

FlexNER: A Flexible LSTM-CNN Stack Framework for Named Entity Recognition

Hongyin Zhu^{1,2}, Wenpeng Hu³, and Yi Zeng^{1,2,4,5,*}

¹ Institute of Automation, Chinese Academy of Sciences, Beijing, China

² University of Chinese Academy of Sciences, Beijing, China

³ School of Mathematical Sciences, Peking University, Beijing, China

⁴ Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences, Shanghai, China

⁵ National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Science, Beijing, China

{zhuhongyin2014,yi.zeng}@ia.ac.cn, wenpeng.hu@pku.edu.cn

Abstract. Named entity recognition (NER) is a foundational technology for information extraction. This paper presents a flexible NER framework⁶ compatible with different languages and domains. Inspired by the idea of distant supervision (DS), this paper enhances the representation by increasing the entity-context diversity without relying on external resources. We choose different layer stacks and sub-network combinations to construct the bilateral networks. This strategy can generally improve model performance on different datasets. We conduct experiments on five languages, such as English, German, Spanish, Dutch and Chinese, and biomedical fields, such as identifying the chemicals and gene/protein terms from scientific works. Experimental results demonstrate the good performance of this framework.

Keywords: Named entity recognition · data augmentation · LSTM-CNN.

1 Introduction

The NER task aims to automatically identify the atomic entity mentions in textual inputs. This technology is widely used in many natural language processing pipelines, such as entity linking, relation extraction, question answering, etc. This paper describes a portable framework that can use different layer stacks and sub-network combinations to form different models. This framework does not rely on language/domain-specific external resources so it can reduce coupling.

While state-of-the-art deep learning models [15,14,3] resolve this problem in the sequence labeling manner, their models are usually trained on a fixed training set where the combination of entity and context is invariant, so the

* Corresponding author

⁶ <https://liftk444.github.io/FLEXNER/>

relationship between entity and context information is not fully exploited. Intuitively, adding diverse training samples [11,28] is helpful to train a better model, but expanding the existing training data is expensive. The Wikipedia entity type mappings [17] or the distant supervision [16] provide a way to augment the data, but these methods rely heavily on outside knowledge resources. The DS-based entity set expansion might introduce noisy instances, which potentially leads to the semantic drift problem [23]. Ideally, we would wish to overcome these two problems, increasing the diversity of training data without external resources, and adapting our approach to any datasets. We solve the first problem by data transformation operations inside the dataset. Our method only uses the ground truth entities in the training set, which naturally reduces the influence of noisy instances. For the second problem, we use a bilateral network to enhance the learning representation, which can achieve better results on different datasets.

The context pattern provides semantics for inferring the entity slot, i.e., from “*Germany imported 47000 sheep from Britain*” we got a context pattern “A imported 47000 sheep from B” which implies that A and B are locations. If this sentence becomes “*America imported 47000 sheep from Britain*”, the appearance of *America* is also reasonable, but a person cannot appear in these placeholders. We refer to these rules as context pattern entailment, which can be emphasized and generalized by increasing entity-context diversity during model training. The data augmentation technique [5] aims to apply a wide array of transformations to synthetically expand a training set. This paper proposes two innovative data augmentation methods on the input stage. Compared with the distant supervision, our approach does not rely on additional knowledge bases since our approach can inherently and proactively enhance low resource datasets.

We conduct experiments on five languages, including the English, German, Spanish, Dutch and Chinese, and biomedical domain. Our bilateral network achieves good performance. The main contributions of this paper can be summarized below.

- (i) We augment the learning representation by increasing entity-context diversity. Our method can be applied to any datasets almost without any domain-specific modification.
- (ii) To improve the versatility of our approach, we present the bilateral network to integrate the baseline and augmented representations of two sub-networks.

2 Related Work

The CNN based [3], LSTM based [14] and hybrid (i.e., LSTM-CNNs [15,1]) models resolve this task in the sequence labeling manner. Yang et. al [26] build a neural sequence labeling framework⁷ to reproduce the state-of-the-art models, while we build a portable framework and also conduct experiments in different languages and domains. Yang et. al [27] use cross-domain data and transfer learning to improve model performance.

