NERIL: Named Entity Recognition for Indian Languages @ FIRE 2013

Pattabhi, R.K Rao

&

Sobha, Lalitha Devi

Computational Linguistics Research Group
AU-KBC Research Centre

Chennai, India

Objectives

- Creation of benchmark data for Evaluation of Named Entity Recognition for Indian Languages
- Encourage researchers to develop Named Entity Recognition (NER) systems for Indian languages.

Motivation

- Over the past decade Indian language content on various media types such as websites, blogs, email, chats has increased significantly.
- Content growth is driven by people from non-metros and small cities.
- 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 identify entities
  - Identification of associations or relation between entities
  - Hence an automatic Named Entity recognizer is required

Motivation

- There is lot of research work going on in NER for Indian languages, such as Workshops NER-SSEA-2008, SANLP 2010, 2011 but,
  - There is lack of bench mark data to compare several existing systems
  - Lack of benchmark data is not encouraging new researchers to work in this area
  - All have been isolated efforts
- Need to create common platform for researchers to interact and share the work
- Need to create bench mark data and evaluation framework

NER Task

- Refers to automatic identification of named entities in a given text document.
- Given a text document, named entities such as Person names, Organization names, Location names, Product names are identified and tagged.
- Identification of named entities is important in several higher language technology systems such as information extraction systems, machine translation systems, and cross-lingual information access systems.
- Several approaches used such as machine learning approach, rule based approach. Techniques such as HMM, CRFs, SVM

Challenges in Indian Language NER

- Indian languages belong to several language families, the major ones being the Indo-European languages, Indo-Aryan and the Dravidian languages.
- The challenges in NER arise due to several factors. Some of the main factors are listed below
  - **Morphologically rich** –
    - Most of the Indian languages are morphologically rich and agglutinative
    - There will be lot of variations in word forms which make machine learning difficult.
  - **No Capitalization feature** –
    - In English, capitalization is one of the main features, whereas that's not there in Indian languages
    - Machine learning algorithms have to identify different features.
  - **Ambiguity** –
    - Ambiguity between common and proper nouns.
    - Eg: common words such as "Roja" meaning Rose flower is a name of a person

Challenges in Indian Language NER

- **Spell variations**
  - One of the major challenges in the web data is that we find different people spell the same entity with differently.
- **Less Resources**
  - Most of the Indian languages are less resource languages.
  - Either there are no automated tools available to perform preprocessing tasks required for NER such as Part-of-speech tagging, chunking.
  - Or for languages where such tools are available they have less performance.
- **Lack of easy availability of annotated data**
  - there are isolated efforts in the development of NER systems for Indian languages,
  - there is no easy availability and access for NE annotated corpus in the community

NERIL – Track @FIRE 2013 -- Description

- **Corpus Collection**
  - Development of benchmark corpus was the main activity
  - Data release for 5 languages
    - Bengali, Hindi, Malayalam, Tamil and English
  - Raw data collected from online sources mainly from Wikipedia and other sources such as blogs, online discussion forums

