A Supplementary Material

A.1 Evaluation

We adopt the standard micro F1 score, recall, precision as the metrics. A correct entity prediction is that the extracted entity matches the ground truth including the type, length and position in the sentence.

A.2 Tuning Hyperparameters

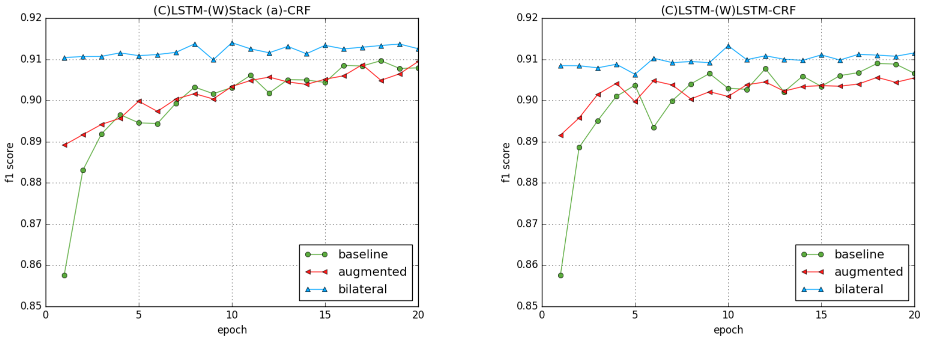
We use the GloVe (200-D) for English and the FastText (300-D) for German, Dutch, and Spanish. Dropout training (0.5) and decayed learning rate (0.9) are helpful to get a stable result. The dimensions of character embeddings are 25 and 100 respectively. We use 300-D LSTM to generate hidden states. For the Chinese language, we randomly initialize the word embeddings (300-D). For the biomedical NER, we trained the 200-D GloVe word embeddings on the PubMed and PMC archives which contains the citations for biomedical literature from MEDLINE, life science journals, and online books. We use the Adam [9] algorithm to update our model parameters. Mask trick also generates a little influence, and we apply the mask to eliminate the influence of padding placeholders in the internal and loss layers. These experiments are run on a Intel(R) Xeon(R) CPU E7-4830 v3 @ 2.10GHz (Mem: 976G) and one GPU Tesla K40c.

A.3 Results of different languages

English As shown in Figure 1, the augmented model and bilateral model both achieve high scores in the first epoch. During the training process, the generated data enhance the generalization performance of the augmented sub-network. For the bilateral model, two sub-networks have already formed their functions so that the fine-tuning process will converge fast.

Table 1 lists some top-performing studies, where the *word, char, gaz, cap, pos, corpus, transfer* represent the word embeddings, character embeddings, gazetteers, capitalization feature, Part-of-speech feature, large corpus and transfer learning respectively. The last row is our bilateral model of Stack (a). Our method achieved a higher result compared with the models using the same features.

Spanish, Dutch, German, and Chinese Table 2 lists the results. We adopt the (C)LSTM-(W)LSTM-CRF model as our Baseline. The $^+$ indicates models trained with external resources. ×2 means that we use two baseline sub-networks. Our system achieved good performances in all four languages. We observe the bilateral models achieved significant improvements to our baseline models on some datasets. For example, the Chinese NER dataset comes from literature text which is more difficult than other domain. Various rhetorical devices pose great challenges. A simple example of personification is that “Hamlett” is a



(a) Results of (C)LSTM-(W)Stack (a)-CRF (b) Results of (C)LSTM-(W)LSTM-CRF

Fig. 1: Results on the test set during model training in Table **??**

Table 1: Results report where the first part uses external data or features, and the other part only use the word and character embeddings

Algorithm	**word**	**char**	**gaz**	**cap**	**pos**	**corpus**	**transfer**	**F1**
[4]	✓		✓	✓				89.59
[17]	✓	✓					✓	91.26
[2]	✓	✓	✓					91.62
[12]	✓	✓				✓		92.22
CVT+Multi [3]	✓	✓				✓		92.60
BERT (Base) [5]	✓					✓		92.40
BERT (Large) [5]	✓					✓		**92.80**
[2]	✓	✓						90.91
[10]	✓	✓						90.94
[11]	✓	✓						91.21
[16]	✓	✓						91.35
this work	✓	✓						91.47

person name but refers to a rabbit$_{\text{[Thing]}}$ [15]. In this framework, the entity of rhetorical devices can be replaced by a more direct entity during model training. This indicates that our approach can be applied to enhance different languages.

For the German language, [13] achieved the highest performance by using a Bi-LSTM based model [10]. This is because they use more pre-trained vectors in the input, such as character- and subword- vectors. These features significantly improve the result on the small dataset, while the methods not using these features, i.e., the Bi-LSTM-CRF [10] and GermaNER achieved 78.76 and 79.37 respectively.

