ESM-IL: Entity Extraction from Social Media Text for Indian Languages @ FIRE 2015 – An Overview

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

Malarkodi CS
AU-KBC Research Centre
MIT Campus of Anna University, Chrompet, Chennai, India
+91 44 22232711
csmalarkodi@au-kbc.org

Vijay Sundar Ram R
AU-KBC Research Centre
MIT Campus of Anna University, Chrompet, Chennai, India
+91 44 22232711
sundar@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
Entity recognition is a very important sub task of Information extraction and find its applications in information retrieval, machine translation and other higher Natural Language Processing (NLP) applications such as co-reference resolution. Entities are real world elements or objects such as Person names, Organization names, Product names, Location names. Entities are often referred to as Named Entities. Entity extraction refers to automatic identification of named entities in a text document. Given a text document, entities such as Person names, Organization names, Location names, Product names are identified and tagged. We observe that in the Indian language scenario there is no social media text corpus which could be used to develop automatic systems. 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. Towards this we have organized the Entity extraction in social media text track for Indian languages (ESM-IL) in the Forum for Information Retrieval Evaluation (FIRE). We present the overview of ESM-IL 2015 track. This paper describes the corpus created for Hindi, Malayalam, Tamil and 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; Twitter; Indian Languages; Tamil; Hindi; Malayalam; English; Named Entity Annotated Corpora for Twitter.

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 Social Media text.
- To develop Named Entity Recognition (NER) systems in Indian language Social Media text.
- To identify the best suiting machine learning techniques.

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 and 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. Thus there is a need to develop systems that focus on social media texts.

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. In the NER sub-task they have used ten tags for annotating the text. 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) **Morphologically rich** – Indian languages are morphologically rich and agglutinative, hence the root word identification is difficult and requires morphological analyzers.
b) **Ambiguity** – Ambiguity between common and proper nouns. Eg: common words such as "Roja" meaning Rose flower is a name of a person.
c) **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".
d) **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.

Example 2 (Malayalam):

ML: ediye ... ente utuppu teechcho? illa

En: hey … my dress ironed? no

ML:chetta ... raavile_tanne engottaa?

En: brother … morning_itself where?

ML: tekkati .... teechchaale parayullo?

En: hey_iron_it … only_after_ironing tell?

(Hey did you iron my dress? No… brother morning itself where are you going? Hey iron it … only after ironing you will tell?)

This is a Malayalam tweet written in spoken form, where the phrase "teekku ati" has been written as "tekkati", spoken form. This makes it resemble a place name and creates ambiguity. This makes understanding difficult.

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.

