

AMRITA_CEN @ FIRE 2015: Extracting Entities for Social Media Texts in Indian Languages

Anand Kumar M

CEN,

Amrita Vishwa Vidyapeetham,
Coimbatore, India

m_anandkumar@cb.amrita.edu

Shriya Se

CEN,

Amrita Vishwa Vidyapeetham,
Coimbatore, India

shriyaseshadrik.r@gmail.com

Soman KP

CEN,

Amrita Vishwa Vidyapeetham,
Coimbatore, India

kp_soman@amrita.edu

ABSTRACT

This contemporary work is done as a slice of the shared task on Entity Extraction from Social Media Text Indian Languages in Forum for Information Retrieval and Evaluation (FIRE2015). Nowadays people are extensively using social media platforms like Face book, Twitter, etc, to exchange their thoughts. The twitter messages are growing rapidly and their style and short nature present a new challenge in language technology field. This extensive amount of textual data is also increases the interest in Information Extraction (IE) on such textual data. Named entity extraction is one of the essential tasks in Information Extraction, aims to extract and classify entities from text. The performance of the present standard language processing tools is severely affected on Tweet messages. Hence, different improvised and non-improvised algorithms are necessary for extracting these entities from the informal text. This paper deals with the extracting the Named Entities from twitter messages of four Indian Languages. The extraction of the Named entity relies mainly on the domain specific features and conventional features. A well known supervised algorithm, Support Vector Machine (SVM) is used to extracting the entities.

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; Support Vector Machine (SVM); supervised algorithm;

1. INTRODUCTION

Named Entity Recognition (NER) is a task of extracting the named entities like people, organization and location etc. from a text. NER is an important task in Natural Language Processing (NLP) system for question answering and information extraction [1]. Twitter, a micro blogging platform is an important source for continuously updating the information. Twitter messages are 140 character length short texts which are noisy, informal nature and seldom contain an informative content. For extracting these information, many algorithms are proposed and in practice too. In general, tweets are based on various products or about different persons. Identifying the named entities in tweets is a difficult task for two reasons [2]. The first one is tweet contains the large number of distinct named entity types (except location and person), so it is difficult to find an adequate size of annotated

examples for each entity type. Another reason is due to the lack of sufficient context, it's challenging to determine the entity type.

Entity extraction task is important for companies to detect their crisis and to response in an earlier stage. By doing entity extraction in twitter data, a company can solve the crisis and can make earlier response which leads to improve their profit. The performance of the present standard language processing tools is severely affected on Tweet messages. So adapting the domain and applying that too for Indian Language is a challenging task. Normalizing the Tweet messages is also not advisable. Hence an improvised machine learning technique is needed for better accuracy. Few researchers treated entity extraction as two level tasks, which include segmentation and disambiguation. Segmentation is used to find out the entity boundaries and disambiguation assist to identify the entity type. There is no prior works on entity extraction for Twitter messages in Indian languages. Widely machine learning approaches are followed for entity recognition. Machine learning algorithm, SVM has been successfully applied to Indian language processing tasks which includes NER, sense disambiguation [3], morphological analyzer [4][5] and clause boundary identification [6]. Entities are extracted and linked with DBpedia source using machine learning algorithms [7]. MaxEnt based Hybrid NER system is tried for Indian languages [8]. SVM and CRF based entity recognition for Indian languages have been done with rich feature set [9] [10].

We converted the given format into sequence labeling format using BIO encoding for representing entities and uses a well known machine learning algorithm, Support vector machine for classification. We include POS tag, nearby context, orthographic and language specific features for all the four languages. In addition to that we used Brown clusters for English language.

The rest of this paper is composed of four sections. In Section 2, we discuss about the dataset descriptions and explained the size of training data and testing data. The methodology and the features are explained in Section 3. We report our experiment results in Section 4 and conclude the paper with future directions in Section 5.

2. DATA SET DESCRIPTIONS

Given the four Indian languages tweets as an input, the task is to extract the named entities and identify its type. The task organizers released the Twitter dataset and its entities for four languages namely Tamil, English, Hindi and Malayalam. There are 22 types of entities present in the training dataset in which the majority of entities from person, location and organization. The size of the training and testing data is shown in the Table.1 in which the most of the tweet contains URL and leading topics are

politics and cinema. The major issue in the dataset is that the size of the training data is lesser than the size of the test data (except for Malayalam) which leads to a drop in the accuracy level of the system. We have converted the tweets into the conventional BIO format and treated them as a sequence labeling problem. Table.1 also describes the token counts of training and testing data of all the four languages. Count of primary entities in the training dataset is shown in the Table.2.

