YNUNLP at SemEval-2023 Task 2:The Pseudo Twin Tower Pre-training Model for Chinese Named Entity Recognition

Jing Li
Yunnan University
18313855625@163.com

Xiaobing Zhou*
Yunnan University
zhouxb@ynu.edu.cn

Abstract

This paper introduces our method of developing a system for SemEval 2023 Task 2: MultiCoNER II Multilingual Complex Named Entity Recognition, Track 9-Chinese. In this task, we need to identify entity boundaries and category labels for the six identified categories. The focus of this task is to detect fine-grained named entities whose data set has a fine-grained taxonomy of 36 NE classes, representing a realistic challenge for NER. We use BERT embedding to represent each character in the original sentence and train CRF-Rdrop to predict named entity categories using the data set provided by the organizer. Our best submission, with a macro average f1 score of 0.5657, ranked 15th out of 22 teams.

1 Introduction

NER (Named Entity Recognition) is a traditional NLP task designed to identify text fragments belonging to predefined categories in free text(Curran and Clark, 2003). NER is an important basic tool for many NLP tasks such as information extraction, question answering system, syntactic analysis and machine translation. Named entities generally refer to entities with specific meaning or strong reference in the text. Academically, it usually includes three categories: entity, time and number, and seven categories: personal name, place name, organization name, time, date, currency and percentage(Chen et al., 2022). NER extracts the above entities from the unstructured input text and can identify more categories of entities based on business requirements.

Chinese named entity recognition(Gui et al., 2019) plays an important role in the field of natural language processing. Compared with English, Chinese NER is more challenging. First of all, Chinese word boundaries are vague, and there are no delimiters, such as Spaces, to clarify word boundaries. If Chinese NER adopts character-level model, there will be semantic loss and boundary information loss. On the other hand, if we use the word-level model, the wrong word segmentation can also degrade performance. There are also more complex properties in Chinese, such as complex combinations, entity nesting, indefinite length, and neologisms on the Internet. In addition, Chinese is not case-sensitive and root-affix, and lacks the expression of a lot of semantic information. Named entity recognition is very important because in many applications we must extract the entities in our vocabulary(Wang et al., 2022).

SemEval 2023 Task 2: MultiCoNER II Multilingual Complex Named Entity Recognition(Fetahu et al., 2023b). This task aims to solve the problem of fine-grained named entity recognition. The mission has 13 tracks, of which tracks 1-12 are monolingual, including English, Spanish, Hindi, Bangla, Chinese, Swedish, Farsi, French, Italian, Portuguese, Ukrainian and German(Malmasi et al., 2022b). Participants trained a model that only worked in one language. Track 13 is a multilingual track, where participants need to use data from 12 languages to train a single multilingual NER model for all languages, which should be able to process single-language data from any language. We took part in the Chinese monolingual program.

2 Related Work

The term named entity recognition first appeared in Message Understanding Conferences (MUC-6), which mainly focuses on information extraction(Hirschberg and Manning, 2015). In addition to information extraction evaluation task, the MUC-6 also introduced a new evaluation task namely named entity recognition task. When a task is first proposed, it defines only a few generic entity categories, such as places, agencies, people, and so on. At present, named entity recognition task has penetrated into various vertical fields, such as medical treatment, finance and so on. The main methods

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 1619–1624
July 13-14, 2023 ©2023 Association for Computational Linguistics

of named entity recognition include dictionary and rule based method, traditional machine learning method and deep learning method(Li et al., 2020).

2.1 Dictionaries and rules based methods

The rules-based NER system(Humphreys et al., 1998) relies on human-made rules. Rules can be designed based on domain-specific gazetteers and syntactic lexical patterns. Better-known systems include LaSIE-II, NetOwl, Facile, and SAR. A rules-based system is a good choice when the vocabulary is exhaustive. However, high accuracy and low recall rates are often observed from such systems in specific areas due to their specific rules and incomplete vocabulary, and these systems cannot be transferred to other areas.