⁷ <https://github.com/jiesutd/NCRFpp>

ELMo [18] and BERT [4] enhance the representations by pre-training language models. BERT randomly masks some words to train a masked language model, while our data augmentation is a constrained entity-context expansion. Our approach aims to retain more entity type information in the context representation. CVT [2] proposes a semi-supervised learning algorithm that uses the labeled and unlabeled data to improve the representation of the Bi-LSTM encoder.

The data augmentation method can be carried out on two stages, the raw input stage [21] and the feature space [5]. Data augmentation paradigm has been well addressed in computer vision research, but receives less attention in NLP. Shi et. al [23] propose a probabilistic Co-Bootstrapping method to better define the expansion boundary for the web-based entity set expansion. Our approach is designed to enhance the entity-context diversity of the training data without changing the entity boundary, which naturally reduces the impact of noisy instances.

The proposed framework is flexible and easy to expand. We can further consider the structural information [7] and build the model in a paradigm of continual learning [8].

3 Methods

3.1 Model Overview

An NER pipeline usually contains two stages, predicting label sequence and extracting entities. Firstly, this model converts the textual input into the most likely label sequence $y^* = \arg \max_{y \in Y(z)} p(y|z)$, where z and $Y(z)$ denote the textual sequence and all possible label sequences. Secondly, the post-processing module converts the label sequence into human-readable entities. The sequence labeling neural network usually contains three components for word representations, contextual representations and sequence labeling respectively.

Word representations. This component projects each token to a d -dimensional vector which is composed of the word embedding and character level representation. The word embedding can be a pre-trained [9] or randomly initialized fixed-length vector. The character level representation can be calculated by a CNN or RNN [26], and the character embeddings are randomly initialized and jointly trained.

Contextual representations. This component can generate contextual representations using CNN or RNN. Besides, our model can use different stack components to extract features, as shown in the left part of Figure 2. The major difference between these stack components is the way they extract local features. In the LSTM-CNN stack, the CNN extracts the local contextual features from the hidden state of Bi-LSTM, while in the CNN-LSTM stack the CNN extracts the local features from the word vectors and the Bi-LSTM uses the context of local features.

Sequence labeling. This component outputs the probability of each token and selects the most likely label sequence as the final result. This paper adopts the conditional random field (CRF) [13] to consider the transition probability between labels.

$$p(y|z; W, b) = \frac{\prod_{i=1}^n \exp(W_{y_{i-1}y_i}^T z_i + b_{y_{i-1}y_i})}{\sum_{y' \in Y(z)} \prod_{i=1}^n \exp(W_{y'_{i-1}y'_i}^T z_i + b_{y'_{i-1}y'_i})} \quad (1)$$

where $\{[z_i, y_i]\}, i = 1, 2 \dots n$ represents the i -th word z_i and the i -th label y_i in the input sequence respectively. $Y(z)$ denotes all the possible label sequences for the input sequence z . W and b are weight matrix and bias vector, in which W_{y_{i-1}, y_i} and b_{y_{i-1}, y_i} are the weight vector and bias corresponding to the successive labels (y_{i-1}, y_i) . $p(y|z; W, b)$ is the probability of generating this tag sequence over all possible tag sequences.

During the training process, the model parameters are updated to maximize the log-likelihood $L(W, b)$. For prediction, the decoder will find the optimal label sequence that can maximize the log-likelihood $L(W, b)$ through the Viterbi algorithm.

$$L(W, b) = \sum_j \log p(y^j | z^j; W, b) \quad (2)$$

$$y^* = \arg \max_{y \in Y(z)} p(y|z; W, b) \quad (3)$$

3.2 Data Augmentation

Sentence-centric augmentation (SCA). As shown in Figure 1(a), this method enhances the context representation by increasing entity diversity. This operation augments the entity distribution of any sample. We generate augmented sentences as follows:

