Named Entity Recognition in Tamil from Code Mix Social Media Text

Pattabhi, R K Rao and Sobha, Lalitha Devi

[1] AU-KBC Research Centre, MIT Campus of Anna University, Chennai, India
pattabhi@au-kbc.org

Abstract. The penetration of smart devices such as mobile phones, tabs has significantly changed the way people communicate. This has led to the growth of usage of social media tools such as twitter, facebook chats for communication. This has led to development of new challenges and perspectives in the language technologies research. The biggest challenge in social media text is code mixing. This paper presents our work on Named Entity Recognition (NER) from the Tweets which have Tamil – English code-mix. In this we describe how the corpus collection and annotation is done. NE system is developed using Conditional Random Fields (CRFs). We have obtained F-measure of 70.93% comparable with the state-of-the-art.

Keywords: Named Entity Recognition (NER), Twitter data, Code mix data, Tamil-English Code mix, Machine Learning, Conditional Random Fields (CRFs)

1 Introduction

In the last decade, Indian language content and especially Tamil on various media types such as websites, blogs, email, chats have 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. The growth of Indian language content is expected to increase by more than 70% every year. Hence there is a 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 recognizes the entities or the associations of them or relation between them. Hence an automatic Entity extraction system is required.

Entity extraction has been actively researched for over 20 years. Most of the research has, however, been focused on resource rich languages, such as English, French and Spanish. The scope of this work covers the task of named entity recognition in social media text (twitter data) for Indian languages. In the past there were events such as Workshop on NER for South and South East Asian Languages (NER-SSEA, 2008), Workshop on South and South East Asian Natural Language Processing (SANLP, 2010&2011) conducted to bring various research works on NER being done on a single platform. NERIL tracks at FIRE (Forum for Information Retrieval and Evaluation) in 2013, 2014 have contributed to the development of benchmark data

and boosted the research towards NER for Indian languages. All these efforts were using texts from newswire data. The user generated texts such as twitter and facebook texts are diverse and noisy. These texts contain non-standard spellings and abbreviations, unreliable punctuation styles. Apart from these writing style and language challenges, another challenge is concept drift (Dredze etal., 2010; Fromreide et al., 2014); the distribution of language and topics on Twitter and Facebook is constantly shifting, thus leading to performance degradation of NLP tools over time.

Some of the main issues in handling of social media texts such as Tweets are i) Spelling errors ii) Abbreviated new language vocabulary such as "gr8" for great iii) use of symbols such as emoticons/emojis iv) use of meta tags and hash tags v) Code mixing.

For example:

Ta: Stamp veliyittu ivaga ativaangi .....
En: stamp released these_people get_beaten ….

Ta: othavaangi .... kadasiya <loc>kovai</loc>
En: get_slapped … at_end kovai

Ta: pooyi pallakaatti kuththu vaangiyaachchu.
En: gone show_tooth punch got

("They released stamp, got slapping and beating … at the end reached Kovai and got punched on the face")

This example is a Tamil tweet where it is written in a particular dialect and also has usage of English words.

The research in analyzing the social media data is taken up in English through various shared tasks. Language identification in tweets (tweetLID) shared task held at SEPLN 2014 had the task of identifying the tweets from six different languages. SemEval 2013, 2014 and 2015 held as shared task track where sentiment analysis in tweets were focused. They conducted two sub-tasks namely, contextual polarity disambiguation and message polarity classification. In Indian languages, Amitav et al (2015) had organized a shared task titled 'Sentiment Analysis in Indian languages' as a part of MIKE 2015, where sentiment analysis in tweets is done for tweets in Hindi, Bengali and Tamil language.

Named Entity recognition was explored in twitter through shared task organized by Microsoft as part of 2015 ACL-IJCNLP, a shared task on noisy user-generated text, where they had two sub-tasks namely, twitter text normalization and named entity recognition for English.

