NCUEE-NLP at SemEval-2022 Task 11: Chinese Named Entity Recognition Using the BERT-BiLSTM-CRF Model

Lung-Hao Lee, Chien-Huan Lu, and Tzu-Mi Lin
Department of Electrical Engineering
National Central University
No. 300, Zongda Rd., Zhongli Dist., Taoyan City 32001, Taiwan
lhlee@ee.ncu.edu.tw, 10952107@ncu.edu.tw,110521087@ncu.edu.tw

Abstract

This study describes the model design of the NCUEE-NLP system for the Chinese track of the SemEval-2022 MultiCoNER task. We use the BERT embedding for character representation and train the BiLSTM-CRF model to recognize complex named entities. A total of 21 teams participated in this track, with each team allowed a maximum of six submissions. Our best submission, with a macro-averaging F1-score of 0.7418, ranked the seventh position out of 21 teams.

1 Introduction

Named Entity Recognition (NER) is a fundamental task in information extraction that locates the mentions of named entities and classifies them (e.g., person, organization and location) in unstructured texts. The NER is a traditional NLP task that has been solved as a sequence labeling problem, where entity boundaries and category labels are jointly predicted. It is difficult to recognize complex named entities like the titles of creative works (e.g., books, songs, movies) that can take the form of any linguistic constituent (Ashwini and Choi, 2014). Syntactic and semantic ambiguity makes it challenging to recognize such complex named entities based on their context.

The SemEval-2022 Task 11 (MultiCoNER) organized a challenge to develop multilingual complex NER system for 11 human languages (Malmasi et al., 2022b). This task focuses on detecting semantically ambiguous and complex entities in short and low-context settings. The languages include: English, Spanish, Dutch, Russian, Turkish, Korean, Farsi, German, Chinese, Hindi, and Bangla. The named entity categories are Person (labeled as PER), Location (LOC), Group (GRP), Corporation (CORP), Product (PROD), and Creative Work (CW). The task evaluation framework is divided in three broad tracks. 1) Multi-lingual (Track 1): participants train a single multi-lingual NER model for all the languages; 2) Mono-lingual (Track 2-12): participants train a model that works for only one language; 3) Code-mixed (Track 13): testing samples include tokens from any of the 11 mentioned languages in the shared task. We only participated the Track 9 for Chinese language.

Chinese NER is more difficult to process than English NER. Chinese language is logographic and provides no conventional features like capitalization. In addition, due to a lack of delimiters between characters, Chinese NER is correlated with word segmentation tasks, and named entity boundaries are also word boundaries. However, incorrectly segmented entity boundaries will cause error propagation in NER. For example, in a short context "這首歌出現在華特迪士尼動畫動物方城市中" (This song appeared in the Walt Disney animation Zootopia), a creative work "動物方城市" (Zootopia) may be incorrectly segmented into three words: "動物" (animal), "方"(square), and "城市" (city). Hence, it has been shown that character-based approaches outperform word-based methods for Chinese NER (He and Wang., 2008; Li et al., 2014; Zhang and Yang., 2018).

Recently, deep learning techniques have been widely used for Chinese NER, mostly with promising results. A character-based LSTM (Long Short-Term Memory)- CRF (Conditional Random Field) model with radical-level features was proposed for Chinese NER (Dong et al., 2016). The ME-CNER model exploited multiple embeddings-based character representation to improve Chinese NER performance (Xu et al., 2019). A joint training objective technique for different types of neural embeddings was adopted for Chinese NER in social media, based on Weibo messages (Peng and Dredze., 2015). The BiLSTM (Bidirectional LSTM)-CRF model was trained based on character-word mixed embeddings to improve the recognition effectiveness of Chinese NER (E and Xiang., 2017). A BiLSTM-CRF model with a self-attention mecha-

Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 1597 - 1602
July 14-15, 2022 ©2022 Association for Computational Linguistics

