Simple Features for Strong Performance on Named Entity Recognition in Code-Switched Twitter Data

Devanshu Jain Maria Kustikova Mayank Darbari
Rishabh Gupta Stephen Mayhew
University of Pennsylvania
{devjain, mkust, mdarbari, rgupt, mayhew}@seas.upenn.edu

Abstract

In this work, we address the problem of Named Entity Recognition (NER) in code-switched tweets as a part of the Workshop on Computational Approaches to Linguistic Code-switching (CALCS) at ACL'18 (Aguilar et al., 2018). Code-switching is the phenomenon where a speaker switches between two languages or variants of the same language within or across utterances, known as intra-sentential or inter-sentential code-switching, respectively. Processing such data is challenging using state of the art methods since such technology is generally geared towards processing monolingual text. In this paper we explored ways to use language identification and translation to recognize named entities in such data, however, utilizing simple features (sans multi-lingual features) with Conditional Random Field (CRF) classifier achieved the best results. Our experiments were mainly aimed at the (ENG-SPA) English-Spanish dataset but we submitted a language-independent version of our system to the (MSA-EGY) Arabic-Egyptian dataset as well and achieved good results.

1 Introduction

Recently, social media texts such as tweets and Facebook posts have attracted attention from the Natural Language Processing (NLP) research community. This content has many applications as it provides clues to analyze sentiments of the masses towards areas ranging from basic electronic products to mental health issues to even national political candidates. These applications have motivated the NLP community to rethink strategies for common tools, such as tokenizers, named entity taggers, POS taggers, dependency parsers, in the context of informal and noisy text.

As access to the internet becomes more and more universal, a linguistically diverse population has come online. Hong et al. (2011) showed that in a collection of 62 million tweets, only a little over 50% of them were in English. This multilingualism has given rise to such interesting patterns as transliteration and code-switching. The multilingual behavior combined with the informal nature of the content makes the task of building NLP tools even harder.

In this paper, we solve the problem of Named Entity Recognition (NER) for code-switched twitter data as a part of the ACL'18 Computational Approaches to Linguistic Code-switching (CALCS) Shared Task (Aguilar et al., 2018). Code-switching is a phenomenon that occurs when multilingual speakers alternate between two or more languages or dialects. This phenomenon can be observed across different sentences, within the same sentence or even in the same word. This shared task is similar to other social media tasks, except that the data is explicitly chosen to contain code-switching. The entities for the task are: Event, Group, Location, Organization, Other, Person, Product, Time, and Title. Below is an example of some code-switched data, switching between English and Spanish:

> My [Facebook]$_{Prod}$, [Ig]$_{Prod}$ & [Twitter]$_{Prod}$ is hellaa dead yall Jk soy yo que has no life!

In this example, there is a combination of English and Spanish words and slang words within a tweet, with 3 entities: Facebook, Instagram (commonly referred to as 'Ig') and Twitter.

Proceedings of The Third Workshop on Computational Approaches to Code-Switching, pages 103–109
Melbourne, Australia, July 19, 2018.

Value / Data	Train	Development	Test
Total number of tweets	50,757	832	15,634
Total number of tokens	616,069	9,583	183,011
Average number of tokens per tweet	12.14	11.52	15.9
Standard deviation of the number of tokens per tweet	7.6	7.12	7.11

Table 1: (ENG-SPA) English-Spanish number of tweets and tokens for train, development, and test data

Value / Data	Train	Development	Test
Total number of tweets	10,103	1,122	1,110
Total number of tokens	204,323	22,742	21,414
Average number of tokens per tweet	20.22	20.27	21.91
Standard deviation of the number of tokens per tweet	6.63	6.76	6.18

Table 2: (MSA-EGY) Modern Standard Arabic-Egyptian number of tweets and tokens for train, development, and test data

2 Related Work

NER is a fundamental part of the Information Extraction pipeline. Most of the available off-the-shelf systems are trained on formal content, and consequently do not generalize well when evaluated on twitter data (Ritter et al., 2011). This can be explained by the fact that such systems rely on hand-crafted standard local features and some background knowledge, which is not reliable in data as noisy as tweets. With only a limited number of characters, people use a variety of creative ways to express their thoughts, including emoticons and novel abbreviations.

There have been few recent workshops and shared-tasks on analysis of such noisy social media data, such as Workshop on Noisy User-Generated Text (WNUT) at EMNLP (2014, 2016, 2017), Workshop on Approaches to Subjectivity, Sentiment and Social Media (WASSA) at NAACL (2016), and Forum for Information Retrieval Evaluation (FIRE: 2015, 2016, 2017).

3 Experimental Setup

Here we describe the data, evaluation, and the model we used.