2.2 Traditional machine learning based methods

The typical method of unsupervised learning is clustering(Gong et al., 2003). Clustering based NER system extracts related entities through context similarity clustering. Collins et al. used only a small amount of seed tagging data and seven characteristics, including spelling, entity context, and entity itself, for entity recognition(Neelakantan and Collins, 2015). Nadeau et al. proposed an unsupervised system for the construction of a local name dictionary and the ambiguity resolution of named entities(Nadeau and Sekine, 2007). The system is based on a simple and efficient heuristic method that combines entity extraction and ambiguity elimination.

With supervised learning, NER can be translated into multiple classification or sequence labeling tasks. Given labeled data samples, carefully designed features can be used to represent each training example. The model is then learned using machine learning algorithms to identify similar patterns from unknown data. Many machine learning algorithms have been applied to supervised NER(Yang et al., 2018), including Hidden Markov Model (HMM), decision tree, maximum entropy model, support vector machine (SVM) and Conditional Random Fields (CRF).

2.3 Deep learning based methods

Word sequence-based model(Sboev et al., 2021): English and most other languages naturally divide words by Spaces. The early Chinese NER model also follows the English NER model of word segmentation before prediction. However, word segmentation errors are unavoidable in the word segmentation stage, which will transfer the errors to the subsequent modules and affect the recognition ability of the model.

Model based on word sequence: In order to avoid word segmentation errors caused by model based on word sequence, Chinese NER model since 2003 has mostly carried out further feature extraction and sequence labeling at the level of word sequence(Mai and Zhou, 2022; Guan and Liu, 2021).

Model integrating external information(Tsai et al., 2022): The simple model based on word sequence only uses the semantic information of characters, which is obviously insufficient in the amount of information. Therefore, the subsequent research considers integrating all kinds of external information into the character sequence, such as peripheral information, pinyin information, dictionary information, etc. Dictionary information is the most widely used external information.

3 Methodology

In this part, the model adopted by us for Chinese program of SemEval-2023 Task 2 is introduced in detail. This model consists of three parts, which we call the BERT-CRF-Rdrop model. Figure 1 shows the architecture of the BERT-CRF-Rdrop model.

3.1 BERT

The full name of BERT's model is: Bidirectional Encoder Representations from Transformer. As can be seen from the name, the goal of BERT(Devlin et al., 2018) model is to use large-scale unannotated corpus to train and obtain representations of text containing rich semantic information, namely, semantic Representation of text, and then fine-adjust the semantic representation of text in a specific NLP task, and finally apply it to the NLP task. In NLP method based on deep neural network, the word/vocabulary in text is usually represented by one-dimensional vector (generally called "word vector"). On this basis, the neural network will take the one-dimensional word vector of each word or vocabulary in the text as input, and after a series of complex transformations, output a one-dimensional word vector as the semantic representation of the text. In particular, we usually hope that semantically similar words/vocabularies are close to each other in the feature vector space, so that the text vector converted from the word/vocabulary vector can contain more accurate semantic information. Therefore, the main input of BERT model is the original word vector of each word/vocabulary in the text, which can be initialized randomly or pre-trained using algorithms such as Word2Vector(Jiang et al., 2018) to take the initial value. The output is the vector representation of each word/vocabulary in the text integrated with the semantic information of the full text.

3.2 CRF

The basic definition of CRF(Huang et al., 2015) is: Let X and Y be random variables and the conditional probability distribution of Y under given X conditions. If the random variable Y forms a Markov random field of an undirected graph, the conditional probability distribution is called CRF. Corresponding to Markov can be understood as, if random variable Y forms an undirected graph, and every variable Y in the graph satisfies Markov (at least satisfies one of global, local and paired Markov), then it is called CRF. Where X is the input variable, namely the observation sequence that needs to be marked, and Y is the output variable, representing the state or marker sequence. In the field of natural language processing, the common input variable X and output variable Y have the same graph structure.

$$p(\mathbf{y} \mid \mathbf{x}) = \frac{1}{Z(\mathbf{x})} \prod_{t=1}^{T} \exp\left\{ \sum_{k=1}^{K} \theta_k f_k\left(y_t, y_{t-1}, \mathbf{x}_t\right) \right\} \tag{1}$$

Generally speaking, CRF formulas have two components: 1. Normalization: The right side of the equation has no probability, but has values and characteristics. However, the expected output is a probability, so normalization is required. The normalization constant Z(X) is the sum of all possible sequences of states, making the total one. 2. Weights and characteristics: This part can be regarded as a Logistic regression formula with weights and corresponding characteristics. Maximum likelihood estimation is used for weight estimation, and features are defined by ourselves.

