AMRITA_CEN@FIRE 2016: Code-Mix Entity Extraction for Hindi-English and Tamil-English Tweets

Remmiya Devi G, Veena P V, Anand Kumar M and Soman K P Centre for Computational Engineering and Networking (CEN) Amrita School of Engineering, Coimbatore Amrita Vishwa Vidyapeetham, Amrita University, India

ABSTRACT

Social media text holds information regarding various important aspects. Extraction of such information serves as the basis for the most preliminary task in Natural Language Processing called Entity extraction. The work is submitted as a part of Shared task on Code Mix Entity Extraction for Indian Languages(CMEE-IL) at Forum for Information Retrieval Evaluation (FIRE) 2016. Three different methodology is proposed in this paper for the task of entity extraction for code-mix data. Proposed systems include approaches based on the Embedding models and feature based model. Creation of trigram embedding and BIO tag formatting were done during feature extraction. Evaluation of the system is carried out using machine learning based classifier, SVM-Light. Overall accuracy through cross validation has proven that the proposed system is efficient in classifying unknown tokens too.

CCS Concepts

•Information systems \rightarrow Information extraction; •Theory of computation \rightarrow Support vector machines;

Keywords

Word embedding, Machine Learning, Support Vector Machine (SVM), Code-Mix, Entity extraction

1. INTRODUCTION

Entity extraction has always been the most primary task in Natural language processing. It is defined as a task of extracting the named entities from any text. Generally, entities fall under the categories of name, person, and organization. It extends to date, time, period, month etc. Entity extraction in social media text is viewed as an information extraction task. Social media text is generally unstructured, yet it is informative. Extracting such informative content from an unorganized text format is the most challenging task. In our task we deal with social media text, specifically code-mix twitter dataset. In a society using multilingual languages, conversation in code mixed language is prevalent. Code-mix language is the combination of English language with any other language. An example for code mixed language from the Hindi-English training data is given below.

shaq hai ki humari notice ke bagair tere ghar ke secret route ki help se you met

In the above example, Hindi words are in italics and English words are in bold. Communication through social networking sites likes Facebook and Twitter are in code-mix language. Dataset given for this task includes two subsets, where Indian languages like Hindi and Tamil were mixed with Roman script. The major task is to develop a method that is applicable to process and extract the entities of the data in code-mix language.

In recent years, a significant number of researches were carried out in the field of data processing using code-mix data. A language identification task was carried out for code-mix social media data [5]. A paper was published on thematic knowledge discovery using topic modeling for chat messages in code mixed language [4]. SVM based classification for entity extraction was carried out previously for Indian languages [3]. Conditional Random Field (CRF) based entity extraction was implemented and Rich features of Indian Languages were also utilized to perform Named Entity Recognition [10] [2]. Entity extraction using Structured skip gram based embedding features was implemented for Malayalam language [9].

Our submission includes three systems. First system is using word embedding features obtained from wang2vec tool [13]. Word embedding features are generally vector representation of words. The second system utilizes word embedding features from word2vec tool [7]. The major difference between wang2vec and word2vec features lies in the Skip gram model used to develop these embedding features. In third system, stylometric features were extracted from the training data. Extracted features from system 1, 2 and 3 are used to develop three separate models using machine learning based classifier SVM-Light [6].

An overview of the task description is given in Section 2. Details regarding the dataset used is given in Section 3. Section 4 discusses on the proposed system we used for the task. Experiments carried out and their corresponding results are discussed in Section 5. The conclusion of the paper is stated in Section 6.

2. TASK DESCRIPTION

The task organizers provided us with dataset obtained from Twitter and other few microblogs. Given training data contains two set of dataset with code mixed tweets - Hindi-English and Tamil-English. The task is to extract the entities from these two dataset. Named entities in the dataset include Person, Location, Organization, Entertainment and so on. With the increasing number of social media platforms

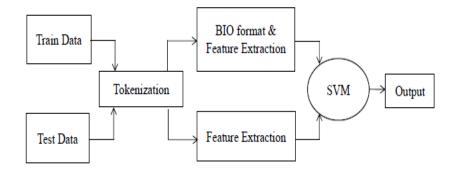


Figure 1: Methodology of the Proposed System for Feature based model

Table 1: Number of Tweets and Average Tokens per tweet for train and test data of Hindi-English and Tamil-English

		Hindi-English	Tamil-English
Train	Tweet count	2700	3200
main	Avg Tokens per Tweet	16.76	11.94
Test	Tweet count	7429	1376
rest	Avg Tokens per Tweet	16.49	12.11

and the use of code-mix language in them, this task holds a significant relevance in today's world.