NERIL – Track @FIRE 2013 -- Description

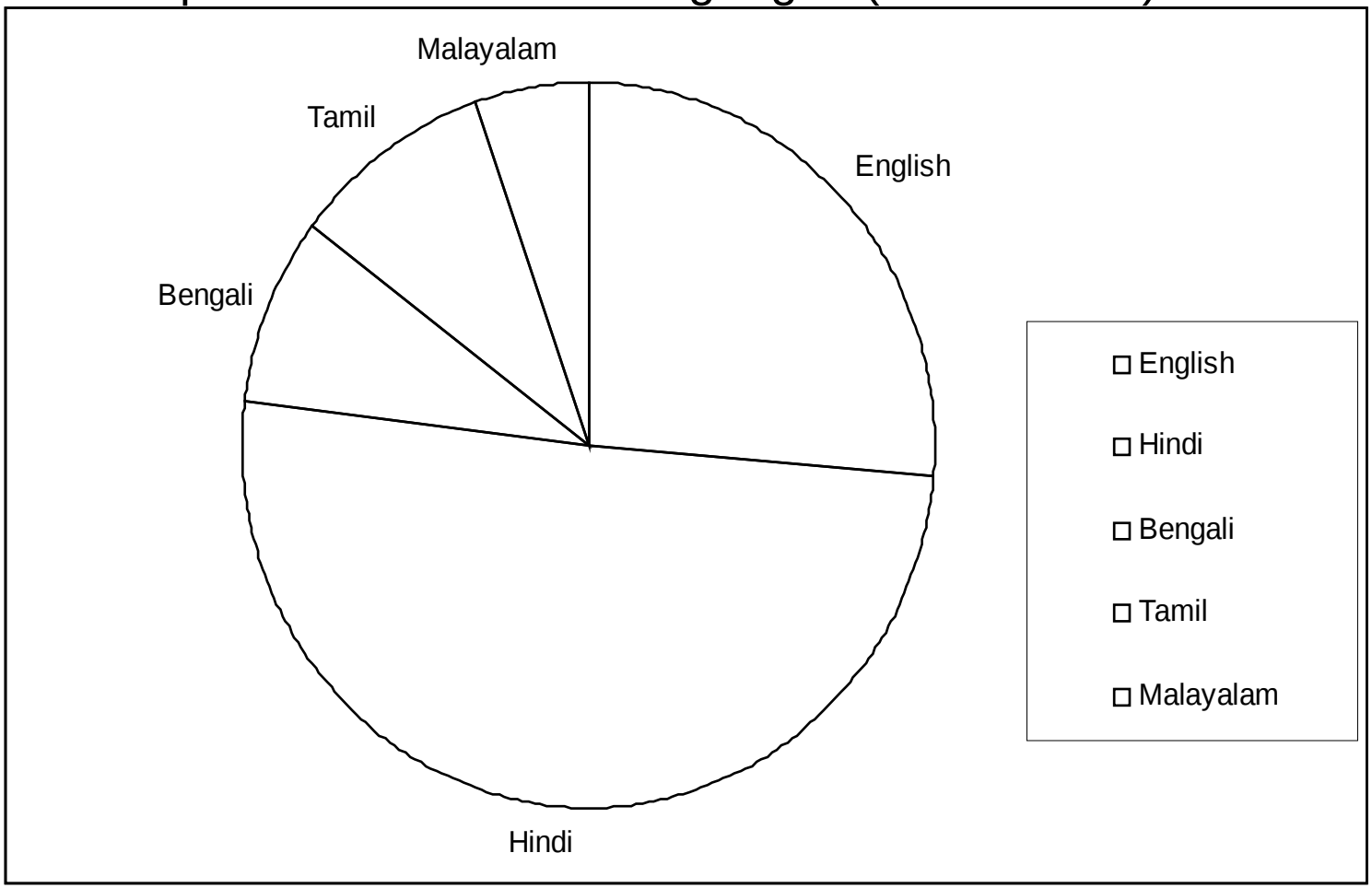
- **Corpus Collection**
  - The raw corpus was cleaned and preprocessed for Part-of-Speech (POS) and chunk information using NLP tools.
  - For the Bengali data preprocessing was not done
  - The corpus was divided into two partitions, training and testing.
  - For the purpose of aiding annotators in the annotation, a graphical user interface (GUI) tool was provided to them.

NERIL – Track @FIRE 2013 -- Description

- **Tagset**
  - Hierarchical tagset developed by AU-KBC Research Centre, and standardized by the MCIT, Govt. of India
  - This tagset is being used in CLIA and IL-ILMT consortium projects
  - The 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.