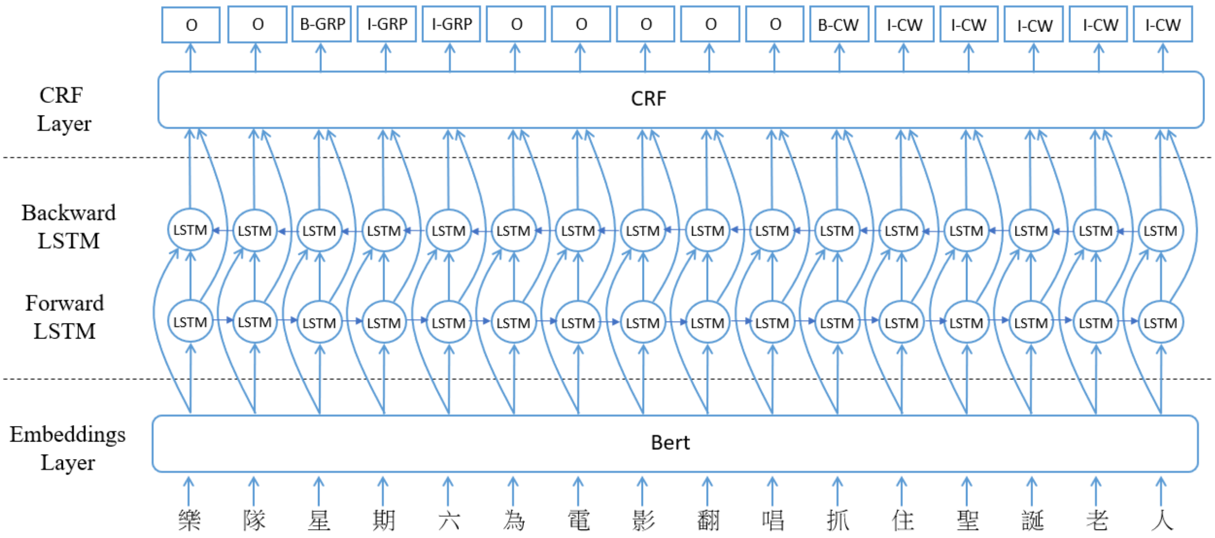

Figure 1: Our NCUEE-NLP system architecture for the Chinese track of SemEval-2022 Task 11.

nism was proposed to integrate part-of-speech labeling information to capture the semantic features of input sequences for Chinese clinical NER (Wu et al., 2019). A residual dilated CNN (Convolution Neural Network) with CRF was also presented to enhance Chinese clinical NER in terms of computational performance and training time (Qiu et al., 2019). An ME-MGNN (Multiple Embeddings enhanced Multi-Graph Neural Network) model was proposed to derive a character representation based on multiple embeddings at different granularities from the radical, character to word levels. Multiple gated graph sequence neural networks, along with standard BiLSTM-CRF, were then used to recognize Chinese named entities in the healthcare domain (Lee and Lu., 2021).

This paper describes the **NCUEE-NLP** (**N**ational **C**entral **U**niversity, Dept. of **E**lectrical **E**ngineering, **N**atural **L**anguage **P**rocessing Lab) system for the Chinese track of SemEval-2022 Task 11. We find that the neural computing approaches based on the BiLSTM-CRF achieved impressive results for Chinese NER. Hence, we follow the investigated results to develop character-based BiLSTM-CRF models.

2 The NCUEE-NLP System

Figure 1 shows our NCUEE-NLP system architecture for the Chinese NER. Our BERT-BiLSTM-CRF model is composed of three main parts: 1) BERT embeddings, 2) Bidirectional LSTM networks, and 3) CRF sequence labeling.

2.1 BERT Embeddings

Word embedding is a type of representation for text analysis that allows words with similar meanings to have similar representations in the form of a real-valued vector (Mikolov et al., 2013). Word embeddings can be obtained using a set of language modeling techniques where words are mapped to a low dimensional vector space of real numbers. Replacing static vectors, such as word2Vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014)), and fastText (Bojanowski et al., 2017), with contextual word representations has led to significant improvements to virtually every NLP task (Ethayarajh., 2019). BERT (Bidirectional Encoder Representations from Transformer) (Devlin et al., 2019), is an encoder-decoder architecture that uses attention mechanisms to incorporate context into word embeddings. Its technical innovation lies in applying the bidirectional training of the transformer with masked language modeling to hide the partial words and infer them using their position information.

Since incorrect Chinese word segmentation may cause error propagation to affect the boundaries of named entities, we only use the last layer of BERT to obtain contextual embedding for each character.