1. Extract the categorical entity glossary $E = \{E_1, E_2, \dots, E_c\}$ based on the original corpus S , where c is class number. An entity may be composed of multiple words, so we need to convert label sequence into complete entities.
2. Resample the sentence $s_i \sim \text{Uniform}(S)$ and light up each entity slot $a_{(i,j)} \sim \text{Bernoulli}(p)$. p (0.5 to 0.9) is chosen according to different datasets.
3. Replace the lighted entities $a_{(i,k)} \in E_j$ with $\hat{a}_{(i,k)} \sim \text{Binomial}(E_j \setminus \{a_{(i,k)}\})$. This is a crossover operation.

Entity-centric augmentation (ECA). In SCA, we can control the augmentation for context, but the entity control is not easy. As shown in Figure 1(b), ECA enhances the entity representation by increasing context diversity. This operation augments the sentence distribution of an entity. We can better control the augmentation for entities.

1. Extract the categorical entity glossary $E = \{E_1, E_2, \dots, E_c\}$ from the training data.

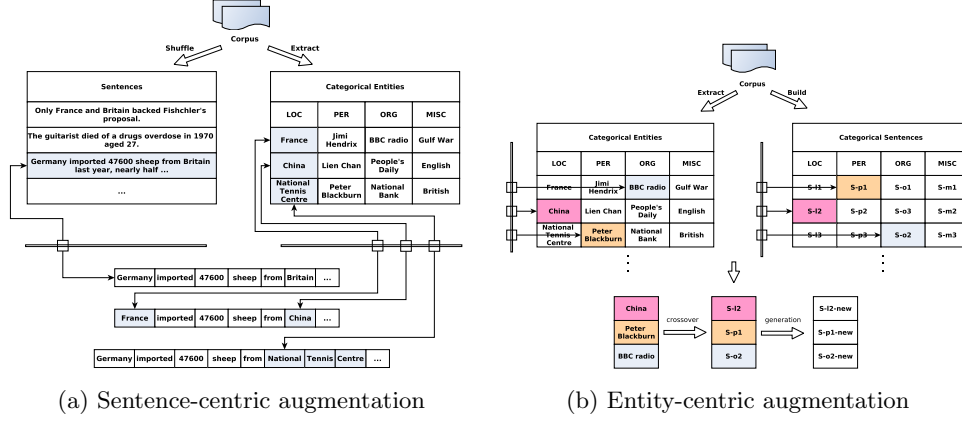


Fig. 1: The schematic diagram of data augmentation operations where the black arrows represent the random selectors

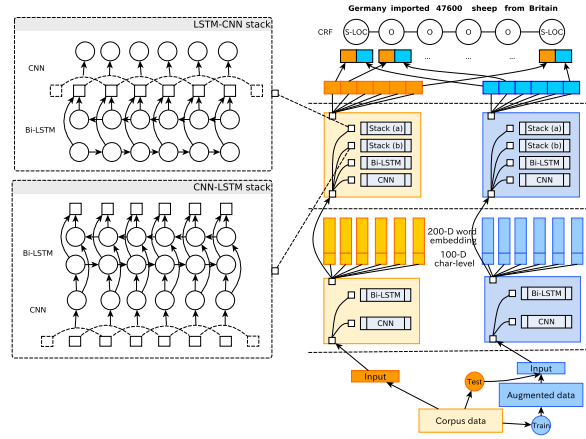


Fig. 2: An overview of bilateral architecture

2. Build the categorical sentence set $S = \{S_1, S_2, \dots, S_c\}$. The main idea is to classify the training samples according to the type mention of entities. Let $e_{(i,1)} \in E_i$ and $e_{i,1} \in s_j$, then $s_j \in S_i$. In Figure 1(b), the S-li denotes the i -th sentence containing at least one l (LOC) entity.

3. Sample an entity $e_{(i,1)}$ with a probability of $p = F(e_{(i,1)})/F(E_i)$ where $F(\cdot)$ is the frequency. Sample a sentence $s_i \sim \text{Uniform}(S_i)$, and then perform the crossover operation.

