Named Entity Recognition from Indian tweets using Conditional Random Fields based Approach

Maithilee L. Patawar[1], M. A. Potey[2]

Abstract—**Task of Named Entity Recognition (NER) refers to identification of entities from text. These entities consist of proper names like person name, location names, temporal entities etc. Significant information always close to such entities which made NER subtask of Information Retrieval (IR) systems. Almost all NLP applications utilize results of NER to achieve better accuracy and precision. Social media is gaining popularity day by day and used by most people to share information, therefore are become source of knowledge. Traditional NER systems which are designed to deal with articles or formal text gives poor results on social media content like tweets due to its short and noisy nature. Normalization of data is required to get over this problem. Multiple Indian languages like Hindi, Marathi are also widely playing role on social media. NLP applications designed to deal with regional language tweets need a dedicated NER system as there is no such system designed yet. In this paper, a CRF based NER system is given for both English and Marathi tweets. Here different language independent features are discussed along with the challenges faced while building the system.**

Index Terms— **Named Entity Recognition, Named Entities, Gazetteer, Tweets, Linguistic Feature, Parts of Speech tags**

I. Introduction

The term Named Entity Recognition (NER) can be considered as subtask of Information Extraction where Named Entities (NE) are extracted from text input. NER seeks to locate and classify elements in text into predefined categories. In the term “Named Entity”, the word Named restricts the task to those entities for which one or many rigid designators stands as referent. This is widely used in Natural Language Processing (NLP).

The task of Named Entity Recognition was formally defined in Message Understanding Conference 6 (MUC6) [1] as the task of identifying the names of all the people, organizations and geographic locations in a text, as well as time, currency and percentage expressions. The task has also been extended to technical domains to recognize domain-specific entities, typically in the domain of biomedical science [11] to recognize domain-specific entities such as gene and protein names. NER is a preprocessing task for many NLP based applications like Question Answering, Event extraction, Relation extraction, etc. In earlier days news-papers and online news articles were the fastest medium of information sharing. Therefore many NLP applications were designed which utilities these for their information need. But later social media has changed this scenario. Today social media is the fastest medium of information sharing. So applications are now switching their focus and trying to obtain information from social networking sites like Facebook, Twitter, etc. Unrestricted style of writing, short nature makes it difficult to extract named entities from social media content. Standard NER systems like Stanford NER [6] designed for news articles has shown poor performance on tweets. Therefore a dedicated system required to deal with social media data and to obtain named entities from it.

Sufficient work has been done in NER system designed for languages like English, Spanish, etc. Additional approaches are proposed to improve accuracy of obtaining named entities (NEs) from English tweets. Different NER systems and approaches were proposed for Indian languages like Punjabi, Hindi, Telugu, Malayalam, etc. But these systems are not able to deal with tweets in their respective languages of task. For Marathi language, very few approaches were proposed and NER system designed have lower accuracy. Furthermore there is no NER system which is designed to get NEs from Marathi tweets. This paper has presented a novel NER system which finds out NEs from Marathi tweets.

II. Literature Survey

Related work is reviewed in two categories: NER for Indian Languages and NER on tweets.

A. *NER on Indian Language*

As earlier said, very few systems were proposed for Indian Language based NER system. As a part of IJCNLP-08 NER Shared Task on South and South East Asian languages (NERSSEAL), multiple NER system utilizing different approaches for Indian languages have been reported [2]. The task of NER for Marathi has been explored by Patel and Ramakrishnan in their paper by including rules in Inductive Logical Programming [4] with highest F-measure 0.82 for person name. A language independent multilingual document clustering approach on comparable corpora was presented by including rules in Inductive Logical Programming [4] with highest F-measure 0.82 for person name. A language independent multilingual document clustering approach on comparable corpora was presented by Kumar and Varma [5]. This approach can be applied for Hindi and Marathi language and it is based on k-means algorithm for clustering. Here

Manuscript received May, 2016.

1. ***Maithilee L. Patawar***, *Computer Engineering, Savitribai Phule Pune University. Pune, India.*

2. ***M. A. Potey***, *Computer Engineering, Savitribai Phule Pune University. Pune, India.*

clusters are formed with identified named entities and unnamed entities. Ekbal et al. has developed NER system [16] for two leading languages: Bengali and Hindi. This system has tested against the gold standard test sets i.e. manually annotated test set of NEs and has shown 0.83 f-score for Hindi language. In addition to supervised approach, language dependent features are used to improve the performance significantly.

