Tackling Code-Switched NER: Participation of CMU

[*,0]Parvathy Geetha [+,0]Khyathi Raghavi Chandu [+]Alan W Black

[*]Electrical and Computer Engineering, [+]Language Technologies Institute
Carnegie Mellon University
{pgeetha, kchandu, awb}@andrew.cmu.edu

Abstract

Named Entity Recognition plays a major role in several downstream applications in NLP. Though this task has been heavily studied in formal monolingual texts and also noisy texts like Twitter data, it is still an emerging task in code-switched (CS) content on social media. This paper describes our participation in the shared task of *NER on code-switched data* for Spanglish (Spanish + English) and Arabish (Arabic + English). In this paper we describe models that intuitively developed from the data for the shared task *Named Entity Recognition on Code-switched Data*. Owing to the sparse and non-linear relationships between words in Twitter data, we explored neural architectures that are capable of non-linearities fairly well. In specific, we trained character level models and word level models based on Bidirectional LSTMs (Bi-LSTMs) to perform sequential tagging. We trained multiple models to identify nominal mentions and subsequently used this information to predict the labels of named entity in a sequence. Our best model is a character level model along with word level pre-trained multilingual embeddings that gave an F-score of 56.72 in Spanglish and a word level model that gave an F-score of 65.02 in Arabish on the test data.

1 Introduction

Named Entity Recognition (NER) is a challenging and one of the most fundamental tasks in NLP. NER not only has stand alone applications including search and retrieval but also aids as a prior step for downstream NLP applications like question answering and dialog state tracking. It has been fairly researched in the community using both supervised (Azpeitia et al., 2014) and semi-supervised (Nadeau, 2007), (Nadeau, 2007) techniques. Moreover, this has also been studied on multiple languages including English (Lample et al., 2016), Spanish (Zea et al., 2016) and Arabic (Shaalan, 2014). The task is projected into an even complex space when there are words from multiple languages interleaved within and between sentences. This phenomenon is commonly known as code switching (CS).

CS is typically used in informal or semi-formal communication and social media stages an accessible platform to interact in this manner. This also comes with additional nuances observed in social media text that can be broadly characterized as noisy text with spelling errors and ungrammatical constructions. Often, the shorthand representations observed in this data are non-standardized and are one to many functions of standard spelling to a non-standard spelling. This makes this task significantly different from NER on formal monolingual texts and the techniques are not directly transferable to the domain of CS text. Supervised techniques to address the task in the domain of noisy texts such as Twitter have been explored (Ritter et al., 2011), (Tran et al., 2017). We leverage these techniques in order to deal with the sparse distribution of entities.

In this paper, we discuss the techniques used from the participation of our team in the shared task of *Named Entity Recognition of Code-switched Data* (Aguilar et al., 2018). We model the problem at both word and character levels along with attempting attention mechanism. We discuss the intuitions from the data that motivate the models. We have also explored ensembling multiple models that cater to identification of the named entity and labeling it with a tag. Our best performing system is the combination of character and word level representations (using pre-trained

[0]Denotes equal contribution

Proceedings of The Third Workshop on Computational Approaches to Code-Switching, pages 126–131
Melbourne, Australia, July 19, 2018.

	Spanglish			Arabish		
Criteria	**Train**	**Dev**	**Test**	**Train**	**Dev**	**Test**
# Tweets	41,024	832	15,634	10,091	1,121	1,110
# Unique Words	57,892	2,559	27,756	44,024	9,800	9,316
# Unique NEs	4,788	156	-	4,435	1,107	-
OOV with Train (%)	0	18.67	52.16	0	30.76	40.79
OOV of NEs with Train (%)	0	62.82	-	0	25.92	-
OOV with MUSE (%)	64.43	20.94	57.23	99.74	99.82	97.44
OOV of NEs with MUSE (%)	17.41	5.76	-	97.74	99.90	-

Table 1: Data Analysis

multilingual embeddings) in a Bi-LSTM that resulted in an F1 score of 56.72 in Spanglish and a word level Bi-LSTM that gave an F1 of 65.02 in Arabish.

2 Related Work

NER is a fairly well researched topic and a lot of literature (Nadeau and Sekine, 2007) is available with regard to this. In this section we focus and present a comprehensive overview of the techniques that lay motivations to our models and experiments.

While traditionally hand crafted features are reliably used (Carreras et al., 2002), neural models have recently been emerging as effective techniques to perform the task. This is owed to the substantial reduction of manual expense in building hand-crafted features for each language. Qi et al. (2009) leverages unannotated sentences to improve supervised classification tasks using Word-Class Distribution Learning. Passos et al. (2014) were among the first to use a neural network to learn word embeddings that leverage information from related lexicon to perform NER. Collobert et al. (2011) used convolution for embeddings with a CRF layer to attain alongside benchmarking several NLP tasks including NER. Lample et al. (2016) achieves the state-of-the-art performance on 4 languages by training models based on BiLSTM and CRF by using word representations from unannotated text and character representations from annotated text. This work has been extended to transfer settings by Bharadwaj et al. (2016) to multiple languages by representing word sequences in IPA. Huang et al. (2015) use a BiLSTM with a CRF layer in addition to making use of explicit spelling and context features along with word embeddings.

