Vira@FIRE 2015: Entity Extraction from Social Media Text Indian Languages (ESM-IL)

Vira Bagiya
Charotar University of Science & Technology,
Changa,Gujarat
India
virabagiya11@gmail.com

Anjana Patel
Charotar University of Science & Technology,
Changa, Gujarat
India
14pgce002@charusat.edu.in

Amit Ganatra
Charotar University of Science & Technology,
Changa, Gujarat
India.
amitganatra.ce@charusat.ac.in

ABSTRACT

In this paper we have tried to identify and extract "Named Entities" from social media text using conditional random field-(CRF) [3]. The paper represents our working methodology and result on Entity Extraction from Social Media Text Indian Languages task of FIRE-2015. We have extracted named entities from two languages Hindi and English. Named Entity Extraction system is implemented based on CRFSuite. CRFSuite [8] is the populer implementation of Conditional Random Fields (CRF). This is a sequential labelling task to achieve the desired tagging output. Conditional random fields (CRF) are a class of statistical modelling method often applied in pattern recognition, machine learning and many natural language processing tasks. We get F1-score of 19.82 and 3.72 for the Hindi and English text respectively.

Keywords

Machine learning; Named Entity Extraction; Named Entity Recognition.

1. INTRODUCTION

Named-entity recognition (NER) (also known as entity identification, entity chunking and entity extraction) is a subtask of information extraction that seeks to locate and classify elements in text into pre-defined categories such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc.

Social Media is vast source of information- from which we can extract lots of important data as per the specified requirements. According to the 8th schedule, India is known to have 22 official Indian languages. NER in Indian languages is still considered to be a budding topic of research in the field of NLP and much of work is needed to be performed in this regard. For English and Hindi languages there are so many NER tagger exists and hence this paper propose a CRF based NER tagger using CRFsuite (Okazaki, 2007) [8]. CRFsuite is an implementation of CRF and it is faster than CRF++ [7]. CRFsuite is an open source software which automatically extract features from the learning.

The paper is organized as follows. Section 2 gives an overview of Task description and approaches applied for NER task and complete description of our system. Furthermore section 3 describes the different issues in development of the system for different Indian languages. In section 4 there is the test result and how its accuracy can be increased. Finally section 5 concludes the paper.

1.1 Task Description

"Entity extraction from social media text in Indian Languages" is a task in which we have provided different tweets. --From this tweets – our work is to annotate and classify these tweets into different named entity tags like Person, Organization, Location, Entertainment etc. In training dataset we have given three columns tweet_id, user_id and tweet_text and in its processed annotated dataset we have given tweet_id, user_id, Named Entity tag(NE tag), Named Entity, index and its length. The Same thing we should perform on the testing dataset provided. Our main task is to identify named entity from testing dataset and apply appropriate tag to it.

1.2 System Architecture

Our Named entity recognition system is developed to classify and tagged named entities into 22 different classes such as Person, Location, Organization, Entertainment etc. We have provided training dataset which is mainly used for learning process.

There are following unique 22 named entity tags.

Table 1: Unique Named entity tags

1. PERSON
2. ORGANIZATION
3. LOCATION
4. ENTERTAINMENT
5. DAY
6. MATERIALS
7. PLANTS
8. PERIOD
9. LOCOMOTIVE
10. YEAR
11. MONEY
12. COUNT
13. FESTIVAL

