HITS@FIRE task 2015: Twitter based Named Entity Recognizer for Indian Languages

Pallavi K P., Srividhya K., Rexiline Ragini John Victor, Ramya M. M.,
Hindustan Institute of Technology and Science,
Padur, Chennai.
rs.pkp0310@hindustanuniv.ac.in, rs.sk0214@hindustanuniv.ac.in,
rs.jrr0913@hindustanuniv.ac.in, mmramya@hindustanuniv.ac.in

ABSTRACT

Natural Language processing (NLP) in its pure sense, is a platform that provides the ability for transforming natural language text to useful information. Named Entity Recognition (NER) is a key task in NLP for classification of named entities in natural languages. Though, there are several algorithms for named entity classification, identifying named entities in twitter data is a demanding task. Loads of information are being shared by people in twitter on a daily basis. This information is unstructured and often contains important information about organizations, politics, disasters, promotional advertisements etc. In this paper, we provide a NER that can effectively classify named entities in twitter data for Indian Languages such as English, Hindi and Tamil. POS, Chunk, Suffix, Prefix information has been used for training in Conditional Random Fields (CRF) based NER Model. CRF is a popular model for labeling and classification in text mining. Performance analysis was done using n-fold validation and F-measure. A maximum precision of 93.82 for English, 92.28 for Hindi and 86.94 for Tamil twitter data was achieved through N fold validation. Results provided by ESM-IL share task in terms of precision for English is 50.48, for Hindi is 81.49 and for Tamil 70.42. The proposed algorithm has a higher classification accuracy and it is achieved through n-fold validation.

CCS Concepts

• **Human centered computing**→ Human machine interaction→ Collaborative and social computing • **Applied Computing** →Document managing and text computing methodologies→ Artificial Intelligence and Machine Learning.

Keywords

Natural Language Processing, Named Entity Recognition, Conditional Random Fields, n-fold validation, Twitter data, English, Hindi, Tamil.

1. INTRODUCTION

NLP is gaining prominence due to the importance given to social media data such as twitter and Facebook. Twitter data are predominantly extracted and monitored by public and private organizations for analysis of various trends in the industry and for opinion mining. For efficient NLP, corpus should be from the same domain of the NLP application [1]. NLP involves a set of computational linguistic tools for interaction between computer and natural languages. NER is one among such tools which identifies and classifies names in a given corpus. NER is backbone for several NLP applications such as language translation, social media analysis and information mining.

NER is important for Indian language twitter data, as there is no clue in identifying named entities in them. Recognizing named entities in social media data is quite challenging due to the unstructured clauses in a sentence and entities are more diverse in nature. There are a variety of approaches for recognizing named entities. Supervised approaches include Hidden Markov Model (HMM), Decision Trees, Maximum Entropy Model (MEnt), Support Vector Machines (SVM) and CRF *etc.*. Here, for the ESM-IL task, CRF has been used to develop NER for English, Hindi and Tamil twitter data [2]. Several research work has been done on Named Entity Recognition [3-5]. Now Named Entity Recognition is gaining popularity for its diverse applications in the real world [6, 7].

2. PROBLEM DEFINITION

Twitter is one of the popular social networking sites where people share their opinions through their tweets. Each tweet or post can contain a maximum of 140 characters, including smilies, hash tags, other symbols and website links. We attempted to develop a twitter based NER for English, Hindi and Tamil data. Twitter training data were provided by ESM-IL for this task. Training data included raw data files and annotated named entities. Before pre-processing, annotated named entities were mapped with the corresponding tweets in the raw data file, to label the named entities. Group of labels is called as Tag set. A total of 22 tags were identified from the training data, namely person, location, organization, date, quantity, money *etc.*. Pre-processing has been done before applying the methodology. It included tokenization and removing noisy data.

3. METHODOLOGY USED

CRF, a probabilistic approach for developing NER is used for classification. CRF is a popular approach for effectively classifying named entities. It takes into account the neighboring samples or the context information of the sentence. But the disadvantage in twitter data is, lack of context information. Tweets in the raw data file were tokenized and pre-processed by removing noisy data website links, hash tag, and smilies etc. was done. Initially, neighboring samples were only considered in developing the model for NER. Later to include lexical information, POS taggers, Chunker and 1st, 2nd, 3rd degree suffixes and prefixes were also added as features. The methodology of the proposed algorithm is shown in Figure 1.

