GHHT at CALCS 2018: Named Entity Recognition for Dialectal Arabic Using Neural Networks

Mohammed Attia
Google Inc.
New York City
NY, 10011
attia@google.com

Younes Samih
Dept. of Computational Linguistics
Heinrich Heine University,
Düsseldorf, Germany
samih@phil.hhu.de

Wolfgang Maier
Independent Researcher
Tübingen, Germany
wolfgang.maier@gmail.com

Abstract

This paper describes our system submission to the CALCS 2018 shared task on named entity recognition on code-switched data for the language variant pair of Modern Standard Arabic and Egyptian dialectal Arabic. We build a a Deep Neural Network that combines word and character-based representations in convolutional and recurrent networks with a CRF layer. The model is augmented with stacked layers of enriched information such pre-trained embeddings, Brown clusters and named entity gazetteers. Our system is ranked second among those participating in the shared task achieving an FB1 average of 70.09%.

1 Introduction

The CALCS 2018 shared task (Aguilar et al., 2018) is about performing named entity recognition (NER) on Modern Standard Arabic (MSA) - Egyptian Arabic (EGY) code-switched tweets. Unlike previous shared tasks on code-switching, the data provided contains no code-switching annotation. Only nine categories of named entities are annotated using BIO tagging. While this makes the task a "pure" NER task, the difficulty is to design a model which can cope with the noise introduced by code-switching, challenging old systems tailored around MSA.

NER is a well-studied sequence labeling problem. Earlier work has applied standard supervised learning techniques to the problem, such as Hidden Markov Models (HMM) (Bikel et al., 1999), Maximum-Entropy Model (ME) (Bender et al., 2003; Curran and Clark, 2003; Finkel et al., 2005), Support Vector Machines (SVM) (Takeuchi and Collier, 2002), and Conditional Random Fields (CRF) (McCallum and Li, 2003). Standard data sets came from the English MUC-6 (Sundheim, 1995) and the multilingual CoNLL-02 (Tjong Kim Sang, 2002) and 03 (Tjong Kim Sang and De Meulder, 2003) shared tasks.

More recent work relies on neural networks. A number of architecture variants have proven to be effective (Huang et al., 2015; Lample et al., 2016; Chiu and Nichols, 2016; Ma and Hovy, 2016; Reimers and Gurevych, 2017). What they have in common is that they use a bidirectional LSTM (bi-LSTM) over vector representations of the input words in order model their left and right contexts. On top of the bi-LSTM, they use a CRF layer to take the final tagging decisions. Other than a softmax layer which would treat tagging decisions independently, the CRF is able to model the linear dependencies between labels. This is essential for NER, where for instance, B-LOCATION cannot be followed by I-PERSON. The architectures differ in their way of obtaining a vector representation for the input words. For instance, in Lample et al. (2016), each word embedding is obtained as a concatenation of the output of a bidirectional LSTM (bi-LSTM) over its characters and a pre-trained word vector. Ma and Hovy (2016) use convolutions over character embeddings with max-pooling for obtaining morphological features from the character level, similar to Chiu and Nichols (2016).

Proceedings of The Third Workshop on Computational Approaches to Code-Switching, pages 98–102
Melbourne, Australia, July 19, 2018.

Our system also relies on the bi-LSTM-CRF architecture. As input representation, we use both word embeddings and a character-level representation based on CNNs. Our system additionally employs a Brown Cluster representation, oversampling, and NE gazetteers.

The remainder of the paper is structured as follows in the following section, we provide a short decription of the task and the data set. Sect. 3 describes our system in detail. Sect. 4 presents our experiments, and Sect. 5 concludes the paper.

2 Task and Data Description

The shared task posed the problem of performing named-entity recognition on code-switched data given nine categories, namely PERSON, LOCATION, ORGANIZATION, GROUP, TITLE, PRODUCT, EVENT, TIME, OTHER.