3.1 Data

In our experiments, we focus primarily on the English-Spanish (ENG-SPA) dataset. However, we submitted our basic system results for Arabic-Egyptian (MSA-EGY) dataset as well.

The organizers provided annotated train and development sets for each language. They also provided an unannotated set of test data, which we annotated with our system, and submitted for evaluation. We never had access to the gold annotated test set, before or after the evaluation.

Tables 1 and 2 provide information about the data in terms of number of tweets and tokens for the (EN-SPA) English-Spanish and (MSA-EGY) Modern Standard Arabic-Egyptian language pairs. Tables 3 and 4 provide statistics of the named entities for both (EN-SPA) English-Spanish and (MSA-EGY) Modern Standard Arabic-Egyptian language pairs, where each cell can be interpreted as *Number (Percentage)* and entity 'O' represents all non-NE tokens. Please note that the data has been tagged using the IOB scheme and data in Tables 3 and 4 is the result of grouping named entities according to the IOB scheme.

3.2 Evaluation

We used the standard harmonic mean F1 score to evaluate the system performance. Additionally, we used surface form F1 score as described in Derczynski et al. (2017). Both of these metrics were a part of the evaluation in the CALCS shared task.

3.3 Method

We used the sklearn implementation of Conditional Random Field (CRF)[1] (McCallum and Li, 2003) as the base model in our NER system.

[1] https://sklearn-crfsuite.readthedocs.io/

Entity	Train Count	Development Count
O	597,526 (97%)	9,361 (97.68%)
Event	232 (0.04%)	4 (0.04%)
Group	718 (0.12%)	4 (0.04%)
Location	2,810 (0.46%)	10 (0.1%)
Organization	811 (0.13%)	9 (0.09%)
Other	324 (0.05%)	6 (0.06%)
Person	4,701 (0.76%)	75 (0.78%)
Product	1,369 (0.22%)	16 (0.17%)
Time	577 (0.09%)	6 (0.06%)
Title	824 (0.13%)	22 (0.23%)

Table 3: (ENG-SPA) English-Spanish named entities counts for train and development data

Entity	Train Count	Development Count
O	181,230 (88.7%)	20,031 (88.08%)
Event	535 (0.26%)	69 (0.3%)
Group	1,799 (0.88%)	191 (0.84%)
Location	3,275 (1.6%)	358 (1.57%)
Organization	1504 (0.74%)	149 (0.66%)
Other	116 (0.06%)	17 (0.07%)
Person	5705 (2.79%)	698 (3.07%)
Product	538 (0.26%)	55 (0.24%)
Time	466 (0.23%)	61 (0.27%)
Title	896 (0.44%)	115 (0.51%)

Table 4: (MSA-EGY) Modern Standard Arabic-Egyptian entities counts for train and development data

System	ENG-SPA	MSA-EGY
Org. Baseline	53.28	62.70
Experiment 1	62.13	67.44
Top System	63.76	71.61

Table 5: (ENG-SPA) and (MSA-EGY) Our best F1 scores on the test datasets compared with the organizer's baseline and the top performing system in the Shared Task.

4 Experiments

This section gives an overview of our experiments. First, we identify various local and global features using a variety of monolingual tweets and Gazetteers and train a CRF-based classifier on the data. Second, we try to improve system recall using a 2-step NER process. Third, we convert the convert the code-mixed data to monolingual data using language identification (using a character-based language model) and translation.

Of the three experiments that we tried, the first method gave the best results. We compare against the best performing system in the shared task as well as the organizer's baseline in Table 5. The baseline was provided by the organizers and used Bi-directional LSTMs followed by softmax layer (trained for 5 epochs) to infer the output labels.

The shared task used Surface Form F1 scores as well, but we omit them from our results as they were the same as harmonic mean F1 in all cases. All scores are reported in Table 6. Detailed scores are available in the appendix.

4.1 Experiment 1

Our first experiment used a standard set of features, augmented with some task-specific ideas, and defined as follows. Given a sequence of words in a sentence: ..., w_{i-2} , w_{i-1} , w_i , w_{i+1} , w_{i+2} , ... and the current word in consideration is w_i , we used the following features:

- If w_i is in the beginning of sentence

		Development Data			Test Data		
		Precision	Recall	F1	Precision	Recall	F1
ENG-SPA	Exp. 1	69.44	32.89	44.64	**72.75**	54.22	**62.13**
	Exp. 2	**71.29**	**47.37**	**56.92**	46.22	**64.66**	53.91
	Exp. 3	66.27	36.18	46.81	71.88	54.00	61.67
MSA-EGY	Exp. 1 (no Gaz)	83.29	73.91	78.32	74.43	61.65	67.44

Table 6: Results on all submissions. Bold indicates best performance for that language.