The ESM-IL track at FIRE 2015 was the first one to come up with the entity annotated benchmark data for the social media text, where the data was in idealistic scenario, where users use only one language. But nowadays we observe that users use code mixing even in writing in the social media platforms. Thus, there is a need to develop systems that focus on social media texts. There have been other efforts on the code mix social media text in the applications of information retrieval (MSIR tracks at

FIRE 2015and 2016). The CMEE-IL track at FIRE 2016 came up with the entity annotated benchmark for code mix twitter data.

The paper is further organized as follows: The next section describes the challenges in named entity extraction from social media texts and in particular Twitter data. Section 3 describes corpus collection, annotation and statistics. Section 4 describes system development methodology. And section 5 discusses results.

2 Challenges in Social Media Named Entity Recognition

The challenges in the development of entity extraction systems for Indian languages from social media text arise due to several factors. One of the main factors being there is no annotated data available for any of the Indian languages, though the earlier initiatives have been concentrated on newswire text. We also find that development of automatic named entity recognition systems for twitter kind of data is difficult due to following reasons:

i) Tweets contain a huge range of distinct named entity types. Almost all these types (except for People and Locations) are relatively infrequent, so even a large sample of manually annotated tweets will contain very few training examples.

ii) Twitter has a 140-character limit; thus, tweets often lack sufficient context to determine an entity's type without the aid of background or world knowledge.

iii) In comparison with English, Indian Languages have more dialectal variations. These dialects are mainly influenced by different regions and communities.

iv) Indian Language tweets are multilingual in nature and predominantly contain English words.

3 Corpus Development and Annotation

The corpus was collected using the twitter API in different time periods. As explained in the above sections, in the twitter data we observe concept drift. Thus to evaluate how the systems handle concept drift we had collected data in two different time periods. Table 1 below shows the corpus statistics.

Table 1. Corpus Statistics

Language	**No. of Tweets**	**No. of NEs**
Tamil-English	4576	2454

The corpus was annotated manually by trained experts. Named Entity Recognition task requires entities mentioned in the document to be detected, their sense to be disambiguated, select the attributes to be assigned to the entity and represent it with a tag. Defining the tag set is a very important aspect in this work. The tag set chosen should be such that it covers major classes or categories of entities. The tag set defined should be such that it could be used at both coarse- and fine-grained level de-

pending on the application. Hence a hierarchical tag set will be the suitable one. Though we find that in most of the works Automatic Content Extraction (ACE) NE tag set has been used, in our work we have used a different tag set. The ACE Tag set is fine grained is towards defense/security domain. Here we have used Government of India standardized tag set which is more generic.

The tag set is a hierarchical tag set. This Hierarchical tag set was developed at AU-KBC Research Centre, and standardized by the Ministry of Communications and Information Technology, Govt. of India.

In this tag set, 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. The tag set consists of three level hierarchies. The top level (or 1st level) hierarchy has 22 tags, the second level has 49 tags and third level has 31 tags. Hence a total of 102 tags are available in this schema.

3.1 Data Format

The data with annotation markup is kept in a separate file called annotation file. The raw tweets as downloaded using the twitter API are kept as it is. The annotation file is a column format file, where each column was tab space separated. It consisted of the following columns:

i) Tweet_ID
ii) User_Id
iii) NE_TAG
iv) NE raw string
v) NE Start_Index
vi) NE_Length

For example:

Tweet_ID:123456789012345678
User_Id:1234567890
NE_TAG:ORGANIZATION
NE Raw String:SonyTV
Index:43
Length:6

Index column is the starting character position of the NE calculated for each tweet and the count starts from ‘0’.

4 Our Methodology

Named Entity Recognition (NER) is defined as the process of automatic identification of proper nouns and classifies the identified entities into predefined categories such as person, location, organization, facilities, products, temporal or numeric expressions etc. Even though named entity recognition is a well-established research filed and lot of research works are available for various languages, not much work has been done towards identification of NEs in code-mix social media text. The system architecture is shown in the below figure 1.