Aguilar et al. (2017) use a character level CNN followed by a word level Bi-LSTM in a multi-task learning setting and also emphasize the importance of gazetteer lists for the task. Multilingual NER on informal text in Twitter was also studied by Etter et al. (2013). Zirikly and Diab (2015) explore the impact of embeddings and representations of words without gazetteer features on NER for social media text in Arabic. Luo et al. (2017) have also shown that attention based Bi-LSTM with additional architecture achieves higher performance than other state-of-the-art techniques to recognize chemical named entities which lean to low resource settings. We hypothesize that CS also belongs to low resource settings and explore the impact of attention. The task of NER becomes harder especially in low resource settings (Tsai et al., 2017), which is similar to CS setting.

3 Data Analysis

Code-switching is more prominently observed in informal communication which is observed in social media platforms. Hence the organizers of the shared task (Aguilar et al., 2018) have provided us with English-Spanish (ENG-SPA) and Arabic-English (MSA-EGY) tweets. In this section, we present an overlap analysis of the tweets from the train and the development set that lead to intuitions of model performance.

An important characteristic of the nature of social media data is that the named entities are very sparse. While table 1 shows that the training data is comprised of 8.27% of unique named entities, we observe that 2.93% of overall surface form distribution belong to named entities. This number is significantly smaller than the number of named entities found in formal texts traditionally used for training this task. For instance, a widely standardized and accepted dataset that is proposed by Tjong Kim Sang and De Meulder (2003) for monolingual English contains 15.04% tagged named entities. This makes the task harder in social media settings.

In order to analyze the distribution of named entities across the different splits in the data, we look at the out of vocabulary (OOV) percentages with respect to different sources. This is performed to estimate the significance of that particular source with respect to the task at hand. There is quite a high OOV percentage of named entities from the training data.

In Section 4 we elaborate on leveraging pre-trained multilingual embeddings MUSE (Multilingual Unsupervised or Supervised word Embeddings) (Conneau et al., 2017) which contain multilingual embeddings based on Fast Text (Bojanowski et al., 2016). Table 1 presents these statistics which helps provide intuitions on the approach that needs to be taken for this data. For Spanglish data, there are 62.82% of named entities that are not present in training data and 5.76% that are not present in the development data.

4 Models and Intuitions

Based on the three main observations in Section 3, we frame the following intuitions to build our model architectures.

- *Sparsity of named entities:* Training a model that classifies nominal entities from their counterparts and using this information to tag them.
- *High OOV with training data:*
  - Character level models that are capable of capturing sequential sub-word level information
  - External knowledge sources like gazetteer lists and/or pre-trained word embeddings such as MUSE.

4.1 Model Architecture

The first architecture is a simple bidirectional LSTM (Bi-LSTM) at word level that captures sequential context information. In addition to this, the second model also needs to learn sub-word level information that is based on characters of words. Soft combinations of character sequences act as a proxy to the valid sequences of phonemes allowed by a sentence. We have not used phonetic features directly as performed by Bharadwaj et al. (2016) due to the noisy nature of the text with multiple instances of shorthand notations. However, we believe that this is an interesting direction and plan to explore this beyond the scope of this paper.

Recurrent Neural Networks (RNNs) model sequential data and are capable of transforming the current sequence into latent space. In our case, the former is a sequence of words and the latter is a sequence of Named Entity tags. While in theory, RNNs are capable of learning dependencies ranging over long distances, in practice this is hindered due to vanishing or exploding gradients. Alternatively, a variant of this model, LSTM (Gers et al., 1999) is used to model the influence of the longer range dependencies since it maintains a memory cell. At this point, we have a couple of options to feed into this network. The first is to directly feed the words into a Bi-LSTM and the second is to include character level information as well.

In each word, each of the characters has a 50 dimensional embedding (let it be e). We pass it through an LSTM to get the latent representation z of the word, over which a $tanh$ non-linearity is applied. This character level modeling of the word is concatenated with the 200 dimensional word lookup embeddings to form the final word level representation. These final modified word embeddings are fed into a Bi-LSTM which computes a hidden left context representation $\overrightarrow{h_t}$ and hidden right context representation $\overleftarrow{h_t}$ which are concatenated. Finally, this is fed into a fully connected layer with a cross entropy loss function to predict the sequence of tags. All the weights in the model are initialized with Xavier distribution. The model is trained with an Adam optimizer for minimum validation loss for 10 epochs.