- If w_i is in the end of sentence
- Lower-case version of w_i
- If w_i is title-cased
- Prefixes and Suffixes of length 4 of w_i
- Brown Clusters[2] (Cluster Size - 40) of w_i
- Word2Vec Clusters: We trained a Word2Vec (Řehůřek and Sojka, 2010) model on the combined tweets dataset (dimension: 100 ; window: 7). Then, we clustered these embeddings into 40 clusters and used cluster IDs as features.
- Gazetteer: We used the Gazetteer (extracted from Wikidata by Mishra and Diesner (2016)) labels as features.
- For each word w_k in a context window of ± 2:
  - The word w_k itself
  - If w_k is upper case
  - Shape and Short shape (where same consecutive characters in the shape are compressed to a single character) of w_k
  - If w_k contains any special symbol like: ,#,$,-,,,etc. or an emoji.
  - If w_k is alphabetic or alphanumeric
  - Emoji Description: We identified the 40 most common emojis present in our dataset and manually labelled them with representative words, such as smile, kiss, sad, etc. These emoji description (sense) of every context word were used as another feature.

We also ran the experiment on the MSA-EGY dataset (without the Gazetteer features).

4.2 Experiment 2

Following the first experiment, our main observation was that the recall was quite low. One reason for this could be the presence of a large amount of tokens tagged as 'O' (∼97%). In contrast, the standard CONLL 2002 Spanish training NER corpus (Tjong Kim Sang, 2002) had ∼87% of the tokens tagged as 'O'.

To solve this issue, we experimented with a 2-step NER process (similar to (Eiselt and Figueroa, 2013)):

1. Train a CRF model to identify whether a token is 'O' or not
2. Train a CRF model to identify the type of named-entity (if identified as non-'O')

As expected, we saw major improvements in recall, but these were offset by a substantial drop in precision. Overall, this led to a lower F1 score than before. In light of these results, we did not use the 2-step approach for any other experiments.

4.3 Experiment 3

In this experiment, we tried to eliminate the code-switching by converting the data to a monolingual form. Our method is to identify the language of each token in the dataset and translate into a common language.

We collected training data for language identification using the Twitter API. We downloaded tweets for English and Spanish and assumed that each word in those tweets belonged to that particular language. The statistics for the downloaded data is shown below:

1. 3000 Spanish tweets (7700 tokens ∼56%)
2. 1900 English tweets (6100 tokens ∼44%)

Then, we trained a character-level RNN-based language model on this data to do language identification. In order to validate, we split our data and used 80% for training and rest for validating, achieving an accuracy of 79% on this validation data. We used this model to identify the language of all the tokens in dataset, then used Google Translate API to translate English tokens to Spanish.

[2] https://github.com/percyliang/brown-cluster

Finally, we used the language identification and the translation as features in our CRF model, in addition to all the features used in experiment 1.

As compared to the results from experiment 1, this improved the recall on both development and test sets, but again, the loss in precision caused a slight overall drop in performance.

5 Conclusion

Our submissions earned 4th place out of 8 submissions in the ENG-SPA task, and 3rd place out of 6 submissions in the MSA-EGY task.

Surprisingly, our simplest NER model, trained without using any language identification or translations, worked best. The other more sophisticated experiments showed promise in improving the recall, but damaged the precision too much to improve the F1 score.

One of the challenges we faced was dissimilarity between development and test dataset. Although some of the techniques that we tried on the development dataset improved the system performance, the same effect was not seen in the test dataset. For example, see the change in performance between Table 7 and Table 8. The F1 score on the development set jumped 12 points, but the score on the test set dropped 9 points. This could be explained by the very small size of the development dataset, where a few errors or successes could change the score dramatically. Without access to the test data, we could not do any qualitative error analysis.

Finally, since the 2-Step NER achieved such a high recall, we believe that creating an ensemble of 1-Step and 2-Step systems could achieve a better overall F1 score.

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 Third Workshop on Computational Approaches to Linguistic Code-Switching*, Melbourne, Australia. Association for Computational Linguistics.

Leon Derczynski, Eric Nichols, Marieke van Erp, and Nut Limsopatham. 2017. Results of the wnut2017 shared task on novel and emerging entity recognition. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 140–147. Association for Computational Linguistics.

Andreas Eiselt and Alejandro Figueroa. 2013. A two-step named entity recognizer for open-domain search queries. In *Sixth International Joint Conference on Natural Language Processing, IJCNLP 2013, Nagoya, Japan, October 14-18, 2013*, pages 829–833. Asian Federation of Natural Language Processing / ACL.

