Multilingual Named Entity Recognition using Hybrid Neural Networks

Yan Shao, Christian Hardmeier, Joakim Nivre

Department of Linguistics and Philology
Uppsala University
{yan.shao, christian.hardmeier, joakim.nivre}@lingfil.uu.se

1. Introduction

Named entity recognition is a significant subtask of information extraction. Most of the high performing NER systems model the task as a sequence labelling or a structured prediction problem (Lafferty et al., 2001; Finkel et al., 2005; Ratinov and Roth, 2009). In recent years, neural network based models obtain remarkable success in a wide range of natural language processing tasks, which outperforms the classical models in terms of accuracy (Collobert et al., 2011; Graves, 2012).

The aim of this study is to systematically evaluate the performances of different neural networks for NER in various languages. We also investigate the impact of additional features and configurations. Three baseline models, including a feedforward network, a standard Bi-LSTM and a window-based Bi-LSTM are extensively tested with different feature and hyper-parameter settings on three data sets. The experimental results indicate that the neural network based models are generally robust and capable of achieving reasonable accuracies across different languages. In addition, the window-based Bi-LSTM is more robust than the standard Bi-LSTM when less information is available for the task. The effectiveness of the features depends on both the architecture of the model and the data set, except that all the models benefit greatly from the pre-trained word embeddings and the Conditional Random Fields (CRF) based interface. Overall, our best performing models are competitive when compared to the state-of-the-art NER systems.

2. Baseline Models

The baseline models do not employ any extra features other than tokens, which are mapped into randomly initialised vectors instead of pre-trained word embeddings. Figure 1 shows the general architectures of the three baseline models with an input instance, in which the target token is *German*.

3. Additional Features and Configurations

We extend the baseline models via adding different features at both the token level and the character level. Moreover, we use a CRF interface at the output layer to handle sequence prediction.

We add a CRF interface after the softmax layer to exploit relevant information for better sequence prediction. We use the probability distribution output by the softmax layer as the only feature for the CRF interface.

Named entities are closely associated with case information in many languages. We follow the approach of Collobert et al. (2011) and make a separate case lookup table to restore the case information in the English and German experiments.

We assume that some extra information that is helpful to NER is encoded at the character level, such as prefixes, suffixes or some special character sequences. We use a one dimensional convolutional layer to capture the relevant information. The output of the convolutional layer is merged with word embeddings and the other word-level features and passed further to the neural networks.

We employ publicly available pre-trained word embeddings (Bengio et al., 2003) as one of our core features. Some types of named entities have specific POS tags that are helpful to NER. In our experiments, we assign predicted POS tags and use them as an extra feature. Furthermore, some Gazetteers are adopted as external lexical resources.

4. Experiments

Word emb size		**Activation functions**	
English	50	Convolution	relu
German	100	LSTM inner	hard sigmoid
Arabic	64	Others	tanh
Other emb size		**Optimiser**	
Character	30	Feedforward	SGD
POS	50	LSTM	RMSprop
Hidden layer size		**Learning rate**	
Feedforward	200	SGD	0.05, 1e-6 decay
LSTM	100	RMSprop	initial: 0.001
Convolution		**No. of epochs**	
Width	3	Feedforward	40
No. of kernels	50	LSTM	80
Loss function		**Weight initialisation**	
Cat. Cross-entropy		uniform	
Batch size		**Dropout rate**	
20		0.2	

Table 1: Parameter settings for the feature experiments

4.1 Experimental Settings

4.2 Data sets

Three data sets in different languages are used in our paper for the experiments. The English data set is from the CoNLL 2003 NER shared task (Tjong Kim Sang and De Meulder, 2003). The German data is from the GermEval 2014 NER shared task (Benikova et al., 2014a). The data set contains annotations of nested named entities (Benikova et al., 2014b). For Arabic, we use ANERcop (Benajiba et