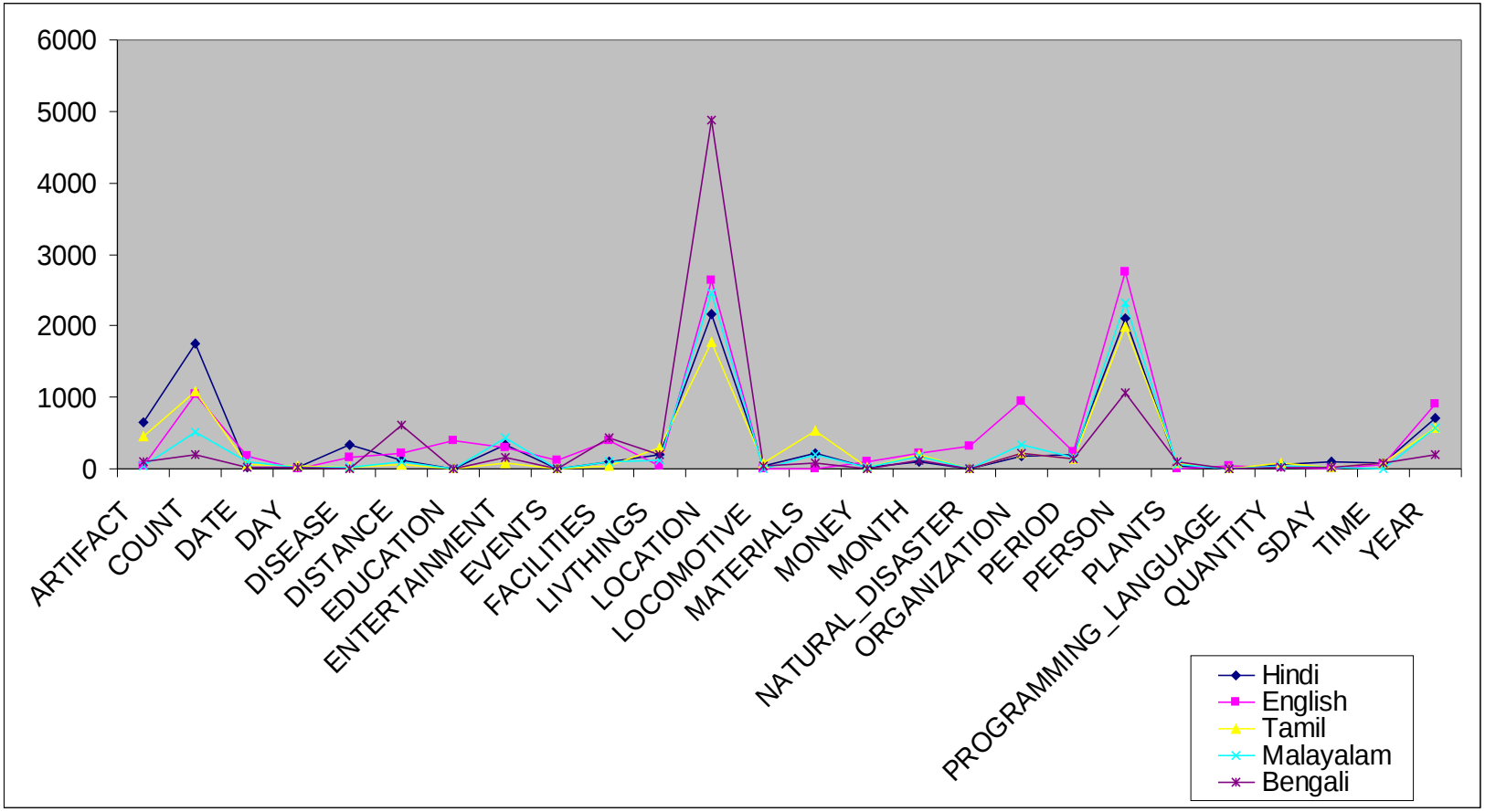
NERIL – Track @FIRE 2013 -- Description

Corpus Size – Different languages (No.of words)