Table 1. Tokens and Tweets Size

Language	Tweets Size		Tokens Size	
	Train	Test	Train	Test
English	5941	9595	100267	152356
Tamil	6000	8222	73640	103003
Malayalam	8426	4121	93311	45776
Hindi	7983	10752	140472	98976

Table 3. Primary Entities Count in Training Dataset

Language	Primary Entities			
	Person	Location	Org	Count
English	3145	2806	2470	572
Tamil	3291	2207	1373	576
Malayalam	2920	3156	919	742
Hindi	3346	1744	1060	565

3. SYSTEM DESCRIPTIONS

This section describes our system submitted in Entity extraction from twitter messages task in FIRE 2015. In the proposed Twitter NER system, feature extraction plays an important role, as the accuracy of the system rely on the extracted features. Generally, for text mining and information extraction, preprocessing is necessary and it is mandatory for the twitter dataset. Preprocessing includes Normalization and Tokenization. Generally the sentences are divided into tokens based on the white spaces. These tokens are further normalized where superficial variations are removed. But, normalization of Twitter messages aims to standardize the non-standard words, spellings variations, expand informal abbreviations (e.g., tmrw for tomorrow), and standardize phonetic substitutions. Here we avoided stemming and Twitter data normalization. Case folding is appropriate for English language where case variations are exist but it is not required for Indian Languages where no such variation are exists.

The methodology of the proposed system is depicted in the Figure.2. The training dataset contains two files which are raw tweets and type annotated extracted entities. The tweet file consists of “Tweet ID”, “UserID” and Tweets. The entity file consists of “Tweet ID”, “UserID”, Entity Type, Entity, Starting Index and length. We have combined these files and converted into the conventional BIO formatted text in which B-XXX tag indicates the beginning word of an entity type XXX and I-XXX is used for the following chunks of an entity. The tag other than named entity is tagged as O. In tokenization the tweets are further

divided into small chunks called as tokens. Training and testing tweets are tokenized. Given named entities and tokenized training tweets are merged to create BIO format in one token-per line fashion. The BIO formatted sequence data is then tagged using POS tagger of the respective languages. For Tamil Language in house Twitter POS tagger and Amrita POS tagger [5] is used. Hindi NLTK POS tagger is applied for Hindi tweets and for Malayalam, Amrita POS tagger is used and for English the tokens is tagged using English Twitter POS tagger developed by Gimbel et.al.[12]. Features are illustrated in Table.4 and explained in Subsection 3.1. Features are extracted for each language and train the system with support vector machine based classifier. We trained our system using the SVM based tool “SvmLight” [13]. Finally, the BIO format tokens are converted into the given annotation format.

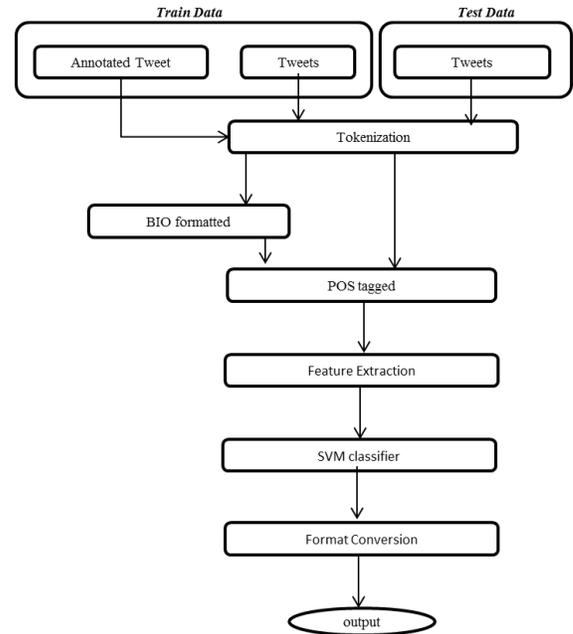


Figure 1. Overview of the system

3.1 Feature Extraction

In this work, feature extraction is given more importance as this decides the accuracy of the machine learning based system. The conventional features like, words, POS tag, prefixes and suffixes of the word, binary feature, shape feature are used. Binary feature and shape feature are the feature if it is present in the tweet then it is marked as ‘0’ or else ‘1’. For prefix and suffix feature maximum up to five characters before and after the current character is taken as feature. The punctuation mark such as question mark, exclamatory marks, comma, and full stop are also taken as feature. For training the system the current token, the token which are above and below are also taken as feature as these will decide the current word. Detailed feature set is illustrated in Table.3. In the following we have discussed about the features we have used to develop the Twitter NER system for Indian languages.