B. NER on Tweets

The era of social media has changed knowledge base of many NLP based application and therefore these applications needed to find NEs from it. At very first, Amazon's Mechanical Tuek service and CrowdFlower [15] were used by Finin et al. to annotate NEs in tweets. Here semi-supervised approach is used to evaluate effectiveness of human labeling. Xiaoha Liu et al. has presented a system [9] which perform task of NER for English tweets. Results are obtained by combining Conditional Random Fields (CRFs) and KNN with F-measure 0.802. Here process of normalization in addition to use of gazetteers has improved the accuracy of NEs significantly. A segmentation based approach was proposed by Chenliang Li et al. [8] which consider global and local context while categorizing and labeling entities from tweets. Tweets are first divided into segments of meaningful phrases using local and global context. Then stickiness score is used to extract NEs.

III. NER for Indian languages

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. But NER work involving Indian languages has started very recently and needed to deal with challenges in order to get better performance. Some of challenges and features of NER system designed for Indian languages are discussed below.

A. Challenges for Entity Extraction from Indian Languages

Indian languages are quite different than English and most of European languages. So while developing NER system, following Challenges are needed to be considered:

- There is no capitalization in Indian languages. This feature lays a very important role to find NEs.
- Many of Indian person's names are kind of common names, verbs and adjectives in the dictionaries (e.g. *Priya* meaning favorite, *Vijay* meaning victory, etc.).
- Indian languages are highly inflected and provide rich and challenging sets of linguistic and statistical features resulting in long and complex word forms.
- There are very less resources like dictionaries, gazetteers available for Indian languages.
- Unavailability of good morphological analyzers, POS taggers with required good quality.
- Indian languages have free-word order.

To overcome issues arising from these challenges, multiple features are used. Some of them are discussed in next section.

B. Language Independent Features of Indian Languages

It is required to consider different combinations of the set of language independent features [7] to select the best set of features for NER build for Indian languages. The following describes the features:

- Context Word Feature: Words preceding and following a particular word can be used as features. This is based on the observation that the surrounding words are very effective in the identification of NEs.
- Word Suffix: Word suffix information is helpful to identify NEs. This is based on the observation that NEs share some common suffixes. This feature can be used in two different ways. The first naive way to use it is to consider a fixed length (say, n) word suffix of the current and/or the surrounding word(s) as features. This is actually the fixed length character strings (i.e, strings of length 1, 2 or 3 etc.) stripped from the word endings. If the length of the corresponding word is less than or equal to n1 the feature values are not defined and denoted by ND. The feature value is also not defined (ND) if the token itself is a punctuation symbol or contains a special symbol or digit. The second and more helpful approach is to use the feature as binary valued. Variable length suffixes of a word are matched with predefined lists of useful suffixes for different classes of NEs. Variable length suffixes belong to the category of language dependent features as they require language specific knowledge for their development.
- Word Prefix: Word prefixes are also helpful and based on the observation that NEs share common prefix strings. This feature has been defined in a similar way as that of the fixed length suffixes.
- Information of Named Entity: The NE tag(s) of the previous word(s) is/are used as the only dynamic feature in the experiment. These tags carry important information in deciding the NE tag of the current word.
- First Word: This is used to check whether the current token is the first word of the sentence or not. Though Indian languages are relatively free-word order languages, the first word of the sentence is most likely a NE as it is the subject most of the time.
- Digit Features: Several binary valued digit features have been defined depending upon the presence and/or the number of digits in a token (e.g., CntDgt [token contains digits], FourDgt [token consists of four digits], TwoDgt [token consists of two digits]).

IV. Hybrid CRF approach

Different approaches can be utilized for NER system. On coarse level these approaches are divided into 2 categories viz: Rule based and Machine Learning (ML) based approaches. Rule based approaches also known as linguistic approaches make use of handcrafted rules wherein ML based approaches utilizes this linguistic rules to infer. Combination of approaches from these categories gives rise to new class of approach i. e. Hybrid approach. Here ML approach is aggregated with rules and exploits advantages of both.

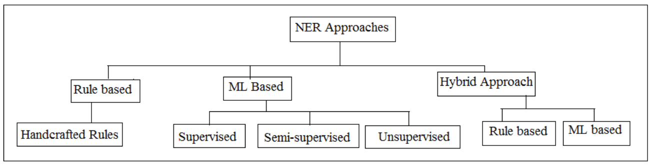


Fig1. Approaches of NER

CRFs are a class of statistical modeling method often applied in pattern recognition and machine learning, they are used for structured prediction. Whereas an ordinary classifier predicts a label for a single sample without regard to "neighboring" samples, a CRF can take context into account. The model uses sequence modeling algorithms which are probabilistic in nature. Sequence labeling is a type of pattern recognition task that involves the algorithmic assignment of a categorical label to each member of a sequence of observed values.