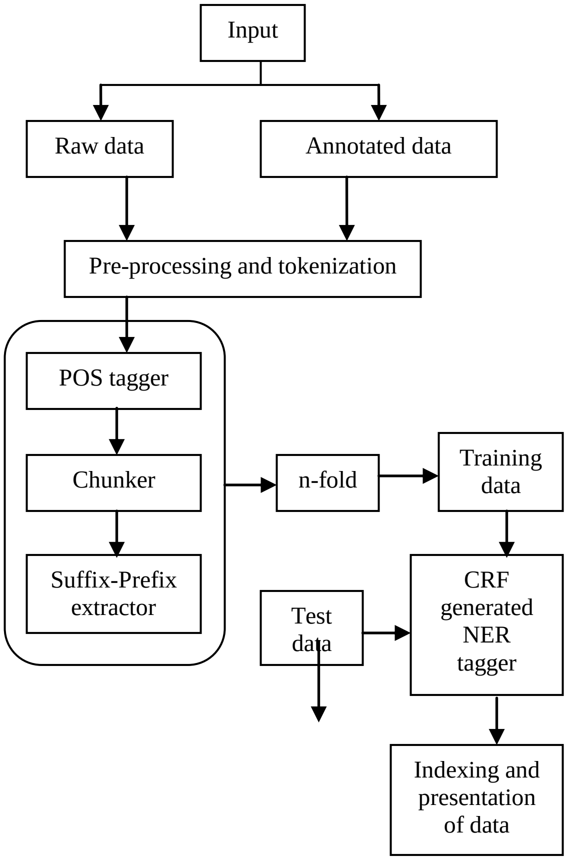

Figure 1: Methodology of the NER

Input considered for NER were raw and annotated data provided by ESM-IL shared task. To avoid misclassification of named entities, labelling was done by assigning the first word of the entity with B-tag (Begining of the tag) and remaining words in the same entity with I-tag (Inside the tag). With the help of the annotated data, named entities in the raw data were mapped. For pre-processing and in order to identify the noisy data rules were framed. For *eg*., detection and removal of website links was performed by checking for the presence of tokens such as http//:, .com.

For English parts of speech tagging and chunking, pattern.en module was used [8]. It is a python based POS tagger cum chunker for twitter data. Training in a twitter based POS tagger provides better accuracy than a normal tagger. So pattern POS tagger was preferred. Due to the unavailability of twitter based POS tagger, for Hindi parts of speech tagging the tagger provided by Society for Natural Language Technology Research was used [9]. It is a CRF based open source software. Apart from the POS and Chunk taggers, other features like suffixes and prefixes were also used. The list of features used are provided in Table 1.

Table1: Features Description

Features	**English**	**Hindi**	**Tamil**
POS tags	Yes	Yes	No
Chunk tags	Yes	No	No
1st, 2nd , 3rdCharacter Suffix	Yes	Yes	Yes
1st, 2nd , 3rd Character Prefix	Yes	Yes	Yes

For n-fold cross validation, the annotated data was randomly portioned into five sub sets. Validation was performed by n rounds of testing the system by providing one set of the annotated data for testing each time and the remaining n-1 subsets for training [10]. Each fold contained equal amount of tokens.

For NER task, CRF++ an open source package was used [11]. Conditional Random Fields is a probabilistic framework for segmenting and labelling sequence data. From the literature it was understood that CRF performs better than other models like Hidden Markov Model by providing conditional probability and Maximum Entropy Markov Models by observation and sequence of labels [1]. Finally, indexing was done, and the data was presented in the prescribed ESM-IL format.

4. PERFORMANCE METRICS

The performance metrics used for analysis are Precision, Recall and F-measure. Precision and Recall are the most effective and frequently used measures in case of information retrieval.

Precision can be defined as

$$Precision\ (P) = \frac{True\ positives}{True\ positives + False\ positives} \qquad (1)$$

Recall can be expressed as

$$Recall\ (R) = \frac{True\ positives}{True\ positives + False\ negatives} \qquad (2)$$

F-measure is defined as the weighted harmonic mean of precision and recall.

$$F - measure\ = \frac{2PR}{P + R} \qquad (3)$$

where, True positives are the total number of NE’s tagged correctly with boundaries.

False positives are the total number of words that are wrongly tagged by the system which are not tagged manually.

False negatives are the total number of untagged words by the system which are manually tagged [12].

5. RESULTS AND DISCUSSION

In this section, the results obtained with the CRF model is discussed. Twitter data for experiment was provided by FIRE 2015. Table 2 compares the results obtained using n-fold validation and the results provided by ESM-IL shared task.

Table 2: Precision, Recall and F measure results for n-fold validation and ESM-IL shared task evaluation.

Languages	**Precision**		**Recall**		**F-measure**	
	n-fold	ESM-IL	n-fold	ESM-IL	n-fold	ESM-IL
English	93.82	50.21	80.53	37.06	86.66	42.64
Hindi	92.28	81.21	76.23	44.57	83.49	57.55
Tamil	86.94	64.52	73.87	22.14	79.87	32.97

As a part of the initial pre-processing website links, hash tags and smilies were removed from the raw data. Further, POS tagging and chunking for English data was done. This was given to the CRF model and a test run was generated. A maximum F-measure of value 80.81 was obtained. To improve the accuracy features like prefixes and suffixes of the NE were included and a maximum F-measure of 86.66 was obtained.

