CMEE-IL: Code Mix Entity Extraction in Indian Languages from Social Media Text @ FIRE 2016 – An Overview

Pattabhi RK Rao
AU-KBC Research Centre
MIT Campus of Anna University, Chrompet, Chennai, India
+91 44 22232711
pattabhi@au-kbc.org

Sobha Lalitha Devi
AU-KBC Research Centre
MIT Campus of Anna University, Chrompet, Chennai, India
+91 44 22232711
sobha@au-kbc.org

ABSTRACT
The penetration of smart devices such as mobile phones, tabs has significantly changed the way people communicate. This has led to the growth of usage of social media tools such as twitter, facebook chats for communication. This has led to development of new challenges and perspectives in the language technologies research. Automatic processing of such texts requires us to develop new methodologies. Thus there is great need to develop various automatic systems such as information extraction, retrieval and summarization. Entity recognition is a very important sub task of Information extraction and finds its applications in information retrieval, machine translation and other higher Natural Language Processing (NLP) applications such as co-reference resolution. Some of the main issues in handling of such social media texts are i) Spelling errors ii) Abbreviated new language vocabulary such as "gr8" for great iii) use of symbols such as emoticons/emojis iv) use of meta tags and hash tags v) Code mixing. Entity recognition and extraction has gained increased attention in Indian research community. However there is no benchmark data available where all these systems could be compared on same data for respective languages in this new generation user generated text. Towards this we have organized the Code Mix Entity Extraction in social media text track for Indian languages (CMEE-IL) in the Forum for Information Retrieval Evaluation (FIRE). We present the overview of CMEE-IL 2016 track. This paper describes the corpus created for Hindi-English and Tamil-English. Here we also present overview of the approaches used by the participants.

CCS Concepts
• **Computing methodologies ~ Artificial intelligence**
• **Computing methodologies ~ Natural language processing**
• **Information systems ~ Information extraction**

Keywords
Entity Extraction; Social Media Text; Code Mixing, Twitter; Indian Languages; Tamil; Hindi; English; Named Entity Annotation Corpora for Code Mix Twitter data.

1. INTRODUCTION
Over the past decade, Indian language content on various media types such as websites, blogs, email, chats has increased significantly. And it is observed that with the advent of smart phones more people are using social media such as twitter, facebook to comment on people, products, services, organizations, governments. Thus we see content growth is driven by people from non-metros and small cities who are mostly comfortable in their own mother tongue rather than English. The growth of Indian language content is expected to increase by more than 70% every year. Hence there is a great need to process this huge data automatically. Especially companies are interested to ascertain public view on their products and processes. This requires natural language processing software systems which recognizes the entities or the associations of them or relation between them. Hence an automatic Entity extraction system is required.

The objectives of this evaluation are:

- Creation of benchmark data for Entity Extraction in Indian language Code Mixed Social Media text.
- To develop Named Entity Recognition (NER) systems in Indian language Social Media text.

Entity extraction has been actively researched for over 20 years. Most of the research has, however, been focused on resource rich languages, such as English, French and Spanish. The scope of this work covers the task of named entity recognition in social media text (twitter data) for Indian languages. In the past there were events such as Workshop on NER for South and South East Asian Languages (NER-SSEA, 2008), Workshop on South and South East Asian Natural Language Processing (SANLP, 2010&2011) conducted to bring various research works on NER being done on a single platform. NERIL tracks at FIRE (Forum for Information Retrieval and Evaluation) in 2013, 2014 have contributed to the development of benchmark data and boosted the research towards NER for Indian languages. All these efforts were using texts from newswire data. The user generated texts such as twitter and facebook texts are diverse and noisy. These texts contain non-standard spellings and abbreviations, unreliable punctuation styles. Apart from these writing style and language challenges, another challenge is concept drift (Dredze etal., 2010; Fromreide et al., 2014); the distribution of language and topics on Twitter and Facebook is constantly shifting, thus leading to performance degradation of NLP tools over time.

