increased from 10k to 35k]* उपमंत्री की बेसिक सैलरी 35 हजार हो गई है। *[The basic salary of deputy minister is 35k]*	SP
	जिमनास्टिक में दीपा 4th पोजिशन पर रहीथीं। *[Deepa came at 4th position in gymnastics]* 11 भारतीय पुरुष जिमनास्ट आजादी के बाद से ओलिंपिक में जाचुकेहैं। *[Since independence 11 male athletes have been to Olympics]*	NP
Tamil	புதுச்சேரியில் 84 சதவீத வாக்குப்பதிவு [*84 percent voting in Puducherry*] புதுச்சேரி சட்டசபை தேர்தலில் 84 சதவீத ஓட்டுப்பதிவானது [*Puducherry assembly elections recorded 84 percent of the vote*]	P
	அப்துல்கலாம் கனவை நிறைவேற்றும் வகையில் மாதம் ஒரு செயற்கைகோள் அனுப்ப திட்டம் [In *order to fulfill Abdul Kalam's dream, planning is to send a satellite per month*] ஒரு செயற்கைகோளை அனுப்ப வேண்டும் என்பது அப்துல்கலாமின் கனவு [Abdul Kalam's dream was to send a satellite]	SP
	அறைகளில் இருந்தும் சிலைகள், ஓவியங்கள் கிடைத்தன [Statues and paintings were found from the rooms] மூன்று நாட்கள் நடத்தப்பட்ட சோதனையில் மொத்தம் 71 கற்சிலைகள் மீட்கப்பட்டுள்ளன [A total of 71 stone statues have been recovered in a three day raid]	NP

3. TASK DESCRIPTION & EVALUATION METRIC

3.1 Task description

There were two subtasks under shared task on Detecting Paraphrase in Indian Languages (DPIL). The description of the subtask are:

Subtask 1: Given a pair of sentences from newspaper domain, the shared task is to classify them as paraphrases (P) or not paraphrases (NP).

Subtask 2: Given a pair of sentences from newspaper domain, the shared task is to identify whether they are paraphrases (P) or semi-paraphrases (SP) or not paraphrases (NP).

The subtask 2 was similar to the subtask 1 except the 3-point scale tag in paraphrases. This makes the shared task even more challenging

3.2 Evaluation metrics

The evaluation metrics used for subtask 1 and subtask 2 were slightly different because of uniqueness of the tasks. To evaluate runs for subtask 1, we used accuracy and f-score values. The $Accuracy$ (1) and $F1-score$ (2) for subtask 1 were calculated as follows:

$$Accuracy = \frac{Number\ of\ correct\ instances}{Total\ number\ of\ instances} \quad (1)$$

$$Precision_P = \frac{Number\ of\ correct\ paraphrases}{Number\ of\ detected\ paraphrases}$$

$$Recall_P = \frac{Number\ of\ correct\ paraphrases}{Number\ of\ reference\ paraphrases}$$

Subsequently, $F1-score$ can be calculated as:

$$F1-score_P = \frac{2 \times Precision_p \times Recall_p}{Precision_p + Recall_p} \quad (2)$$

The subscript p refers to paraphrase (P) class for the subtask 1. Similarly, $Accuracy$ and $F1-score$ for non-paraphrase class could be calculated.

To evaluate runs for subtask 2, we used $Accuracy$, $micro-F\ score$ and $macro-F\ score$. Since it is a multiclass classification task, $Accuracy$ and $micro-F\ measure$ gives identical scores. The $macro-F\ score$ (3) could be computed as:

$$Macro - P = \frac{Precision_P + Precision_{NP} + Precision_{SP}}{Number\ of\ classes}$$

$$Macro - Re = \frac{Recall_P + Recall_{NP} + Recall_{SP}}{Number\ of\ classes}$$

$$Macro - F1\ score = \frac{2 \times Macro - P \times Macro - R}{Macro - P + Macro - R} \quad (3)$$

Where $Macro - P$ and $Macro - R$ are the macro precision and macro recall, which is used to calculate $Macro - F1\ score$.