Word Features: The current word is used as feature. Additionally, to provide context information, the previous and next words of a current word are also used as features (w_{-1} , w_0 , w_{+1}).

Named Entities: Named Entity tag of the previous word, next words are used as features (t_{-1} , t_{+1}).

POS Tags: POS tags contain the essential information for identifying and disambiguating named entities. The current word’s part-of-speech tag, the previous and next words (p_{-1} , p_0 , p_{+1}) part-of-speech tags are used as features.

We have used the Twitter POS tagger developed by Gimbel et.al.[12] for English language. Twitter POS tagger does not exist for Indian languages, so we have used the standard POS taggers except for Tamil system (Run-2 and Run-3). Malayalam POS tags are retrieved from the in house Malayalam POS tagger. NLTK Hindi POS tagger is used for tagging the Hindi tweets. For Tamil language (Run-1), we have used our standard Tamil POS tagger [5] and Twitter POS tagger. We have developed a Twitter POS tagger for Tamil with 400 manually annotated tweets collected from SAIL-2015 neutral dataset [14].

Table 3. Feature set Details

Features	Symbols
Word Features	w_{-1}, w_0, w_{+1}
Named entities	t_{-1}, t_{+1}
POS tags	p_{-1}, p_0, p_{+1}
Prefixes and Suffixes	$P_3, P_4, P_5, S_3, S_4, S_5$
End features	$., ! ?$
Hash, Numbers, Punctuation and Acronym with periods	H, N, PU, AP
Language Specific features	LS
Length	l
Position	pos
Brown Cluster(Only for English)	CL
Capital, All Capital, Contains Apostrophe(Only for English)	C, AC, AP

Language	Hindi			Tamil			Malayalam			English		
	P	R	F	P	R	F	P	R	F	P	R	F
Run1	71.56	54.09	61.61	55.23	11.03	18.39	51.18	40.29	45.08	58.78	40.73	48.11
Run2	-	-	-	61.55	19.82	29.98	-	-	-	-	-	-
Run3	-	-	-	60.82	19.42	29.44	-	-	-	-	-	-

5. CONCLUSIONS AND FUTURE SCOPE

In this paper, we have presented an approach to identify and extract the named entities on the Twitter messages in Indian languages. Due to the huge volumes of rapidly increasing tweets, text mining applications like entity extractions are immediately required to accomplish language processing tasks in informal short texts. This entity extraction is highly helpful in the field of information extraction and trend analysis. Here we have extracted the different types of named entities, such as person, location, organization, etc. We have converted the training data into sequence labeling format and we extracted rich language specific

Prefixes and Suffixes: Prefix and suffix information are helpful to identify the named entities, particularly for unknown words. The first and last n characters are used as features; characters are ranging from 3 to 5 ($P_3, P_4, P_5, S_3, S_4, S_5$).

End features: Here, we have used four different kind of features ($., ! ?$). If the token ends with dot or comma or exclamatory or question mark, then the corresponding feature is set to 1. Otherwise, it is set to 0. For example, for the token “*help!*” the feature is “0010”.

Non-alphabet features: If a token contains hash (#) or any digit or punctuation or acronym with periods, then the corresponding feature is set to 1. Otherwise, it is set to 0. For example, for the token “*B.J.P*” the feature is “0011”.

Other features: Length of the token and the position in current tweet is considered as features for all the four languages. Language Specific features are used only for Tamil and English.

For Tamil, we check whether the token contains ‘ ஹ ’ or ‘ ஹ .’

followed by digit and any digit followed by - ஹ , - ஹ , ஹ , ஹ . For English, we consider the digit followed by “*st*” “*th*”. In addition to that for English language we have used brown cluster (size of 10) and orthographic features like Starts with capital, all capital and contains apostrophe as features.