Some of the main issues in handling of such texts are i) Spelling errors ii) Abbreviated new language vocabulary such as "gr8" for great iii) use of symbols such as emoticons/emojis iv) use of meta tags and hash tags v) Code mixing.
For example:

"Muje kabi bhoolen gy to nhi na? :(Want ur sweet feedback about my FC ? mai dilli jaa rahi hoon".

The research in analyzing the social media data is taken up in English through various shared tasks. Language identification in tweets (tweetLID) shared task held at SEPLN 2014 had the task of identifying the tweets from six different languages. SemEval 2013, 2014 and 2015 held as shared task track where sentiment analysis in tweets were focused. They conducted two sub-tasks

namely, contextual polarity disambiguation and message polarity classification. In Indian languages, Amitav et al (2015) had organized a shared task titled 'Sentiment Analysis in Indian languages' as a part of MIKE 2015, where sentiment analysis in tweets is done for tweets in Hindi, Bengali and Tamil language.

Named Entity recognition was explored in twitter through shared task organized by Microsoft as part of 2015 ACL-IJCNLP, a shared task on noisy user-generated text, where they had two sub-tasks namely, twitter text normalization and named entity recognition for English.

The ESM-IL track at FIRE 2015 was the first one to come up with the entity annotated benchmark data for the social media text, where the data was in idealistic scenario, where users use only one language. But nowadays we observe that users use code mixing even in writing in the social media platforms. Thus there is a need to develop systems that focus on social media texts. There have been other efforts on the code mix social media text in the applications of information retrieval (MSIR tracks at FIRE 2015and 2016).

The paper is organized as follows: section 2 describes the challenges in named entity recognition on Indian languages. Section 3 describes the corpus annotation, the tag set and corpus statistics. And section 4 describes the overview of the approaches used by the participants and section 5 concludes the paper.

2. CHALLENGES IN INDIAN LANGUAGE ENTITY EXTRACTION

The challenges in the development of entity extraction systems for Indian languages from social media text arise due to several factors. One of the main factors being there is no annotated data available for any of the Indian languages, though the earlier initiatives have been concentrated on newswire text. Apart from the lack of annotated data, the other factors which differentiate Indian languages from other European languages are the following:

a) **Ambiguity** – Ambiguity between common and proper nouns. Eg: common words such as "Roja" meaning Rose flower is a name of a person.
b) **Spell variations** – One of the major challenges is that different people spell the same entity differently. For example: In Tamil person name -Roja is spelt as "rosa", "roja".
c) **Less Resources** – Most of the Indian languages are less resource languages. There are no automated tools available to perform preprocessing tasks required for NER such as part-of-speech tagging, chunking which can handle social media text.

Apart from these challenges we also find that development of automatic entity recognition systems is difficult due to following reasons:

i) Tweets contain a huge range of distinct named entity types. Almost all these types (except for People and Locations) are relatively infrequent, so even a large sample of manually annotated tweets will contain very few training examples.

ii) Twitter has a 140 character limit, thus tweets often lack sufficient context to determine an entity's type without the aid of background or world knowledge.

iii) In comparison with English, Indian Languages have more dialectal variations. These dialects are mainly influenced by different regions and communities.

iv) Indian Language tweets are multilingual in nature and predominantly contain English words.

The following examples illustrate the usage of English words and spoken, dialectal forms in the tweets.

Example 1 (Tamil):

Ta: Stamp veliyittu ivaga ativaangi .....

En: stamp released these_people get_beaten ….

Ta: othavaangi .... kadasiya <loc>kovai</loc>

En: get_slapped … at_end kovai

Ta: pooyi pallakaatti kuththu vaangiyaachchu.

En: gone show_tooth punch got

("They released stamp, got slapping and beating … at the end reached Kovai and got punched on the face")

This example is a Tamil tweet where it is written in a particular dialect and also has usage of English words.

Similarly in Hindi we find lot of spell variations. Such as for the words "mumbai", "gaandhi", "sambandh", "thanda" there are atleast three different spelling variations.