14. DATE
15. QUANTITY
16. FACILITIES
17. DISEASE
18. ARTIFACT
19. MONTH
20. TIME
21. LIVTHINGS
22. SDAY

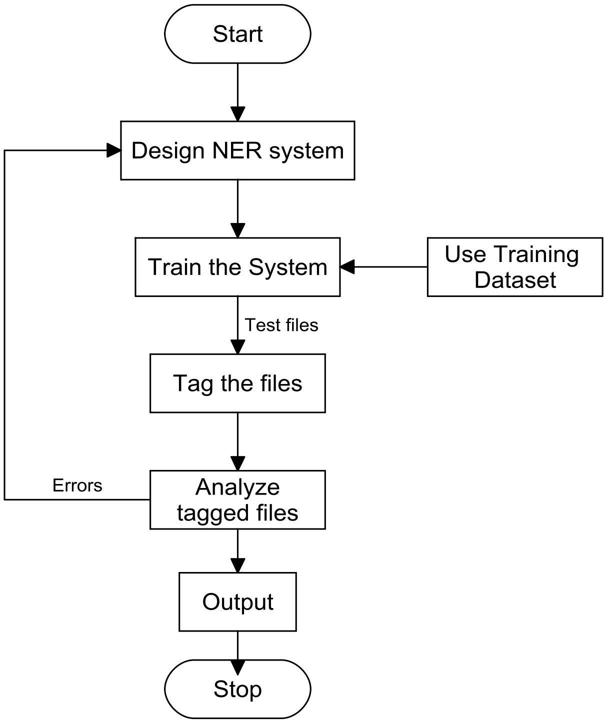

Figure 1: System Architecture

We have used the supervised learning, as we are given training dataset. We used this training dataset to train our system for tagging named entities and kept these tags in a space separated files. CRFsuite generate the model based on training – learning provided. Later on, system uses these model for generate output (named entity tagging). The training dataset is primary focus for our training. Figure 1 - flow-chart is showing the basic flow of our system in detail.

As we have used CRFsuite to implement our NER system. In which features can be easily extracted for labeling entities based on the provided training datasets. Hence We can easily add our own features by modifying some line of codes as it is an open source software. Features can be generated for unigram as well as bigrams.

2. APPROACHES FOR NER

There are basically two approaches that are employed in Named Entity Recognition. These include:

a. Rule Based Approach

b. Machine Learning Based Approach

In rule based approach there are Handcrafted or automatically generated rules or patterns. Machine learning techniques are used for statistical modeling which can be either unsupervised, semi-supervised or supervised mode of learning. Unsupervised and semi-supervised mode of learning are used when there is a scarcity of annotated data for training but the best performance is obtained by using supervised mode of learning which requires a large amount of good quality annotated corpus.

We have used machine learning based approach. This approach is also known as automated approach or statistical approach. Machine learning based approach is more efficiently and frequently used as compared to the Rule based approach.

We have developed a system to perform NER in English and Hindi and submitted the same. We use the open-source software, CRFsuite[8] which is one of the popular implementations of Conditional Random Fields (CRF)[3] for training a model based on the training dataset and then use the model to generate tags for the test dataset.

2.1 Extracted features from learning

For English and Hindi languages there are following features extracted from learning

CRFsuite automatically extract some features from learning. Other features are added for tagging different different named entities.

- **Gazetter**(specified as list-look up table): Gazetter of location names has been created and applied to identify different locations in India. Same as - Plant names, Festival names, Entertainment , Locomaotives , Livthings tagged using the different different gazetters as a feature.
- **Suffixes**: In hindi person name identified using suffix "जी".Means word ending with "ji" can be the person name and it can be specified as the feature.

  for example: Modiji is a person name

  If a word followed by "ko" then this word can be specified as the person name for eample:

  "Gita ko haridwar jana hain" – 'Gita' is aperson name which is followed by 'ko'.
- **Prefixes** :There are so many prefixes can be used to identify named entities. For example, Named entity followed by Mr. or Miss or Mrs would be a person name
- **Word Context**: Context of the word of window size four is used which takes two words before and two words after the word as feature. This helps modeling the language structure about how where and with which words entities are used in a sentence. There are total seven feature values for word context which includes the word itself, two words before it, and two words after it, and pairing of word with its previous and next word.

- **POS tag**: Parts of Speech (POS) tag of a word is also considered as a feature because all the entities are nouns.
- **Regular Expressions**: We have used different regular expressions to identify temporal based named entity like Date,Month,Year, Period,Day and Time.

2.2 Pre-processing

Social media text is noisy in nature. People use shorthand and ungrammatical text for saving their time. Thus capitalization is not properly applied as well as Spellings are not correct. This data becomes hard to handle in the aspect of Information-extraction.

First from the given testing dataset we have removed all the links presented in the tweets. Then tokenizing, Part-of-speech tagging and chunking is done. For English language we have used Stanford Part-Of-Speech tagger [5]. For Hindi language we have used RDRPostagger [6].

Input for CRFsuite(NER Tagger):

As there is space separated 4 fields input to the CRFsuite, we have combined output from tokenizer , POS tagger and make one space separated file for training. Then this same process applied for testing dataset.