The training set contains 10,100 tweets and 204,286 tokens, with an average tweet length of 20.2 tokens and 91.5 characters. 11.3% of all tokens are labeled as named entities. The most frequent category is PERSON with 4.3% of all tokens, followed by LOCATION (2.2%), GROUP and ORGANIZATION (1.3% each), as well as TITLE (1%). All other categories cover less than 1% of all tokens each, the least frequent category being OTHER (0.06%).

The validation set contains 1,122 tweets and 22,742 tokens, and exhibits similar average tweets lengths, as well as a similar distribution of labels.

3 System Description

We used a DNN model which is mainly suited for sequence tagging. It is a variant of the bi-LSTM-CRF architecture proposed by Ma and Hovy (2016); Lample et al. (2016); Huang et al. (2015).[1] It combines a double representation of the input words by using word embeddings and a character-based representation (with CNNs). The input sequence is processed with bi-LSTMs, and the output layer is a linear chain CRF. The model uses the following.

Word-level embeddings allow the learning algorithms to use large unlabeled data to generalize beyond the seen training data. We explore randomly initialized embeddings based on the seen training data and pre-trained embedding.

[1]Our implementation is mostly inspired by the work of Reimers and Gurevych (2017).

We train our word embeddings using word2vec (Mikolov et al., 2013) on a corpus we crawled from the web with total size of 383,261,475 words, consisting of dialectal texts from Facebook posts (8,241,244), Twitter tweets (2,813,016), user comments on the news (95,241,480), and MSA texts of news articles (from Al-Jazeera and Al-Ahram) of 276,965,735 words.

Character-level CNNs have proven effective for various NLP tasks due to their ability to extract sub-word information (ex. prefixes or suffixes) and to encode character-level representations of words (Collobert et al., 2011; Chiu and Nichols, 2016; dos Santos and Guimarães, 2015).

Bi-LSTM Recurrent neural networks (RNN) are well suited for modeling sequential data, achieving ground-breaking results in many NLP tasks (e.g., machine translation).

Bi-LSTMs (Hochreiter and Schmidhuber, 1997; Schuster and Paliwal, 1997) are capable of learning long-term dependencies and maintaining contextual features from both past and future states while avoiding the vanishing/exploding gradients problem. They consist of two separate bidirectional hidden layers that feed forward to the same output layer.

CRF is used jointly with bi-LSTMs to avoid the output label independence assumptions of bi-LSTMs and to impose sequence labeling constraints as in Lample et al. (2016).

Brown clusters (BC) Brown clustering is an unsupervised learning method where words are grouped based on the contexts in which they appear (Brown et al., 1992). The assumption is that words that behave in similar ways tend to appear in similar contexts and hence belong to the same cluster. BCs can be learned from large unlabeled texts and have been shown to improve POS tagging (Owoputi et al., 2013; Stratos and Collins, 2015). We test the effectiveness of using Brown clusters in the context of named entity recognition in a DNN model. We train BCs on our crawled code-switched corpus of 380 million words (mentioned above) with 100 Brown Clusters.

Named Entity Gazetteers We use a large collec-

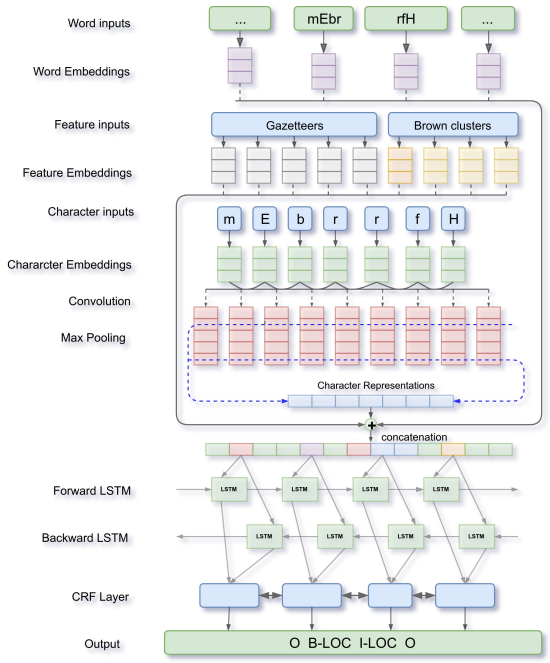

Figure 1: DNN Architecture.

tion of named entity gazetteers of 40,719 unique names from Attia et al. (2010), who collected named entities from the Arabic Wikipedia, and Benajiba et al. (2007), who annotated a corpus as part of a named entity recognition system.

