Entity Extraction from Social Media Text Indian Languages (ESM-IL)

Chintak Mandalia
LDRP Institute of Technology & Research Center,
Gandhinagar, Gujarat, India
chintak.soni75@gmail.com

Memon Mohammed Rahil
LDRP Institute of Technology & Research Center,
Gandhinagar, Gujarat, India
rmemon122@gmail.com

Manthan Raval
LDRP Institute of Technology & Research Center,
Gandhinagar, Gujarat, India
manthanraval249@gmail.com

Sandip Modha
LDRP Institute of Technology & Research Center,
Gandhinagar, Gujarat, India
sjmodha@gmail.com

ABSTRACT

This paper shows the implementation of named entity recognition (NER) which is one of the applications of Natural Language Processing and is regarded as the subtask of information retrieval. NER is the process to detect Named Entities (NEs) in a document and to categorize them into certain Named entity classes such as the name of organization, person, location, sport, river, city, country, quantity etc. There are lots of work have been accomplished in English related to NER. But, at present, still we have not been able to achieve much of the success pertaining to NER in the Indian languages. The following paper discusses about NER, the various approaches of NER, Performance Metrics, the challenges in NER in the Indian languages and finally some of the results that have been achieved by performing NER in Hindi by aggregating approaches such as Rule based CRF suite and for tagging RDRpostagger and geniatagger. The paper shows working methodology and its result on named entity extraction from social media text of fire 2015.

CCS Concepts

• **Theory of computation~Support vector machines**
• **Computing methodologies~Natural language processing**
• *Information systems~Information extraction* • *Human-centered computing~Social tagging systems*

Keywords

Entity Extraction; Features; Social Media text; Machine Learning; Conditional Random Fields (CRFs); supervised algorithm;

1. INTRODUCTION

Social media is vast source of information from which we can extract lots of important data as per the specific requirement. This paper presents a technique for named entity recognition from English and Hindi text data. Our main task is to extract name entity from social media tweets in Indian language (Hindi and English) and classify these tweets in named entity tags as people, location etc., which is around 22 classes to be tagged. We used machine learning algorithm CRF (Conditional Random Field)[5] to identify Named Entities in corpus. CRF algorithm is implemented using CRFSuite[5] tool. CRFsuite[5] is an implementation of Conditional Random Fields for labeling sequential data which provides Fast training and tagging, Linear-chain CRF, etc.

Supervised learning is used for training dataset. We have used this training dataset to train out system for tagging named entities. CRFsuite[5] generate model based on the supervised learning provided.

2. CONDITIONAL RANDOM FIELDS (CRFs)

Given Conditional Random Field is a type of discriminative probabilistic model used for the labeling sequential data such as natural language text. Conditional Random Fields (CRFs) is mainly used as a class of statistical modeling method which is applied in pattern recognition and machine learning. CRFs are undirected graphical models, a special case of which correspond to conditionally-trained finite state machines. In the special case in which the output nodes of the graphical model are linked by edges in a linear chain, CRFs[5] make first order markov assumption and can viewed as a conditionally trained probabilistic finite automata. CRFs model consists of F=<f1,…,fk>, a vector of feature functions, **θ** = <θ1,…,θk> a vector of weights for each feature function. Let O=<o1,…,ot> be an observed sentence.

$$P(y \vee O) = \frac{\exp\big(\theta \cdot F(y, O)\big)}{\sum_{y'} \exp\big(\theta \cdot F(y', O)\big)}$$

3. METHODOLOGY

We use two different methods for identifying Named-Entity form given text. In one method we use Handcrafted or automatically generated rules for NER. In second method or approach we use machine learning technique for modeling. Also we have different machine learning technique i.e. supervise learning, semi-supervised learning, unsupervised learning for modeling.

Supervised learning gives best performance but it requires large amount of good quality annotated data. Unsupervised and semi-supervised learning is used when there is scarcity of annotated data in training.

We have used Machine learning based approach to perform NER task for given data, because it is more efficient than rule-based approach and it is more frequently used.

3.1 Pre Processing

The given task requires prediction of named entities from social media, so first task is to tag the word from the whole sentence. Therefore we have to split into word by doing these we get 'The' 'brown' 'cat' for both English and Hindi. Next step is to give part of speech(POS)[2] to text here we have used RDR POS Tagger for both the languages which identifies noun, verb, adverb from the given text. We used genia tagger for chunking in English. Genia tagger tag words with relevant IOB chunking tag. For example:

“The brown cat” will get chunk tag as the: B-NP, brown: I-NP, cat: I-NP.

We were provided with NER tagged data for training by FIRE-2015. We prepared a file with tag word and its pos tag, chunk tag and NER tag for training purpose.
For example:
Location India NNP B-NP

3.2 Training

We have used the open-source tool, CRFsuite[5] which is one of the popular implementations of CRF (Conditional Random Fields) for training data and also for tagging test data. CRFsuite[5] internally generates features from attributes in a data set. In general, this is the most important process for machine-learning approaches because a feature design greatly affects the labeling accuracy.