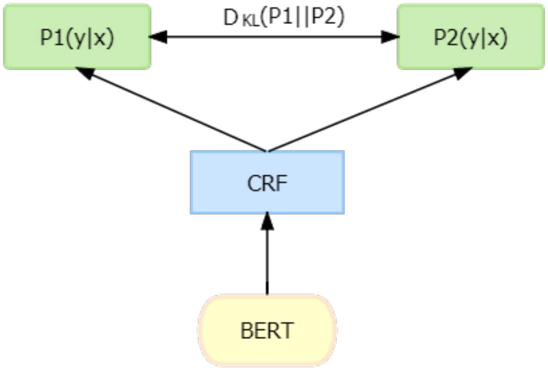

Figure 1: Our basic model structure

3.3 R-drop

R-drop(Zhuang and Zhang, 2022) is a regularization strategy proposed on the basis of Dropout. Its main idea is to make the fractional model generated by dropout have uniformly distributed outputs for the same input data. Specifically, for each training sample, R-Drop minimizes the KL-divergence between the output distributions of the two submodels generated by dropout. Due to the randomness of the Dropout, it is possible to approximate the path network that the input X walks through twice as two slightly different models. Theoretical analysis shows that R-Drop can reduce the degree of freedom of model parameters, supplement the loss, thus reducing the complexity of model space and improving the generalization of the model. One of the loss functions is the regular cross entropy, training data is (x_i, y_i), the model is P_θ(y|x), the cross entropy of each sample is L_i = - log $P_\theta(y_i|x_i)$, and in the case of "Dropout twice," we can think of the sample as having gone through two slightly different models, $P_\theta^{(1)}(y_i|x_i)$, $P_\theta^{(2)}(y_i|x_i)$, so one part of the loss function is:

$$\mathrm{L1} = -\log \mathrm{P}_\theta^{(1)}\left(\mathrm{y_i} \mid \mathrm{x_i}\right) - \log \mathrm{P}_\theta^{(2)}\left(\mathrm{y_i} \mid \mathrm{x_i}\right) \tag{2}$$

The other part of the loss function is the KL divergence, in order to make the two outputs as consistent as possible.

$$\begin{aligned} \mathrm{L2} = \frac{1}{2}[&\mathrm{KL}(\mathrm{P}_\theta^{(1)}\left(\mathrm{y_i} \mid \mathrm{x_i}\right) \mid \mathrm{P}_\theta^{(2)}\left(\mathrm{y_i} \mid \mathrm{x_i}\right)) \\ &+\mathrm{KL}(\mathrm{P}_\theta^{(2)}\left(\mathrm{y_i} \mid \mathrm{x_i}\right) \mid \mathrm{P}_\theta^{(1)}\left(\mathrm{y_i} \mid \mathrm{x_i}\right))] \end{aligned} \tag{3}$$

The final loss of Rdrop is the weighted sum of L1 and L2.

$$\mathrm{L} = \mathrm{L1} + \alpha \mathrm{L2} \tag{4}$$

4 Dataset

The SemEval-2023 Task 2 asks participants to develop complex named entity recognition systems for 12 languages, and we have experimented with Chinese(Fetahu et al., 2023a).

In this task, we used official raw data to train and test our model. The data set for this semeval task consists mainly of three sources: Low-Context Wikipedia(Meng et al., 2021), MS-MARCO Question(Bajaj et al., 2016)and ORCAS Search Query(Craswell et al., 2020).

