Effective Context and Fragment Feature Usage for Named Entity Recognition

Nargiza Nosirova, Mingbin Xu, Hui Jiang
Department of Electrical Engineering and Computer Science
Lassonde School of Engineering, York University
4700 Keele Street, Toronto, Ontario, Canada
{nana, xmb, hj}@cse.yorku.ca

Abstract

In this paper, we explore a new approach to named entity recognition (NER) with the goal of learning from context and fragment features more effectively, contributing to the improvement of overall recognition performance. We use the recent fixed-size ordinally forgetting encoding (FOFE) method to fully encode each sentence fragment and its left/right contexts into a fixed-size representation. Next, we organize the context and fragment features into groups, and feed each feature group to dedicated fully-connected layers. Finally, we merge each group's final dedicated layers and add a shared layer leading to a single output. The outcome of our experiments show that, given only tokenized text and trained word embeddings, our system outperforms our baseline models, and is competitive to the state-of-the-arts of various well-known NER tasks.

1 Introduction

Named entity recognition is the task of identifying proper nouns in a given text, and categorizing them into various types of entities. It is a fundamental problem in NLP, and its usefulness extends to tasks such as summarization and question answering (Aramaki et al., 2009; Ravichandran and Hovy, 2002). Traditional NER methods involve using hand-crafted features, such as conditional random fields (CRFs) . For example, McCallum and Li (2003) use a CRF model with a web-based lexicon as a feature enhancement, while Che et al. (2013) and Krishnan and Manning (2006) show the benefits of using non-local features. Over the recent years, researchers have turned to neural network architectures using non hand-crafted features. For example, Collobert et al. (2011) proposed a neural architecture that learns from word embeddings and requires little feature engineering. However, in his use of feed-forward neural networks (FFNNs), the context used around a word is restricted to a fixed-size window, which could result in the loss of potentially relevant information between words that are further apart. Xu et al. (2017) has recently proposed a non-sequence labelling method for NER with FOFE features, which can encode any variable-length sequence of words into a fixed-size representation. This method alleviates the limitations of Collobert's (2011) FFNN model, since the encoding uses the whole context around a word within the sentence, without settling for a fixed-size window. Our main contribution lies in extending the model suggested by Xu et al. (2017). In this paper, we propose a FOFE-based neural network model dedicating separate initial layers for fragment and context features and merging them into a shared layer to perform a unified prediction. Experimental results have shown that this method yields competitive results compared to the state-of-the-arts while increasing recall compared to our baseline models.

2 Model

Our neural network model is inspired by the work of Xu et al. (2017), where we use a local detection approach relying on the FOFE method to fully encode each sentence fragment and its contexts. Instead of using consecutive fully-connected layers that handle both context and fragment features, we propose to dedicate the initial fully-connected layers of the network to each feature kind, and subsequently combine the layers into a single shared layer that leads to a single output.

2.1 Fixed-Size Ordinally Forgetting Encoding (FOFE)

In this section, we describe the FOFE method. Given a vocabulary V, each word can be represented by a 1-of-$|V|$ one-hot vector. FOFE mimics bag-of-words but incorporates a forgetting factor to capture positional information. It encodes any variable length sequence composed of words in V. Let $S = w_1 \cdots w_N$ denote a sequence of N words from V, and denote $\boldsymbol{e_n}$ to be the one-hot vector of the n-th word in S, where $1 \leq n \leq N$. Assuming $\boldsymbol{z_0} = \mathbf{0}$, FOFE generates the code using a simple recursive formula from word w_1 to w_n of the sequence as follows:

$$\boldsymbol{z_n} = \alpha \cdot \boldsymbol{z_{n-1}} + \boldsymbol{e_n}$$

where α is a constant forgetting factor. Hence, $\boldsymbol{z_n}$ can be viewed as a fixed-size representation of the subsequence $\{w_1, w_2, \cdots, w_n\}$.