For example: PERSON Gitika NNP B-NP

Each tweets is preprocessed according to the requirement of CRF suite which needs a file in which each line has a single word and its NER tag separated with a white space, A new line represents end of a sentence Two processed files were created, one with BIO tags which shows multiword entities (for English language) and another without it (for hindi language).

2.3 Post-processing

Output of the NER tagger would be only NE tag. So we have combined it with its named entity, tweet_id and user_id. Then find the length of the named entity and its index-means position of the named entity.

And as per given format we have arranged such as:

Tweet ID:623472520352636928 User Id:241166752
NETAG:PERSON NE:Ali Index:104
Length:3

3. ANALYSIS

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. Though still this Indian language content is only a fraction of the English content. The growth of Indian language content is expected to increase by more than 70% every year.

Hence there is 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 identify entities, identification of associations or relation between entities. Hence an automatic Entity extraction system is required.

Named Entity Recognition (NER) is one of the most important information extractions techniques being developed in the NLP and IR communities. Considerable success has been achieved in English with extraction of multiple entities as per domain of interest. However, the area poses considerable challenges when tried in other languages and particularly Indian Languages. Such as - There is no capitalization available in Indian languages.

There is lot of research work going on in NER for Indian languages, such as Workshops NERSSEA-2008, SANLP 2010, 2011 but, there is lack of bench mark data to compare several existing systems. There is no common evaluation methods exists to judge any researchers' work.

4. RESULTS AND DISCUSSION

4.1 Evaluation metrics

Two standard measures, Precision (P) and Recall (R) are used for evaluation of the Named Entity (NE) tagger, where precision is the measure of the number of entities correctly identified over the number of entities identified and recall is the measure of number of entities correctly identified over actual number of entities. F measure is calculated which is the harmonic mean of precision and recall

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

When β = 1, F measure is called F1 measure or simply F1 score.

4.2 Test results

Table1: test result

Languages	Precision	Recall	F1-Score
Hindi	25.65	16.14	19.82
English	4.13	3.39	3.72

5. CONCLUSION

CRF models are appropriate for the highly inflective Indian languages and perform better than other systems like HMM, MEMM etc. (Vijay Sundar Ram R, 2011). CRFsuite generate model based on the learning and provides output (NE tag) as per the generated model. But Problem is, NER system learned using CRF takes more time for training the model. The parts-of-speech tag is the important feature for NER to identify the named entity chunk. Incorrect parts-of-speech tag for the token may result in reducing the accuracy of NER system. Achieving a high performing NER system requires more study and deeper understating of linguistic features. Various permutation and combination of feature sets can be used and tested for getting high recall value and eventually higher F1-scores.

6. FUTURE WORK

In English and Hindi both language we will try to get more accurate results in identifying and tagging named entities. For that we will optimize our features sets.

7. ACKNOWLEDGEMENT

We would like to thanks Prof. Prasenjit Majumder sir for his guidance throughout the work. Additionally, we would like to thank FIRE 2015 for providing a great opportunity to work under this task and facilitating the support.

8. REFERENCES

[1] Asif Ekbal, R.H.: Language Independent Named Entity Recognition in Indian Languages. In: IJCNLP, pp. 33–40 (2008)

[2] David Nadeau, S.S. (n.d.).: A survey of named Entity recognition and classification. National Research Council Canada/ New York University

[3] J. Lafferty, A. McCallum, and F. Pereira, “Conditional random fields: Probabilistic models for segmenting and labeling sequence data,” in Proceedings of ICML, pp. 282–289, (2001)

[4] Vipul Garg, Nikit Saraf, and Prasenjit Majumder: Named Entity Recognition for Gujarati: A CRF Based Approach

[5] standford POStagger

http://www-nlp.stanford.edu/software/tagger.shtml

[6] RDRPostagger

http://rdrpostagger.sourceforge.net/

[7] Kudo, Taku. "CRF++: Yet another CRF toolkit." Software available at http://crfpp. sourceforge. net (2005)

[8] Okazaki, N.: CRFsuite: A fast implementation of Conditional Random Fields, CRFs (2007), retrieved from http://www.chokkan.org/software/crfsuite/