Twitter Named Entity Recognition for Indian Languages

Malarkodi C.S., Sobha Lalitha Devi

AU-KBC Research Center, MIT Campus of Anna University, Chromepet, Chennai, India

{csmalarkodi,sobha}@au-kbc.org

Abstract. Social media is considered as the great source of communication, as millions of users share their opinions about the products, celebrities, movies etc... in the social media sites such as twitter, face-book and other discussion forums. Besides that, nowadays people are communicating in their mother tongue to exchange their ideas. This has lead to a rise of tweets in Indian languages. There is a great demand from business and commercial perspective, to extract potential information from such informal, noisy and unstructured data. In order to transform this informal content into useful information, NLP applications like entity extraction, relation extraction, and sentiment analysis requires to be developed for social media content. This paper deals with one such task of NLP, named entity recognition for Indian languages and English for social media text. The corpus utilized for our work is obtained from FIRE 2015 shared task. The challenges of entity extraction in Indian languages for twitter messages are discussed in detail with examples. We have developed an entity extraction system for Indian languages Hindi, Tamil, Malayalam and English using machine learning technique CRFs. The performance of the proposed work is compared with the systems submitted in the FIRE NER shared task 2015.

Keywords: NER, named entity recognition, twitter ner, Indian Language tweets, CRF, NER social media

1 Introduction

Named Entity Recognition (NER) or entity extraction refers to the task of identify the names of entity mentions and classify it into predefined categories such as person, location, organization, products etc... Entity identification is the core component of Information Extraction (IE), Machine Translation, Question & Answering (Q&A) systems. The association between named entities has been identified in the relation extraction system, sentiment analysis identifies the opinion about the entities such as person, organization or the product and in Q&A systems named entities are the answer strings to the 'WH' questions. Hence identification of named entities acts as a fundamental task in NLP applications. Though the research about named entities is well known for the past two decades, entity extraction in social media content has gained attention only in the recent days. With the emerging trend of social media platform, nowadays millions of people exchange their opinions; share the posts in the various social networking sites such as twitter, face-book, LinkedIn, microblogging and other forums. Across the social media sites, users are discussing their opinions about the product that has been purchased, the strength and weakness of the political parties, economical growth and current affairs. The public and private organizations need to extract the potential information that is helpful in improving their decision making, productivity from such informal, noisy and unstructured social media content.

Over the past two decades, research works and shared task about NER has conducted on newswire documents. Recently researchers have started to focus on social media text for NER and other NLP tasks. In contrast to Indian Languages, several works on social media content has been performed in English. In order to concentrate on social media text, various shared task has conducted in English language. The shared task for language identification in Tweets (tweetLID) has held as part of SEPLN 2014. The distributed dataset consists of tweets belongs to six languages [13]. The shared task for sentiment analysis of tweets has conducted as part of SemEval shared task [7]. The shared task for analyzing sentiments' on IL tweets has organized as part of MIKE 2015. The languages used are Hindi, Bengali and Tamil [5]. Akthar et.al. has participated in the ACL 2015 workshop on noisy user-generated text. They have developed a Differential Evolution (DE) based Named Entity Recognition (NER) system for twitter text. They have used CRF as a baseline classifier and used DE based technique to determine the relevant features and context information [1].

In order to contribute a benchmark corpus for NE on IL social media text and motive researchers to work on IL informal text, Entity Extraction from Social Media Text in Indian Languages (ESM-IL) has organized as part of Forum for Information Retrieval Evaluation (FIRE) 2015 workshop. The shared task focused on 3 Indian Languages namely Hindi, Tamil, Malayalam and English. The corpus consists of 106 hierarchical NE tags. There were seven teams participated in the task. Except one team, rest of them is participated in English and Hindi languages and 3 teams submitted the test runs Tamil and 2 for Malayalam [6].

