Named Entity Recognition on Code-Switched Data Using Conditional Random Fields

Utpal Kumar Sikdar[1], **Biswanath Barik**[2] and **Björn Gambäck**[2]
[1]Flytxt, Thiruvananthapuram, India
utpal.sikdar@gmail.com
[2]Dpt. of Computer Science, Norwegian University of Science and Technology
{biswanath.barik,gamback}@ntnu.no

Abstract

Named Entity Recognition is an important information extraction task that identifies proper names in unstructured texts and classifies them into some pre-defined categories. Identification of named entities in code-mixed social media texts is a more difficult and challenging task as the contexts are short, ambiguous and often noisy. This work proposes a Conditional Random Fields based named entity recognition system to identify proper names in code-switched data and classify them into nine categories. The system ranked fifth among nine participant systems and achieved a 59.25% F1-score.

1 Introduction

With the increasing usage of social media, micro blogs and chats in various socio-economical classes, ethnicities and genres in the global society, a new category of informal short texts has evolved in recent years. One of the important phenomena that can appear in such texts is code-mixing or *code-switching* (CS), where bi-lingual users often switch back and forth between their common languages during interactions. Processing of such texts by automatic means encounters several challenges due to the usage of mixed vocabulary, misspellings, abbreviations, transliterations, emojis, and many more. Furthermore, it is in many cases difficult to interpret the texts because of the short contexts.

The Natural Language Processing and text mining communities have taken necessary initiatives to encourage researchers through organizing various workshops and shared-tasks, and by opening mainstream research tracks to develop resources and novel approaches to processing code-mixed texts efficiently and for extracting valuable information from such messy contents. In this direction, the CALCS 2018 Shared Task (Aguilar et al., 2018) focused on identifying a predefined set of nine Named Entity (NE) types: `Person`, `Location`, `Organization`, `Group`, `Title`, `Product`, `Event`, `Time`, and `Other`. The NE identification task addressed code-mixed texts of Spanish-English (SPA-ENG) and Modern Standard Arabic-Egyptian (MSA-EGY); here we will look at the first pair (SPA-ENG) only.

Previously, several machine learning techniques have been applied to the NE recognition problem such as Hidden Markov Models (HMM) (Bikel et al., 1997), Maximum Entropy models (Borthwick, 1999), Conditional Random Fields (CRF) (Lafferty et al., 2001), and Support Vector Machines (SVM) (Isozaki and Kazawa, 2002), as well as deep neural network-based Long Short-Term Memories (LSTM) (Limsopatham and Collier, 2016), Convolutional Neural Networks (CNN) (Santos and Guimaraes, 2015), or hybrid combinations (Chiu and Nichols, 2016).

In this work, the named entity recognition task is considered as a sequence labeling problem, for which CRF is a natural choice to identify entity mentions from code-switched data and classify them to one of the nine aforementioned NE categories. With initial named entity token and language identification, a wide range of features (described in Section 3) are explored for this purpose. As per the overall ranking of the submitted systems under the shared task, our approach is reasonably effective.

The paper is organized as follows: The shared task datasets are presented in Section 2. The named entity recognition system is described in Section 3. Results are presented in Section 4, with error analysis reported in Section 5. Section 6 addresses future work and concludes.

Proceedings of The Third Workshop on Computational Approaches to Code-Switching, pages 115–119
Melbourne, Australia, July 19, 2018.

Dataset	#Tweets	#Named Entities
Training	50,238	12,365
Development	828	151
Test	15,634	-

Table 1: Code-switched dataset statistics

2 Datasets

The shared task organizers provided three different datasets: training, development and test sets. The statistics of the datasets are reported in Table 1, with the total number of tweets and total number of named entities. No gold standard annotation of the test data was made available.

3 Named Entity Recognition

To identify and classify each token from the code-switched data into nine categories (Person, Location, Organization, Group, Title, Product, Event, Time and Other), a supervised CRF-based (Lafferty et al., 2001) approach was used. Different features were extracted from external sources and applied to recognize the target entities.

In a first step, each token was identified as either being a named entity (called a mention) or not. All the beginning and intermediate parts of named entities (for all nine entity categories) were converted into 'B-mention' and 'I-mention', respectively, and a CRF-based model was applied to identify the mentions.

In the next step, the identified mentions ('B-mention' and 'I-mention') were used as features along with other features described in subsections 3.1 and 3.2 to classify each token into one of the nine categories. The 'BIO'[1] notation was used to represent the named entities.

The CRF-based mention and named entity identification models were implemented using CRFsuite (python-crfsuite),[2] which allows for fast training by utilizing L-BFGS (Liu and Nocedal, 1989), a limited memory quasi-Newton algorithm for large scale numerical optimization. The classifier was trained both on features retrieved from external resources and on features directly extracted from the training data, as detailed in the following two subsections.

[1]Here 'B' represents the beginning of, 'I' inside, and 'O' outside of a named entity.