Lichan Hong, Gregorio Convertino, and Ed H. Chi. 2011. Language matters in twitter: A large scale study. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, pages 519–521.

Andrew McCallum and Wei Li. 2003. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 188–191. Association for Computational Linguistics.

Shubhanshu Mishra and Jana Diesner. 2016. Semi-supervised named entity recognition in noisy-text. In *Proceedings of the 2nd Workshop on Noisy User-generated Text*, pages 203–212. The COLING 2016 Organizing Committee.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA.

Alan Ritter, Sam Clark, Mausam, and Oren Etzioni. 2011. Named entity recognition in tweets: An experimental study. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, 27-31 July 2011, John McIntyre Conference Centre, Edinburgh, UK, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1524–1534. ACL.

Erik F. Tjong Kim Sang. 2002. Introduction to the conll-2002 shared task: Language-independent named entity recognition. In *Proceedings of the 6th Conference on Natural Language Learning - Volume 20*, COLING-02, pages 1–4, Stroudsburg, PA, USA. Association for Computational Linguistics.

Appendices

A ENG-SPA detailed results

We show detailed results for ENG-SPA experiments in the following tables.

	Development Data			Test Data		
	Precision	Recall	F1	Precision	Recall	F1
Event	0.00	0.00	0.00	41.67	11.11	17.54
Group	100.00	25.00	40.00	68.89	31.96	43.66
Location	66.67	40.00	50.00	72.16	68.16	70.1
Organization	100.00	11.11	20.00	51.11	22.77	31.51
Person	73.33	44.00	55.00	83.33	69.5	75.79
Product	58.33	43.75	50.00	66.41	45.19	53.79
Time	50.00	50.00	50.00	18.10	12.58	14.84
Title	100.00	4.55	8.70	47.57	22.17	30.25
Other	0.00	0.00	0.00	0.00	0.00	0.00
Overall	69.44	32.89	44.64	72.75	54.22	62.13

Table 7: (ENG-SPA) Results for Experiment 1: simple features and gazetteers

	Development Data			Test Data		
	Precision	Recall	F1	Precision	Recall	F1
Event	50.00	25.00	33.00	18.60	17.78	18.18
Group	100.00	25.00	40.00	25.34	38.14	30.45
Location	50.00	50.00	50.00	57.16	71.38	63.48
Organization	50.00	11.11	18.18	36.31	30.20	32.97
Person	74.58	58.67	65.67	60.19	80.70	68.95
Product	62.50	62.50	62.50	50.64	51.17	50.90
Time	100.00	100.00	100.00	13.19	64.90	21.92
Title	66.67	9.09	16.00	28.23	31.67	29.85
Other	100.00	33.33	50.00	5.56	5.17	5.36
Overall	71.29	47.37	56.92	46.22	64.66	53.91

Table 8: (ENG-SPA) Results for Experiment 2: 2-step NER

	Development Data			Test Data		
	Precision	Recall	F1	Precision	Recall	F1
Event	0.00	0.00	0.00	45.45	11.11	17.86
Group	100.00	25.00	40.00	68.18	30.93	42.55
Location	55.56	50	52.63	70.97	67.80	69.35
Organization	50	11.11	18.18	48.78	19.80	28.17
Person	74.51	50.67	60.32	83.19	69.43	75.69
Product	53.85	43.75	48.28	65.54	45.45	53.68
Time	50.00	50.00	50.00	18.10	13.91	15.73
Title	0.00	0.00	0.00	45.37	22.17	29.79
Other	0.00	0.00	0.00	0.00	0.00	0.00
Overall	66.27	36.18	46.81	71.88	54.00	61.67

Table 9: (ENG-SPA) Results for Experiment 3: Language Identification + Translation

B MSA-EGY detailed results

We show detailed results for the one MSA-EGY experiment in the following table.

	Development Data			Test Data		
	Precision	Recall	F1	Precision	Recall	F1
Event	66.67	43.48	52.63	67.57	35.71	46.73
Group	86.63	78.01	82.09	69.92	73.50	71.67
Location	87.14	75.70	81.02	76.64	57.95	66.00
Organization	74.24	65.77	69.75	68.75	61.60	64.98
Person	85.28	79.66	82.37	79.34	64.70	71.27
Product	79.17	69.09	73.79	66.67	54.55	60.00
Time	74.60	77.05	75.81	68.00	68.00	68.00
Title	77.11	55.65	64.65	26.32	50.00	34.48
Other	92.86	76.47	83.87	100.00	50.00	66.67
Overall	83.29	73.91	78.32	74.43	61.65	67.44

Table 10: (MSA-EGY) Results for Experiment 1 (without Gazetteer features)