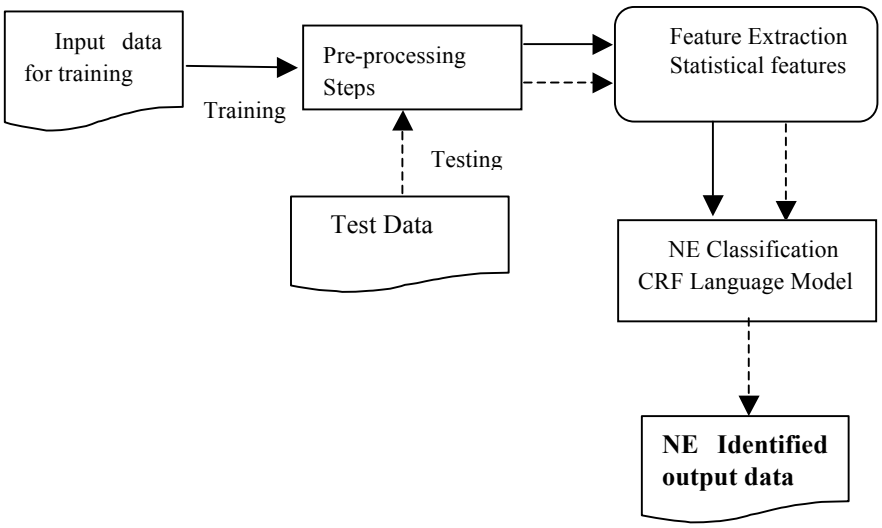

Fig 1. System Architecture – Process Flow Diagram

We analyzed the corpus to arrive at the most suitable word level features for identifying the NE which can be used for machine learning purposes. In Tamil we have POS tagger available but it is trained on Newswire text and not suitable for our task here. Since it is not suitable for our text, we have not used any syntactic processing in this work. We have used only statistical suffixes as features. And we have taken a window of three words for the training. The NER engine is developed using Conditional Random Fields (CRFs), a machine learning technique.

Conditional random fields (CRFs) are a probabilistic framework which is suitable for sequence prediction problem. It selects the label sequence y which maximizes the conditional probability of p(y|x) to the observation sequence x. The probability of a label sequence y given an observation sequence x is given below

$$P(y|x,\lambda) = \frac{1}{z(x)} \exp \sum_{i=1}^{n} \sum_{j} \lambda_j f_j(y_{i-1}, y_i, x, i)$$

$$z(x) = \sum_{i=1}^{n} \sum_{j} \lambda_j f_j(y_{i-1}, y_i, x, i)$$

Where x is the data sequence to be labelled and y is the label sequence. For example x is the range over sentences and y is the range over named entity tag, z is normalization factor, fj ($_{Yi}$-1, $_{Y}$i, x,i) is a state transition feature function of an observation sequence and the labels at position i and i − 1. For example, our objective is to assigning the named entity tag or label y "LOCATION" to the sentence x "He is in Finland", then the transition function fj($_{yi}$-1,$_{y}$i,x,i) = 1 if y_i="LOCATION" and the suffix of i_{th} word is "land"; otherwise 0; If the weight λ_j associated with the above feature is large and positive, then the words ending with the suffix "land" are labelled as NE type "LOCATION" (Lafferty, 2001; Wallach, 2004).

We have used CRF++ toolkit for this work. The system can learn the NE patterns in the training data with the help of features provided and the language model is generated. Named entities are identified in the test data by the named entity model file.

5 Results and Discussion

Entities share common prefixes and suffixes for a particular type of Named entity. For example, the words ending with "Ur" most likely denotes the location name such as "porur", "thanjavur" in Tamil. Hence, we consider bigrams and trigrams of prefix and suffix information as features. The features applied here are frequency based and is generic. Though the corpus is tagged with 3 level hierarchical tag set, we have only used the first level tags consisting of 22 tags and developed the NER engine in this work.