[2]www.chokkan.org/software/crfsuite/

3.1 Features from external sources

The following features were extracted from other external resources:

3.1.1 Language identification

The language identification data from the previous code-switching workshop (Diab et al., 2016) was collected and converted into 'lang1', 'lang2' and 'other' (with 'other' grouping the labels 'mixed', 'ne', 'fw' and 'unknown'). If any token of the 'other' categories was followed by 'lang1', it was assigned to 'lang1'. If the token was followed by 'lang2', it was assigned to 'lang2'. A model described by Sikdar and Gambäck (2016) was built using the converted language identification data and applied to the current shared task's (Aguilar et al., 2018) training and development sets to get language information ('lang1', 'lang2' and 'other') for each token. This language information was then used as a feature for named entity identification in the current shared task.

3.1.2 Named entity token identification

Only the tweets containing named entities were extracted from the data from the previous code-switched workshop, and a CRF based model was built using these tweets with different features (local context, suffix, prefix, all-upper-case, starts-with-upper-case, and hash symbol) and applied to the current shared task's training, development and test data to get named entity information for each token.

3.1.3 Part-of-speech information

The Stanford tagger[3] was used to extract part-of-speech (POS) information for training, development and test data. First, the English version of the Stanford tagger was applied to get English POS tags, and then the Spanish version of the tagger was applied. For tokens belonging to 'lang1' or 'other', the English POS tag was considered. For tokens belonging to 'lang2', the Spanish POS was picked. The POS information for a word together with its two preceding and two following tokens' part-of-speech tags (i.e., a -2 to +2 window) were used as features.

In addition, the first two characters of the current word's POS tag and those of the previous and next two words' POS tags (-2 to +2 tokens) were used as features.

[3]https://nlp.stanford.edu/software/tagger.shtml

3.1.4 Stem

The stem of each token was identified using the Stanford parser.[4]

3.1.5 Noisy data named entity recognizer

The named entities of the current workshop's datasets were identified using the model for named entity recognition in noisy user generated texts described by Sikdar and Gambäck (2017).

3.2 Features from training data

The following features were extracted from the training data.

- word itself: the current word.
- word in lower case: all alphabetic characters in the word converted to lower-case.
- local context of word in lower-case (with a -2 to +2 window, i.e., from two preceding to two following tokens).
- all-upper-case: binary feature checking whether the current token only has upper-case letters or not.
- starts-with-upper-case: binary feature checking whether the current token starts with a capital letter or not.
- word-length: binary feature set if the length of a word is greater than a threshold (> 5).
- suffix: n-grams of the last 1, 2 or 3 characters.
- prefix characters: n-grams of the first 1, 2 or 3 characters.
- is-digit: binary feature checking whether the current word contains any digit or not.
- two-digit: binary feature set if the current word contains two digits.
- is-alphanumeric: current word contains both digits and letters.
- is-special-characters: binary feature set if the current word contains either '#' or '@'.
- is-stop-word: the current word is on NLTK's[5] stop word list.
- most-frequent-word: after removing all stop words, a list was prepared based on high frequency of words (1000 words from the training data). The feature is set if the current word belongs to this high frequency word list.
- word-normalization: the current word with all lower-case letters replaced with 'a', all upper-case letters replaced with 'A', all digits replaced with '0', and all other characters left unaltered.
- Pair-wise-mutual-information-score: PMI calculated based on the number of times the current word belongs to each NE category divided by the word's total number of occurrences in training data.
- beginning-of-the-word: binary feature checking whether the current token belongs to beginning of the sentence or not.
- ending-of-the-word: binary feature checking whether the current token belongs to end of the sentence or not.

To identify the mentions, the above features were used together. To identify named entities, the predicted mentions along with contexts consisting of the previous two and the next two tokens were used as features, in addition to the other features described in subsections 3.1 and 3.2.

[4] https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html

[5] https://www.nltk.org/

Data	Precision	Recall	F-score
5-fold	80.64	71.82	75.95
Dev_Data	81.10	50.20	62.00

Table 2: Mention identification results (%)

Data	F-score
5-fold	59.19
Dev_Data	41.70
Test	59.25

Table 3: Named entity recognition results (%)

Team	F-score
IIT BHU	63.76
CAiRE++	62.76
FAIR	62.66
Linguists	62.13
Flytxt	**59.25**
semantic	56.72
WallyGuzman	54.16
Fraunhofer FKIE	53.65
Baseline	53.28

Table 4: Comparison with other systems (%)

	EVENT	GROUP	LOC	ORG	OTHER	PER	PROD	TIME	TITLE	O
EVENT	1	0	0	0	0	3	0	0	0	2
GROUP	0	2	1	0	0	0	0	0	0	2
LOC	2	0	7	0	0	1	0	0	0	6
ORG	0	0	0	0	0	4	5	0	0	1
OTHER	0	0	0	0	1	0	0	0	0	6
PER	0	0	1	0	0	52	0	0	0	42
PROD	0	0	0	0	0	2	11	0	0	8
TIME	0	0	0	0	0	0	0	6	0	3
TITLE	0	0	2	0	0	6	0	0	2	40
O	0	0	6	2	1	4	1	2	0	9348

Table 5: Confusion matrix for NER on the development data

4 Results

The supervised learning approach was applied to identify mentions. Identified mentions were taken as features along with the other features mentioned in Section 3 to recognize named entities. The classifiers were learned from the training data and tested on the development data. 5-fold cross-validation (CV) was applied to the training data.