Pallavi et.al used CRFs to build the NE language model. The features used are POS, chunk, unigram, bigram and trigrams of statistical Suffixes and prefixes. The data was first pre-processed to remove URLs and emoticons and then for POS and chunk tagging. They have used open source NLP tools such as "patter.en" for English and nltr for Hindi. They have worked on 3 languages namely Hindi, Tamil and English and submitted 3 test runs for Hindi and 2 test runs for English and Tamil [4]. Sarkar et.al built the NE system using machine learning technique HMM. The data was preprocessed with POS information and applied it as a feature for training the system. Gazetteer lists created by semi-manual efforts was also used in the work. They had participated only in English language [9]. Shriya team applied SVM for NE identification. The data was preprocessed with POS and chunk information. They have used in house POS and chunk tagger for Malayalm and Tamil languages. The features used are word window of 3, POS, chunk, capitalization, statistical suffixes and prefixes and brown clusters. They have participated in all the languages [10].

Sanjay et.al used CRFs for training the NE system. The data was preprocessed with POS and Chunk information; they have used in house tools for Tamil and Malayalam languages and for English and Hindi open source NLP tools were used for preprocessing task [8]. Chintak et.al. Applied machine learning technique CRFs for system development and used Genia tagger for preprocessing task. The features used are POS, chunk and gazetteer information and other heuristic information [2]. Vira et.al also used CRFs for the NE identification. The data was preprocessed with Stanford POS tagger. The features used are word window of 5, POS information, statistical suffixes and prefixes [13]. Sombuddha et.al. used four ML techniques and developed four systems for English language. They have used word, POS information, capitalization, numeric, hash tags and dictionary as features for their system development [11].

2 Challenges of twitter text in NE Identification

In this section we have discussed the common challenges of NE identification for tweet data across languages. The challenges specific to Indian language tweets are also discussed in the sub section.

Less contextual Information.

Due to the 140 character size limitation, people can post the tweets as much shorter with less contextual information, which creates more ambiguous entities and makes the entity identification difficult.

Repeated Characters.

Social media allows users to post in free flow of language which leads to repeated characters in between or at the end of the word. For example, “sooo gooood”

Punctuation marks.

Though there are few tokenizers, pos tagger available for English twitter data, still twitter messages becomes a great challenge for other languages due to the improper punctuation marks. Tweets consist of emoticons, emphasis markers and other symbols in between named entities. Besides that, two or more named entities can occur together without space due to size restriction.

Short forms and misspellings.

The size limitation makes user to discover many non-standard short forms which cause many variants of entities. Tweets also has spelling errors, as users can type as they wish which a human can understand, but it is hard for the system to recognize.

Partially Completed NEs or drop out NEs.

Some of the NEs are mentioned partially in the twitter messages, which cause many variants of the particular entity.

Challenges specific to Indian Language Tweets.

Code Mixing.

Code-mixing refers to the mixing of two or more languages in the text or conversation. Indian language tweets are multilingual in nature. We have observed two types of code-mixing in Indian Languages. English words occurring in IL tweets and in some other instances English words are transliterated in IL.

Spoken form and Dialectal variation.

In comparison with English, Indian languages has lot of variations between written and spoken form. Spoken language affects the grammar, vocabulary and pronunciation also. From our observation 75-80% of the IL tweets are in spoken language.

3 Corpus Description

In 2015 AU-KBC have organized the Entity extraction in social media text track for Indian languages (ESM-IL) in the Forum for Information Retrieval Evaluation (FIRE). The main objective of the shared task was the creation of benchmark data for Entity Extraction in Indian language Social Media text. The corpus utilized for this work is obtained from the shared task. The training and test corpus statistics is given in the Table 1 & 2.

Table 1. Training data Statistics

Language	No. of Tweets	No. Of words	Total No. of NEs
English	5,941	1,38,049	11,561
Hindi	7,983	1,64,045	8,905
Malayalam	8,426	1,23,545	10,503
Tamil	6,000	1,03,930	9,818

Table 2. Test data Statistics

Language	No. of Tweets	No. of words	Total No. of NEs
English	4,259	1,07,648	10,192
Hindi	8,281	1,75,601	10,300
Malayalam	4,121	63,008	5,739
Tamil	3,905	71,315	7,794

4 Our Methodology

We have used machine learning technique CRFs for NE identification. CRF is a graphical model which defines a log-linear distribution over labelling sequence given the corresponding observation sequence. It is a probabilistic framework conditioned on the random variable X which refers to the observation sequence. CRF is well suited for sequence labelling task because of its advantages over HMM and MEMM. HMM has dependency problem, as the current one depends on the previous label. MEMM suffered from labelling bias problem. While CRFs has overcome both the labelling bias and dependency issue. By using the contextual information represented in the training data, the unknown entities can identified by CRFs. The features are learned from the training data and the model file was generated.

