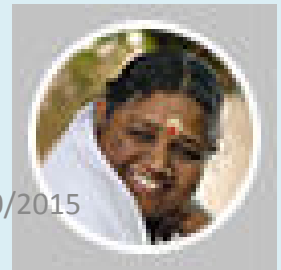
DPIL@FIRE 2016: Overview of Shared Task on Detecting Paraphrases in Indian Languages (DPIL)

M. Anand Kumar, Shivkaran Singh, Kavirajan B, and Soman K P
Center for Computational Engg and Networking,
Amrita Vishwa Vidyapetham,
Coimbatore



AMRITA VISHWA VIDYAPEETHAM UNIVERSITY

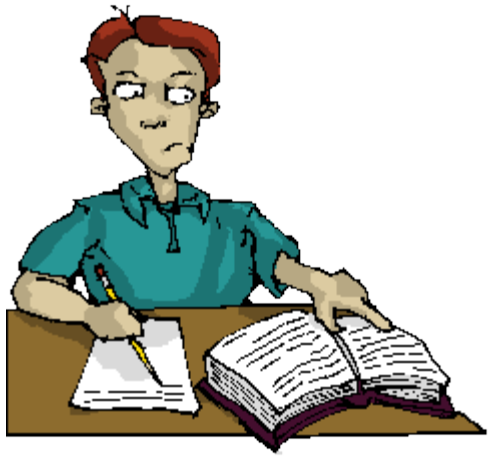
Outline

- Paraphrase Detection
- Motivation
- Task Descriptions
- DPIL Dataset
- Applications
- Participants
- Methodologies and Features
- Results
- Conclusion and Future Scope

Paraphrase Detection

- *Paraphrase detection "find out whether the given two sentences convey the same meaning or not"*.

- Four Indian languages (Hindi, Punjabi, Tamil and Malayalam).



- Since there are ***no annotated corpora*** or automated semantic interpretation systems available for Indian languages .
- 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.

Task description

- There were two subtasks under shared task on Detecting Paraphrase in Indian Languages (DPIL).
  - ***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).***

Given: A pair of Sentences S1 = { w1,w2,..wm} and S2={w1,w2,..wn} in same language.

Task1: Classify whether s1 and S2 are P or NP

Task2: Classify whether S1 and S2 are P or NP or SP

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

Applications of Paraphrase Detection

- Paraphrase identification is strongly connected with ***generation*** and ***extraction*** of paraphrases.
- ***Evaluation*** of Machine Translation system.
- Question answering system
- Automatic ***short answers grading*** is another interesting application which needs semantic similarity for providing grades to the short answers.

Evaluation Metrics

$$Accuracy = \frac{Number\ of\ correct\ instances}{Total\ number\ of\ instances}$$

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

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

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

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

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

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

$$Macro - F1\ score = \frac{2 \times Macro - P \times Macro - R}{Macro - P + Macro - R}$$