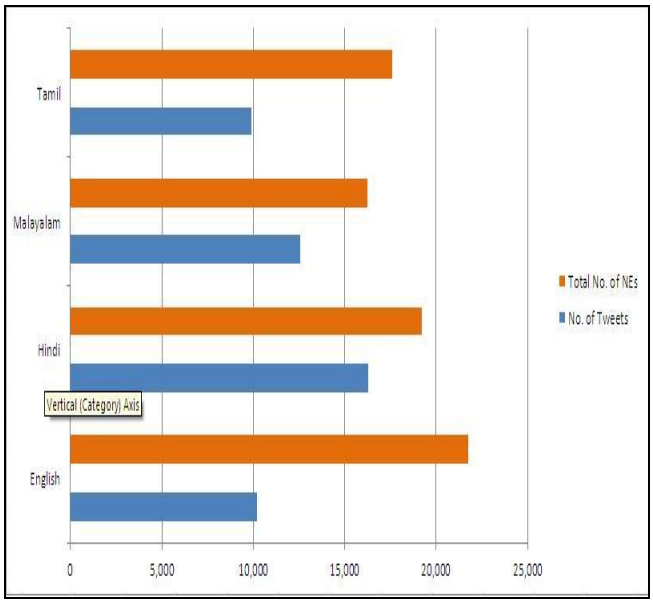

Figure 1. Corpus Statistics – No.of Tweets and Entities in each language

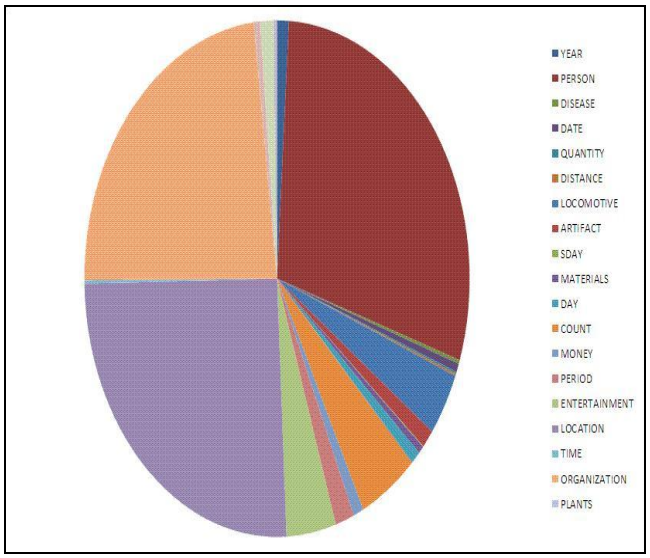

Figure 2. Entity distribution - for English

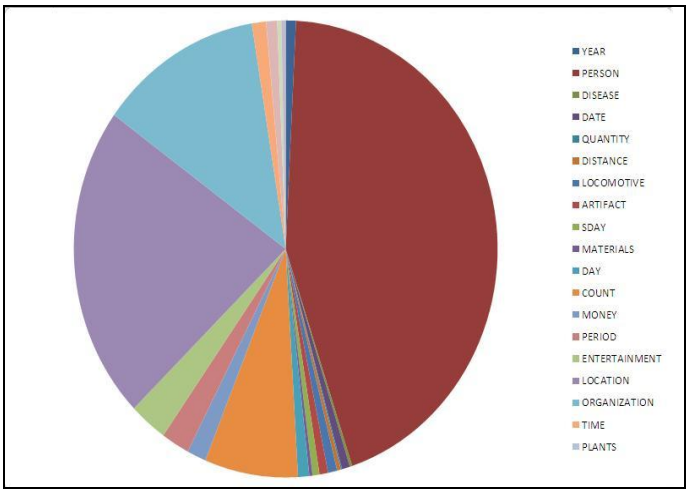

Figure 3. Entity Distribution – for Hindi

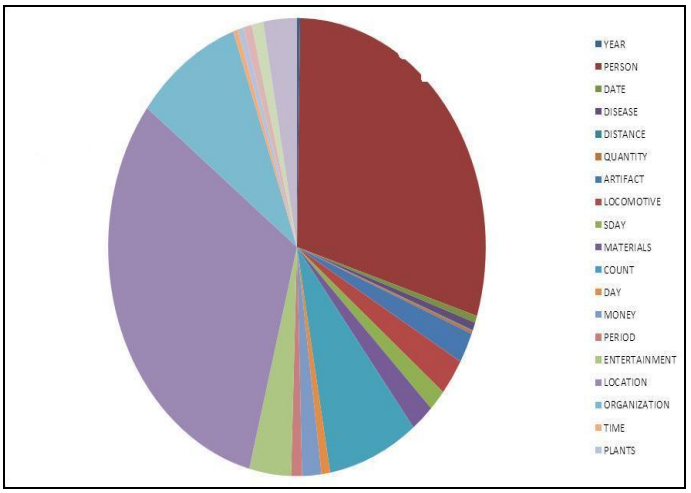

Figure 4. Entity Distribution – for Malayalam

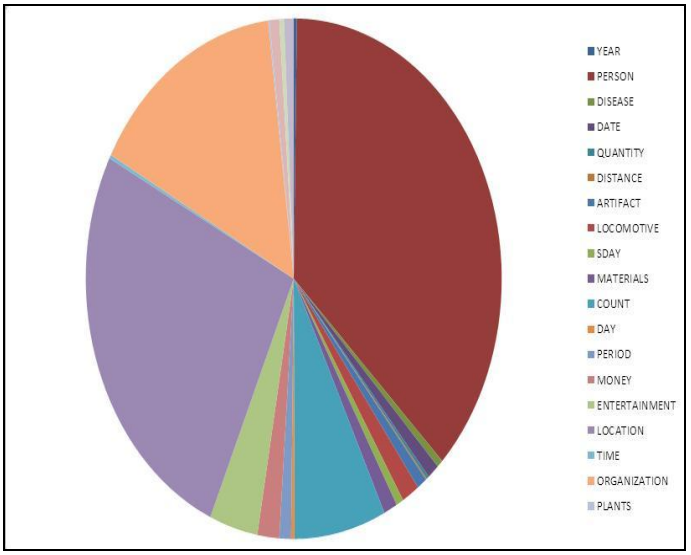

Figure 5. Entity Distribution – for Tamil

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.