4.1 Feature Selection

We have used the contextual information and affix information as the features for entity identification. The features we have used is generic for all the languages.

Lexical level Features.

- Context word: Current word and the contextual words in the window of five is considered as a feature.
- Affix Information: Prefix and suffix information helps to identify the entities. So we have considered the statistical prefixes and suffixes up to length five as feature.

Digit Features.

We have used the two digits related features namely presence of digits and presences of four digits.

- Presence of digits: If the token consists of digits then the boolean value 1 is assigned to the respective feature.
- Presence of four digits: If the token is a 4 digit number, then the feature is triggered otherwise it is set to 0.

Orthographic features.

- Acronyms: If the token consists of period in between words, the respective feature is set to 1. This feature helps to find the political parties and other organization names.
- Presence of hash tag: If the token consists of hash tag at the beginning, then the feature is assigned to the Boolean value 1. If the tweet is about the particular per-

son, location or the organization then the token with hashtag might be a named entity.

5 Experiments Results & Discussion

The data we have used is obtained from 2015 NER shared task. The training and development data provided in the task are used for system development. The features used are lexical level information such as word, statistical suffixes and prefixes of 5,4, and 3 characters. The performance of the system is tested with the test data provided in the task and the results for different combination of features for each language are given in Table 3.

The system has obtained 40.56 f-m for English with contextual word information, addition of affixes improves the performance by 7%, digit features increases it further by 3% and orthographic features raised the f-m to 52.17%. The baseline system for Hindi scored the f-m value of 47.1%, inclusion of affixes leads to the improvement of F-M value by 10%, the digit and orthographic features boosted the performance by 3%. For Malayalam, use of context word yields 31.24% f-m, inclusion of affixes, digit and orthographic features improves it further by 18%. The baseline system for Tamil achieved the f-m value of 19.05% and the affix information along with the content word scored f-m value of 37.98% and incorporation of digit and orthographic features improved the results and achieved the f-m value of 41.33%. From our observation, affix information boost the performance by 7% to 10% for all the languages. The usage of digit and orthographic features raised the accuracy by 1% to 3% for IL and English.

Table 3. Feature-wise performance

Language	Features	Precision	Recall	F-Measure (F-M)
English	A=Contextual word	71.54	28.01	40.56
	B=A+Affixes	72.20	36.00	48.04
	C=A+B+Digit	73.11	39.20	51.03
	A+B+C+Orthographic	74.33	40.19	52.17
Hindi	A=Contextual word	73.05	34.81	47.1
	B=A+Affixes	74.30	47.21	57.73
	C=A+B+Digit	75.21	49.26	59.52
	A+B+C+Orthographic	76.20	50.13	60.48
Malayalam	A=Contextual word	62.58	20.82	31.24

	B=A+Affixes	64.32	36.40	46.49
	C=A+B+Digit	65.12	39.02	48.79
	A+B+C+Orthographic	66.75	39.27	49.45
Tamil	A=Contextual word	56.50	11.46	19.05
	B=A+Affixes	62.32	27.32	37.98
	C=A+B+Digit	64.11	29.23	40.15
	A+B+C+Orthographic	65.09	30.28	41.33

We have compared the performance of the present work with the highest scores reported in FIRE task. The F-score value obtained by our system and FIRE task are given in Table 3. For English we have obtained 52.17% f-m which is 4% highest than the results reported by [9]. They have used the machine learning technique HMM and applied POS tag and gazetteer list as features. The results reported by Shriyaet.al are 1% highest than the present system. The features used for SVM learning are POS, Chunk, Statistical Suffixes and prefixes, capitalization, Gazetteer list and Shape features. The present work scored better than the performance reported by Sanjay et.al. They have used CRF for the system development. The features used by [4] for CRF training are POS, chunk, statistical prefixes and suffixes, the results reported by them is 8% lesser than the present work. Except [4] rest of the systems achieved highest score in FIRE task has used gazetteer list as one of the features. Without any external resources and list our system achieved the state-of-art performance.