2.2 Bidirectional LSTM Networks

In traditional neural network architectures such as multilayer perceptron, all the inputs and outputs are mutually independent. To address this issue, Recurrent Neural Networks (RNN) create networks with

Data Source		Sent.	All NE	#PER	#LOC	#GRP	#CORP	#PROD	#CW
Official	Training	15,300	23,717	2,221	6,984	710	3,756	4,818	5,228
	Validation	800	1,273	129	378	26	189	271	280
	Test	151,661	N/A	N/A	N/A	N/A	N/A	N/A	N/A
External	MSRA	16,302	36,088	10,378	25,710	-	-	-	-
	Weibo	490	891	648	243	-	-	-	-
	PD	9,281	21,238	6,256	14,982	-	-	-	-
	Boson	566	4,280	2,479	1,801	-	-	-	-
	CLUENER	2,305	3,136	-	-	-	-	-	3,136
	LG	6,135	16,266	-	-	-	-	-	16,266

Table 1: Data statistics of Chinese NER.

loops called Long Short-Term Memory (LSTM) to remember all information over time. Bidirectional LSTM (BiLSTM) (Graves et al., 2013) combines two independent RNNs that allows the networks to obtain both forward (from left to right) and backward (from right to left) information about the character sequence at every time step.

2.3 CRF Sequence Labeling

The learned feature representations of characters in the BiLSTM layer are then fed to a standard Conditional Random Field (CRF) (Lafferty et al., 2001), following the character order in the original sentence to predict the sequence of labels.

During the model training phase, a sentence represented in terms of a character sequence, along with the corresponding named entity labels, are used to train the BERT-BiLSTM-CRF model. We adopt the commonly used BIO (Beginning, Inside, and Outside) format. The B-prefix before a tag indicates that the character is the beginning of a named entity and an I-prefix before a tag indicates that the character is inside a named entity. An O tag indicates a character belongs to no named entity. For example, a sample sentence “樂隊星期六為電影翻唱抓住聖誕老人” (The band The Saturdays covered Christmas Wrapping for the movie.) in Figure 1, “星期六” (The Saturdays) are a British-Irish girl band that belongs to the Group (labeled as GRP) category. The corresponding named entity labels are “B-GRP,” “I-GRP,” and “I-GRP” for individual character “星”, “期”, and “六”. Similarly, “抓住聖誕老人” (Christmas Wrapping) is a song belonging to the Creative Work (CW) category, so we have the named entity labels “B-CW,” “I-CW”, “I-CW”, “I-CW”, “I-CW”, and “I-CW”.

During the testing phase, our trained BERT-BiLSTM-CRF model is used to predict the named entity label of each character for performance evaluation.

3 Experiments and Results

3.1 Data

Table 1 shows detailed statistics for mutually exclusive datasets. The experimental datasets were mainly provided by the task organizers (Malmasi et al., 2022a). The evaluation test set is about 10 times larger than original training dataset. In addition, according to an FAQ on the task website, the training and test data have dissimilar label distributions, though we have yet to obtain the real distribution in the test set. Hence, we also collected external data, including MSRA (Levow., 2006), Weibo (Peng and Dredze., 2015), People Daily (PD)[1] , Boson[2] , CLUENER (Xu et al., 2020), and LG [3] to train our model. The former four datasets consist of sentences annotated with the named entity categories Person (PER) and Location (LOC), while the latter two datasets were converted to contribute instances for the Creative Work (CW) category. We did not find sentences annotated with appropriate labels for the named entity categories Group (GRP), Corporation (CORP), and Product (PROD) as defined in this task.

[1] https://github.com/OYE93/Chinese-NLP-Corpus/tree/master/NER/People’s%20Daily

[2] https://static.bosonnlp.com/dev/resource

[3] https://github.com/LG-1/video_music_book_datasets

Models		Precision	Recall	F1
Embedding	Data Usage			
BERT	Official (training)	0.8856	**0.8759**	**0.8807**
	Official + External	0.8778	0.8579	0.8677
RoBERTa	Official (training)	**0.8867**	0.8618	0.8741
	Official + External	0.8647	0.8532	0.8589
MacBERT	Official (training)	0.8817	0.8610	0.8712
	Official + External	0.8759	0.8594	0.8676

Table 2: Results of our NER models on the validation set.