3. DATASET DESCRIPTION

The task contains two code-mix dataset, Hindi-English and Tamil-English. The training data contains three fields-Tweet ID, User ID and the tweets. Each training file holds a corresponding annotation file. Annotation file contains Tweet ID, User ID, Length, Index and Entity tag of the entities present in the train data. Hindi-English training data includes tweets entirely in code-mix language but Tamil-English dataset includes some tweets in pure Tamil language. Since we proposed embedding based methodology, we were in need of additional dataset to train our word embedding model. The additional dataset for Hindi-English were collected from Mixed Script Information Retrieval 2016 (MSIR) [1], International Conference on Natural Language Processing (ICON) 2015 POS Tagging task [12] and some twitter data. Dataset provided by Sentiment Analysis in Indian Languages (SAIL-2015) [8] [11] were used for Tamil-English. The total number of tweets in the training data, testing data and average tokens per tweet is tabulated in Table 1. The additional dataset collected for Hindi-English is 20671 and for Tamil-English is 1625.

4. METHODOLOGY

Our submission for the task of entity extraction in codemix language includes three systems.

- System 1: Wang2vec based embedding features
- System 2: Word2vec based embedding features
- System 3: Stylometric features

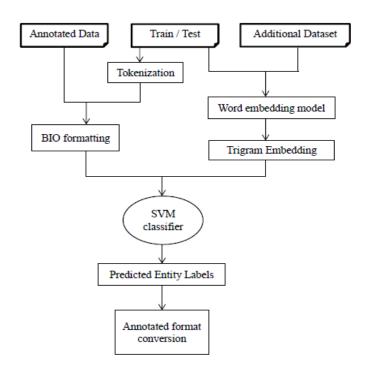


Figure 2: Methodology of the Proposed System for word embedding models

Social media text is generally subjected to various preprocessing tasks. Given dataset includes Twitter data which is subjected to tokenization. Each token from this tokenized dataset is converted to conventional BIO format. This results in BIO tag information for each word in the training data. BIO tag is defined as Beginning, Inside, Outside tag of entities. For example, consider the sentence, "Pranab Mukherjee is the President of India". In general, the en-

Features	Representation
Lower case	Represent word in lowercase
Prefix-Suffix	P3/P4: first $3/4$ characters
I Tellx-Sullix	S3/S4: last $3/4$ characters
Starts with Hash, apostrophe	1 if word starts with $\#$, 'symbol
Numbers, apostrophe, punctuation	Marks 1/0 if present/absent
Length & Index	No of chars & Position of word
Contain HTTP	1 if HTTP present
First character Uppercase	1 if first char is in uppercase
Full character Uppercase	1 if entire word is in uppercase
Contain 4-digit numbers	1 if Token is a 4-digit number
Gazetted Features	Location, Person, Organization, Entertainment

Table 2: Features extracted from train and test data for System 3

Table 3: Cross Validation Accuracy for Hindi-English and Tamil-English

	H	Iindi-Englis	h	Tamil-English			
	System1	System2	System3	System1	System2	System3	
Known	92.9893	91.1001	94.2576	97.2717	97.378	97.4953	
Ambiguous Known	83.0998	78.3239	86.5626	83.8063	83.839	85.9711	
Unknown	91.0318	90.9519	86.9385	93.6683	93.4368	92.4647	
Overall Accuracy	92.4688	91.0278	92.3718	96.1491	96.2534	95.9847	

tity 'Pranab Mukherjee' indicates PERSON and 'India' indicates LOCATION. Since Pranab Mukherjee has two parts, it is tagged as beginning and inside. Words other than entities are tagged as O i.e. Outside. So using BIO tag, the proposed system labels Pranab as B-PERSON, Mukherjee as I-PERSON and India as LOCATION. This BIO tag information is utilized in the three systems proposed in the paper.

Illustration of proposed feature based method is shown in Figure 1 and word embedding based models are shown in Figure 2.