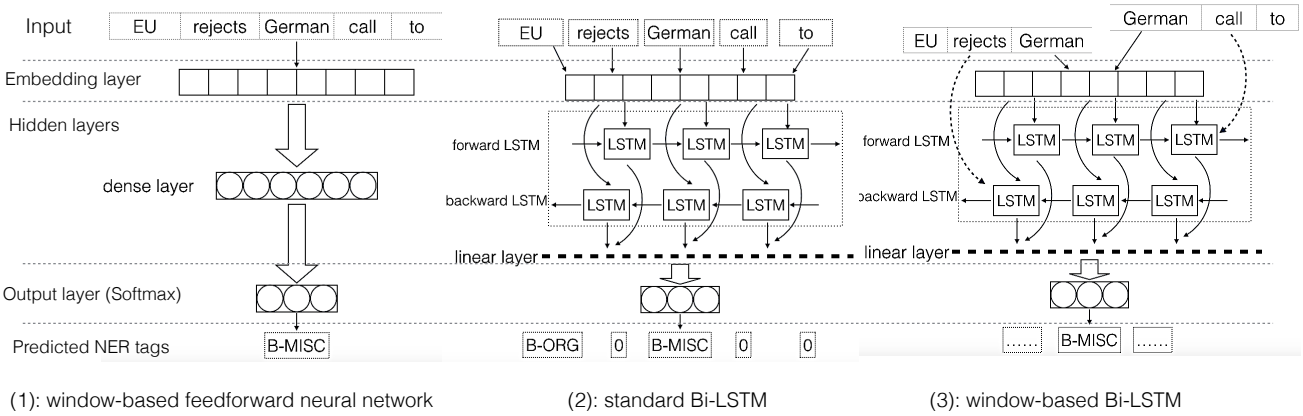

Figure 1: Architecture of the three baseline models

Data set	Models	Baseline	+CRF	+Case	+Char.	+Emb	+POS	+Gazett.
English	Feedforward	73.65	76.46	81.67	81.43	87.26	87.24	87.59
	Bi-LSTM	67.48	73.52	78.30	79.71	87.34	88.15	**88.49**
	Win-Bi-LSTM	72.04	76.27	82.33	80.47	87.41	87.59	88.31
German	Feedforward	55.24	56.79	60.16	62.43	74.58	74.92	**75.54**
	Bi-LSTM	51.90	56.43	60.61	58.50	74.01	74.49	74.59
	Win-Bi-LSTM	55.20	54.64	59.78	58.39	74.47	74.70	74.94
Arabic	Feedforward	56.41	57.30	—	58.43	69.28	69.94	70.28
	Bi-LSTM	52.98	53.77	—	53.30	68.62	**69.95**	53.14
	Win-Bi-LSTM	52.43	53.63	—	54.08	68.49	69.84	70.57

Table 2: Results of the cumulative feature experiments

Data set	Models	Full Conf.	-CRF	-Case	-Char.	-Emb.	-POS	-Gazett.
English	Feedforward	87.49	86.19	87.36	88.30	85.59	**88.70**	87.33
	Bi-LSTM	88.44	82.60	86.99	86.30	81.80	87.25	87.58
	Win-Bi-LSTM	88.07	86.90	87.23	**88.79**	81.82	87.88	88.20
German	Feedforward	75.39	72.91	75.35	75.15	66.20	75.15	75.25
	Bi-LSTM	74.44	71.52	72.77	74.49	66.36	73.21	**74.66**
	Win-Bi-LSTM	**74.88**	72.72	73.75	74.69	62.53	74.00	74.82
Arabic	Feedforward	**70.39**	67.35	—	69.79	63.37	69.69	69.28
	Bi-LSTM	54.18	50.42	—	54.54	38.02	47.17	69.56
	Win-Bi-LSTM	70.64	68.52	—	**71.52**	59.42	69.06	70.03

Table 3: Results of the ablation feature experiments

al., 2007) and follow the same splits as described in Benajiba et al. (2007).

First, we systematically investigate the impact of the additional features on our employed NER models. Two sets of experiments are carried out under consistent parameter settings. In the first set, we start from the very basic models and incrementally add the extra configuration and information. In the second set, we ablate one of the extra features at each time from the models that are equipped with full configuration and feature sets. Table 1 shows the parameters that are used for the feature experiments.

Based on the results of the feature experiments, for each model on each data set, we choose the best feature configurations and experiment further with two hyper parameters, namely the size of the hidden layer and the dropout rate.

4.3 Experimental Results

Table 2 and Table 3 show the results of the feature experiments in terms of F1-score.

For the baseline models, the feedforward neural networks significantly outperform the LSTM models when no extra features are employed. Comparatively, the LSTM models, especially the standard Bi-LSTM relies more on the extra information to achieve better performance. The window-based Bi-LSTM has no clear advantage in accuracy over the standard Bi-LSTM. However, it is more robust without or with fewer extra features thanks to the information provided by the context window.

Data set	Models	100		200		300	
		0.2	0.5	0.2	0.5	0.2	0.5
English	Feedforward	87.84	88.00	88.16	88.10	87.76	87.67
	Bi-LSTM	87.69	87.93	88.58	88.18	88.33	88.25
	Win-Bi-LSTM	87.55	87.87	88.70	88.44	88.82	**88.91**
German	Feedforward	**76.12**	75.21	75.81	75.19	75.33	75.25
	Bi-LSTM	74.63	74.28	75.19	75.00	75.14	75.01
	Win-Bi-LSTM	74.78	74.31	75.43	75.19	75.34	75.22
Arabic	Feedforward	70.21	71.20	70.24	70.85	70.54	71.03
	Bi-LSTM	69.06	70.41	70.56	70.85	70.96	70.59
	Win-Bi-LSTM	71.22	71.02	71.18	**71.34**	71.06	71.02