4. PARAPHRASE CORPUS FOR INDIAN LANGUAGES

A paraphrase is a linguistic phenomenon. It has many applications in the field of language teaching as well as computational linguistics. Linguistically, paraphrases are defined in terms of meaning. According to Meaning-Text Theory [11], if one or more syntactic construction retains semantic evenness, those are addressed as paraphrases. The exchangeability of semantic alikeness between the source text and paraphrased version mark the range of semantic alikeness between them. A paraphrase is a very fine mechanism to shape various language models. Different linguistic units like synonyms, semi-synonyms, figurative meaning and metaphors are considered as the basic elements for paraphrasing. Paraphrasing is closely related with synonyms. Paraphrasing is not only found in lexical level but another linguistic level such as phrasal and sentential level also. Different levels of paraphrasing disclose the diversified forms of paraphrases and the semantic relationship to its source text. In paraphrase typologies, *Lexical paraphrasing* is the most popular forms of paraphrasing found in the literature. For example: If a source text is, "The two ships were *acquired* by the navy after the war", then possible paraphrased versions are: "The two ships were *conquered* by the navy after the war" and "The two ships were *won* by the navy after the war". There are even more paraphrases possible for the given sentence. Here the source verb 'acquire' is paraphrased with its exact synonyms. The source and paraphrases show the same syntactic structural phenomena. These types of paraphrase are the best examples for exact paraphrases. Some of the other common paraphrase typologies are; *approximate paraphrases*, *sentential level paraphrases*, *adding extra linguistic units*, *changing the order* etc.

The shared task on Detecting Paraphrases in Indian Languages (DPIL)[3] required participants to identify sentential paraphrases in four Indian languages, namely Hindi, Tamil, Malayalam, and Punjabi. The corpora creation task for these Indian languages started with collecting news articles from various web-based news sources. The collected dataset was further cleaned from any noise or informal information. Apart from cleaning, some sentences required spelling corrections and text transformations. After removing all the irregularities, the dataset was annotated according to the paraphrases phenomena (Paraphrase, Non-Paraphrase, Semi-Paraphrase) present in each sentence pair. The annotation tags used were P, SP and NP corresponding to Paraphrase, Semi-Paraphrase and Non-Paraphrase. These annotations were done by language experts for each language. The annotated files were further proofread by a linguistic expert and then again by language expert (Two-step Proofreading). Additionally, the annotated dataset proofread by linguistic expert was converted to Extensible Markup Language (XML) format.

[3]http://nlp.amrita.edu/dpil_cen/

Table 1 includes examples of Paraphrase, Semi-Paraphrase, and Non-Paraphrase for Hindi and Punjabi Language. Where H stands for Hindi and P stand for Punjabi and P, SP and NP are the tags used for Paraphrase, Semi-Paraphrase, and Non-Paraphrase. English translation for each sentence pairs is given for the non-native speakers to understand the meaning. It can be seen that Paraphrased sentence pairs contain the same information, Semi paraphrased sentence pair's lacks additional information and Non-Paraphrases conveys totally different information.

4.1 Corpora statistics

The paraphrase corpus was further analysed for certain parameters such as number of sentence pairs for each class (P, NP, and SP), average number of words per sentence per task, and overall vocabulary size. The statistics for number of sentence pairs in testing and training phase for each subtask is given in Table 2.

Table 2. Statistic for sentence pairs in Subtask 1 & 2

Language	Subtask1 (in pairs)		Subtask2 (in pairs)	
	Train	Test	Train	Test
Tamil	2500	900	3500	1400
Malayalam	2500	900	3500	1400
Hindi	2500	900	3500	1400
Punjabi	1700	500	2200	750

The average number of words per sentence along with average pair length for subtask 1 and subtask 2 is given in Table 3 & Table 4.

Table 3. Average number of words per sentence for Subtask 1

Language	Subtask - 1		
	Sentence 1	Sentence 2	Pair
Hindi	16.058	16.376	16.217
Tamil	11.092	12.044	11.568
Malayalam	9.253	9.035	9.144
Punjabi	19.485	19.582	19.534

Table 4. Average number of words per sentence for Subtask 2

Language	Subtask - 2		
	Sentence 1	Sentence 2	Pair
Hindi	17.78	16.48	17.130
Tamil	11.097	11.777	11.437
Malayalam	9.414	8.449	8.932
Punjabi	20.994	19.699	20.347

The overall vocabulary size (Subtask 1 & Subtask 2) for training as well as testing for all the languages is shown in the form of line chart in Figure 1.Notably, vocabulary size for Hindi & Punjabi languages is less than Tamil and Malayalam. This is because, like other Dravidian languages (Kannada & Telugu), Tamil and Malayalam are agglutinative in nature. Due to this phenomenon, Dravidian languages end up having more unique words and hence larger vocabulary.