4.1 System 1: Wang2vec based embedding features

Wang2vec model is the modified version of word2vec with an improvement in the structure of skip gram model. This modification made wang2vec better than word2vec. The major difference in these two embedding models is that the skip gram model in word2vec becomes Structured skip gram model in wang2vec. The significant modification in this model is the fact that the word order information is taken into consideration. Wang2vec features are the word vectors obtained using wang2vec model. The size of the vector n is fixed during the training of wang2vec. Thus each word in the training data holds a vector of size n, which is set as 50. The resultant vectors are the word embedding features of the given dataset. From these vectors, the left context and the right context features were extracted and appended to the original embedding features. This resulted in a feature set of size 150. Integrating the context features to the original features forms Trigram embedding features. Thus embedding features along with BIO tag information is integrated and given to train the SVM classifier. Hence a SVM model corresponding to system 1 is obtained. Similar procedure is followed for extraction of wang2vec embedding features for test data. These features are used to form the trigram embedding feature set, which is given to the SVM classifier for testing.

4.2 System 2: Word2vec based embedding features

Word2vec model provides the vector representation for each word. Input for word2vec is sentences, as the major advantage of this model is that it provides vectors for each word based on the context. Vector representation for each word in the training data is obtained through Skip gram model in word2vec. These vectors are used to develop a entity extraction system. Similar to system 1, this system also includes trigram embedding feature set of word2vec embedding vectors.

Each word from the training data is combined with its corresponding BIO tag information and the trigram embedding feature set. This combined feature set is given for training machine learning based classifier, SVM-Light. After training using SVM model, the test data is appended with the trigram embedding features and is given for testing. SVM classifier uses the knowledge acquired from training data and performs recognition of entities in testing data.

4.3 System 3: Stylometric features

Our third system is implemented using stylometric feature extraction. Features such as length, position, numbers, hash tag, punctuation are considered to be the stylometric features. The list of features used in our system is tabulated in Table 2. Stylometric features of each word in the training data is extracted. This feature set is integrated with BIO tag information and used for developing a SVM model. These features are also extracted for test data and given for testing the system.

The proposed methodology results with three SVM models for the three systems using wang2vec features, word2vec features and stylometric features.

5. EXPERIMENTS AND RESULTS

Experiments for system 1 and 2 are similar in case of extracting word embedding features. Major difference be-

Table 1. Repair by Chill II Table Organizers for Himar English									
TEAM	RUN 1			RUN 2			RUN 3		
LIANI	Precision	Recall	F-measure	Precision	Recall	F-measure	Precision	Recall	F-measure
Irshad-IIT-Hyd	80.92	59	68.24	-	-	-	-	-	-
Deepak-IIT-Patna	81.15	50.39	62.17	-	-	-	-	-	-
Amrita_CEN	75.19	29.46	42.33	75	29.17	42.00	79.88	41.37	54.51
NLP_CEN_Amrita	76.34	31.15	44.25	77.72	31.84	45.17	-	-	-
Rupal-BITS_Pilani	58.66	32.93	42.18	58.84	35.32	44.14	59.15	34.62	43.68

Table 4: Result by CMEE-IL Task Organizers for Hindi-English

Table 5: Result by CMEE-IL Task Organizers for Tamil-English

	RUN 1			RUN 2			RUN 3		
TEAM	Precision	Recall	F-measure	Precision	Recall	F-measure	Precision	Recall	F-measure
Deepak-IIT-Patna	79.92	30.47	44.12	-	-	-	-	-	-
Amrita_CEN	77.38	8.72	15.67	74.74	9.93	17.53	79.51	21.88	34.32
NLP_CEN_Amrita	77.7	15.43	25.75	79.56	19.59	31.44	-	-	-
Rupal-BITS_Pilani-R2	55.86	10.87	18.2	58.71	12.21	20.22	58.94	11.94	19.86
CEN@Amrita	47.62	13.42	20.94	-	-	-	-	-	-

tween these two features lies in the fact that system 1 will use wang2vec embedding features which is retrieved using structured skip gram model that takes the word order into consideration. System 2 will use word2vec features retrieved using skip gram model without the word order consideration. In order to train word embedding model, additional dataset is required. Input data for word embedding models i.e., word2vec and wang2vec will be the combination of training data and additional dataset. The size of vector to be generated is set to 50. From these 50 vectors, trigram embedding feature set of size 150 has been extracted. The training dataset is tokenized based on whitespace and is converted to BIO-formatted data. For each word in the training data, its BIO-tag along with the 150 embedding features are given as input to the SVM classifier. Trigram embedding feature set of test data is also extracted in the same manner. After tokenization of test data, the trigram embedding feature vectors of size 150 are given to classifier for testing.