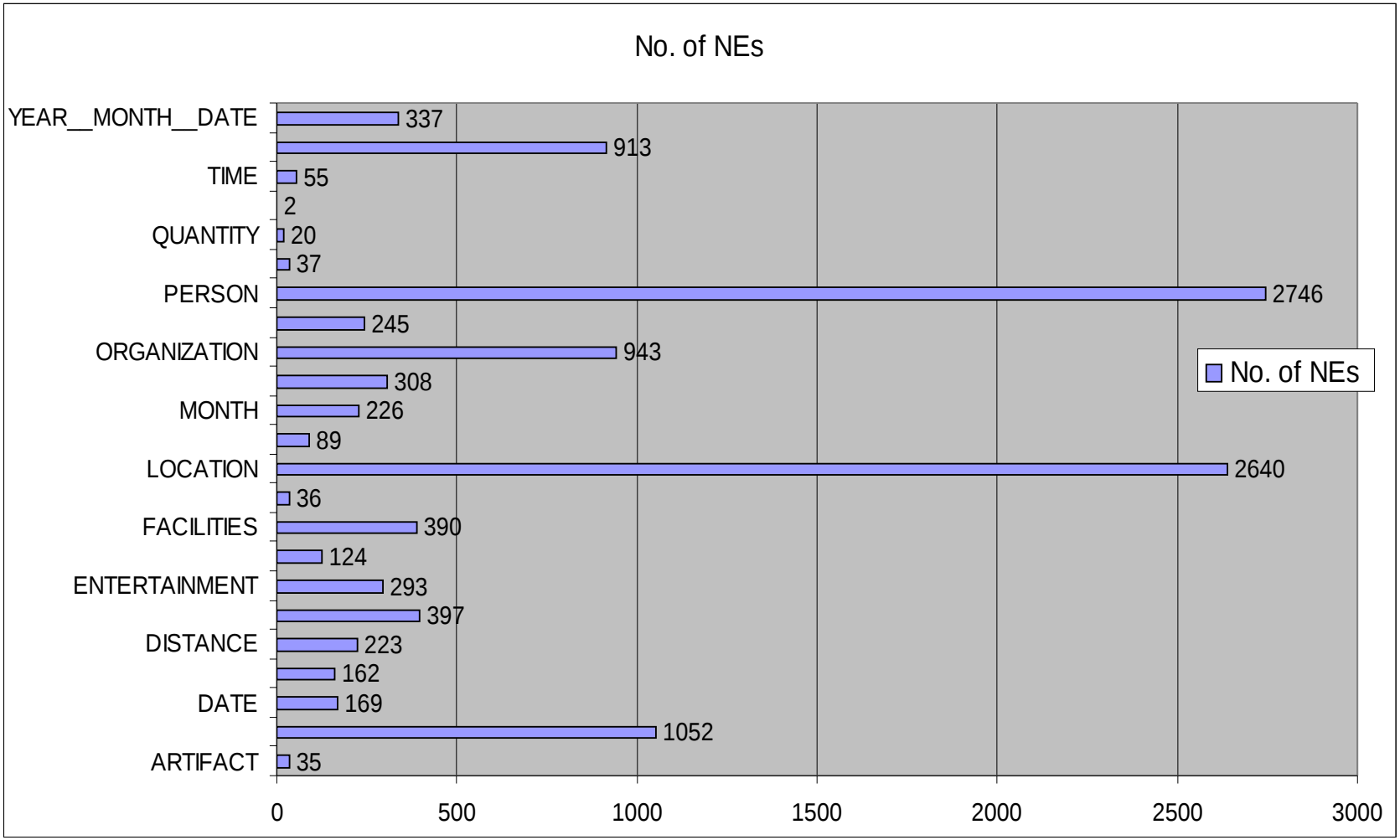
NERIL – Track @FIRE 2013 -- Description

NE distribution in the corpus – For All Languages



NERIL – Track @FIRE 2013 -- Description

NE distribution in the corpus – For English



Evaluation - Submission Overviews

- The evaluation metrics used is Precision, Recall and F-measure.
- Data for five 5 languages were released
- participation was for three languages viz., Bengali, Hindi and English.
- Five teams participated by submitting 9 systems,
  - 4 for English, 3 for Hindi and 2 for Bengali.

Evaluation - Submission Overviews

Team	Languages & Sys Submissions	Approaches Used	Resources/Features Used
TRDCC	English, Hindi – 2 Eng & 1 Hindi submissions	CRFs - Machine Learning	WordNet, Suffix information, Gazetteer
ISM, Dhanbad	English – 2 submissions	List Based search	Gazetteer list
ISI, Kolkata	Bengali – 2 submissions	Rule Based and CRFs -Machine Learning	POS, Chunk, associated verb, token id and gazetteer information
IITB	Hindi – 1 submission	CRFs -Machine Learning	Bigram and trigrams of words, Bigram and unigrams of POS and chunk information, four character suffixes of the words.
MNIT	Hindi – 1 submission	List Based search	Gazetteer list

Evaluation - Submission Results

Language	System	Precision (%)	Recall (%)	F-Measure (%)
Bengali	**ISI Kolkata Sys 1**	**23.69**	**28.02**	**25.68**
	ISI Kolkata Sys 2	28.61	16.09	20.59
English	TRDDC Sys 1	64.79	67.23	65.99
	TRDCC Sys 2	**64.92**	**68.63**	**66.73**
	ISM Sys 1	14.89	32.02	20.33
	ISM Sys 2	39.33	34.46	36.74
Hindi	TRDCC	47.51	68.35	56.06
	IITB	**83.68**	**74.14**	**78.62**
	MNIT	01.72	04.82	02.53

Conclusion

- Benchmark data for 5 languages created
- Available to research community
- Data is generic
- 8 teams registered
- 5 could complete their system development in the available time of 45 days
- Future plan to hold the track with new languages and new type of data

Acknowledgments

We thank Prof. Sudeshna Sarkar and her team from Indian Institute of Technology, Kharagpur (IIT-Kgp) for providing us with the corpus annotation for the Bengali corpus.

Thank You