DPIL Dataset

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

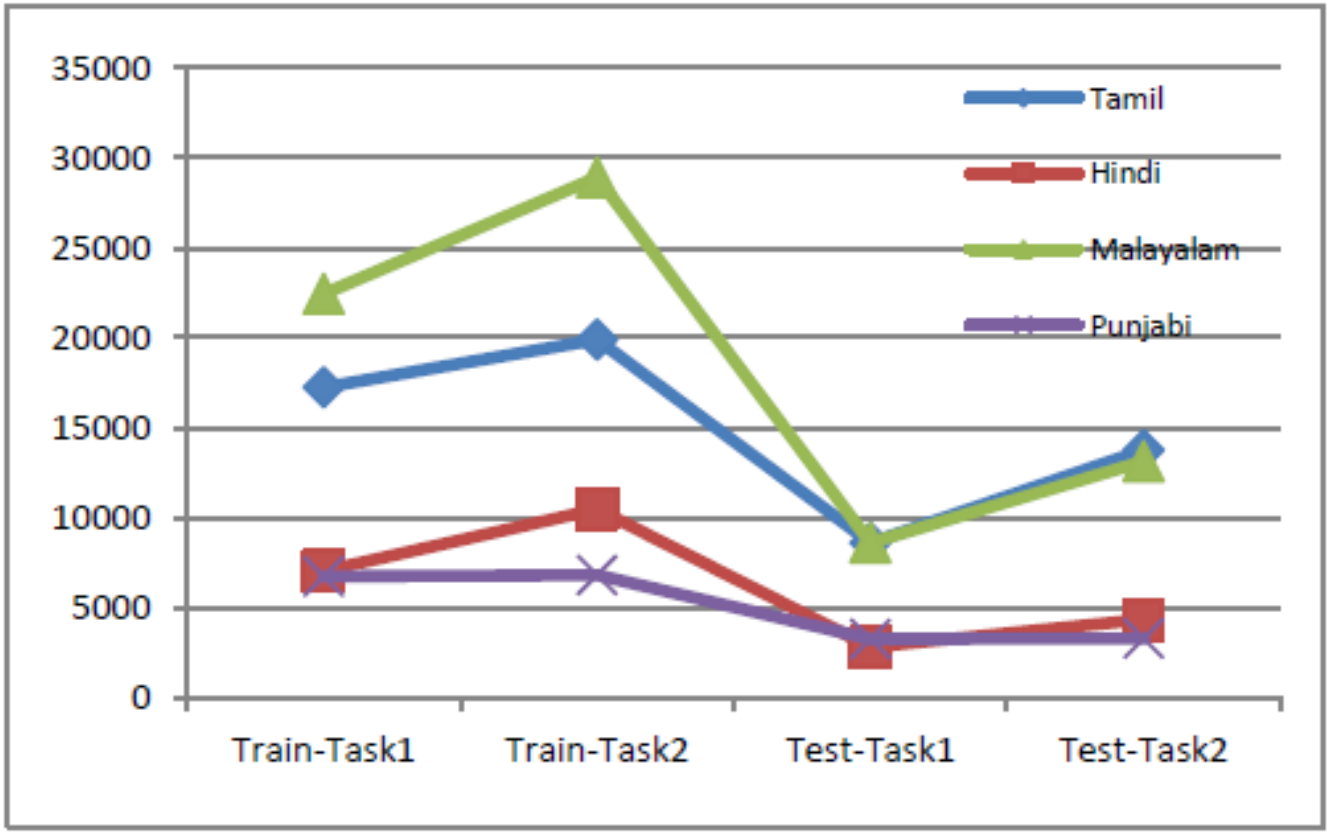
Average Number of Words per Sentence

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

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

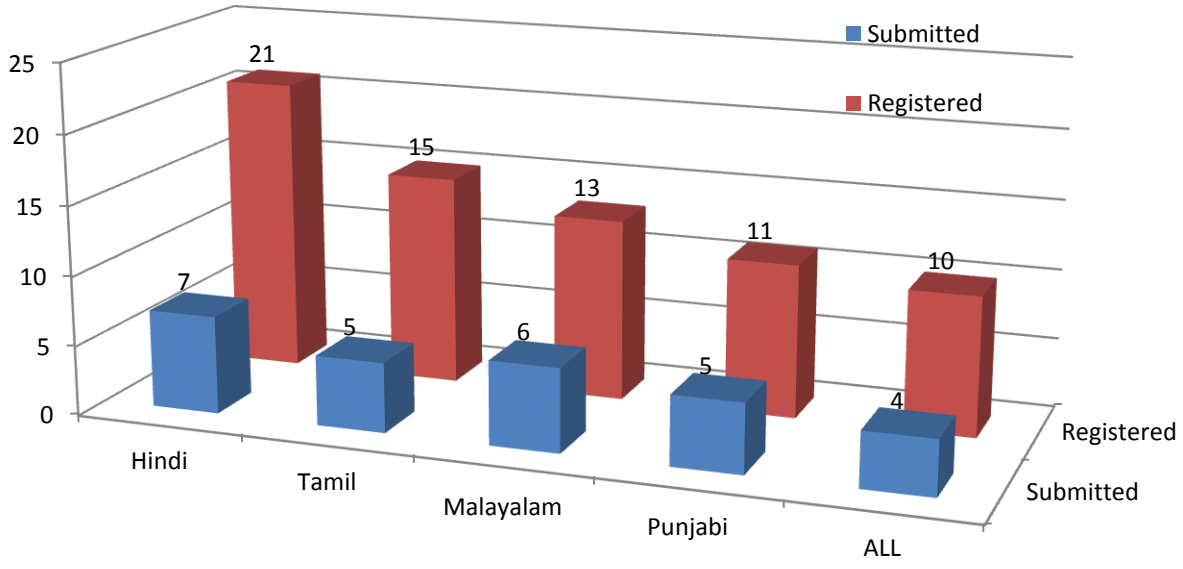
Vocabulary Size vs Tasks

- Vocabulary size for Hindi & Punjabi languages is less than Tamil and Malayalam. Tamil and Malayalam are highly ***agglutinative*** in nature



Participants

- 35 teams registered -11 teams successfully submitted their runs – Working notes 10.



Methodologies

- Two teams used the ***threshold based method*** to detect the paraphrases, remaining teams used the machine learning based approaches.
- 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 used the ***word2vec*** embeddings.
- 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.