3. CORPUS DESCRIPTION

The corpus was collected using the twitter API in two different time periods. The training partition of the corpus was collected during May – June 2015. And the test partition of the corpus was collected during Aug – Sep 2015. As explained in the above sections, in the twitter data we observe concept drift. Thus to evaluate how the systems handle concept drift we had collected data in two different time periods. In this present initiative the corpus is available for three Indian languages Hindi, Malayalam and Tamil. And we have also provided the corpus for English, so that it would help researchers to compare their efforts with respect to English vis-à-vis the respective Indian languages. The following figures show different aspects of corpus statistics.

3.1 ANNOTATION TAGSET

The corpus for each language was annotated manually by trained experts. Named Entity Recognition task requires entities mentioned in the document to be detected, their sense to be disambiguated, select the attributes to be assigned to the entity and represent it with a tag. Defining the tag set is a very important aspect in this work. The tag set chosen should be such that it covers major classes or categories of entities. The tag set defined should be such that it could be used at both coarse and fine grained level depending on the application. Hence a hierarchical tag set will be the suitable one. Though we find that in most of the works Automatic Content Extraction (ACE) NE tag set has been used, in our work we have used a different tag set. The ACE Tag set is fine grained is towards defense/security domain. Here we have used Government of India standardized tag set which is more generic.

The tag set is a hierarchical tag set. This Hierarchical tag set was developed at AU-KBC Research Centre, and standardized by the Ministry of Communications and Information Technology, Govt. of India. This tag set is being used widely in Cross Lingual Information Access (CLIA) and Indian Language – Indian Language Machine Translation (IL-IL MT) consortium projects.

In this tag set, named entity hierarchy is divided into three major classes; Entity Name, Time and Numerical expressions. The Name hierarchy has eleven attributes. Numeral Expression and

time have four and three attributes respectively. Person, organization, Location, Facilities, Cuisines, Locomotives, Artifact, Entertainment, Organisms, Plants and Diseases are the eleven types of Named entities.

Numerical expressions are categorized as Distance, Money, Quantity and Count. Time, Year, Month, Date, Day, Period and Special day are considered as Time expressions. The tag set consists of three level hierarchies. The top level (or 1st level) hierarchy has 22 tags, the second level has 49 tags and third level has 31 tags. Hence a total of 102 tags are available in this schema. But the data provided to the participants consisted of only the 1st level in the hierarchy that is consisting of only 22 tags. The other levels of tagging were hidden. This was done to make it little easier for the participants to develop their systems using machine learning methods.

The data statistics are as follows:

Table 1. Corpus Statistics

Language	No. of Tweets	No. of NEs
Hindi-English	10129	7573
Tamil-English	4576	2454

The NE distribution in both language datasets has been found to be having majority of Person, Location, and Entertainment. This shows that majority of people communication has been on the topics movies and persons.

3.2 DATA FORMAT

The participants were provided the data with annotation markup in a separate file called annotation file. The raw tweets were to be separately downloaded using the twitter API. The annotation file is a column format file, where each column was tab space separated. It consisted of the following columns:

i) Tweet_ID
ii) User_Id
iii) NE_TAG
iv) NE raw string
v) NE Start_Index
vi) NE_Length

For example:

Tweet_ID:123456789012345678
User_Id:1234567890
NE_TAG:ORGANIZATION
NE Raw String:SonyTV
Index:43
Length:6

Index column is the starting character position of the NE calculated for each tweet and the count starts from '0'. The participants were also instructed to provide the test file annotations in the same format as given for the training data.

4. SUBMISSION OVERVIEWS

In this evaluation exercise we have used Precision, Recall and F-measure, which are widely used for this task. A total of 21 teams had registered for participation in this track. Later 9 teams were able to submit their systems for evaluation. A total of 25 test runs were submitted for evaluation. All the teams had participated for Hindi-English language pair and 5 teams participated for Tamil-English language pair. We had developed a base system without any pre-processing of the data and use of any lexical resources. We had developed this base system by just using the raw data as such without any other features. We used Conditional Random Fields (CRFs) for developing the base system. This base line system was developed so that it would help in making a better comparative study. And it was observed that all the teams had outperformed the base line system. In the following paragraphs we would be briefly explaining the approaches used by each team. All the teams' results are given in Table 3 and 4.