For Hindi, after initial pre-processing, POS tagging was done. An efficient chunker is not available for Hindi data, and therefore chunking was not done. This was supplied as input to the CRF model and a maximum F-measure of 83.49 was obtained. Prefixes and suffixes of NEs were also combined with the existing POS tags and 78.47 F-measure value was obtained. When considering only prefixes and suffixes of NE as feature, the model resulted in an F-measure of value 73.12. The classification accuracy reduced for Hindi when compared to English due to the unavailability of twitter based POS tagger and chunker.

For Tamil data, since there is no proper POS tagger and chunker available of twitter and generic data, features like suffixes and prefixes of NE were considered and the model obtained a maximum F-measure 79.87. Excluding the suffixes and prefixes, considering only the tokens as a feature, the model gave an F-measure of 62.95. The classification accuracy reduced for Tamil data similar to Hindi due to the unavailability of Twitter based POS tagger and chunker. Table 3 shows the examples for the results obtained from CRF generated model for English, Hindi and Tamil twitter data. Sample tweets contain tweet-id, user-id and the tweet.

	Sample Tweets
1.	624133739023446016 917553836 As outrage builds, Karnataka Vikas Grameena Bank looks at rescuing farmers: Much of the farming community's outrage... http://dlvr.it/BcS9Kh
2.	623537594798739456 3016932104 विश्व टी20 फाइनल की मेजबानी करेगा ईडन गार्डन्स http://hindi.webdunia.com/latest-cricket-news/twenty20-world-cup-115072100078_1.html ... pic.twitter.com/pqzycmT9b8
3.	621569431932530688 317752766 நெல்லை மாவட்டம் கல்லிடைக்குறிச்சியில் திறக்கப்பட்ட பாலத்தால் பொதுமக்கள் மகிழ்ச்சி: நெல்லை... http://goo.gl/fb/Efqp0N
	NEs identified in tweets
1.	Karnataka B-ORGANIZATION Vikas I-ORGANIZATION Grameena I-ORGANIZATION Bank I-ORGANIZATION
2.	विश्व B-ENTERTAINMENT टी20: I-ENTERTAINMENT ईडन B-LOCATION गार्डन्स I-LOCATION
3.	நெல்லை B-LOCATION மாவட்டம் I-LOCATION கல்லிடைக்குறிச்சியில் B-LOCATION

Table 3: Sample results obtained for English, Hindi and Tamil twitter data.

6. CONCLUSION AND FUTURE WORK

Named Entity Recognizers for English, Hindi and Tamil were developed for the twitter data. A CRF based model was generated by POS tagging, chunking and applying other feature information to the given data. To test the accuracy of the CRF model n fold validation was done. N fold experiment for training data gave a maximum precision of 93.82 for English, 92.28 for Hindi and 86.94 for Tamil twitter data. ESM-IL evaluation for English, Hindi and Tamil data resulted in competitive precisions of 81.49, 70.42 and 50.21 respectively. Due to the increase in the percentage of recall, F-measure was reduced. This increase in recall can be reduced by providing pre-processed rules based on lexical resources.

7. ACKNOWLEDGEMENTS

The authors thank Hindustan Institute of Technology & Science for their continuous support.

8. REFERENCES

1. Pallavi, K. P., and Anitha S. Pillai. 2015, Kannpos-Kannada Parts of Speech Tagger Using Conditional Random Fields. Emerging Research in Computing, Information, Communication and Applications. Springer India, (Aug 2015), 479-491.

2. Lafferty, John, Andrew McCallum, and Fernando CN Pereira., 2001, Conditional random fields: Probabilistic models for segmenting and labeling sequence data., In *Proceedings of the 18th International Conference on Machine Learning 2001,* (Jun 2001), 282-28

3. Malarkodi, C. S., & Pattabhi, R. K. Rao and Sobha, Lalitha Devi, 2012,Tamil NER–Coping with Real Time Challenges, *In Proceedings of the Workshop on Machine Translation and Parsing in Indian Languages*, (Dec 2012), 23-38.

4. Biswas, S., et al., 2010, A Two Stage Language Independent Named Entity Recognition for Indian Languages, *IJCSIT International Journal of Computer Science and Information Technologies*, 1.4 , 285-289.

5. Saha, Sujan Kumar, et al.,2008, Named entity recognition in Hindi using maximum entropy and transliteration., *Research journal on Computer Science and Computer Engineering with Applications*, (Jul 2008), 33-41.

6. Jung, Jason J., 2012, Online named entity recognition method for microtexts in social networking services: A case study of twitter." Expert Systems with Applications, 39,9, (Jan 2012), 8066-8070.

7. Zirikly, Ayah, and Mona Diab. "Named entity recognition for arabic social media." Proceedings of naacl-hlt. (Jun 2015), 176-185.

8. www.clips.ua.ac.be/pages/pattern-en

9. nltr.org/snltr-software/

10. John J. Hutton, Pediatric Biomedical Informatics: Computer Applications in Pediatric Research, Springer Publications, 2012.

11. https://taku910.github.io/crfpp/

12. Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval, Cambridge University Press. 2008.