5. SYSTEM DESCRIPTION AND RESULTS

A total of 35 teams registered for the organized shared task and out of those, 11 teams successfully submitted their runs. A brief description about the methodologies used by each team is given in the following subsection.

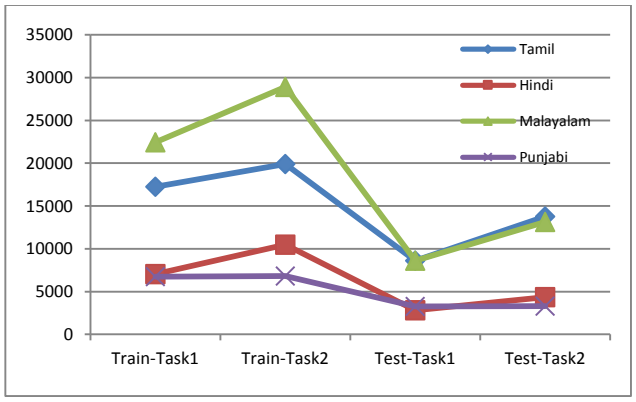

Figure 1. Overall vocabulary size

5.1 Participants System Description

The brief description of the techniques used by all the teams to submit the runs for the shared task are as follows:

ANUJ: This team participated only for the Hindi language. They pre-process the sentences using stemmer, soundex, synonym handler. After that, they extracted the features using overlapping words and normalized IDF scores. Finally, the Random forest classifier is used for classification.

ASE: This team participated only for Hindi Language. They extracted the features using POS tags and stemming information. Semantic similarity metric is employed which extracts the word synonyms from WordNet to check whether the compared words are synonyms. Finally, decision tree classifier is used to detect the paraphrases.

BITS_PILANI: This team participated for Hindi language only. They attempted paraphrase detection with different classifiers and finally used Logistic Regression for Subtask-1 and Random Forest for Subtask2.

CUSAT-TEAM: This team participated only for the Malayalam Language. They stemmed the words and calculated the sentence vector using Bag of Words model and find out the similarity score between sentences. Finally, they set a threshold for determining the appropriate class.

CUSAT_NLP: This team participated only in the Malayalam Language. They used identical tokens, matching lemmas and synonyms for finding the similarity between sentences. They also utilized in-house Malayalam Wordnet to replace the synonyms. Finally, the similarity score is compared and a threshold is fixed to identify the exact class.

HIT2016: This team participated in all the four languages. Cosine Distance, Jaccard Coefficient, Dice Distance and METEOR features are used and classification is done based on Gradient Boosting Tree. They experiment various aspects of the classification method for detecting paraphrases.

JU_NLP: This team competed in all the four languages. They used similarity based features, word overlapping features and scores from the machine translation evaluation metrics to find out the similarity scores between pair of sentences. They tried with three different classifiers namely Naïve Bayes, SVM and SMO.

KEC@NLP: This team participated in Tamil language only. They used existing Tamil Shallow parser to extract the morphological features and utilizing Support Vector Machine and Maximum Entropy for classifying paraphrases.

KS_JU: This team participated in all the four languages. They used different lexical and semantic level (Word embeddings) similarity measures for computing features and used multinomial logistic regression model as a classifier.

NLP-NITMZ: This team also participated in all the four languages. They used features based on Jaccard Similarity, length normalized Edit Distance and Cosine Similarity. Finally, these feature-set are trained using Probabilistic Neural Network (PNN) to detect the paraphrases.

5.2 Overall Results

As announced during the shared task, we are giving **Sarwan award** for top performers in each languages. The name of the top performing team in each language is given in Table 5.The overall results of all the participating teams can be seen in Table 6. For representation purpose we have truncated the evaluation measures (Precision, Recall, and Accuracy) to two digits[4].

Table 5. Top performers in each language

Punjabi	Hindi	Malayalam	Tamil	Rank
0.932 (HIT)	***0.907 (Anuj)***	***0.785 (HIT)***	***0.776 (HIT)***	***First****
0.922 (JU_KS)	0.896 (HIT)	0.729 (JU_KS)	0.741 (KEC)	Second
0.913 (JU)	0.876 (JU_KS)	0.713 (NIT-MZ)	0.727 (NIT-MZ)	Third

6. DISCUSSIONS

Out of the 11 teams which submitted their runs, 10 teams successfully submitted their working notes. There were four teams which participated in all the four languages and rest of the teams (3-Hindi, 2-Malayalam and 1-Tamil) participated in only one language. Two out of ten teams used the threshold based method to detect the paraphrases, remaining teams used the machine learning based approaches. The different types of feature set used by the participant teams are illustrated in Table 7. Most of the teams used the common similarity based features like cosine, Jaccard, and only two teams used the Machine Translation evaluation metrics, BLEU and METEOR as features. Very few teams used the synonym replacement and Wordnet features. For Tamil language, team KEC@NLP used the morphological information as features to the machine learning based classifier. KS_JU team created the word2vec embeddings with the help of additional in-house unlabeled data and found out the semantic similarity features which were used as features in the classifier. The top performing team (HIT-2016) for the three languages used the character n-gram based features and they experimented the results for different n-gram size.