The theoretical properties that show FOFE code uniqueness are as follows:

Theorem 1. *If the forgetting factor α satisfies $0 < \alpha \leq 0.5$, FOFE is unique for any countable vocabulary V and any finite value N .*

Theorem 2. *For $0.5 < \alpha < 1$, given any finite value N and any countable vocabulary V, FOFE is almost unique everywhere, except only a finite set of countable choices of α.*

When $0.5 < \alpha < 1$, uniqueness is not guaranteed. However, the odds of ending up with such scenarios is small. Furthermore, it is rare to have a word reappear many times within a near context. Thus, we can say that FOFE can uniquely encode any sequence of variable length, providing a fixed-size lossless representation for any sequence. The proof for those theorems can be found in Zhang et al. (2015).

2.2 FOFE Context & Fragment Features

Fragment Features At word level, we extract the bag-of-words of the sentence fragment in both cased and uncased forms. Since we can view the fragment as a cased character sequence, it can be encoded with FOFE. We encode the sequence from left to right as well as from right to left. The encodings are then projected into a trainable character embedding matrix. For a fair comparison, we also use character CNNs to generate additional character-level features (Kim et al., 2015).

Context Features We convert the contexts of the fragment within the sentence to FOFE codes at word-level in cased and uncased forms, once containing the fragment, and once without. Those codes are then projected to lower-dimensional dense vectors using projection matrices. Those projection matrices are pre-trained using word2vec (Mikolov et al., 2013) and allowed to be modified during training.

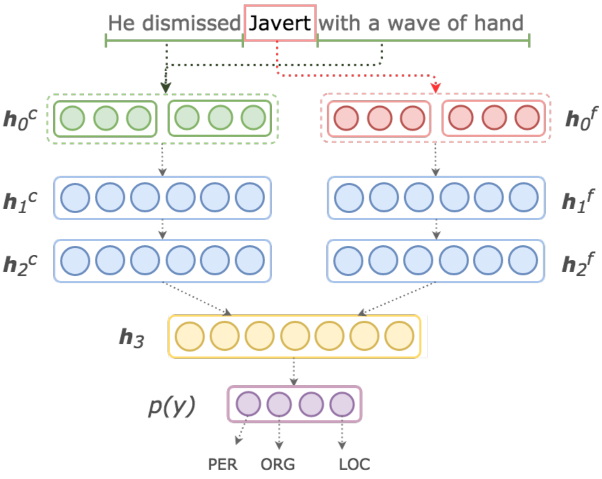

Figure 1: Illustration of an example structure of out model using FOFE codes. The window currently examines the fragment *Javert*.

2.3 Effective Context & Fragment Feature Usage for NER

We aim to consider influences between contexts and their corresponding fragments. If the context of a named entity fragment is not indicative of the fragment as being such, we can resort to learning the morphology of the fragment itself and grasp patterns that could lead us to believe that the fragment is indeed a named entity. By dedicating layers to each feature kind, we ensure that the context-based layer signals are tuned to identify entities based on the surrounding context, while the fragment-based layer signals identify them based on the morphology of the fragment itself. Since the layers merge into a shared layer, this permits the model to have a higher chance of predicting entities that would be hard to recognize based on the context, but self-evident based on the fragment. Furthermore, our model structure provides the flexibility of modeling information based on multiple sources of knowledge. Figure 1 illustrates an example of our neural architecture:

1. The context and fragment features are extracted from the text based on section 2.2 and concatenated within their categories resulting in $\mathbf{h}_0^w \big| w \in \{f, c\}$.
2. Two hidden layers $\mathbf{h}_1^w$ and $\mathbf{h}_2^w$ are fully connected to each category's embedding layer

$\mathbf{h}_0^w$.

3. A shared hidden layer $\mathbf{h}_3$ is fully connected to each $\mathbf{h}_2^w$.
4. The final layer is a softmax layer which outputs the probability distribution over classes, $p(y)$.

Each layer $h_{j,j>0}$ consists of ReLUs (Nair and Hinton, 2010) and are initialized based on a uniform distribution following Glorot et al. (2011).