We performed a 10-fold experiment and obtained an average precision of 78.56% and a recall of 64.66%, which is comparable with the state of the art for NER in code mix social media text. Table 2 gives results summary.

Table 2. System Results – Average Scores

Language	Precision	Recall	F-measure
Tamil-English	78.56 %	64.66%	70.93%

Some of the main errors we found were as follows:
1) Named entities occurring in adjacent positions are tagged as single entity (two NEs combined as one NE)
2) one named entity with multiple tokens is tagged as two entities (Single NE split as two NEs)
3) NE boundary is not identified properly, beginning of an entity is tagged by intermediate tag, part of an entity is tagged by the system (as BIO format of tagging is followed by the system.
4) Organization names are not identified correctly.

6 Conclusion

We have presented a named entity system which can be used for identifying named entities in a code mix twitter data of Tamil – English. In our method, we have used generic statistical suffixes feature. The results obtained show that our feature is well suited. Error analysis shows there is need to include syntactic features of POS tagger and chunker to overcome boundary problems. In future, we plan to extend this work towards the development of POS tagger and Chunker suitable for Social media text.

References

1. José Ramom Pichel Campos, Iñaki Alegría Loinaz, Nora Aranberri, Aitzol Ezeiza, Víctor Fresno. 2014 TweetLID@SEPLN 2014, Girona, Spain, September 16th, 2014. CEUR Workshop Proceedings 1228, CEUR-WS.org 2014
2. Mark Dredze, Tim Oates, and Christine Piatko. 2010. "We're not in kansas anymore: detecting domainchanges in streams". In *Proceedings of the 2010 Conferenceon Empirical Methods in Natural LanguageProcessing*, pages 585–595. Association for Computational-Linguistics.
3. Hege Fromreide, Dirk Hovy, and Anders Søgaard.2014. "Crowdsourcing and annotating ner for twitter#drift". *European language resources distributionagency.*
4. H.T. Ng, C.Y., Lim, S.K., Foo. 1999. "A Case Study on Inter-Annotator Agreement for Word Sense Disambiguation". In *Proceedings of the {ACL} {SIGLEX} Workshop on Standardizing Lexical Resources {(SIGLEX99)}*. Maryland. pp. 9-13.
5. Preslav Nakov and Torsten Zesch and Daniel Cer and David Jurgens. 2015. Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015).
6. Nakov, Preslav and Rosenthal, Sara and Kozareva, Zornitsa and Stoyanov, Veselin and Ritter, Alan and Wilson, Theresa. 2013. SemEval-2013 Task 2: Sentiment Analysis in Twitter. *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*
7. Rajeev Sangal and M. G. Abbas Malik. 2011. Proceedings of the 1st Workshop on South and Southeast Asian Natural Language Processing (SANLP)
8. Aravind K. Joshi and M. G. Abbas Malik. 2010. *Proceedings of the 1st Workshop on South and Southeast Asian Natural Language Processing* (SANLP). (http://www.aclweb.org/anthology/W10-36)
9. Rajeev Sangal, Dipti Misra Sharma and Anil Kumar Singh. 2008. *Proceedings of the IJCNLP-08 Workshop on Named Entity Recognition for South and South East Asian Languages*. (http://www.aclweb.org/anthology/I/I08/I08-03)
10. Pattabhi RK Rao, CS Malarkodi, Vijay Sundar R and Sobha Lalitha Devi. 2014. Proceedings of Named-Entity Recognition Indian Languages track at FIRE 2014. http://au-kbc.org/nlp/NER-FIRE2014/
11. Wallach, H.M. (2004). Conditional random fields: An introduction Technical Reports (CIS), *MSCIS-04-21*
12. Lafferty, J., McCallum, A. & Pereira, F. (2001). Conditional Random Fields for segmenting and labeling sequence data. *ICML-01*. 1, 282-289.