The architecture of our model is shown in Figure 1. For each word in the sequence, the CNN computes the character-level representation with character embeddings as inputs. Then the character-level representation vector is concatenated with both word embeddings vector and feature embedding vectors (Brown Clusters and Gazetteers) to feed into the bi-LSTM layer. Finally, an affine transformation followed by a CRF is applied over the hidden representation of the bi-LSTM to obtain the probability distribution over all the named entity labels. Training is performed using stochastic gradient descent with momentum of 0.9 and batch size equal to 150. We employ dropout (Hinton et al., 2012) and early-stopping (Caruana et al., 2000) (with patience of 35) to mitigate overfitting. We use the hyper-parameters detailed in Table 1.

The only preprocessing operation we conducted on the data was to convert it into Buckwalter transliteration (a character-to-character mapping) in order to avoid the complexity of dealing with UTF-8 characters.

Layer	Hyper-Parameters	Value
Characters CNN	window size	4
	number of filters	40
Bi-LSTM	state size	100
Dropout	dropout rate	0.5
Word Emb.	dimension	300
Characters Emb.	dimension	100
Clustering Emb.	dimension	100
Gazetteer Emb.	dimension	2
	batch size	150

Table 1: Parameter fine-tuning

4 Experiments

We conduct five experiments with different layers stacked on top of each other, making use of word embeddings, character representation, and other features. The experiments are as follows:

Experiments	f-score	f-score macro
Baseline	95.70	66.49
Word+Chars	96.06	69.60
Word+Chars +Embed	96.92	72.38
Word+Chars +Embed+BC	96.99	72.30
Word+Chars +Embed+BC+OS	96.92	73.05
Word+Chars +Embed+BC +OS+GZ	97.33	77.97
Results on Test set	–	70.09

Table 2: DNN experiments and Results

Baseline. We use word representations only with randomly-initialized embeddings. It is to be mentioned that the shared task baseline for the test set is 62.71%.

Word+Chars. We add character representations in a one-dimensional CNN layer.

Word+Chars+Embed. We use pre-trained embeddings for words trained on a corpus of about 380 million words (described above) consisting of dialectal Egyptian and MSA data.

Word+Chars+Embed+BC. We add Brown Clusters (BC) to the network.

Word+Chars+Embed+BC+OS. We add oversampling (OS) to the network. We conduct oversampling by heuristically making 10-fold repetitions of sentences containing minority labels, in this case all classes other than the "O" label.

Word+Chars+Embed+BC+GZ. We further add a new layer for the named entity gazetteer (GZ).

Label	Total	% of data	Accuracy %
O	20031	88.08	99.20
B-PER	705	3.10	92.34
I-PER	408	1.79	89.71
B-LOC	358	1.57	88.83
I-LOC	116	0.51	79.31
B-GROUP	191	0.84	81.68
I-GROUP	112	0.49	76.79
B-ORG	149	0.66	79.19
I-ORG	114	0.50	80.70
B-TITLE	115	0.51	69.57
I-TITLE	143	0.63	81.12
B-PROD	55	0.24	76.36
I-PROD	26	0.11	61.54
B-EVENT	69	0.30	43.48
I-EVENT	52	0.23	51.92
B-TIME	61	0.27	85.25
I-TIME	18	0.08	38.89
B-OTHER	17	0.07	82.35
I-OTHER	2	0.01	50.00

Table 3: Results breakdown on the validation set

The results in Table 2 are reported on the validation set (except for the last row), and they show that the DNN model is incrementally improving by adding more features and external resources. The best result is obtained with the aggregation of all features.