Training At each training step, we randomly choose a training sample represented as a one of the feature forms and forward pass. Next, we backpropagate the loss of the current instance through the shared and feature dedicated layers and update the model parameters. For predicting models relative to the ground truth, we use categorical cross entropy loss. For optimization, we use mini-batch SGD with momentum of 0.9 (Bottou, 2010) and learning rates decaying exponentially by a factor of $1/16$. The mini-batch size is set to 128 for all experiments. Grid search is used for the other hyper-parameters, tuned against the task's development set with early stopping. The FOFE forgetting factor for all models are set to $\alpha_w = 0.5$ for words, and $\alpha_c = 0.8$ for characters. We apply dropout (Srivastava et al., 2014) to all layers with 0.5 probability. The same post-processing and decoding steps are followed as in Xu et al. (2017). Detailed hyper-parameter settings used in our experiments are given in Appendix A.

3 Experiments

We experiment with four diverse NER tasks of different languages: CoNLL-2003 English, OntoNotes 5.0 English and Chinese, trilingual KBP 2016 (English, Chinese and Spanish), and CoNLL-2002 Spanish. For the CoNLL-2003 task, we use cased and uncased word embeddings of size 256 trained on the Reuters RCV1 corpus. The remaining tasks use cased and uncased word embeddings of size 256 trained on the English (Parker et al., 2011), Spanish (Mendonca et al., 2009) and Chinese (Graff and Chen, 2005) Gigaword for the corresponding models evaluated in that language.

Dataset Description *CoNLL-2003 ENG*: The CoNLL-2003 (Tjong Kim Sang and De Meulder, 2003) dataset consists of newswire data from the Reuters RCV1 corpus. It has four entity types: person, location, organization and miscellaneous. *OntoNotes 5.0 ENG and ZH*: The OntoNotes dataset is built from sources such as broadcast conversation and news, newswire, telephone conversation, magazine and web text. It is tagged with eighteen entity types, some of which are: person, facility, organization, product and so forth. The dataset was assembled by Pradhan et al. (2013) for the CoNLL-2012 shared task, and specifies a standard train, validation, and test split followed in our evaluation.

KBP 2016: The KBP 2016 trilingual EDL task (Ji and Nothman, 2016) consists of identifying named entities (including nested) from a collection of recent news article and discussion forum documents in three languages, and their classification to the following named and nominal entity types: person, geo-political entity, organization, location and facility. We use an in-house dataset that consists of 10k English and Chinese documents labelled manually using KBP 2016 format. Since KBP 2016 does not contain any training and development data, we use our in-house data as training and validation data with a 90:10 split. We also make use of the KBP 2015 dataset as additional data for training.

CoNLL-2002 SPA: The CoNLL-2002 (Tjong Kim Sang, 2002) named entity data is tagged similarly to CoNLL-2003. We only make use of Spanish files for our experiments.

Baselines Our baseline models are from Xu et al. (2017). We use the author's findings for CoNLL-2003 and KBP 2016, and apply the implementation released by the author to train the model with OntoNotes 5.0 and CoNLL-2002 tasks.

4 Results and Discussion

The results for the trilingual KBP 2016 task are presented in Table 1, where our system outperforms the baseline by 3.2 F_1 points for English and 4.3 F_1 points for Chinese. It also outperforms the best KBP 2016 English system by 1 F_1 point. It is worth considering that the best 2016 system uses 5-fold cross-validation. The CoNLL-2003 results in Table 2 show that our model is nearly on par with the state-of-the-arts compared to both models that used the dev-set to train the model, and to those who used training data only. The OntoNotes English and Chinese task results are presented in Tables 3 and 4, and the CoNLL-2002 results in Table 5. We do not observe significant improvement

LANG	Xu et al. (2017)			Our model			2016 Best		
	P	R	F1	P	R	F1	P	R	F1
ENG	0.836	**0.680**	**0.750**	0.812	**0.756**	**0.782**	0.846	0.710	**0.772**
CMN	0.789	0.625	0.698	0.797	0.693	**0.741**	0.789	0.737	0.762
SPA	0.835	0.602	0.700	0.848	0.608	0.708	0.839	0.656	0.736
ALL	0.819	0.639	0.718	0.815	0.693	**0.749**	0.802	0.704	**0.756**

Table 1: Comparison of our model to the baseline models in Xu et al. (2017) as well as to the best system for the KBP 2016 task.