Irshad team had used Neural Networks, to develop their system. They had used external resource of Wiki data for creating word embedding. They had not done any cleaning work such as removal of URLs, emoticons from tweets. And NLP pre-processing of the text was done. This team had participated only in Hindi- English and submitted 1run.

Deepak team had used CRFs. Here they have preprocessed the data for tokenization. They had also used gazetteer lists for disease names. And this team had submitted results for both Hindi-English and Tamil-English.

Veena team had used machine learning method SVM. They have used word2vec for feature engineering and extraction. Here they have used other external corpus from MSIR 2016 and ICON 2015 track data sets. They had submitted 3 run each for both Hindi-English and Tamil-English. This team had also used stylometric features, suffixes and prefixes, gazetteers in run 3. Here it is interesting to note that though many kinds of features and resources, the system performance was not significantly higher than other runs where all of these features were not used.

Barathi team, have submitted 2 runs each for Hindi-English and Tamil-English. They have used CRFs and Random Forest Tree. Their run 1 was based upon lexical features and CRF algorithm. Along with the Run 1 features an additional binary feature (entity or not) decided by the Random Forest Tree is added in Run 2.

Rupal team had decision trees and extremely randomized tree algorithms. The precision obtained is comparatively lower than other new ML methods used by earlier teams. They had cleaned the data for emojis, urls as the first step of processing.

The team lead by Somnath had used CRFs and used the popular CRF++ tool. The system performance was relatively lower. Probably this could be attributed to lack of proper feature extraction and feature engineering.

One interesting observation is that the team led by Nikhil had also used neural networks similar to another team, but have not used any external resource for training. This shows that the data size needs to be improved for better machine learning.

The team lead by Srinidhi, had used SVM with context based character embedding as feature engineering. This team had used several external unlabeled datasets such as MSIR 2016, ICON 2015 shared task datasets.

The different methodologies used by different teams have been summarized in Table 2.

Evaluation metrics used are precision, recall and f-measure. All the systems have been evaluated automatically by comparing the

gold annotations. The results obtained by participant systems have been shown in table 3 and 4.

5. CONCLUSION

The main objective of creating benchmark data representing some of the popular Indian languages has been achieved. And this data has been made available to research community for free for research purposes. The data is user generated data and is not any genre specific. Efforts are still going on to standardize this data and make it perfect data set for future researchers. We observe that the results obtained for Hindi-English data has been more than Tamil-English. This is due to data being noisier and size is less compared to Hindi-English. We hope to see more publications in this area in the coming days from these different research groups who could not submit their results. Also we expect more groups would start using this data for their research work.

This CMEE-IL track is one of the first efforts towards creation of entity annotated user generated code mixed social media text for Indian languages. In this CMEE-IL annotation tag set we have made use of a hierarchical tag set. Thus this annotated data could be used for any kind of applications. This tag set is very exhaustive and has finer tags. The applications which require fine grain tags could use the data with full annotation. And for applications which do not require fine grain, the finer tags could be suppressed in the data. The data being generic, this could be used for developing generic systems upon which a domain specific system could be built after customization.

6. ACKNOWLEDGMENTS

We thank the FIRE 2016 organizers for giving us the opportunity to conduct the evaluation exercise.

7. REFERENCES

[1] Arkaitz Zubiaga, Iñaki San Vicente, Pablo Gamallo, José Ramom Pichel Campos, Iñaki Alegría Loinaz, Nora Aranberri, Aitzol Ezeiza, Víctor Fresno. 2014 TweetLID@SEPLN 2014, Girona, Spain, September 16th, 2014. CEUR Workshop Proceedings 1228, CEUR-WS.org 2014

[2] Mark Dredze, Tim Oates, and Christine Piatko. 2010. "We're not in kansas anymore: detecting domainchanges in streams". In *Proceedings of the 2010 Conferenceon Empirical Methods in Natural LanguageProcessing*, pages 585–595. Association for ComputationalLinguistics.

[3] Hege Fromreide, Dirk Hovy, and Anders Søgaard.2014. "Crowdsourcing and annotating ner for twitter#drift". *European language resources distributionagency.*

[4] H.T. Ng, C.Y., Lim, S.K., Foo. 1999. "A Case Study on Inter-Annotator Agreement for Word Sense Disambiguation". In *Proceedings of the {ACL} {SIGLEX} Workshop on Standardizing Lexical Resources {(SIGLEX99)}*. Maryland. pp. 9-13.