If we iteratively substitute entities once a time in SCA, it behaves like ECA to some extent, but the distribution of entity and context is different. The basis of the two augmentation methods is different, because ECA can be extended to

augment the vertex representation of a knowledge graph, while SCA focuses on text-level expansion. Here we use a random selector to simplify the augmentation process and make this work widely available. But this may lead to some noisy samples, e.g., “Germany imported 47600 sheep from national tennis centre”. Quality control is critical when faced with data in a specific domain and is reserved for future work. Our approach actively extracts entities from within the dataset, thus not relying on the external resources. Our approach can be generally applied to augment the entity-context diversity for any datasets.

3.3 Bilateral architecture

As shown in Figure 2, the bilateral architecture is composed of a baseline network on the left side and an augmented network on the right side. In the word and contextual representation layers, each sub-network can optionally use the Bi-LSTM, CNN and mixed stack layers to form flexible network combinations. This strategy can generate at least 64 $((2 \times 4)^n, n \geq 2$, where n is the number of sub-networks) types of bilateral networks. We compare different layer stack networks and their combinations in experiments. The data augmentation operations are only activated during the model training.

The inputs to the left and right sub-networks are different, so they form the function that adapts to different patterns. The right sub-network is more generalized, but the weakness is that it may generate noisy samples. This bilateral network also supports the joint training, but the inputs to both sides are the same. The bilateral network is a case of the multi-lateral network which also can be easily extended by adding more sub-networks. This paper adopts the IOBES scheme [19]. The outputs of bilateral sub-networks are concatenated as the input to the final CRF layer.

3.4 Training Procedure

We introduce two training methods, the separate training and the joint training. The separate training contains three steps.

(1) We train the left sub-network of Figure 2 (freezing the right sub-network) using the human-annotated data. The left and the right sides represent the baseline and augmented models respectively. The outputs of the two sub-networks share the same CRF layer. In this step, the CRF layer only accepts the output of the baseline network. The output and the gradient of the augmented network are masked so the parameters of the augmented network are not updated. The baseline network learns the original features of the training data.

(2) Then, we train the right sub-network (freezing the left sub-network) with the human-annotated and the augmented data. The augmented data is generated dynamically based on the algorithm in subsection Data Augmentation. Contrary to step (1), the CRF layer only accepts the output of the augmented network. This step also updates the weights of the full connection layer before the CRF layer. The augmented network enhances representations by increasing entity-context diversity.

(3) We retrain the last CRF layer (freezing all the components before the CRF layer) with the human-annotated data to fuse the representation. In this step, the functions of two sub-networks are kept, so the outputs of them are concatenated to form the rich representation from different perspectives.

We refer to the step (1) and (2) as the pre-training and step (3) as the fine-tuning. The separate training method can form two functional sub-networks, each of which retains its own characteristics.

For the joint training, we input the same sample into two sub-networks. Although the joint training can update the parameters simultaneously, the separate training achieved better results. This is because the separate training accepts different sentences in different sub-networks. The separate training retains the functionality of each sub-network, and features can be extracted independently from different perspectives, while the joint training processes the same task and focuses on extending layer width, so it did not fully extract diverse features.

4 Experiments

4.1 Dataset and Evaluation

Different languages. For different languages, we adopt the CoNLL-2002 [22] and CoNLL-2003 [24] datasets which are annotated with four types of entity, location (LOC), organization (ORG), person (PER), miscellaneous (MISC) in English, German, Dutch, Spanish. The Chinese dataset [25] is a discourse-level dataset from hundreds of Chinese literature articles where seven types of entities (Thing, Person, Location, Time, Metric, Organization, Abstract) are annotated. **Biomedical field.** For the biomedical NER, we use the SCAI corpus which is provided by the Fraunhofer Institute for Algorithms and Scientific Computing. We focused on the International Union of Pure and Applied Chemistry (IUPAC) names (e.g., adenosine 3',5'-(hydrogen phosphate)) like the ChemSpot [20]. The second dataset is the GELLUS corpus [12] which annotates cell line names in 1,212 documents drawn from the biomedical literature in the PubMed and PMC archives.