Features used

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

Team Name	Language	Subtask 1		Subtask 2	
		Accuracy	F1 Score	Accuracy(Micro-F1)	Macro-F1 Score
Anuj	Hindi	**0.92**	**0.91**	**0.90142**	**0.90001**
ASE	Hindi	0.35888	0.34	0.35428	0.3535
ASE-1[$]	Hindi	0.8922	0.89	0.666	0.667
BITS-PILANI	Hindi	0.89777	0.89	0.71714	0.71226
CUSAT NLP	Malayalam	0.76222	0.75	0.52071	0.51296
CUSATTEAM	Malayalam	0.80444	0.76	0.50857	0.46576
DAVPBI*	Punjabi	0.938	0.94	0.74666	0.7274
HIT2016	Hindi	0.89666	0.89	0.9	0.89844
HIT2016	Malayalam	**0.83777**	**0.81**	**0.74857**	**0.74597**
HIT2016	Punjabi	0.944	0.94	**0.92266**	**0.923**
HIT2016	Tamil	0.82111	**0.79**	**0.755**	**0.73979**
JU-NLP	Hindi	0.8222	0.74	0.68571	0.6841
JU-NLP	Malayalam	0.59	0.16	0.42214	0.3078
JU-NLP	Punjabi	0.942	0.94	0.88666	0.88664
JU-NLP	Tamil	0.57555	0.09	0.55071	0.4319
KS_JU	Hindi	0.90666	0.9	0.85214	0.84816
KS_JU	Malayalam	0.81	0.79	0.66142	0.65774
KS_JU	Punjabi	**0.946**	**0.95**	0.896	0.896
KS_JU	Tamil	0.78888	0.75	0.67357	0.66447
NLP@KEC	Tamil	0.82333	**0.79**	0.68571	0.66739
NLP-NITMZ	Hindi	0.91555	**0.91**	0.78571	0.76422
NLP-NITMZ	Malayalam	0.83444	0.79	0.62428	0.60677
NLP-NITMZ	Punjabi	0.942	0.94	0.812	0.8086
NLP-NITMZ	Tamil	**0.83333**	**0.79**	0.65714	0.63067

Sarwan Award Winners

Punjabi ***0.932*** ***(HIT)***	**Hindi** ***0.907*** ***(Anuj)***	**Malayalam** ***0.785*** ***(HIT)***	**Tamil** ***0.776*** ***(HIT)***	**Rank** ***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

Conclusion and Future Scope

- Tamil and Malayalam language ***accuracy is low*** as compared to the accuracy obtained by Hindi and Punjabi language.
- ***Discrepancies*** can be found in manually annotated paraphrase corpus .
- Extend the task to analyze the performance of ***cross-genre*** and ***cross-lingual paraphrases*** for more Indian languages.
- Detecting paraphrases in social media content and ***code-mixed text*** of Indian languages.
- Role of ***Morpho-Syntactic knowledge with Recursive Auto Encoders*** in Paraphrase Detection in Indian Languages.
- Applying to Machine Translation Evaluation.

References

- Dolan, W.B. and Brockett, C., 2005, October. Automatically constructing a corpus of sentential paraphrases. In *Proc. of IWP.*
- Xu, W., Callison-Burch, C. and Dolan, W.B., 2015. SemEval-2015 Task 1: Paraphrase and semantic similarity in Twitter (PIT). *Proceedings of SemEval.*
- 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.*
- Socher, Richard, Eric H. Huang, Jeffrey Pennin, Christopher D. Manning, and Andrew Y. Ng. "Dynamic pooling and unfolding recursive autoencoders for paraphrase detection." In *Advances in Neural Information Processing Systems*, pp. 801-809. 2011.
- 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.
- 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.*
- Rus, V., Banjade, R. and Lintean, M.C., 2014. On Paraphrase Identification Corpora. In *LREC (pp. 2422-2429).*


thank you
merci
danke
謝謝
gracias
teşekkür ederim
ngiyabonga
obrigado
dziękuję
bedankt
go raibh maith agat
sukriya
kop khun krap
terima kasih
grazie
arigatô
dakujem
tapadh leat
hvala
mauruuru
sagolun
takk
taiku
спасибо
рахмат
Баярлалаа
faafetai lava
vinaka
спасибі
blagodaram
dank je
misaotra
matondo
paldies
grazzi
mahalo
хвала
asante
manana
obrigada
tenki
murakoze
chokrane
djiere dieuf
tau
mochchakkeram
дякую
mamnun
akun
dankon
ačiū
chnorakaloutioun
gratias ago
gràcies
sulpáy
ありがとう
tanemirt
rahmet
najis tuke
kam sah hamnida
rahmat
তোমাকে ধন্যবাদ
감사합니다
xièxie
ευχαριστώ
diolch
dhanyavadagalu
shukriya
merci
мерси
trugarez
didi madloba
mèsi
sobodi
děkuji
enkosi
bayarlalaa
gràçie
dhanyavad
kiitos
dankie
nanni
nandri
köszönöm
welalin
tack
barka
kia ora
spas
mersi