System 3 implements stylometric feature extraction for code-mix data. Training data of Hindi-English and Tamil-English are subjected to the preprocessing task, tokenization. For these tokenized words, features listed in Table 2 are extracted from training data. BIO tag information of these words is combined to the extracted features and thus forms stylometric feature set of training data. As far as testing is concerned, the tokenized words and its corresponding feature set is integrated and given to the classifier.

Cross validation results for Hindi-English and Tamil-English dataset using system 1, 2, 3 are tabulated in Table 3. The System 1 which uses wang2vec based features have shown better results in case of unknown tokens.

According to the results provided by CMEE-IL organizers, for Hindi-English we have acquired third place and for Tamil-English we have acquired second place. The Precision, Recall, F-measure of the top five teams for Hindi-English and Tamil English is tabulated in Table 4 and Table 5 respectively.

6. CONCLUSION

The work is submitted as a part of Shared task on Code Mix Entity Extraction for Indian Languages in FIRE 2016. The use of native languages using Roman script in social media platforms is commonly seen today and extracting entities like person, location or organization from them is a challenging task. The task organizers provided us with data from Twitter and other few microblogs. Three systems were submitted for the task. The first two systems uses the word embedding features of word2vec and wang2vec for entity extraction task. The training data along with some additionally collected dataset were used for training the word embedding models. The third system uses only stylometric features for classification. The three systems were trained and tested using machine learning based classifier, Support Vector Machine. As future work, instead of SVM based classifier we are planning to use regression based methods.

7. ACKNOWLEDGMENT

We would like to thank organizers of Forum for Information Retrieval Evaluation 2016 for organizing the task. We would also like to thank the organizers of the CMEE-IL task.

8. REFERENCES

- Shared Task on Mixed Script Information Retrieval, https://msir2016.github.io, 2016.
- [2] N. Abinaya, N. John, H. Barathi Ganesh, M. Anand Kumar, and K. P. Soman. AMRITA-CEN@FIRE-2014: Named entity recognition for Indian languages using rich features. ACM International Conference Proceeding Series, 05-07-Dec-2014:103-111, 2014.
- [3] M. Anand Kumar, S. Shriya, and K. P. Soman. AMRITA-CEN@FIRE 2015: Extracting entities for social media texts in Indian languages. *CEUR Workshop Proceedings*, 1587:85–88, 2015.
- [4] K. Asnani and J. D. Pawar. Discovering thematic knowledge from code-mixed chat messages using topic model. 2016.
- [5] U. Barman, A. Das, J. Wagner, and J. Foster. Code mixing: A challenge for Language Identification in the Language of Social Media. *EMNLP 2014*, 13, 2014.
- [6] Joachims and Thorsten. Svmlight: Support vector machine. SVM-Light Support Vector Machine

http://svmlight. joachims. org/, University of Dortmund, 19(4), 1999.

- [7] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems, pages 3111–3119, 2013.
- [8] B. G. Patra, D. Das, A. Das, and R. Prasath. Shared Task on Sentiment Analysis in Indian Languages (SAIL) Tweets-An Overview. In *International Conference on Mining Intelligence and Knowledge Exploration*, pages 650–655. Springer, 2015.
- [9] G. Remmiya Devi, P. V. Veena, M. Anand Kumar, and K. P. Soman. Entity Extraction for Malayalam Social Media Text using Structured Skip-gram based Embedding Features from Unlabeled Data. *Proceedia Computer Science*, 93:547–553, 2016.
- [10] S. Sanjay, M. Anand Kumar, and K. P. Soman. AMRITA-CEN-NLP@FIRE 2015:CRF based named entity extraction for Twitter microposts. *CEUR* Workshop Proceedings, 1587:96–99, 2015.
- [11] S. Shriya, R. Vinayakumar, M. Anand Kumar, and K. P. Soman. AMRITA-CEN[®] SAIL2015: Sentiment Analysis in Indian Languages. In *International Conference on Mining Intelligence and Knowledge Exploration*, pages 703–710. Springer, 2015.
- [12] Y. Vyas, S. Gella, J. Sharma, K. Bali, and M. Choudhury. POS Tagging of English-Hindi Code-Mixed Social Media Content. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 974–979. Association for Computational Linguistics, October 2014.
- [13] L. Wang, D. Chris, B. Alan, and T. Isabel. Two/too simple adaptations of word2vec for syntax problems. In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies., 2015.