Model	**P**	**R**	**F1**
Collobert et al. (2011)	–	–	89.59
Huang et al. (2015)	–	–	90.10
Strubell et al. (2017)	–	–	90.54
Yang et al. (2016)	–	–	**90.94**
Luo et al. (2015) [1]	91.50	91.40	91.2
Lample et al. (2016)	–	–	90.94
Chiu and Nichols (2016)	91.39	91.85	**91.62**
Xu et al. (2017)	93.29	**88.27**	90.71
Xu et al. (2017) + dev set + 5-fold	92.58	89.31	**90.92**
Our model	91.81	**89.85**	**90.82**
Our model + dev set	92.02	90.30	**91.15**

Table 2: Results on the CoNLL-2003 ENG evaluation task. The three sections, in order, are models: trained with training set only, trained with both training and dev set, our baselines and our models.

Model	**P**	**R**	**F1**
Strubell et al. (2017)	–	–	**86.84**
Chiu and Nichols (2016)	86.04	86.53	86.28
Durrett and Klein (2014)	85.22	82.89	84.04
Xu et al. (2017)	86.84	84.94	85.88
Our model	86.95	85.44	**86.19**

Table 3: A comparison with the state-of-the-art results for the OntoNotes 5.0 ENG evaluation task.

Model	**P**	**R**	**F1**
Che et al. (2013)	74.38	65.78	69.82
Pappu et al. (2017)	–	–	67.2
Xu et al. (2017)	72.91	70.78	71.83
Our model	76.20	68.96	**72.40**

Table 4: A comparison with published results for the OntoNotes 5.0 ZH evaluation task.

Model	**P**	**R**	**F1**
dos Santos and Guimarães (2015)	82.21	82.21	82.21
Gillick et al. (2016)	–	–	82.95
Lample et al. (2016)	–	–	85.75
Yang et al. (2016)	–	–	**85.77**
Xu et al. (2017)	84.20	82.26	83.22
Our model	85.06	82.31	**83.66**

Table 5: A comparison with the state-of-the-arts results for the CoNLL-2002 SPA evaluation task.

on the CoNLL-2002 task. Both of our baseline and proposed models outperform the other published results for the OntoNotes Chinese task.

5 Related Work

In recent years, deep learning methods have gained much success in the NLP community. In view of the perceived limitations of FFNNs, recent methods involve more powerful neural networks, such as recurrent neural networks (RNNs), since they can process sequences of variable length (Huang et al., 2015). As for character-level modeling, studies have turned to Convolutional neural networks (CNNs). dos Santos and Guimarães (2015) have employed CNNs to extract character-level features for Spanish and Portuguese, and obtained successful NER results. Chiu and Nichols (2016) introduce a bi-directional LSTM-CNN architecture, and achieve state-of-the-art results in CoNLL2003. Strubell et al. (2017) uses Iterated dilated CNNs to reach state-of-the-art results in OntoNotes 5.0 English. Similarly, there have been conducted a few studies aiming at solving the problem of low recall in NER tasks (Mao et al., 2007; Kuperus et al., 2013).

6 Conclusion

We present a new neural model that can achieve near state-of-the-art results on a range of NER tasks by learning better representations of fragment and context features in NER systems. We use simple yet powerful neural networks with an effective context/fragment training approach.

References

Eiji Aramaki, Yasuhide Miura, Masatsugu Tonoike, Tomoko Ohkuma, Hiroshi Mashuichi, and Kazuhiko Ohe. 2009. Text2table: Medical text summarization system based on named entity recognition and modality identification. In *Proceedings of the BioNLP 2009 Workshop*, pages 185–192. Association for Computational Linguistics.

Léon Bottou. 2010. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186, Heidelberg. Physica-Verlag HD.

Wanxiang Che, Mengqiu Wang, Christopher D. Manning, and Ting Liu. 2013. Named entity recognition with bilingual constraints. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 52–62, Atlanta, Georgia. Association for Computational Linguistics.

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

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *J. Mach. Learn. Res.*, 12:2493–2537.