Models		Precision	Recall	F1
Embedding	Data Usage			
BERT	Official (training + validation)	0.7701	**0.7299**	**0.7418**
	Official+ External	0.7506	0.6991	0.7055
RoBERTa	Official (training + validation)	0.7629	0.7015	0.7207
	Official+ External	0.7477	0.6883	0.7008
MacBERT	Official (training + validation)	**0.7727**	0.7186	0.7351
	Official+ External	0.7553	0.7037	0.7151

Table 3: Results of our NER models on the test set.

3.2 Settings

For character representations, in addition to BERT[4] (Devlin et al., 2019), we also adopted RoBERTa[5] (Liu et al., 2019) and MacBERT [6] (Cui et al., 2020) to compare the performance of different embeddings. We downloaded these pre-trained models from HuggingFace and continuously trained their language models using official data including training, validation and test datasets. The hyper-parameter values for our embedding training were embedding size 768; batch size 64; epoch 20; and learning rate 4e-5.

We trained the BiLSTM-CRF model based on official data provided by task organizers and their variants with our collected external data to confirm performance differences. The hyper-parameter values for our model implementation were optimized as follows: batch size 256; epoch 40; learning rate 0.004; LSTM hidden size 1024; and LSTM dropout rate 0.1.

The evaluation metrics of this shared task are standard precision, recall, and F1-score, which are the most typically used metrics for NER systems at a character level. For each track, the task participants are allowed to have maximum 6 submissions. The final ranking is determined from the best submission based on macro-averaging F1-score.

3.3 Results

Tables 2 and 3 respectively show the results of our submissions on the validation and test sets. We obtained closely consistent results on both datasets. Comparing the embedding effects with RoBERTa and MacBERT, although these two models are modified to improve the BERT model, we did not obtain NER performance improvements when using them as the embedding representation usage. Surprisingly, including external data to train BiLSTM-CRF does not improve the overall F1 performance. The architecture of the BERT-BiLSTM-CRF model using official data training only obtained the best F1-score of 0.7418 on the test set. Table 4 further shows the detailed results per named entity category. The class LOC obtained the best F1-score, followed by PER. In our observations, these two classes are most commonly categories with relatively clear definitions that may not cause recognition confusion. Both combinations of class GRP with CORP and class PROD with CW are usually difficult to distinguish even with manual annotation if insufficient annotation training is provided.

A total of 21 teams participated in the Chinese track of SemEval-2022 MultiCoNER Task, each submitting at least one entry. Our best submission

[4] https://huggingface.co/hfl/chinese-bert-wwm-ext

[5] https://huggingface.co/hfl/chinese-roberta-wwm-ext-large

[6] https://huggingface.co/hfl/chinese-macbert-large

BERT-BiLSTM-CRF	Precision	Recall	F1
Class			
PER	0.8072	0.7356	0.7698
LOC	0.7704	**0.853**	**0.8096**
GRP	**0.8468**	0.5045	0.6323
CORP	0.7641	0.7712	0.7677
PROD	0.755	0.7594	0.7572
CW	0.6768	0.7558	0.7141

Table 4: Detailed results of our BERT-BiLSTM-CRF model on the test set.

achieved an F1 score of 0.7418, ranking in the seventh position out of 21 teams.

4 Conclusion

This study describes the NCUEE-NLP system in the Chinese track of SemEval-2022 MultiCoNER task, including system design, implementation and evaluation. We used the BERT embedding to represent each character in the original sentences and trained BiLSTM-CRF using datasets provided by the organizers to predict the named entity categories. Our best submission had a marco-averaging F1-score of 0.7418, ranking in the 7th position among a total of 21 participating teams.

Acknowledgements

This study is partially supported by the Ministry of Science and Technology, Taiwan, under the grant MOST 108-2218-E-008-017-MY3.

References

Sandeep Ashwini and Jinho D. Choi. 2014. Targetable named entity recognition in social media. *CoRR, arXiv:1408.0782.*

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for chinese natural language processing. *In Findings of the Association for Computational Linguistics: EMNLP 2020. Association for Computational Linguistics*, pages 657–668.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: pre-training of deep bidirectional transformers for language understanding. *In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1:4171–4186.