Table 4: Results of using different hidden layer sizes and dropout rates

Data set	Systems	System Description	F1 Score
English	Florian et al. (2003)	Combinations of multiple classifiers	88.76
	Collobert et al. (2011)	Conv. Feedforward Neural Network with Gazetteer	89.59
	Huang et al. (2015)	Bi-BLSTM-CRF with hand crafted features	90.10
	Chiu and Nichols (2015)	Bil-LSTM-CNNs with large lexical resources	91.62
	This paper	Win-Bi-LSTM, Hidden Layer: 300, Dropout: 0.5	88.91
German	Hänig et al. (2014)	Ensemble of classifiers with lexical lists	76.38
	Reimers et al. (2014)	Feedforward Neural network with Wiki-definitions	75.09
	Agerri and Rigau (2016)	Multiple induced features from different sources	76.43
	This paper	Feedforward, Hidden Layer: 100, Dropout: 0.2	76.12
Arabic	Benajiba and Rosso (2008)	CRF with multiple features	79.21
	This paper	Win-Bi-LSTM, Hidden Layer: 200, Dropout: 0.5	71.34

Table 5: Comparison with some state-of-the-art NER systems

From the results, we see that pre-trained word embeddings are the most crucial feature for NER. Substituting the random embeddings with the pre-trained word embeddings leads to drastic improvements in all the experiments and vice versa. The GloVe embeddings work better than SENNA with our experiments.

The CRF interface is another important boost for the NER models. We can see a drastic decrease when it is removed in the ablation experiments. As expected, the capitalisation feature is much more helpful for English than for German due to the fact that the predominant majority of the English named entities starts with capital letters while in German, this is the case for all the nouns.

The character level information appears to be helpful, but no significant improvement is obtained. Sometimes it even brings detrimental effects.

The improvement gained by using predicted POS tags is rather limited. Employing the gazetteers leads to some slight improvements in general, but we also see that the standard Bi-LSTM model on Arabic is consistently and drastically under performing when the gazetteer is used, which indicates that the Bi-LSTM model has potential robustness problems.

Based on the evaluation results in Table 2 and Table 3, we select the best performing feature settings whose scores are highlighted in the tables and test with two hyper parameters, the hidden layer size and the dropout rate. The results are presented in Table 4. Overall, we see that all the models perform reasonably well despite of different values of the two parameters.

4.4 Comparison with the State of the Art

We compare our best performing models with some of the state-of-the-art NER systems as in Table 5. Our model is comparable to the best performing NER system on English and German data sets even though extra lexical resources such as Wikipedia definitions are not used. For Arabic, our system is notably behind Benajiba and Rosso (2008). They preprocess the data via tokenisation, which appears to be helpful while it is not employed in our model. Their POS information also gives a great boost (around 7%) to the baseline. If the POS information is subtracted in their model, the accuracy is comparable to ours.

5. Conclusion

In this paper, we experiment with three neural network based models for NER on three data sets in English, German and Arabic. The experimental evaluations indicate that the feedforward network is robust and efficient in training while the Bi-LSTM is better in accuracy except for the German data. Our proposed window-based Bi-LSTM is both robust and accurate when less features are employed thanks to the combination of the LSTM based recurrent network and context information. The effectiveness of the features depends on both the architecture of the neural network and the data set. In general, the accuracies of our best performing models are comparable to the state-of-the-art systems.

References

Rodrigo Agerri and German Rigau. 2016. Robust multilingual named entity recognition with shallow semi-supervised features. *Artificial Intelligence*, 238:63–82.

Yassine Benajiba and Paolo Rosso. 2008. Arabic named entity recognition using conditional random fields. In *Proceedings of Workshop on HLT & NLP within the Arabic World, LREC*, volume 8, pages 143–153. Citeseer.

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.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of Machine Learning Research*, 3(Feb):1137–1155.

Darina Benikova, Chris Biemann, Max Kisselew, and Sebastian Padó. 2014a. Germeval 2014 named entity recognition shared task: companion paper. *Organization*, 7:281.

Darina Benikova, Chris Biemann, and Marc Reznicek. 2014b. Nosta-d named entity annotation for german: Guidelines and dataset. In *LREC*, pages 2524–2531.

Jason PC Chiu and Eric Nichols. 2015. Named entity recognition with bidirectional lstm-cnns. *arXiv preprint arXiv:1511.08308*.

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

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.

Radu Florian, Abe Ittycheriah, Hongyan Jing, and Tong Zhang. 2003. Named entity recognition through classifier combination. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 168–171. Association for Computational Linguistics.

Alex Graves. 2012. Neural networks. In *Supervised Sequence Labelling with Recurrent Neural Networks*, pages 15–35. Springer.

Christian Hänig, Stefan Bordag, and Stefan Thomas. 2014. Modular classifier ensemble architecture for named entity recognition on low resource systems. In *Workshop Proceedings of the 12th Edition of the KONVENS Conference*, pages 113–116.

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

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Lev Ratinov and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, pages 147–155. Association for Computational Linguistics.

Nils Reimers, Judith Eckle-Kohler, Carsten Schnober, Jungi Kim, and Iryna Gurevych. 2014. Germeval-2014: Nested named entity recognition with neural networks. In *Proceedings of the KONVENS GermEval Shared Task on Named Entity Recognition*.

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.