Each line of text in the data set of SemEval-2023 Task 2 belongs to a sample of languages(Fetahu et al., 2023a): English, Spanish, Hindi, Bangla, Chinese, Swedish, Farsi, French, Italian, Portuguese, Ukrainian and German. It consists of 6 entity types: Location(LOC), Creative Work(CW), Group(GRP), Person (PER), Product (PROD), Medical (MED). Participants must use their systems to accurately detect entities and submit predictions of tasks(Malmasi et al., 2022a). Track 9 - Chinese provides 9,759 training data, 506 validation sets and at least 100,000 final test data.

5 Results

We use accuracy, recall rate and F1 to evaluate the performance of our proposed model. Precision measures the model's ability to present only correct entities, while Recall measures the model's ability to identify all entities in the data set. F1 is the harmonic average of Precision and Recall, which can be calculated as follows:

$$\text{Precision } = \frac{TP}{TP+FP} \tag{5}$$

$$\text{Recall} = \frac{TP}{TP+FN} \tag{6}$$

$$F1 = 2 * \frac{\text{Precision } * \text{ Recall}}{\text{Precision } + \text{ Recall}} \tag{7}$$

Table 1 shows the details of the Bert-Crf-Rdrop model hyper-parameters. The total training cycle of this model was set to 58, and the batch size was set to 4. The initial learning rate is set to 1 x 105. Taking into account the average length of sentences in the data set, the sequence length is 256. We also use Dropout in our approach for better performance in the open domain.

Due to the limitation of time and computing resources, we only give the results of the Chinese model. Table 2 shows the prediction results of the test set based on BERT model, BERT-CRF model and BERT-CRF-Rdrop model.

It can be seen that the BERT-CRF-Rdrop model achieves better performance in the response part of the test set compared with BERT and BERT-CRF, and significantly improves precision, recall rate and F1 score. A total of 22 teams participated in the China leg of the SemEval-2023 MultiCoNER Task, and each team submitted at least one entry. Our best submission achieved an F1 score of 0.5657, placing 15th out of 22 teams.

Parameter	value
sequence length	256
batch size	4
learning rate	1e-5
dropout	0.1
epoch	58

Table 1: Hyper-parameter of the model

Model	precision	recall	F1
BERT	0.5002	0.5813	0.5421
BERT-CRF	0.5106	0.6041	0.5552
BERT-CRF-Rdrop	0.5288	0.6203	0.5657

Table 2: Results of each model on the Chinese test set

6 Conclusion

This study presents the BERT-CRF-Rdrop system submitted for the SemEval2023 Task 2: MultiCoNER Chinese track, including system design, implementation and evaluation. To address the challenges of complex, fuzzy, and emerging entity problems, we used BERT embedding to represent each character in the original sentence and trained CRF-Rdrop to predict named entity categories using the data set provided by the organizer. The experimental results show that our pre-training language model is effective in named entity recognition. In future work, we will further improve performance by developing more complex integration strategies and more diverse models.

References

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, et al. 2016. Ms marco: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268.*

Beiduo Chen, Jun-Yu Ma, Jiajun Qi, Wu Guo, Zhen-Hua Ling, and Quan Liu. 2022. Ustc-nelslip at semeval-2022 task 11: gazetteer-adapted integration network for multilingual complex named entity recognition. *arXiv preprint arXiv:2203.03216.*

Nick Craswell, Daniel Campos, Bhaskar Mitra, Emine Yilmaz, and Bodo Billerbeck. 2020. Orcas: 20

million clicked query-document pairs for analyzing search. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 2983–2989.

James R Curran and Stephen Clark. 2003. Language independent ner using a maximum entropy tagger. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003*, pages 164–167.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Besnik Fetahu, Zhiyu Chen, Sudipta Kar, Oleg Rokhlenko, and Shervin Malmasi. 2023a. MultiCoNER v2: a Large Multilingual dataset for Fine-grained and Noisy Named Entity Recognition.

Besnik Fetahu, Sudipta Kar, Zhiyu Chen, Oleg Rokhlenko, and Shervin Malmasi. 2023b. SemEval-2023 Task 2: Fine-grained Multilingual Named Entity Recognition (MultiCoNER 2). In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*. Association for Computational Linguistics.