Chuanhai Dong, Jiajun Zhan, Chengqing Zong, Masanori Hattri, and Hui Di. 2016. Character-based lstm-crf with radical-level features for chinese named entity recognition. *In Proceedings of International Conference on Computer Processing of Oriental Languages. Springer Link*, pages 239–250.

Shijia E and Yang Xiang. 2017. Chinese named entity recognition with character-word mixed embedding. *In Proceedings of the 2017 ACM Conference on Information and Knowledge Management. Association for Computing Machinery*, pages 2055–2058.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. *In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Association for Computational Linguistics*, pages 55–65.

Alex Graves, Abdel rahman Mohamed, and Geoffrey Hinton. 2013. Speech recognition with deep recurrent neural networks. *In Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. Institute of Electrical and Electronics Engineers*, pages 6645–6649.

Jingzhou He and Houfeng Wang. 2008. Chinese named entity recognition and word segmentation based on character. *In Proceedings of the 6th SIGHAN Workshop on Chinese Language Processing. Association for Computational Linguistics*, pages 128–132.

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 18th International Conference on Machine Learning. Association for Computing Machinery*, pages 282–289.

Lung-Hao Lee and Yi Lu. 2021. Multiple embeddings enhanced multi-graph neural networks for chinese healthcare named entity recognition. *IEEE Journal of Biomedical and Health Informatics*, 25(7):2801–2810.

Gina-Anne Levow. 2006. The third international chinese language processing bakeoff: word segmentation and named entity recognition. *In Proceedings* *of the 5th SIGHAN Workshop on Chinese Language Processing.Association for Computational Linguistics*, pages 108–117.

Haibo Li, Masato Hagiwara, Qi Li, and Heng Ji. 2014. Comparison of the impact of word segmentation on name tagging for chinese and japanese. *In Proceedings of the 9th International Conference on Language Resources and Evaluation. European Language Resources Association*, pages 2532–2536.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: a robustly optimized bert pretraining approach. *arXiv:1907.11692.*

Shervin Malmasi, Anjie Fang, Besnik Fetahu, Sudipta Kar, and Oleg Rokhlenko. 2022a. Multiconer: a large-scale multilingual dataset for complex named entity recognition. *In Proceedings of the 16th International Workshop on Semantic Evaluation. Association for Computational Linguistics.*

Shervin Malmasi, Anjie Fang, Besnik Fetahu, Sudipta Kar, and Oleg Rokhlenko. 2022b. Semeval-2022 task 11: Multilingual complex named entity recognition (multiconer). *In Proceedings of the 16th International Workshop on Semantic Evaluation. Association for Computational Linguistics.*

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Gorrado, and Jeffery Dean. 2013. Distributed representation of words and phrases and their compositionality. *In Proceedings of the 27th Conference on Neural Information Processing Systems*, pages 3111–3119.

Nanyun Peng and Mark Dredze. 2015. Named entity recognition for chinese social media with jointly trained embedding. *In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics*, pages 548–554.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: global vectors for word representation. *In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics*, pages 1532–1543.

Jiahui Qiu, Yangming Zhou, Qi Wang, Tong Ruan, and Ju Gao. 2019. Chinese clinical named entity recognition using residual dilated convolutional neural network with conditional random field. *IEEE Transactions on NanoBioscience*, 18(3):306–315.

Guohua Wu, Guangen Tang, Zhongru Wang, Zhen Zhang, and Zhen Wang. 2019. An attention-based bilstm-crf model for chinese clinic named entity recognition. *IEEE Access*, 7:113942–113949.

Canwen Xu, Feiyang Wang, Jialong Han, and Chenliang Li. 2019. Exploiting multiple embeddings for chinese named entity recognition. *In Proceedings of the 28th ACM International Conference on Information and Knowledge Management. Association for Computing Machinery*, pages 2269–2272.

Liang Xu, Yu Tong, Qianqian Dong, Yixuan Liao, Cong Yu, Yin Tian, Weitang Liu, Lu Li, Caiquan Liu, and Xuanwei Zhang. 2020. Cluener2020: Fine-grained named entity recognition dataset and benchmark for chinese. *arXiv:2001.04351.*

Yue Zhang and Jie Yang. 2018. Chinese ner using lattice lstm. *In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 1:1554–1564.