Table 2: Results of four languages

(a) Spanish NER results

Algorithms	**F1**
BTS$^+$ [6]	82.95
[10]	85.75
this work (Baseline)	86.60
this work (Baseline×2)	86.56
this work (Augment)	86.27
this work (Bilateral)	**87.08**

(b) Dutch NER results

Algorithms	**F1**
[10]	81.74
BTS$^+$ [6]	82.84
this work (Baseline)	87.39
this work (Baseline×2)	87.42
this work (Augment)	87.60
this work (Bilateral)	**88.30**

(c) German NER results

Algorithms	**F1**
StanfordNER	73.33
BTS$^+$ [6]	76.22
[10]	78.76
GermaNER [1]	79.37
[13]	**82.99**
this work (Baseline)	78.90
this work (Baseline×2)	78.89
this work (Augment)	79.58
this work (Bilateral)	80.31

(d) Chinese NER results

Algorithms	**F1**
Bi-LSTM [15]	65.05
CRF+Features [15]	72.03
this work (Baseline)	70.81
this work (Baseline×2)	70.83
this work (Augment)	73.12
this work (Bilateral)	**73.96**

A.4 Results on Biomedical NER

In biomedical domain, one of the challenges is the limited size of training data. However, expanding biomedical datasets is more challenging because annotators need to design and understand domain-specific criteria, which complicates the process. There are many feature-based systems, but they cannot be used in different areas. Automatically expanding datasets is a promising way to enhance the use of deep learning models. To evaluate our model in the biomedical field, we also conduct experiments with the following two corpora, using the (C)LSTM-(W)LSTM-CRF model as the Baseline. As shown in Table 3, in the GELLUS corpus, the augmented and the bilateral models improve 5.11% and 6.08% F1 sore than our baseline model. This means that our approach will be a good choice in the biomedical field.

Table 3: Results of IUPAC Chemical terms and Cell lines on the SCAI chemicals corpus and the GELLUS corpus respectively

Algorithm	**SCAI**	**GELLUS**
OSCAR4 [8]	57.3	—
ChemSpot [14]	68.1	—
CRF [7]	—	72.14
LSTM-CRF [7]	—	73.51
this work (Baseline)	69.08	78.78
this work (Baseline×2)	69.06	78.80
this work (Augment)	**69.98**	83.89
this work (Bilateral)	69.79	**84.86**

A.5 Representation Test

To evaluate the quality of learning representation, we design a mask experiment. We also adopt the (C)LSTM-(W)LSTM-CRF as the Baseline model.

This experiment aims to test the reasoning capacity of the bilateral model. We hide all the entity names in the test set and replaces the entity word with "UNK" token, i.e. *"Germany imported 47000 sheep from Britain"* becomes *"UNK imported 47000 sheep from UNK"* . As shown in Table 4, the augmented network achieves a better result than the baseline model, while the bilateral model forms a trade-off. This scenario is more subtle than the case of out-of-vocabulary (OOV) words because entities cannot differentiate each other from morphology. This model needs to infer the entity type based on the context information, which is a way to test the contextual representation. The result demonstrates that the augmented model can extract more features from context to infer the entity type.

Table 4: Results of mask test on the CoNLL-2003 English test set

Model	Recall	Precision	F1
Baseline	32.68	36.09	34.40
Augment	**41.36**	**41.51**	**41.43**
Bilateral	36.13	38.71	37.38

A.6 Case Study

It is instructive to analyze the type and length of entities in the prediction results. We select some samples from the CoNLL-2003 English dataset. The Baseline model adopts the (C)LSTM-(W)LSTM-CRF model in Table **??**. As shown in Table 5, in sentence 1, the *Costa* is an ambiguity entity since the *Costa* represents a location in most cases but the context decides the *Costa* to be the person. This suggests that the baseline model focuses more on the word sense from the statistic level. The augmented model makes a correct prediction because it reduces the entity specificity by extracting more information from the context pattern. Sentence 4 demonstrates the same issue.

Table 5: Examples on the CoNLL-2003 English test set where the blue and red label represent the correct and incorrect predictions respectively and [...]$_{\text{miss}}$ is the not recognized entity