We then extended the model to explore the effect of attention over the Bi-LSTM model but it did not show any improvements over the base model.

Classifying Nominalization:

To deal with the problem of sparse distribution of named entities, we model the problem in 2 phases. The first phase is a binary classification of named entities in a sequence of words. The second phase is to add additional features based on the prediction of the first network to the embeddings in the second network to label the tags. We intentionally used the same network architecture excepting for the final transformation layer to predict the tags. This is because we intend to pose this as a Multi Task Learning (MTL) problem (Collobert and Weston, 2008), where we can share the bottom layers so the network can generalize better with sparse distribution of tags. This idea is similar to the work by Aguilar et al. (2017) but we restrict to predicting the named entities since we do not have POS information of the words. We present the results of hierarchical phase formulation of this method in Table 2 and leave the end to end MTL training (where the first task is predicting whether it is an NE and the second task is predicting the tag of NE which are jointly trained) for future work.

Pre-trained multilingual Embeddings: Since the data is too sparse, we leveraged pre-trained multilingual word embeddings that are trained based on fastText embeddings (Conneau et al., 2017) and are aligned across multiple languages.

Models/Metrics	Spanglish		Arabish	
	Entity	Surface Form	Entity	Surface Form
Word Bi-LSTM	52.34	51.34	**73.05**	**60.80**
Char Bi-LSTM + Word Bi-LSTM	50.22	50.95	73.95	61.38
Pre-trained MUSE + Char Bi-LSTM + Word Bi-LSTM	**54.47**	**53.27**	64.38	47.23
Attention + Word Bi-LSTM	36.50	35.19	68.11	53.86
NE v non-NE + Char Bi-LSTM + Word Bi-LSTM	49.48	49.61	70.70	10.87

Table 2: F scores of different models motivated by intuitions from the data

This boosted the F score by 2 points which is comparatively better performing model in our space of models.

5 Results and Discussion

We have tried different models based on the intuitions from this domain of data that are explained in Section 4. The F1 scores of these different architectures are presented in table 2 for both Spanglish and Arabish. As it is observed from the data, the model that performed best is the character level model with pre-trained MUSE embeddings (Conneau et al., 2017) and a word level Bi-LSTM for Spanglish data. However, this is not the case with Arabish data where a simple word level Bi-LSTM performed better. This can be explained from Table 1 as there are 99.82% of vocabulary that is not present in the MUSE embeddings.

Based on automatic as well as a brief manual analysis of the entity wise scores on the development set, we identify that our models do not perform very well on TITLE entities. One interesting challenge for this category is that the word level composition of the entities comprise of several common terms. Examples of this include *'High School Musical'*, *'Oh My God'* etc., which are very hard to be identified as named entities. This category can co-occur in similar contexts of other named entities. For example *'Keep calm and enjoy your GYPSY SUMMER'*, where *'GYPSY SUMMER'* is a named entity (which could have easily been *'drink'*).

We annotated the development data to understand and motivate the need to build an NER for CS contexts as opposed to using monolingual NERs. The annotation is done in the perspective of whether the words belong to one of the following 4 categories: *English*, *Spanish*, *Mixed* and *Ambiguous*, which are 156, 54, 4 and 5 respectively. This might give a naive impression that an NER trained on English is sufficient to perform reasonably well for this data as well. This in in contrary to the results that Stanford NER (Finkel et al., 2005) performed on this data by giving an Entity F1 of 10.89 and Surface F1 of 11.96. Hence we need to train the models explicitly for the switched language by treating it as a new language or by transferring learning from both the individual languages.

As described in Section 4, we experimented with combining multiple neural models performing different tasks (predicting a binary named entity or not, and labeling the sequence). This model did not improve the performance on development set. The binary model predicts 42 named entities correctly that the best model is unable to capture in comparison to 16 by the character model. However, the binary model gets a lot of false positives in the sense that 39 tokens are predicted as named entities incorrectly while this number for the embedding model is 7. The possible solution to leverage this model more accurately is either thresholding the softmax scores of the binary model to only get the predictions of named entities with high confidence or perform MTL where weights are updated by the loss from both the tasks.

The huge gap between entity and surface form for the Arabish data that is observed by the character model along with the binary features (based on the predictions of whether it is an NE or not), is due to a large number of invalid sequences.

Among the true named entities that are wrongly predicted in Spanglish data, 154 of them are occurring in training data. This implies that the context information can be leveraged better to improve the models since the contexts in which these entities are embedded are very broad.

6 Conclusion and Future Work

Developing intuitions from the data to build models is necessary for domains that do not have other NLP tools such POS taggers, parsers etc,. Based on these intuitions, a character level model along with pre-trained multilingual word embeddings from MUSE with a Bi-LSTM has given an F score of 56.72 on Spanglish and word level Bi-LSTM that gave an F score of 65.02 on Arabish. We believe that there is a lot of potential in exploring the attention model in synergy with predicting whether a term is named entity or not as a Multi Task Learning problem.