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. The figures below show various aspects of corpus statistics.

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 10 teams had registered for participation in this track. Later 7 teams were able to submit their systems for evaluation. A total of 17 test runs were submitted for evaluation. All the teams had participated for English and Hindi languages, except for one team which had only participated in English language. And three teams had participated in Tamil, and two teams had participated in Malayalam. 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 CRFs for developing the base system. This base system was developed so that it would help in making a better comparative study. In the following paragraphs we would be briefly explaining the approaches used by each team. All the teams results along with the bas system results are given in Table 2.

Pallavi team, had used CRFs, a machine learning technique to develop their system. They had used features such as POS, Chunk, Statistical Suffixes and prefixes (unigram, bigram and trigrams). They had first cleaned the provided training data to remove URLs and emoticons from tweets and pre-processed the text for POS and chunks. For the preprocessing purpose they have used open source NLP tools, "patter.en" for English and for Hindi nltr. This team had participated in three languages Hindi, Tamil and English. They had submitted 3 runs for Hindi and 2 runs each for English and Tamil.

Sarkar team, had used HMM for the development. Here they have preprocessed the data for POS and used POS tag as one of the states for HMM training. They had also used gazetteer lists. These lists were collected using semi-manual efforts. And this team had only submitted results for English only.

Shriya team had used machine learning method SVM. They have used open source preprocessing tools for POS tagging and Chunking. For Tamil and Malayalam they had developed in house POS tagger and chunker by manually annotating small data of the training corpus. The have also used an external resource brown clusters as one of the features in training SVM. Other main features used in training are 3-word window, POS tags, heuristic features such as capitalization, statistical suffixes and prefixes up to 3 characters. This is one of the teams that has participated in all four languages.

Sanjay team has used CRFs for their system development. This is another team which has participated in all four languages. This team also preprocessed the data for POS and chunking. For English and Hindi they have open source tools for this purpose and whereas for Tamil and Malayalam in house they have developed these pre-processing tools.

Chintak team had used CRFs and had pre-processed data for POS tagging and chunking. For this purpose they have used open source tools Genia tagger, which is tuned towards biological domain. And we feel this could have resulted in very lower recall values. They had also used features such as POS, Chunk, heuristic features. They had also used gazetteers as one of the features in their machine learning.

The team led by Vira had also used CRFs. They had used Stanford preprocessing tools. They have used window of 5 words in the features for training along with POS tag, statistical suffixes and prefixes.

The team lead by Sombuddha had four different ML methods and submitted four different runs for English. They had also submitted runs for Hindi, but since the test submission did not conform to the format specified as per the task guidelines, it was disqualified. The features used are POS tag, window of words, heuristic features such as Capitalisation, presence of numeric, hash tags. They had also used dictionary as binary feature.

The different methodologies used by different teams have been summarizeed in Table 1.

We observe that some of the participant systems have not performed well in comparison with the base system though several features were used for training. And most of the systems have almost the same precision scores as obtained in the base system. There is significant improvement in the recall of the systems with respect to base system. A deeper analysis of the results obtained by the participant systems has to be done.

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 response from the participants for Hindi language has been more than other languages. 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 ESM-IL track is one of the first elaborate efforts towards creation of entity annotated user generated social media text for Indian languages. In this ESM-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 2015 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 1. Participant Team Overview - Summary