We calculated F1-measure and accuracy for evaluating the submissions of the teams. The accuracy of the Task-2 is comparably low with the accuracy of Task-1 due to complexity of the task. In general, the accuracy obtained by runs submitted for Tamil and Malayalam language is low as compared to the accuracy obtained by Hindi and Punjabi language. This is due to the agglutinative nature of the Dravidian languages.

[4] It does not affect the result of the participating teams

Table 6. Overall result for Subtask 1 & Subtask 2

Team Name	Language	Subtask 1				Subtask 2			
		Precision	Recall	Accuracy	F1 Score	Precision	Recall	Accuracy	F1 Score
Anuj	Hindi	0.95	0.90	0.9200	0.91	0.90	0.90	0.9014	0.90
ASE	Hindi	0.41	0.35	0.3588	0.34	0.35	0.35	0.3542	0.35
ASE*	Hindi	0.82	0.97	0.8922	0.89	0.68	0.67	0.6660	0.67
BITS-PILANI	Hindi	0.91	0.90	0.8977	0.89	0.72	0.72	0.7171	0.71
CUSAT NLP	Malayalam	0.83	0.72	0.7622	0.75	0.52	0.52	0.5207	0.51
CUSATTEAM	Malayalam	0.79	0.88	0.8044	0.76	0.51	0.50	0.5085	0.46
DAVPBI[φ]	Punjabi	0.95	0.92	0.9380	0.94	0.77	0.76	0.7466	0.73
HIT2016	Hindi	0.97	0.84	0.8966	0.89	0.90	0.90	0.9000	0.89
HIT2016	Malayalam	0.84	0.87	0.8377	0.81	0.74	0.75	0.7485	0.74
HIT2016	Punjabi	0.95	0.94	0.9440	0.94	0.95	0.95	0.9226	0.92
HIT2016	Tamil	0.82	0.87	0.8211	0.79	0.75	0.75	0.7550	0.73
JU-NLP	Hindi	0.75	0.99	0.8222	0.74	0.68	0.68	0.6857	0.68
JU-NLP	Malayalam	0.58	0.99	0.5900	0.16	0.42	0.42	0.4221	0.30
JU-NLP	Punjabi	0.95	0.94	0.9420	0.94	0.91	0.91	0.8866	0.88
JU-NLP	Tamil	0.57	1.00	0.5755	0.09	0.55	0.55	0.5507	0.43
KS_JU	Hindi	0.94	0.89	0.9066	0.90	0.85	0.85	0.8521	0.84
KS_JU	Malayalam	0.83	0.82	0.8100	0.79	0.66	0.66	0.6614	0.65
KS_JU	Punjabi	0.95	0.94	0.9460	0.95	0.92	0.92	0.8960	0.89
KS_JU	Tamil	0.79	0.85	0.7888	0.75	0.67	0.67	0.6735	0.66
NLP@KEC	Tamil	0.82	0.87	0.8233	0.79	0.68	0.68	0.6857	0.66
NLP-NITMZ	Hindi	0.92	0.92	0.9155	0.91	0.78	0.78	0.7857	0.76
NLP-NITMZ	Malayalam	0.8	0.94	0.8344	0.79	0.62	0.62	0.6243	0.60
NLP-NITMZ	Punjabi	0.95	0.94	0.9420	0.94	0.83	0.83	0.8120	0.80
NLP-NITMZ	Tamil	0.8	0.92	0.8333	0.79	0.66	0.66	0.6571	0.63