Language/Metrics		Event	Group	Location	Org	Other	Person	Product	Time	Title
Spanglish	**Entity**	0.00	33.33	57.14	30.77	0.00	69.57	60.00	28.57	0.00
	Surface Form	0.00	33.33	57.14	36.36	0.00	68.29	55.56	33.33	0.00
Arabish	**Entity**	51.85	71.73	74.97	57.61	68.75	81.89	64.22	66.67	56.74
	Surface Form	42.11	58.43	57.71	48.73	42.86	71.55	53.57	59.70	52.76

Table 3: F scores of best models for Spanglish and Arabish

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.

Gustavo Aguilar, Suraj Maharjan, Adrian Pastor López Monroy, and Thamar Solorio. 2017. A multi-task approach for named entity recognition in social media data. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 148–153.

Andoni Azpeitia, Montse Cuadros, Seán Gaines, and German Rigau. 2014. Nerc-fr: supervised named entity recognition for french. In *International Conference on Text, Speech, and Dialogue*, pages 158–165. Springer.

Akash Bharadwaj, David Mortensen, Chris Dyer, and Jaime Carbonell. 2016. Phonologically aware neural model for named entity recognition in low resource transfer settings. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1462–1472.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606.*

Xavier Carreras, Lluis Marquez, and Lluís Padró. 2002. Named entity extraction using adaboost. In *proceedings of the 6th conference on Natural language learning-Volume 20*, pages 1–4. Association for Computational Linguistics.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167. ACM.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(Aug):2493–2537.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087.*

David Etter, Francis Ferraro, Ryan Cotterell, Olivia Buzek, and Benjamin Van Durme. 2013. Nerit: Named entity recognition for informal text. *The Johns Hopkins University, the Human Language Technology Center of Excellence, HLTCOE 810Wyman Park Drive Baltimore, Maryland 21211, Tech. Rep.*

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd annual meeting on association for computational linguistics*, pages 363–370. Association for Computational Linguistics.

Felix A Gers, Jürgen Schmidhuber, and Fred Cummins. 1999. Learning to forget: Continual prediction with lstm.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991.*

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360.*

L Luo, Z Yang, P Yang, Y Zhang, L Wang, H Lin, and J Wang. 2017. An attention-based bilstm-crf approach to document-level chemical named entity recognition. *Bioinformatics (Oxford, England).*

David Nadeau. 2007. *Semi-supervised named entity recognition: learning to recognize 100 entity types with little supervision.* Ph.D. thesis, University of Ottawa.

David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

Alexandre Passos, Vineet Kumar, and Andrew McCallum. 2014. Lexicon infused phrase embeddings for named entity resolution. *arXiv preprint arXiv:1404.5367.*

Yanjun Qi, Ronan Collobert, Pavel Kuksa, Koray Kavukcuoglu, and Jason Weston. 2009. Combining labeled and unlabeled data with word-class distribution learning. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 1737–1740. ACM.

Alan Ritter, Sam Clark, Oren Etzioni, et al. 2011. Named entity recognition in tweets: an experimental study. In *Proceedings of the conference on empirical methods in natural language processing*, pages 1524–1534. Association for Computational Linguistics.

Khaled Shaalan. 2014. A survey of arabic named entity recognition and classification. *Computational Linguistics*, 40(2):469–510.

Erik F Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 142–147. Association for Computational Linguistics.

Van Cuong Tran, Ngoc Thanh Nguyen, Hamido Fujita, Dinh Tuyen Hoang, and Dosam Hwang. 2017. A combination of active learning and self-learning for named entity recognition on twitter using conditional random fields. *Knowledge-Based Systems*, 132:179–187.

Chen-Tse Tsai, Stephen Mayhew, Yangqiu Song, Mark Sammons, and Dan Roth. 2017. Illinois ccg lorehlt 2016 nmed entity recognition nd sitution frme systems. *Machine Translation*, pages 1–13.

Jenny Linet Copara Zea, Jose Eduardo Ochoa Luna, Camilo Thorne, and Goran Glavaš. 2016. Spanish ner with word representations and conditional random fields. In *Proceedings of the Sixth Named Entity Workshop*, pages 34–40.

Ayah Zirikly and Mona Diab. 2015. Named entity recognition for arabic social media. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 176–185.

7 Supplementary Information

The Table 3 shows category wise F scores for the experiments that gave the best results for Spanglish (the model is trained using pre-trained MUSE embeddings with a character level Bi-LSTM and a word level Bi-LSTM) and Arabish (a simple word level Bi-LSTM), which are discussed in detail in the paper.