Because of the strong ability to integrate any kind of features which plays an important role during training, CRFs becomes one of the key factors affecting the NER performance. The features of CRFs based NER include not only the internal features from context, such as character information, POS and boundary, but also the external features based on the statistical results such as surname that the prefix of family names, the suffix of location and organization and so on. In addition, the feature template is also found to play an important role in CRF based NER.

V. Proposed system

Proposed system is designed to deal with tweets. Here the CRF based hybrid NER system[9] is studied which has shown excellent performance for English tweets. As the CRF approach is suitable for Indian languages, proposed approach can be referred to build Marathi tweet based NER system. Problem definition, system architecture, algorithm and mathematical model for given system is explained in next sections.

A. Problem Definition

As discussed earlier, there is no NER system yet designed for Marathi tweets. Standard NER systems like Stanford NER and tweets dedicated systems like TwiNER shows poor performance for Marathi tweets. Morphologically rich nature of Marathi makes it difficult to get NE for these systems. So a dedicated NER system for Marathi tweets required to develop.

B. System Design

Fig. 2. represents the system architecture of proposed system. All the intermediate steps of the systems are clearly mentioned in this architecture. Here first input file or string is taken from the user.

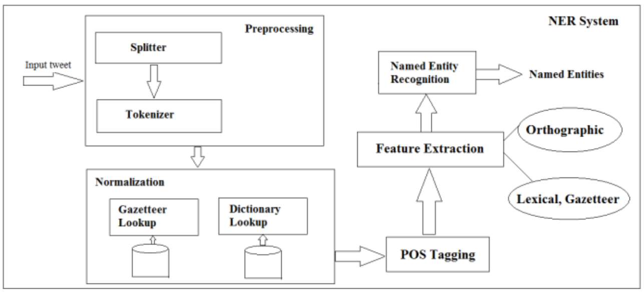


Fig. 2. Proposed NER system for Tweets

The whole system is divided into three parts: Pre-processing, Normalization and NER where the input is given to preprocessing which splits the sentences and tokenizes the words. These words are provided as input to Normalization model where dictionary and gazetteer lookup helps to normalize words. Here general named entities like week of days, month names are identified with the help of Gazetteers. Dictionary search is used to replace misspelled words. This data is then eligible for NER where named entities are identified and labels are assigned based on CRF approach. Results are then displayed to user.

As there is lack of availability of annotated corpora for Indian languages and especially for Marathi language, POS tags are used instead. Many POS taggers for Indian languages have proposed [18][19]. In proposed system, an OpenNLP POS tagger is used for tweets for assigning tags.

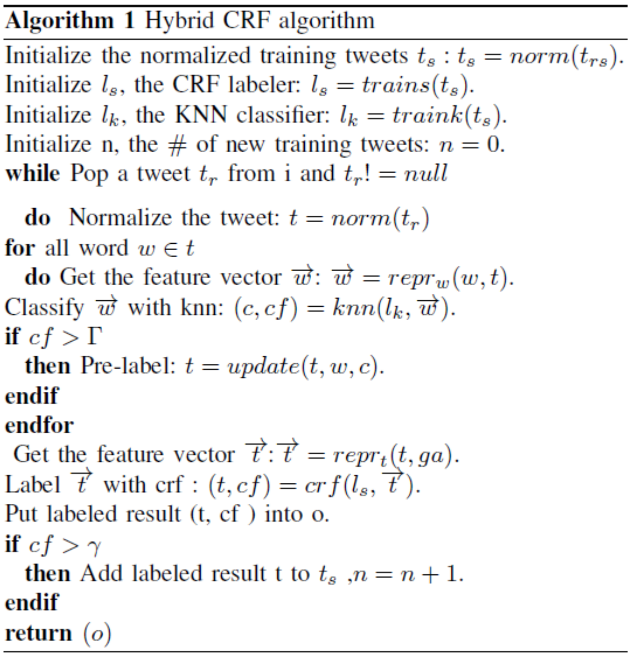
C. Algorithmic Steps

While extracting NEs from tweets, it is required to normalize them first. In normalization process, ill formed words, abbreviated words are replaced with corrected words. After that confidence value obtained from KNN classifier for each word is checked to assign label and decide class. Initially this confidence value is set to 0.1 and incremented as word appeared more than once. Then based on learning weight assigned for each feature function, CRF labeler calculates probability. This value is used to finally assigned label for word. Algorithm 1 explains the procedure clearly [9]. From multiple variants of CRFs, here a linear chain CRF is used. Therefore algorithm has a time complexity quadratic in number of labels i.e. O (n^2) where 'n' represents number of labels. Here t_s are tweets after normalizing them while norm represents process of normalization. Confidence value obtained from KNN is represented by cf and o represents tweets with labels i.e. output.