Shiaoching Gong, Chen Zheng, Martin L Doughty, Kasia Losos, Nicholas Didkovsky, Uta B Schambra, Norma J Nowak, Alexandra Joyner, Gabrielle Leblanc, Mary E Hatten, et al. 2003. A gene expression atlas of the central nervous system based on bacterial artificial chromosomes. *Nature*, 425(6961):917–925.

Zhengyi Guan and Renyuan Liu. 2021. Yunnan-deep at ehealth-kd challenge 2021: Deep learning model for entity recognition in spanish documents. volume 2943, pages 731 – 736, Virtual, Malaga, Spain. Convolutional neural network;Ehealth;Entity recognition;F1 scores;Learning models;Modeling architecture;Pre-training;Subtask;Test sets.

Tao Gui, Ruotian Ma, Qi Zhang, Lujun Zhao, Yu-Gang Jiang, and Xuanjing Huang. 2019. Cnn-based chinese ner with lexicon rethinking. In *ijcai*, pages 4982–4988.

Julia Hirschberg and Christopher D Manning. 2015. Advances in natural language processing. *Science*, 349(6245):261–266.

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

Kevin Humphreys, Robert Gaizauskas, Saliha Azzam, Christian Huyck, Brian Mitchell, Hamish Cunningham, and Yorick Wilks. 1998. University of sheffield: Description of the lasie-ii system as used for muc-7. In *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29-May 1, 1998.*

Mingyi Jiang, Rui Liu, and Fei Wang. 2018. Word network topic model based on word2vector. In *2018 IEEE Fourth International Conference on Big Data Computing Service and Applications (BigDataService)*, pages 241–247. IEEE.

Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2020. A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, 34(1):50–70.

Hanjie Mai and Xiaobing Zhou. 2022. Clinical text entity recognition based on pretrained model and bigru-crf.

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 29th International Conference on Computational Linguistics*, pages 3798–3809, Gyeongju, Republic of Korea. International Committee on 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 (SemEval-2022)*, pages 1412–1437, Seattle, United States. Association for Computational Linguistics.

Tao Meng, Anjie Fang, Oleg Rokhlenko, and Shervin Malmasi. 2021. GEMNET: Effective gated gazetteer representations for recognizing complex entities in low-context input. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1499–1512.

David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

Arvind Neelakantan and Michael Collins. 2015. Learning dictionaries for named entity recognition using minimal supervision. *arXiv preprint arXiv:1504.06650*.

Alexander Sboev, Sanna Sboeva, Ivan Moloshnikov, Artem Gryaznov, Roman Rybka, Alexander Naumov, Anton Selivanov, Gleb Rylkov, and Viacheslav Ilyin. 2021. An analysis of full-size russian complexly ner labelled corpus of internet user reviews on the drugs based on deep learning and language neural nets. *arXiv preprint arXiv:2105.00059*.

Ming-Chun Tsai, Shu-Ping Lin, and Ching-Chan Cheng. 2022. A comprehensive quality improvement model: integrating internal and external information. *Total Quality Management & Business Excellence*, 33(5-6):548–565.

Xinyu Wang, Yongliang Shen, Jiong Cai, Tao Wang, Xiaobin Wang, Pengjun Xie, Fei Huang, Weiming Lu, Yueting Zhuang, Kewei Tu, et al. 2022. Damo-nlp at semeval-2022 task 11: A knowledge-based system for multilingual named entity recognition. *arXiv preprint arXiv:2203.00545*.

Yaosheng Yang, Wenliang Chen, Zhenghua Li, Zhengqiu He, and Min Zhang. 2018. Distantly supervised ner with partial annotation learning and reinforcement learning. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2159–2169.

Yan Zhuang and Yanru Zhang. 2022. Yet@ smm4h'22: Improved bert-based classification models with rdrop and polyloss. In *Proceedings of The Seventh Workshop on Social Media Mining for Health Applications, Workshop & Shared Task*, pages 98–102.