[5] Preslav Nakov and Torsten Zesch and Daniel Cer and David Jurgens. 2015. *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015).*

[6] Nakov, Preslav and Rosenthal, Sara and Kozareva, Zornitsa and Stoyanov, Veselin and Ritter, Alan and Wilson, Theresa. 2013. SemEval-2013 Task 2: Sentiment Analysis in Twitter. *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*

[7] Rajeev Sangal and M. G. Abbas Malik. 2011. *Proceedings of the 1st Workshop on South and Southeast Asian Natural Language Processing (SANLP)*

[8] Aravind K. Joshi and M. G. Abbas Malik. 2010. *Proceedings of the 1st Workshop on South and Southeast Asian Natural Language Processing* (SANLP). (http://www.aclweb.org/anthology/W10-36)

[9] Rajeev Sangal, Dipti Misra Sharma and Anil Kumar Singh. 2008. *Proceedings of the IJCNLP-08 Workshop on Named Entity Recognition for South and South East Asian Languages*. (http://www.aclweb.org/anthology/I/I08/I08-03)

[10] Pattabhi RK Rao, CS Malarkodi, Vijay Sundar R and Sobha Lalitha Devi. 2014. Proceedings of Named-Entity Recognition Indian Languages track at FIRE 2014. http://au-kbc.org/nlp/NER-FIRE2014/

Table 2. Participant Team Overview - Summary

Team	Languages & System Submissions	Approaches (ML method) Used	Pre-Processing Step	Lexical Resources Used	Open Source NLP Tools Used	Variation Between Runs
Barathi –AmrithaT2	i)Hindi – English: 2 ii) Tamil– English: 2	Run1: Conditional Random Field Run2: Conditional Random Field + Random Forest Tree	Run1: Tweet Preprocessor alone used to eliminate http links, emoticons Run2: Tweet Preprocessor alone used to eliminate http links, emoticons		Run1:Tweet Preprocessor, Scikit – Learn, sklearn – crfsuite, nltk Run2: Tweet Preprocessor, Scikit – Learn, sklearn – crfsuite, nltk	Along with the run 1 features binary feature (outcome of random forest tree) utilized in run 2
Deepak-IITPatna	i)Hindi – English: 1 ii)Tamil– English: 1	Machine learning(CRFs)+Rule based system	Tokenization by CMU tagger + Token Encoding (IOB)	1, Dictionary of Disease name, Living Things & Special days	CMU ark tagger, CRF++	
(Irshad-IIIT-Hyd)	i) Hindi – English: 1	Simple Feed Forward Neural Network with 1 hidden layer of 200 nodes, Activation function - Rectifier, Learning rate - 0.03, Dropout - 0.5, Learning rule - adagrad, Regularization L2, Mini-batch - 200, Trained for 25 iterations.	Converted the given data to BIO format	2, English wiki corpus to develop word-embeddings using Gensim Word2Vec	Gensim Word2Vec	
(Nikhil_BITSHyd) Nikhil Bharadwaj Gosala BITS Pilani, Hyderabad Campus	i) Hindi – English: 2	Run1: seq2seq LSTM network was used with 3 layers and 192 nodes in each layer Run2: seq2seq LSTM network was used with 4 layers and 256 nodes in each layer	1, Replacement of HTML Escape Characters 2, Tokenize Tweets 3, Stop Word Removal 4, Rule Tagging 5, Mapping Common Misspellings	NLTK Stop Words	NLTK Word Tokenizer and NLTK Stop Words	Run 1: 3 hidden layers with each hidden layer having 192 nodes. Run 2: 4 hidden layers with each hidden layer having 256 nodes.
(Rupal_BITSPiliani) Rupal Bhargava	**i) Hindi – English:** 3 **ii) Tamil – English:** 3	**Run1:** Decision Tree (Hindi-English) Decision Tree (Tamil-English) **Run2:** Extremely Randomized Tree (Hindi-English)	Convert to lowercase, remove links and tokenize	Pyenchant (a Python English Dictionary; Gazetteer Lists were created from the annotations file.		Differences are in the machine-learning technique used.