Table 3 shows a breakdown of our system performance (in terms of accuracy) on the validation set. It also shows the number of instances and the ratio percentage for each label. As the table shows, the category "other" accounts for 88% of the entire data, while all other tags combined make up the remaining 12% which shows an imbalance in the representation of the other categories. Our system performs best with 'B-PER', 'I-PER', 'B-LOC' and 'B-TIME'.

Our system is ranked second among those participating in the shared task achieving an FB1 average of 70.09% with the first scoring 71.62%, which is a difference of about 1.5% absolute.

5 Conclusion

We have presented a description of our system participating in the Shared Task on "Named Entity Recognition on Code-switched Data". We build a deep neural network with multiple layers for accommodating various features, such as pre-trained word embeddings, Brown Clustering and named entity gazetteers. We have not relied on any linguistic rules, morphological analyzers or PoS taggers. We also make the different layers as optional plug-ins, which makes our system more adaptable and scalable for languages that do not have similar external resources.

References

Gustavo Aguilar, Fahad AlGhamdi, Victor Soto, Diab Mona, 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.

Mohammed Attia, Antonio Toral, Lamia Tounsi, Monica Monachini, and Josef van Genabith. 2010. An automatically built named entity lexicon for arabic. In *European Language Resources Association*, pages 3614–3621, Valletta, Malta.

Yassine Benajiba, Paolo Rosso, and José Miguel Benedíruiz. 2007. Anersys: An arabic named entity recognition system based on maximum entropy. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 143–153. Springer.

Oliver Bender, Franz Josef Och, and Hermann Ney. 2003. Maximum entropy models for named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 148–151.

Daniel M. Bikel, Richard Schwartz, and Ralph M. Weischedel. 1999. An algorithm that learns what's in a name. *Mach. Learn.*, 34(1-3):211–231.

Peter F Brown, Peter V Desouza, Robert L Mercer, Vincent J Della Pietra, and Jenifer C Lai. 1992. Class-based n-gram models of natural language. *Computational linguistics*, 18(4):467–479.

Rich Caruana, Steve Lawrence, and Lee Giles. 2000. Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping. In *NIPS*, pages 402–408.

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

R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12:2493–2537.

James Curran and Stephen Clark. 2003. Language independent ner using a maximum entropy tagger. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*.

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*, ACL '05, pages 363–370, Stroudsburg, PA, USA. Association for Computational Linguistics.

Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov. 2012. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. *CoRR*, abs/1508.01991.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California. Association for Computational Linguistics.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074, Berlin, Germany. Association for Computational Linguistics.

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.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A Smith. 2013. Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of NAACL-HLT 2013*, pages 380–390. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2017. Reporting score distributions makes a difference: Performance study of lstm-networks for sequence tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 338–348, Copenhagen, Denmark.

Cicero dos Santos and Victor Guimarães. 2015. Boosting named entity recognition with neural character embeddings. In *Proceedings of the Fifth Named Entity Workshop*, pages 25–33, Beijing, China. Association for Computational Linguistics.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.

Karl Stratos and Michael Collins. 2015. Simple semi-supervised pos tagging. In *VS@ HLT-NAACL*, pages 79–87.

Beth M. Sundheim. 1995. Overview of results of the muc-6 evaluation. In *Proceedings of the 6th Conference on Message Understanding*, MUC6 '95, pages 13–31, Stroudsburg, PA, USA. Association for Computational Linguistics.

Koichi Takeuchi and Nigel Collier. 2002. Use of support vector machines in extended named entity recognition. In *Proceedings of the 6th Conference on Natural Language Learning - Volume 20*, COLING-02, pages 1–7, Stroudsburg, PA, USA. Association for Computational Linguistics.

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.

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*, pages 142–147.