Team	Languages &System Submissions	Approaches (ML method) Used	Features Used	Resources/Tools used
Pallavi et al., Hindustan Institute of Technology and Science, Chennai	English – 2 runs, Hindi – 3 runs, Tamil – 2 runs	CRFs – CRF++ tool kit	POS, Chunk, Statistical suffixes and prefixes	Cleaned data to remove URLs, emoticons For English preprocessing open source tool 'pattern.en' For Hindi used open source tool nltr.org
K Sarkar Jadavpur University, Kolkata	English – 1 run	HMM	POS tag	POS Tagger – Monty Tagger – open source tool Gazetteer List
Shriya et al., Amritha Vishwa Vidyapeetam, Coimbatore	English – 1 run, Hindi – 1 run, Malayalam – 1 run Tamil – 3 runs	SVM - Machine Learning	POS, Chunk, Statistical Suffixes, Statistical prefixes, Heuristics such as capitalization information, Gazetteer list, Shape feature	Gazetteer list, For preprocessing used NLTK, Gimpel POS tagger for English. NLTK for Hindi Developed in house tools for POS and Chunking for Tamil and Malayalam Brown Cluster for English
Sanjay et al., Amritha Vishwa Vidyapeetam, Coimbatore	English – 2 Runs, Hindi – 1 run, Malayalam – 1 run Tamil – 2 Runs	CRFs – CRF++ tool was used	POS, Chunk, Statistical Suffixes, Statistical prefixes, Heuristics such as capitalization information, Gazetteer list, Shape feature	Gazetteer list, For preprocessing used NLTK, Gimpel POS tagger for English. NLTK for Hindi Developed in house tools for POS and Chunking for Tamil and Malayalam Brown Cluster for English
Chintak et al., LDRP Institute, Gujarat	English – 2 runs Hindi – 2 runs	CRFs – CRFSuite tool was used	POS, Chunk, Gazetteer information, heuristics	Gazetteer list, POS Tagger – RDR open source tool, Chunker – Genia tagger
Vira et al, Charotar University of Science and Technology, Gujarat	English – 1 run, Hindi – 1 run	CRFs – CRFSuite tool was used	Word structures, statistical suffixes and prefixes, heuristic features using postpositions	English – Stanford NLp tool Hindi – RDR tool
Sombuddha et al., Jadavpur University	English – 4 runs	CRFs, Naïve Bayes, MIRA, Decision tree- J-48	POS, Window of Words, heuristics features	POS tagger open source – ark-tweet-nlp tool

Table 2. Evaluation Results

Language		**Hindi**			**Tamil**			**Malayalam**			**English**		
Teams		P	R	F	P	R	F	P	R	F	P	R	F
Base System		**73.05**	**34.81**	**47.10**	**56.50**	**11.46**	**19.05**	**62.58**	**20.82**	**31.24**	**73.54**	**28.01**	**40.56**
Shriya - Amritha	Run1	71.56	54.09	61.61	55.23	11.03	18.39	51.18	40.29	45.08	58.78	40.73	48.11
	Run2	-	-	-	61.55	19.82	29.98	-	-	-	-	-	-
	Run3	-	-	-	60.82	19.42	29.44	-	-	-	-	-	-
Sanjay - Amritha	Run1	74.65	5.26	9.83	70.11	19.81	30.89	60.05	39.94	47.97	46.78	24.90	32.50
	Run2	-	-	-	54.87	18.91	28.13	-	-	-	46.88	25.64	33.15
Chintak - LDRP	Run1	67.11	0.76	1.51	-	-	-	-	-	-	7.30	4.17	5.31
	Run2	74.73	46.84	57.59	-	-	-	-	-	-	5.35	5.67	5.50
KSarkar – JU	Run1	-	-	-	-	-	-	-	-	-	61.96	39.46	48.21
Vira - Charotar Univ	Run1	25.65	16.14	19.82	-	-	-	-	-	-	4.13	3.39	3.72
Pallavi - HITS	Run1	81.21	44.57	57.55	70.42	14.13	23.54	-	-	-	50.48	32.03	39.19
	Run2	80.86	44.25	57.20	64.52	22.14	32.97	-	-	-	50.21	37.06	42.64
	Run3	81.49	41.58	55.06	-	-	-	-	-	-	-	-	-
Sombuddha – JU **	Run1	-	-	-	-	-	-	-	-	-	46.92	32.41	38.34
	Run2	-	-	-	-	-	-	-	-	-	58.09	31.85	41.15
	Run3	-	-	-	-	-	-	-	-	-	49.10	31.59	38.45
	Run4	-	-	-	-	-	-	-	-	-	58.09	31.85	41.15

**** Though this team had submitted Hindi runs, these were disqualified due to data format not confirming with the task guidelines.**