3.3 Testing

The untagged test data are given for testing with its POS tag[2] and Chunk tag. POS tagging and chunk tagging is done with help of RDR POS [2] tagger and genia tagger. After that this untagged test data with its POS tag and chunk tag are given as input to our model to get test result.

3.4 Feature Set

Feature set which is used for CRF [5] based NER System which includes Prefix or Suffix of word, length of word, Capitalization, POS tag, Chunking etc. we created two different model for both Hindi and English using different feature sets.

Table 1. Feature Set Usage description

Features	**Eng model (1)**	**Eng model (2)**	**Hin model (1)**	**Hin model (2)**
POS Tag	Yes	Yes	Yes	Yes
Chunk Tag	Yes	Yes	-	-
Prefix & Suffix	Yes	Yes	Yes	Yes
Capit-alize	Yes	Yes	-	-
Token Shape	Yes	Yes	-	-
Token Type	Yes	Yes	Yes	Yes
Length	Yes	Yes	Yes	Yes
Dot(.)	Yes	Yes	Yes	Yes
Comma(,)	Yes	Yes	Yes	Yes
Hyphen(-)	Yes	Yes	Yes	Yes
Colon(:)	Yes	-	Yes	-
Apostrophe(')	Yes	-	Yes	-
Back Slash	Yes	Yes	Yes	Yes
Two Digit Number	Yes	Yes	Yes	Yes
Four Digit Number	Yes	Yes	Yes	Yes
All Uppercase	Yes	Yes	Yes	Yes
All Digit	Yes	Yes	Yes	Yes
$ or Rs	Yes	-	Yes	-
POS Tag- NNP or QC	Yes	-	Yes	-
Gazzaters	Yes	-	Yes	-

Also we have included more features in hindi like जी , बजे, etc. in CRFsuite training.
For example:
मोदी जी का मिशन है
कार्यवाही 12 बजे तक स्थगित

So this kind of feature words are used in training model.

3.5 Post Processing

CRFsuite [5] gives only NE tag as output. So we combined output with its named entity. Then we prepared output as given format in training file by adding relevant information like tweet_id, user_id, Index, length of word. For example:

Tweet ID:618698235092152320 User ID:2922444438
NETAG:LOCATION NE:india Index:122
Length:5

4. RESULTS

4.1 Evaluation

There are two standard measures used for evaluation of NE tagger. (I) Precision(P) is the measure of the number of entities correctly identified over the number of entities identified. (II) Recall(R) is the measure of number of entities identified correctly over actual number of entities. Both precision and recall are therefore based on an understanding and measure of relevance. Harmonic mean of precision and recall which is F measure is calculated.

$$F = \frac{(\beta^2+1)PR}{\beta^2 R+P}$$

4.2 Test Result

Table 2. Test results of our system.

Language	Precision(P)	Recall(R)	F1-Score
Hin run-1	67.11	0.76	1.51
Hin run-2	74.73	46.84	57.59
Eng run-1	7.30	4.17	5.31
Eng run-2	5.35	5.67	5.50

5. CONCLUSION

Conditional random field(CRF) [5] are better for Indian languages than other models like HMM, MEMM etc. NER learned using CRFs takes more time for training. As part of Speech (POS) and Chunking is part of training, incorrect tagging also reduce the accuracy of the Recognized Named Entity. For achieving high performance and accuracy of NER system more study and deeper understanding of linguistic features are required.

6. ACKNOWLEDGMENTS

We thank Mr. Sandip Modha and other faculties of college for helpful input. This work is part of ESM-IL (Entity Extraction from Social Media Text - Indian Language).

7. REFERENCES

[1] Andrew McCallum, Wei Li: Named Entity Recognition with Conditional Random Fields, Feature Induction and Web-Enhanced Lexicons

[2] RDR Postagger http://rdrpostagger.sourceforge.net/

[3] Alan Ritter, Sam Clark, Mausam and Oren Etzioni. Named Entity Recognition in Tweets

[4] John Lafferty,Andrew McCallum, and Fernando Pereira. 2001.Conditional random fields: Probabilistic models for segmenting and labeling sequence data

[5] Naoaki Okazaki's (CRF Suit): Implementation of Conditional Random Fields (CRFs) http://www.chokkan.org/software/crfsuite/

[6] Dr.Rakesh ch. Balabantaray,Suprava Das,Kshirabdhi Tanaya Mishra IIIT, BBSR(2013): CRF++ based approach

[7] Yassine Benajiba and Paolo Rosso:Arabic name entity recognition using conditional Random Fields

[8] Genia tagger http://www.nactem.ac.uk/GENIA/tagger/

[9] CRF++ CRF++: Yet Another CRF toolkit CRF++ a simple, customizable, and open source implementation of Conditional Random Fields (CRFS)