		Decision Tree (Tamil-English) **Run3:** Extremely Randomized Tree (Hindi-English) Extremely Randomized Tree (Tamil-English)				
(ShivkaranAMU3) Srinidhi Skanda V CEN@Amrita	i) Hindi – English: 1 ii) Tamil – English: 1	Context Based Character Embedding	1, Tokenizing data into each token per-line 2, Special tag is added to identify end of each tweet 3, Converting input datasets to IOB format	Hindi-English: unlabeled datasets from Mixed Script Information Retrieval 2016 (MSIR) and International Conference on Natural Language Processing (ICON) 2015 POS Tagging task, external twitter data collected using web scrapping. Tamil-English: systems unlabeled datasets from Sentiment Analysis in Indian Languages (SAIL-2015)	1, Word2vec Model 2, SVM-Light	
(SomnathJU) Somnath Banerjee Jadavpur University	i) Hindi – English: 1	Conditional Random Fields	Clean links and emoticons		CRF++	
(VeenaAMU1) Anand Kumar M Amrita Vishwa Vidyapeetham	i) Hindi – English: 3 ii) Tamil – English: 3	Run1: Wang2vec based embedding features Run2:Word2vec based embedding features Run3:Stylometric features	Tokenization, BIO formatting	MSIR 2016 & ICON 2015, SAIL 2015, Twitter dataset	wang2vec, word2vec, SVM-Light	Run 1 –Structured Skip-gram based embedding features. Structured skip gram model takes the word position into consideration and extracts the features. Run 2 – neural network based word2vec embedding features. Run 3 –stylometric features - prefix, suffix, punctuation, hash tags, gazetted features, index, length, etc.

Table 3. Evaluation Results for Hindi-English

Team	Run1			Run2			Run3			Best Run		
	Precision	Recall	F-Measure	Precision	Recall	F-Measure	Precision	Recall	F-Measure	Precision	Recall	F-Measure
Irshad-IIIT-Hyd	80.92	59	68.24			NA			NA	**80.92**	**59.00**	**68.24**
Deepak-IIT-Patna	81.15	50.39	62.17			NA			NA	81.15	50.39	62.17
Veena-Amritha-T1	75.19	29.46	42.33	75	29.17	42.00	79.88	41.37	54.51	79.88	41.37	54.51
Barathi-Amritha-T2	76.34	31.15	44.25	77.72	31.84	45.17			NA	77.72	31.84	45.17
Rupal-BITS-Pilani	58.66	32.93	42.18	58.84	35.32	44.14	59.15	34.62	43.68	58.84	35.32	44.14
Somnath-JU	37.49	40.28	38.83			NA			NA	37.49	40.28	38.83
Nikhil-BITS-Hyd	59.28	19.64	29.50	61.8	26.39	36.99			NA	61.80	26.39	36.99
Shivkaran-Amritha-T3	48.17	24.9	32.83			NA			NA	48.17	24.90	32.83
AnujSaini	72.24	18.85	29.90			NA			NA	72.24	18.85	29.90

Table 4. Evaluation Results for Tamil-English

Team	Run1			Run2			Run3			Best Run		
	Precision	Recall	F-measure	Precision	Recall	F-measure	Precision	Recall	F-measure	Precision	Recall	F-Measure
Deepak-IIT-Patna	79.92	30.47	44.12			NA			NA	**79.92**	**30.47**	**44.12**
Veena-Amritha-T1	77.38	8.72	15.67	74.74	9.93	17.53	79.51	21.88	34.32	79.51	21.88	34.32
Barathi-Amritha-T2	77.7	15.43	25.75	79.56	19.59	31.44			NA	79.56	19.59	31.44
Rupal-BITS-Pilani-R2	55.86	10.87	18.20	58.71	12.21	20.22	58.94	11.94	19.86	58.71	12.21	20.22
Shivkaran-Amritha-T3	47.62	13.42	20.94			NA			NA	47.62	13.42	20.94