Algorithm 1 is used to implement the system. Here inputs to system are labeled tweets and gazetteers. These are used to train system. After training, a POS tagger is used to assign tags and used to generate rules based on tags assigned by tagger. Problem of lack of availability of labeled tweets is partially solved using these tagged tweets. Based on features extracted by systems named entities are assigned and result is given back to users.

Here two main functions are used. First function is used to obtained confidence value for KNN algorithm. This function can be given as:

$$cf = \frac{\sum_{w' \epsilon\, nb} \delta(w',c').\cos(w,w')}{\sum_{w' \epsilon\, nb} \cos(w,w')}$$



```
Algorithm 1 Hybrid CRF algorithm
Initialize the normalized training tweets $t_s$ : $t_s = norm(t_{rs})$.
Initialize $l_s$, the CRF labeler: $l_s = trains(t_s)$.
Initialize $l_k$, the KNN classifier: $l_k = traink(t_s)$.
Initialize n, the # of new training tweets: $n = 0$.
while Pop a tweet $t_r$ from i and $t_r! = null$

  do  Normalize the tweet: $t = norm(t_r)$
for all word $w \in t$
  do Get the feature vector $\overrightarrow{w}$: $\overrightarrow{w} = repr_w(w, t)$.
Classify $\overrightarrow{w}$ with knn: $(c, cf) = knn(l_k, \overrightarrow{w})$.
if $cf > \Gamma$
  then Pre-label: $t = update(t, w, c)$.
endif
endfor
 Get the feature vector $\overrightarrow{t}$:$\overrightarrow{t} = repr_t(t, ga)$.
Label $\overrightarrow{t}$ with crf : $(t, cf) = crf(l_s, \overrightarrow{t})$.
Put labeled result (t, cf ) into o.
if $cf > \gamma$
  then Add labeled result t to $t_s$ ,$n = n + 1$.
endif
return $(o)$
```

Later CRF function is used to assign labels based on probability and it is expressed as:

$$P(l|t) = \frac{1}{z} \exp \sum_{i=1}^{l} \lambda_k f_k (s_{i-1}, s_i, t, i)$$

VI. Results

Results of given hybrid NER approach is shown in next table. Here results of rule based approach are compared with proposed CRF approach. Total 1000 political tweets (English) are extracted for experiment from which 800 tweets are used for training and 200 used for testing. Along this gazetteers of locations in India, Indian person's names and political parties are used while labeling. Based on this results of English tweets, same approach is implemented for limited Marathi tweets. While obtaining results for Marathi tweets gazetteers are not utilized due to unavailability of Marathi gazetteers. This result can be improved by incorporating different Marathi gazetteers.

System	Precision	Recall	F- measure
NER using CRF for English tweets	78.4	72.6	75.38
NER using rules for English tweets	65.4	60.8	63.01
NER using CRF for Marathi tweets	53.7	48.2	50.8

Parameters for calculating results of NER are the same that of Information Extraction (IE) i.e. precision, recall and F-measure. Usually these values are evaluated against gold standard for the task of NER. Gold standards are manually annotated data which contains correct label for each word in input. Given results are compared against gold standard manually. Value of F-measure is calculated from precision and recall value using formula:

$$\frac{2 * Precision * Recall}{Precision + Recall}$$

VII. Conclusion

In this paper an effective and robust NER system for Indian Tweets (English and Marathi) is proposed. To overcome the problem of tagged training dataset, a semi-supervised algorithm is implemented by combining CRF approach with KNN classifier. We first normalized tweets with gazetteers and dictionaries and then obtained named entities which gave better results. So this normalization acts as a preprocessing step for this system.