4.2 Results on NER for different languages

We use (C) and (W) to represent the character level input and word level input respectively, i.e., the (C)CNN and (W)LSTM denote this model use the CNN to accept the input of character embeddings and the Bi-LSTM to accept the input of word embeddings respectively.

English Table 1 shows the results of 24 models on the English NER. This paper reproduces the baseline models and tests the stack models. To eliminate the influence of random factors we ran the experiments three times. ^a and ^b denote the models of (C)LSTM-(W)LSTM-CRF [14] and (C)CNN-(W)LSTM-CRF [15] models respectively. In some cases, our baseline models slightly underperform (0.1~0.2% F1) than the corresponding prototypes, but the bilateral

Table 1: Results of different network combinations on the CoNLL-2003 English dataset

| Layers | Models | | | | | |
|-------------------|-------------------------|------------|-------------------|-------------------------|------------|------------|
| Character input | LSTM | | | CNN | | |
| Word input | LSTM | CNN | Stack (a) | LSTM | CNN | Stack (b) |
| Baseline | 91.00±0.04 ^a | 89.89±0.06 | 90.99±0.06 | 90.95±0.06 ^b | 90.13±0.04 | 90.15±0.04 |
| Augment | 90.87±0.11 | 90.01±0.08 | 91.10±0.05 | 91.06±0.06 | 90.32±0.06 | 89.67±0.07 |
| Baseline+Baseline | 91.02±0.05 | 89.85±0.03 | 90.99±0.05 | 90.98±0.04 | 90.09±0.03 | 90.19±0.04 |
| Baseline+Augment | 91.36±0.08 | 90.24±0.09 | 91.47±0.06 | 91.14±0.07 | 90.52±0.06 | 90.54±0.05 |

models (Baseline+Augment) achieve better performances than the prototypes of [14,1,3]. Concatenating two separate baseline models (Baseline+Baseline) almost did not change results, while the Baseline+Augment model produces better results. This demonstrates data augmentation is helpful to enhance the representations to generate better results. The entity-based data transformation paradigm has great potential to improve the performance of other tasks. More details could be found in the supplementary material⁸.

4.3 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. Our method achieves good performance in the following two corpora. In the GELLUS corpus, the augmented and the bilateral models improve 5.11% and 6.08% F1 score than our baseline model. This means that our approach will be a good choice in the biomedical field.

Table 2: Results of IUPAC Chemical terms and Cell lines on the SCAI chemicals corpus and the GELLUS corpus respectively

| Algorithm | SCAI | GELLUS |
|------------------------|--------------|--------------|
| OSCAR4 [10] | 57.3 | — |
| ChemSpot [20] | 68.1 | — |
| CRF [6] | — | 72.14 |
| LSTM-CRF [6] | — | 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 |

⁸ <https://github.com/liftkkkk/FLEXNER/blob/master/pic/appendix.pdf>

Due to space constraints, extensive discussions and case studies will be introduced in the supplementary material⁸.

5 Conclusion

This paper introduces a portable NER framework FlexNER which can recognize entities from textual input. We propose a data augmentation paradigm which does not need external data and is straightforward. We augment the learning representation by enhancing entity-context diversity. The layer stacks and sub-network combinations can be commonly used in different datasets to provide better representations from different perspectives.

It seems effortless to extend this framework to the multilingual NER research since we can use different sub-networks to learn different languages and then explore the interaction among them. Data quality control is an important task that seems to improve the learning process. Besides, this method is potential to be used in low-resource languages and may benefit in other entity related tasks. In the future, we also plan to apply this system to biomedical research, i.e., extracting the functional brain connectome [29] or exploring the relations between drugs and diseases.

6 Acknowledgement

This study is supported by the Strategic Priority Research Program of Chinese Academy of Sciences (Grant No.XDB32070100).