Table 4: comparison with FIRE shared task 2015

Language	Our F-score	FIRE TASK Highest F-score
English	52.17	48.21
Hindi	60.48	61.61
Malayalam	49.45	47.97
Tamil	41.33	32.91

6 Error Analysis

In order to know where our system fails to identify entities, we went through the system output of all the languages. From our analysis the errors in the social media text are arisen due to the following reasons.

1. Space drop out

2. Code-mixing
3. Less Contextual Information
4. Nonstandard acronyms

6.1 Space drop out

In few instances there is no space between NE strings. In twitter text two or more words in a single NE can come together without space.

Example 1.
Ta: benkaloor vijaymakkaliyakkam caarpil, vijay avarkalin katavutukku
En: bangalore vijaypeoplemovement on behalf of vijay's cutout
Ta: 1000kg lilli malaril aeladukku maalai aNivikkappattathu
En: 1000kg lily flower seven layers garland worn
(On behalf of Bangalore's Vijay people movement, seven layer garland made of 1000kg Lily flowers was worn to Vijay's cutout)
In the above example "vijaymakkaliyakkam (vijay fans club)" is a three token string, but it was mentioned as a single token. As "vijay" is the person name the system wrongly tagged the entire string as "PERSON".

Example 2.
Ta: aNNAvaimuthalilcanthithu napikaL nAyakam vilYYAvilthaan-karunaanithi
En: Annaafirstmet napikal nayakam function-karunanithi
(First met Anna only in Prophet Muhammad's function-Karunanidhi)

In example 2, the string "aNNAvaimuthalilcanthithu (First met Anna)" consists of 3 words where "aNNA" is a name of the person, due to space drop out issue, 3 individual words joined together and occurred as a single token. Hence the system failed to identify the entity "aNNA".

6.2 Code-mixing:

Mixing of two or more languages in a tweet is known as code mixing.

Example 3.
Ta: ataiyAr theosophical societyyin Annie Besantkku inru piRantha nALAm.
En: Adayar theosophical society Annie Besant+has today birthday
(Today Annie Besant of Adayar theosophical society has birthday)

In example 3, the entities "theosophical society" and "Annie Besant" are written in English and rest of the words is in Tamil. The code-mixed entities are not identified by the system.

6.3 Less Contextual Information

Example 4.
Ta: mOdi inthiyAvin puli-thamilYYasai

En: Modi India's Tiger- Tamilisai
(Modi is India's Tiger-Tamilisai)

Example 5.
Ta: Puli padaiththa caathanai
En: Tiger done achivement
(Tiger's greatest achievement)

Due to the character limitation, tweets are more ambiguous and posted with less contextual information. In example 3, “mOdi” Modi is being compared to Tiger “puli”, but in example 4, “puli” is not a name of an animal, it refers to a movie name. But there is no context information in the tweet to identify “puli” is a movie name. Hence less contextual information in twitter text create more ambiguity.

Example 6.
Ml: ediye … ente utuppu teechcho? illa
En: hey … my dress ironed? no
Ml: chetta … raavile_tanne engottaa?
En: brother … morning_itself where?
Ml: tekkati …. teechchaale parayullo?
En: hey_iron_it … only_after_ironing tell?
(Hey did you iron my dress? No… brother morning itself where are you going? Hey iron it … only after ironing you will tell?)

In example 6, the tweet is posted in spoken form. In the example, “tekkati” refers to “iron it”, but “tekkati” is also a place name. Due to the less contextual information, system wrongly tagged it as location.