Results of ML based approach i.e. CRF is compared with rule based showing better performance of earlier one. CRFs used for implementation allows to utilize features like suffixes, prefixes easily and thus increases accuracy of labeling. These features also help to get NEs from Indian languages like Marathi. In addition to this, use of multiple manually created gazetteers has improved accuracy of our system. In future work, this task can be extended to extract events from Marathi tweets. Additionally it can be used by NLP applications like actor identification, relation extraction from Marathi books. Language or data in this case will be more formal and less effort will therefore required extracting entities.

References

[1] Sundheim, Beth M. "Overview of results of the MUC-6 evaluation." Proceedings of a workshop on held at Vienna, Virginia: May 6-8, 1996. Association for Computational Linguistics, 1996.

[2] Pingli, Prasad. "A Hybrid Approach for Named Entity Recognition in Indian Languages." IJCNLP, 2008.

[3] Sil, Avirup, and Alexander Yates. "Re-ranking for joint named-entity recognition and linking." Proceedings of the 22nd ACM international conference on Conference on information & knowledge management. ACM, 2013.

[4] Patel, Anup, Ganesh Ramakrishnan, and Pushpak Bhattacharya. "Relational learning assisted construction of rule base for Indian language NER." Proceedings of ICON 2009.

[5] Kumar, N. Kiran, G. S. K. Santosh, and Vasudeva Varma. "A languageindependent approach to identify the named entities in under-resourced languages and clustering multilingual documents." Multilingual and multimodal information access evaluation. Springer Berlin Heidelberg, 2011. 74-82.

[6] Finkel, Jenny Rose, Trond Grenager, and Christopher Manning. "Incorporating non-local information into information extraction systems by gibbs sampling." Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2005.

[7] Tkachenko, Maksim, and Andrey Simanovsky. "Named entity recognition: Exploring features." Proceedings of KONVENS. Vol. 2012. 2012.

[8] Li, Chenliang, et al. "Tweet Segmentation and its Application to Named Entity Recognition." Knowledge and Data Engineering, IEEE Transactions on 27.2, 2015. 558-570.

[9] Liu, Xiaohua, et al. "Named entity recognition for tweets." ACM Transactions on Intelligent Systems and Technology (TIST) 4.1, 2013.

[10] Ilina, Elena, et al. "Social event detection on twitter." Web Engineering. Springer Berlin Heidelberg, 2012. 169-176.

[11] Keretna, Sara, et al. "Classification ensemble to improve medical Named Entity Recognition." Systems, Man and Cybernetics (SMC), 2014 IEEE International Conference on. IEEE, 2014.

[12] Duan, Huanzhong, and Yan Zheng. "A study on features of the CRFsbased Chinese Named Entity Recognition." International Journal of Advanced Intelligence 3.2 2011. 287-294.

[13] Nadeau, David, and Satoshi Sekine. "A survey of named entity recognition and classification." Linguistic Investigations 30.1, 2007. 3-26

[14] Irmak, Utku, and Reiner Kraft. "A scalable machine-learning approach for semi-structured named entity recognition." Proceedings of the 19th international conference on World wide web. ACM, 2010.

[15] Finin, Tim, et al. "Annotating named entities in Twitter data with crowdsourcing." Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk. Association for Computational Linguistics, 2010.

[16] Ekbal, Asif, and Sivaji Bandyopadhyay. “A conditional random field approach for named entity recognition in Bengali and Hindi.” Linguistic Issues in Language Technology (LiLT) 2.1, 2009. 1-44.

[17] Srivastava, Shilpi, Mukund Sanglikar, and D. C. Kothari. “Named entity recognition system for Hindi language: a hybrid approach.” International Journal of Computational Linguistics (IJCL) 2.1, 2011.

[18] Singh, Jaskirat, Niranjan Joshi, and Iti Mathur. “Development of Marathi part of speech tagger using statistical approach.” Advances in Computing, Communications and Informatics (ICACCI), 2013 International Conference on. IEEE, 2013.

[19] Kumar, Dinesh, and Gurpreet Singh Josan. “Part of speech taggers for morphologically rich indian languages: a survey.” International Journal of Computer Applications (09758887) Volume, 2010. 1-9.

[20] Hakimov, Sherzod, Salih Atilay Oto, and Erdogan Dogdu. “Named entity recognition and disambiguation using linked data and graph-based centrality scoring.” Proceedings of the 4th international workshop on semantic web information management. ACM, 2012.

[21] Wu, Xixin, et al. “Adaptive named entity recognition based on conditional random fields with automatic updated dynamic gazetteers.” Chinese Spoken Language Processing (ISCSLP), 2012 8th International Symposium on. IEEE, 2012.