Table 7. various Features used by the participants

Features	Anuj	ASE	BITS-PILANI	CUSAT NLP	CUSAT TEAM	HIT2016	JU-NLP	KS_JU	NLP@KEC	NLP-NITMZ
POS			✓	✓					✓	
Stem/Lemma	✓	✓	✓	✓	✓			✓		
Stopwords	✓	✓			✓					
Word Overlap	✓						✓	✓		
Synonym	✓	✓		✓						
Cosine				✓	✓	✓	✓	✓		✓
Jaccord						✓	✓			✓
Levinstin			✓							✓
METEOR/BLEU						✓	✓			
Others	IDF		Soundex	WordNet	BoW	N-gram	Dice	word2vec	Morph	
Classifier	Random Forest	J 48	Log Reg/ Random Forest	Threshold	Threshold	Gradient Tree Boosting	SMO	Multi-nomial Log Reg	Maximum Entropy	Prob NN

** Due to some formatting issues, this participant re-submitted the system after deadline.*

φ This participant didn't submitted the working notes.

7. CONCLUSIONS AND FUTURE SCOPE

In this overview paper, we explained the paraphrase corpus details and evaluation results of subtask-1 and subtask-2 of Detecting Paraphrases in Indian Languages (DPIL) shared task held at the 8th Forum for Information Retrieval (FIRE) Conference - 2016. A total number of 35 teams registered in which 11 teams submitted their runs successfully. The corpora developed for the shared task is the first publicly available paraphrase detection corpora for Indian languages. Detecting paraphrases and semantic similarity in Indian languages is a challenging task because the morphological variations and the semantic relations in Indian languages are more crucial to understand. Discrepancies can be found in manually annotated paraphrase corpus, to revise the annotations feedbacks are welcome and appreciated. Our detailed experiment analysis provides fundamental insights into the performance of paraphrase identification in Indian languages. Overall, HIT-2016 (HeiLongJiang Institute of Technology) got the first place in Tamil, Malayalam, and Punjabi languages and Anuj (Sapient Global Markets) got the first place in Hindi. As a future work, we plan to extend the task to analyze the performance of cross-genre and cross-lingual paraphrases for more Indian languages. Detecting paraphrases in social media content of Indian languages, plagiarism detection and use of paraphrases in Machine Translation Evaluation are also interesting areas for further study.

8. ACKNOWLEDEMENT

First, we would like to thank FIRE 2016 organizers for giving us an opportunity to organize the shared task on Detecting Paraphrases for Indian Languages (DPIL). We would like to extend our gratitude to the advisory committee members Prof. Ramanan, RelAgent Pvt. Ltd, and Prof. Rajendran S, Computational Engineering and Networking (CEN) for actively supporting us throughout the track. We would like to thank our PG students at CEN for helping us in creating the paraphrase corpora.

9. REFERENCES

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

[2] Xu, W., Callison-Burch, C. and Dolan, W.B., 2015. SemEval-2015 Task 1: Paraphrase and semantic similarity in Twitter (PIT). *Proceedings of SemEval*.

[3] Xu, W., Ritter, A., Callison-Burch, C., Dolan, W.B. and Ji, Y., 2014. Extracting lexically divergent paraphrases from Twitter. *Transactions of the Association for Computational Linguistics*, *2*, pp.435-448.

[4] Pronoza, E., Yagunova, E. and Pronoza, A., 2016. Construction of a Russian paraphrase corpus: unsupervised paraphrase extraction. In Information Retrieval (pp. 146-157). Springer International Publishing.

[5] Potthast, M., Stein, B., Barrón-Cedeño, A. and Rosso, P., 2010, August. An evaluation framework for plagiarism detection. In *Proceedings of the 23rd international conference on computational linguistics: Posters* (pp. 997-1005). Association for Computational Linguistics.

[6] Rus, V., Banjade, R. and Lintean, M.C., 2014. On Paraphrase Identification Corpora. In *LREC* (pp. 2422-2429).

[7] Kothwal, R. and Varma, V., 2013. Cross lingual text reuse detection based on keyphrase extraction and similarity measures. In *Multilingual Information Access in South Asian Languages* (pp. 71-78). Springer Berlin Heidelberg.

[8] 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

[9] Idicula, S.M., 2015, December. Fingerprinting based detection system for identifying plagiarism in Malayalam text documents. In *2015 International Conference on Computing and Network Communications (CoCoNet)* (pp. 553-558). IEEE.

[10] Mathew, D. and Idicula, S.M., 2013, December. Paraphrase identification of malayalam sentences-an experience. In *2013 Fifth International Conference on Advanced Computing (ICoAC)* (pp. 376-382). IEEE.

[11] Kahane, S., 2003. The meaning-text theory. *Dependency and Valency. An International Handbook of Contemporary Research*, *1*, pp.546-570.