Greg Durrett and Dan Klein. 2014. A joint model for entity analysis: Coreference, typing, and linking. *Transactions of the Association for Computational Linguistics*, 2:477–490.

Dan Gillick, Cliff Brunk, Oriol Vinyals, and Amarnag Subramanya. 2016. Multilingual language processing from bytes. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1296–1306, San Diego, California. Association for Computational Linguistics.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Deep sparse rectifier neural networks. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pages 315–323, Fort Lauderdale, FL, USA. PMLR.

David Graff and Ke Chen. 2005. Chinese gigaword. *LDC Catalog No.: LDC2003T09, ISBN*, 1:58563–58230.

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

Heng Ji and Joel Nothman. 2016. Overview of tac-kbp2016 tri-lingual edl and its impact on end-to-end cold-start kbp. In *Proceedings of the Nineth Text Analysis Conference (TAC2016)*.

Yoon Kim, Yacine Jernite, David Sontag, and Alexander M. Rush. 2015. Character-aware neural language models. *CoRR*, abs/1508.06615.

Vijay Krishnan and Christopher D. Manning. 2006. An effective two-stage model for exploiting non-local dependencies in named entity recognition. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, ACL-44, pages 1121–1128, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jasper Kuperus, Cor J. Veenman, and Maurice van Keulen. 2013. Increasing ner recall with minimal precision loss. In *Proceedings of EISIC*, pages 106–111.

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. Association for Computational Linguistics.

Gang Luo, Xiaojiang Huang, Chin-Yew Lin, and Zaiqing Nie. 2015. Joint entity recognition and disambiguation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 879–888. Association for Computational Linguistics.

Xinnian Mao, Xu Wei, Yuan Dong, He Saike, and Haila Wang. 2007. Using non-local features to improve named entity recognition recall. In *Proceedings of the 21st Pacific Asia Conference on Language, Information and Computation*, pages 303–310.

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 - Volume 4*, CONLL '03, pages 188–191, Stroudsburg, PA, USA. Association for Computational Linguistics.

Angelo Mendonca, David Andrew Graff, and Denise DiPersio. 2009. *Spanish gigaword second edition*. Linguistic Data Consortium.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pages 3111–3119, USA. Curran Associates Inc.

Vinod Nair and Geoffrey E. Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference* *on International Conference on Machine Learning*, ICML'10, pages 807–814, USA. Omnipress.

Aasish Pappu, Roi Blanco, Yashar Mehdad, Amanda Stent, and Kapil Thadani. 2017. Lightweight multilingual entity extraction and linking. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, WSDM '17, pages 365–374, New York, NY, USA. ACM.

Robert Parker, David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2011. English gigaword. *Linguistic Data Consortium*.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Bjrkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. 2013. Towards robust linguistic analysis using ontonotes. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 143–152, Sofia, Bulgaria. Association for Computational Linguistics.

Deepak Ravichandran and Eduard Hovy. 2002. Learning surface text patterns for a question answering system. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 41–47, Stroudsburg, PA, USA. Association for Computational Linguistics.

Cícero Nogueira dos Santos and Victor Guimarães. 2015. Boosting named entity recognition with neural character embeddings. *CoRR*, abs/1505.05008.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958.

Emma Strubell, Patrick Verga, David Belanger, and Andrew McCallum. 2017. Fast and accurate entity recognition with iterated dilated convolutions. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 2670–2680.

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 - Volume 4*, CONLL '03, pages 142–147, Stroudsburg, PA, USA. Association for Computational Linguistics.

Mingbin Xu, Hui Jiang, and Sedtawut Watcharawittayakul. 2017. A local detection approach for named entity recognition and mention detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1237–1247, Vancouver, Canada. Association for Computational Linguistics.

Zhilin Yang, Ruslan Salakhutdinov, and William W. Cohen. 2016. Multi-task cross-lingual sequence tagging from scratch. *CoRR*, abs/1603.06270.

Shiliang Zhang, Hui Jiang, Mingbin Xu, Junfeng Hou, and Li-Rong Dai. 2015. A fixed-size encoding method for variable-length sequences with its application to neural network language models. *CoRR*, abs/1505.01504.