3.2 Run Submissions

We have participated in all the four Indian languages and submitted our outputs to organizers.

Run-1: All the four languages with aforesaid features and SVM training without Validation.

Run-2: Only Tamil system has been submitted. In feature extraction, instead of standard POS tagger, we have used our own Twitter POS tagger which is developed with small amount of manually annotated examples.

Run3: Only Tamil system has been submitted. Tamil Twitter POS tagger has been used. In addition to that, in order to tune the SVM parameters, we have taken first 1000 tokens (from BIO format) as development set and train the system accordingly.

4. RESULTS

features for training the SVM classifier. We would also like to perform a through error analysis to understand the issues, so we can easily address them in future. We want to invest time in crafting the informal text features to improve Twitter NER for Indian languages. Other directions includes, using Twitter POS tagger for Indian languages, adding entity specific prefix suffix features and word embeddings with unlabeled Twitter data.

6. ACKNOWLEDGMENTS

First and foremost, we would like to thank Forum of Information Retrieval and Evaluation (FIRE 2015) organizers for organizing a wonderful research evaluation workshop and giving opportunities for researchers to present their work on Text Analytics and Natural Language Processing (NLP). We also like to thank

Computational Linguistics Research Group (CLRG), AU-KBC Research Centre, for organizing the Entity Extraction from Social Media Text Indian Languages (ESM-IL) Task.

7. REFERENCES

- [1] Toh, Zhiqiang, Bin Chen, and Jian Su. 2015. Improving twitter named entity recognition using word representations Shared tasks of the 2015 workshop on noisy user-generated text: Twitter lexical normalization and named entity recognition. *ACL-IJCNLP 2015*.
- [2] Ritter, Alan, Sam Clark, and Oren Etzioni. "Named entity recognition in tweets: an experimental study." *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2011.
- [3] Anand Kumar, M., Rajendran, S., Soman, K.P. Tamil word sense disambiguation using support vector machines with rich features (2014) *International Journal of Applied Engineering Research*, 9 (20), pp. 7609-7620.
- [4] Anand Kumar, M., Rajendran, S., & Soman, K. P. AMRITA@ FIRE-2014: Morpheme Extraction for Tamil using Machine Learning. Working notes in FIRE 2014 – MET Task, 2014
- [5] Anand Kumar, M., Dhanalakshmi, V., Soman, K.P., Rajendran, S. Factored statistical machine translation system for English to Tamil language (2014) *Pertanika Journal of Social Science and Humanities*, 22 (4), pp. 1045-1061
- [6] Dhivya, R., Dhanalakshmi, V., Anand Kumar, M., Soman, K.P. Clause boundary identification for Tamil language using dependency parsing (2012) *Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering*, 62 LNICST, pp. 195-197.
- [7] Barathi Ganesh, H.B., Abinaya, N., Anand Kumar, M., Vinayakumar, R., Soman, K.P. AMRITA - CEN@NEEL : Identification and linking of twitter entities, (2015) *CEUR Workshop Proceedings*, 1395, pp. 64-65.
- [8] Saha, Sujan Kumar, et al. "A hybrid approach for named entity recognition in indian languages." *Proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian Languages*. 2008.
- [9] Abinaya.N, Neethu John, Anand Kumar.M and Soman.K.P. Amrita@fire-2014: Named entity recognition for Indian languages. Working notes in FIRE 2014 – NER Task, 2014.
- [10] Abinaya.N et.al., 2014. AMRITA@FIRE-2014 Named entity recognition for Indian languages using rich features. In *Proceedings of the 6th 2014 Forum on Information Retrieval Evaluation (FIRE '14)*, ACM.
- [11] Malarkodi, RK Rao Pattabhi 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, COLING 2012, Mumbai, India*.
- [12] Kevin Gimpel , et.al, , Part-of-speech tagging for Twitter: annotation, features, and experiments, *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers*, June 19-24, 2011, Portland, Oregon
- [13] T. Joachims. SvmLight: Support vector machine. 1999, <http://svmlight.joachims.org/>, University of Dortmund, 19(4).
- [14] Braja Gopal Patra, Dipankar Das, Amitava Das, and Rajendra Prasath. Shared Task on Sentiment Analysis in Indian Languages (SAIL) Tweets - An Overview. In the proceeding of the Mining Intelligence and Knowledge Exploration (MIKE-2015).