6.4 *Nonstandard acronyms.*

Example 7.
Ta: ooththikichu upA ., kapilcipAl
En: Flop+v UP+N kapilsibal+N
(Flop Uttar Pradesh ., Kapilsipal)

Non standard acronyms in the twitter text, makes the system hard to identify entities. In example 7, “upA,” denotes the state name “Uttar Pradesh”. Spelling mistakes and non-standard abbreviations creates several forms for the same entity which poses a greater challenge for NE identification.

7 Conclusion

In this work we have developed a NE identification system for IL twitter text and also for English. The challenges of classify named entities across languages and challenges specific to Indian languages are also discussed. The features used are generic to all the languages. The performance of the present system is compared with the best systems reported in FIRE-2015 NER shared task. No other external resources or gazetteer list are used for the system development. The point to be noted is without using any list, other resources and language specific features we have achieved state-of-art performance. In future we will focus on code mixing tweets.

References

1. Akhtar, M.S., Sikdar, U.K., Ekbal, A.: Iitp: Multiobjective differential evolution based twitter named entity recognition. In: ACL-IJCNLP 2015, pp. 61, (2015)
2. Chintak, M., Memon, M.R., Manthan, R., Sandip M.: Entity Extraction from Social Media Text Indian Languages (ESM-IL). In FIRE Workshops, pp. 100-102, (2015)
3. Kudo, T.: CRF++, an open source toolkit for CRF. http://crfpp.sourceforge.net, 2005.
4. Pallavi, K.P., Srividhya, K., Victor, R.R.J., and Ramya, M.M.: HITS@ FIRE task 2015: Twitter based Named Entity Recognizer for Indian Languages. In: FIRE Workshops, pp. 81-84, (2015)
5. Patra, Braja G., Dipankar, D., Amitava, D., Rajendra P.: Shared task on sentiment analysis in Indian languages (sail) tweets-an overview. In: International Conference on Mining Intelligence and Knowledge Exploration, Springer, pp. 650-655, (2015)
6. Pattabhi, R.K., Malarkodi, C. S., Vijay S.R., Lalitha Devi, S.: ESM-IL: Entity Extraction from Social Media Text for Indian Languages@ FIRE 2015-An Overview. In: FIRE Workshops, pp. 74-80, (2015)
7. Preslav, N., Sara, R., Zornitsa, K., Veselin, S., Alan, R., Theresa, W.: SemEval-2013 Task 2: Sentiment Analysis in Twitter. In: seventh international workshop on semantic evaluation (semeval 2013), pp. 312-320, (2013).
8. Sanjay, S.P., Anand K.M, and Soman, K.P.: AMRITA_CEN-NLP@ FIRE 2015: CRF Based Named Entity Extractor For Twitter Microposts. In: FIRE Workshops, pp. 96-99, (2015)
9. Sarkar, K.: A hidden markov model based system for entity extraction from social media English text at fire 2015. In: FIRE Workshops, arXiv preprint arXiv: 1512.03950, (2015)
10. Shriya, S., Soman, K.P.: AMRITA_CEN@ FIRE 2015: Extracting Entities for Social Media Texts in Indian Languages, In: FIRE Workshops, (2015)
11. Sombuddha, C., Somnath, B., Sudip, K.N, Paolo, R., Sivaji, B.: Entity Extraction from Social Media using Machine Learning Approaches. In FIRE Workshops, pp. 103-106. (2015)
12. Usbeck, R., Röder, M., Ngonga Ngomo, A. C., Baron, C., Both, A., Brümmer, M., and Ferragina, P.: GERBIL: general entity annotator benchmarking framework. In: 24th International Conference on World Wide Web, pp. 1133-1143, (2015)
13. Vira, B., Anjana P., Amit G.: Vira@ FIRE 2015: Entity Extraction from Social Media Text Indian Languages (ESM-IL). In: FIRE Workshops. (2015)
14. Zubiaga, A., San Vicente, I., Gamallo, P., Campos, J. R. P., Loinaz, I. A., Aranberri, N., Fresno-Fernández, V.: Overview of TweetLID: Tweet Language Identification at SEPLN 2014. In: TweetLID@ SEPLN, pp. 1-11, (2014)