Model	Sentence
Baseline	[Weah]$_{\text{PER}}$ has admitted head butting [Costa]$_{\text{LOC}}$ but said he reacted to racist taunts .
Augment	[Weah]$_{\text{PER}}$ has admitted head butting [Costa]$_{\text{PER}}$ but said he reacted to racist taunts .
Bilateral	[Weah]$_{\text{PER}}$ has admitted head butting [Costa]$_{\text{PER}}$ but said he reacted to racist taunts .
Baseline	SOCCER - [ENGLISH]$_{\text{MISC}}$ [F.A. CUP]$_{\text{MISC}}$ SECOND ROUND RESULT .
Augment	SOCCER - [ENGLISH F.A. CUP]$_{\text{MISC}}$ SECOND ROUND RESULT .
Bilateral	SOCCER - [ENGLISH F.A. CUP]$_{\text{MISC}}$ SECOND ROUND RESULT .
Baseline	(Corrects headline from [NBA]$_{\text{ORG}}$ to [NHL]$_{\text{ORG}}$ and corrects team name in second result from [La Clippers]$_{\text{ORG}}$ to [Ny Islanders]$_{\text{LOC}}$.
Augment	(Corrects headline from [NBA]$_{\text{ORG}}$ to [NHL]$_{\text{miss}}$ and corrects team name in second result from [La Clippers]$_{\text{ORG}}$ to [Ny Islanders]$_{\text{ORG}}$.
Bilateral	(Corrects headline from [NBA]$_{\text{ORG}}$ to [NHL]$_{\text{ORG}}$ and corrects team name in second result from [La Clippers]$_{\text{ORG}}$ to [Ny Islanders]$_{\text{ORG}}$.
Baseline	High-flying Italy topped the league in a week of meagre returns on government bonds , [Salomon Brothers]$_{\text{PER}}$ said on Friday .
Augment	High-flying Italy topped the league in a week of meagre returns on government bonds , [Salomon Brothers]$_{\text{ORG}}$ said on Friday .
Bilateral	High-flying Italy topped the league in a week of meagre returns on government bonds , [Salomon Brothers]$_{\text{ORG}}$ said on Friday .

In sentence 2, the *English* is a “S-MISC” entity in most scenarios, but the *ENGLISH F.A. CUP* is a whole phrase here. This implies that the baseline model sometimes makes a greedy prediction about the current word. The augmented model enhanced the representations by considering more context information.

In sentence 3, from a syntax level we know the pattern “from {1} to {2}” means that the {1} and {2} are more likely in the same type. The baseline model predicted the “from ORG to LOC”. The augmented model fix the last pair to "from ORG to ORG" but missed the *NHL* entity in the other pair. The bilateral model made a correct prediction in both pairs. These results indicate that the bilateral model helps to improve the model robustness through enhancing the representations.

References

1. Benikova, D., Muhie, S., Prabhakaran, Y., Biemann, S.C.: C.: Germaner: Free open german named entity recognition tool. In: In: Proc. GSCL-2015. Citeseer (2015)
2. Chiu, J.P., Nichols, E.: Named entity recognition with bidirectional lstm-cnns. arXiv preprint arXiv:1511.08308 (2015)
3. Clark, K., Luong, M.T., Manning, C.D., Le, Q.V.: Semi-supervised sequence modeling with cross-view training. arXiv preprint arXiv:1809.08370 (2018)
4. Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., Kuksa, P.: Natural language processing (almost) from scratch. Journal of Machine Learning Research **12**(Aug), 2493–2537 (2011)
5. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
6. Gillick, D., Brunk, C., Vinyals, O., Subramanya, A.: Multilingual language processing from bytes. arXiv preprint arXiv:1512.00103 (2015)
7. Habibi, M., Weber, L., Neves, M., Wiegandt, D.L., Leser, U.: Deep learning with word embeddings improves biomedical named entity recognition. Bioinformatics **33**(14), i37–i48 (2017)
8. Jessop, D.M., Adams, S.E., Willighagen, E.L., Hawizy, L., Murray-Rust, P.: Oscar4: a flexible architecture for chemical text-mining. Journal of cheminformatics **3**(1), 41 (2011)
9. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
10. Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., Dyer, C.: Neural architectures for named entity recognition. arXiv preprint arXiv:1603.01360 (2016)
11. Ma, X., Hovy, E.: End-to-end sequence labeling via bi-directional lstm-cnns-crf. arXiv preprint arXiv:1603.01354 (2016)
12. Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. arXiv preprint arXiv:1802.05365 (2018)
13. Riedl, M., Padó, S.: A named entity recognition shootout for german. In: Proceedings of ACL 2018 (Volume 2: Short Papers). vol. 2, pp. 120–125 (2018)
14. Rocktäschel, T., Weidlich, M., Leser, U.: Chemspot: a hybrid system for chemical named entity recognition. Bioinformatics **28**(12), 1633–1640 (2012)
15. Xu, J., Wen, J., Sun, X., Su, Q.: A discourse-level named entity recognition and relation extraction dataset for chinese literature text. arXiv preprint arXiv:1711.07010 (2017)
16. Yang, J., Liang, S., Zhang, Y.: Design challenges and misconceptions in neural sequence labeling. In: Proceedings COLING 2018 (2018), https://arxiv.org/pdf/1806.04470.pdf
17. Yang, Z., Salakhutdinov, R., Cohen, W.W.: Transfer learning for sequence tagging with hierarchical recurrent networks. arXiv preprint arXiv:1703.06345 (2017)