References

1. Chiu, J.P., Nichols, E.: Named entity recognition with bidirectional lstm-cnns. arXiv preprint arXiv:1511.08308 (2015)
2. 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)
3. 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)
4. 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)
5. DeVries, T., Taylor, G.W.: Dataset augmentation in feature space. arXiv preprint arXiv:1702.05538 (2017)
6. 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)
7. Hu, W., Chan, Z., Liu, B., Zhao, D., Ma, J., Yan, R.: Gsn: A graph-structured network for multi-party dialogues. In: *Proceedings of IJCAI 2019* (2019)

8. Hu, W., Lin, Z., Liu, B., Tao, C., Tao, Z., Ma, J., Zhao, D., Yan, R.: Overcoming catastrophic forgetting for continual learning via model adaptation. In: Proceedings of ICLR 2019 (2019)
9. Hu, W., Zhang, J., Zheng, N.: Different contexts lead to different word embeddings. In: Proceedings of COLING 2016 (2016)
10. 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)
11. Jiang, L., Meng, D., Yu, S.I., Lan, Z., Shan, S., Hauptmann, A.: Self-paced learning with diversity. In: Proceedings of NeuIPS 2014 (2014)
12. Kaewphan, S., Van Landeghem, S., Ohta, T., Van de Peer, Y., Ginter, F., Pyysalo, S.: Cell line name recognition in support of the identification of synthetic lethality in cancer from text. *Bioinformatics* **32**(2), 276–282 (2015)
13. Lafferty, J., McCallum, A., Pereira, F.C.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data (2001)
14. Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., Dyer, C.: Neural architectures for named entity recognition. arXiv preprint arXiv:1603.01360 (2016)
15. Ma, X., Hovy, E.: End-to-end sequence labeling via bi-directional lstm-cnns-crf. arXiv preprint arXiv:1603.01354 (2016)
16. Mintz, M., Bills, S., Snow, R., Jurafsky, D.: Distant supervision for relation extraction without labeled data. In: Proceedings of ACL 2009 (2009)
17. Ni, J., Florian, R.: Improving multilingual named entity recognition with wikipedia entity type mapping. arXiv preprint arXiv:1707.02459 (2017)
18. 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)
19. Ratnikov, L., Roth, D.: Design challenges and misconceptions in named entity recognition. In: Proceedings of CoNLL 2009. pp. 147–155 (2009)
20. Rocktäschel, T., Weidlich, M., Leser, U.: Chemsport: a hybrid system for chemical named entity recognition. *Bioinformatics* **28**(12), 1633–1640 (2012)
21. Saito, I., Suzuki, J., Nishida, K., Sadamitsu, K., Kobashikawa, S., Masumura, R., Matsumoto, Y., Tomita, J.: Improving neural text normalization with data augmentation at character-and morphological levels. In: Proceedings of IJCNLP 2017 (2017)
22. Sang, E.F.T.K.: Introduction to the conll-2002 shared task: Language-independent named entity recognition. In: Proceedings of CoNLL 2002 (2002)
23. Shi, B., Zhang, Z., Sun, L., Han, X.: A probabilistic co-bootstrapping method for entity set expansion. In: Proceedings of COLING 2014. pp. 2280–2290 (2014)
24. Tjong Kim Sang, E.F., De Meulder, F.: Introduction to the conll-2003 shared task: Language-independent named entity recognition. In: Proceedings of HLT-NAACL 2003. pp. 142–147 (2003)
25. 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)
26. Yang, J., Liang, S., Zhang, Y.: Design challenges and misconceptions in neural sequence labeling. In: Proceedings COLING 2018 (2018)
27. Yang, Z., Salakhutdinov, R., Cohen, W.W.: Transfer learning for sequence tagging with hierarchical recurrent networks. arXiv preprint arXiv:1703.06345 (2017)
28. Zhou, L., Hu, W., Zhang, J., Zong, C.: Neural system combination for machine translation. arXiv preprint arXiv:1704.06393 (2017)

29. Zhu, H., Zeng, Y., Wang, D., Xu, B.: Brain knowledge graph analysis based on complex network theory. In: Proceedings of BI 2016. Springer (2016)