The mention identification results are shown in Table 2. The average precision, recall and F1-score values of 5-fold CV on the training data were 80.64%, 71.82% and 75.95%, respectively. The F1-score on the development data was 62.00% due to a significant drop in recall.

The system was applied to named entity recognition and results are shown in Table 3. The average F1-score of 5-fold cross-validation was 59.19%. When tested on the development data, the system achieved an F-score of 41.70%.

The system was then applied to the unseen test data and achieved an F1-score of 59.25%, which is similar to the 5-fold CV F1-score.

Comparing our system ('Flytxt') to the other systems participating in the shared task, Table 4 reports the results and shows that the system secured fifth position and achieved clearly better scores than the baseline system ('Baseline').

5 Error Analysis

When analyzing the output on the development data for named entity recognition, it is clear that many of the named entities are not identified at all by the system. This might be due to the word itself and/or some the contexts word not occurring in the training data.

Furthermore, some named entities are misclassified into other categories, plausibly since those words occur in both named entity categories.

The confusion matrix for named entity recognition is reported in Table 5, for each of the nine classes ('EVENT', 'GROUP','LOC', 'ORG', 'OTHER', 'PER', 'PROD', 'TIME', 'TITLE'). The matrix was built using relaxed match, with the 'B-' and 'I-' distinctions ignored for each named entity class.

6 Conclusion

This paper proposed a Conditional Random Field based approach to identifying and classifying named entities. Compared to the baseline, the proposed system achieved better results.

To investigate the effectiveness of the external features, a feature ablation study should be the next step. Most of the features have been extracted directly from training data, but the features could have been further optimized using grid search and evolutionary approaches.

As an alternative to the feature-based classifier, deep learning-based approaches such as LSTM (Long Short-Term Memory), stack-based LSTM and CNN (Convolution Neural Network) can be explored to classify the proper names into the nine categories.

Acknowledgements

Thanks to the organizers of the 2016 and 2018 code-switching workshops for providing and annotating the training and test data. Thanks also to the three anonymous reviewers for comments that helped improve the paper and for suggestions that can be useful in the future.

References

Gustavo Aguilar, Fahad AlGhamdi, Victor Soto, Mona Diab, Julia Hirschberg, and Thamar Solorio. 2018. Overview of the CALCS 2018 shared task: Named entity recognition on code-switched data. In *Proceedings of the 3rd Workshop on Computational Approaches to Linguistic Code-Switching*, Melbourne, Australia. ACL.

Daniel M. Bikel, Scott Miller, Richard Schwartz, and Ralph Weischedel. 1997. Nymble: a high-performance learning name-finder. In *Proceedings of the 5th Conference on Applied Natural Language Processing*, pages 194–201, Washington, DC, USA. ACL.

Andrew Borthwick. 1999. *A Maximum Entropy Approach to Named Entity Recognition*. Ph.D. thesis, Computer Science Department, New York University, New York, NY, USA.

Jason P.C. Chiu and Eric Nichols. 2016. Named entity recognition with bidirectional LSTM-CNNs. *Transactions of the Association for Computational Linguistics*, 4:357–370.

Mona Diab, Pascale Fung, Mahmoud Ghoneim, Julia Hirschberg, and Thamar Solorio, editors. 2016. *Proceedings of the 2nd Workshop on Computational Approaches to Code Switching@EMNLP 2016*. ACL, Austin, Texas, USA.

Hideki Isozaki and Hideto Kazawa. 2002. Efficient support vector classifiers for named entity recognition. In *Proceedings of the 19th International Conference on Computational Linguistics*, volume 1, pages 1–7, Taipei, Taiwan. ACL.

John Lafferty, Andrew McCallum, and Fernando C.N. Pereira. 2001. Conditional Random Fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning*, pages 282–289, Williamstown, MA, USA. IMIS.

Nut Limsopatham and Nigel Henry Collier. 2016. Bidirectional LSTM for named entity recognition in Twitter messages. In *Proceedings of the 2nd Workshop on Noisy User-generated Text, WNUT@COLING2016*, pages 145–152, Osaka, Japan. ACL.

Dong C. Liu and Jorge Nocedal. 1989. On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, 45(1):503–528.

Cicero Nogueira dos Santos and Victor Guimaraes. 2015. Boosting named entity recognition with neural character embeddings. In *Proceedings of the 5th Named Entity Workshop (NEWS 2015)*, Beijing, China. ACL.

Utpal Kumar Sikdar and Björn Gambäck. 2016. Language identification in code-switched text using Conditional Random Fields and Babelnet. In *Proceedings of the 2nd Workshop on Computational Approaches to Code Switching@EMNLP 2016*, pages 127–131, Austin, Texas, USA. ACL.

Utpal Kumar Sikdar and Björn Gambäck. 2017. A feature-based ensemble approach to recognition of emerging and rare named entities. In *Proceedings of the 3rd Workshop on Noisy User-generated Text, NUT@EMNLP*, pages 177–181, Copenhagen, Denmark. ACL.