CASIA at SemEval-2022 Task 11: Chinese Named Entity Recognition for Complex and Ambiguous Entities

Jia Fu[1,2], **Zhen Gan**[1,3], **Zhucong Li**[1,2], **Sirui Li**[4], **Dianbo Sui**[1],**Yubo Chen**[1,2],
Kang Liu[1,2], **Jun Zhao**[1,2]
[1] National Laboratory of Pattern Recognition, Institute of Automation,
Chinese Academy of Sciences, Beijing, China
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences,Beijing, China
[3] Beijing University of Chemical Technology, Beijing, China
[4] Department of Computer Science, Emory University, Altlanta, GA, USA
{zhucong.li, dianbo.sui,yubo.chen, kliu, jzhao}@nlpr.ia.ac.cn,
fujia2021@ia.ac.cn,ganzhen@mail.buct.edu.cn, sirui.li@emory.edu

Abstract

This paper describes our approach to develop a complex named entity recognition system in SemEval 2022 Task 11: MultiCoNER Multilingual Complex Named Entity Recognition,Track 9 - Chinese. In this task, we need to identify the entity boundaries and category labels for the six identified categories of CW, LOC, PER, GRP, CORP, and PORD.The task focuses on detecting semantically ambiguous and complex entities in short and low-context settings. We constructed a hybrid system based on Roberta-large model with three training mechanisms and a series of data augmentation. Three training mechanisms include adversarial training, Child-Tuning training, and continued pre-training. The core idea of the hybrid system is to improve the performance of the model in complex environments by introducing more domain knowledge through data augmentation and continuing pre-training domain adaptation of the model. Our proposed method in this paper achieves a macro-F1 of 0.797 on the final test set, ranking second.

1 Introduction

SemEval-2022 Task 11: MultiCoNER Multilingual Complex Named Entities Recognition(Malmasi et al., 2022b). This task aims to address the problem of complex and ambiguous named entities in practical and open domain environments .

The task has 13 tracks, with track 1 being a multilingual track where participants need to train a multilingual model using data from 11 languages. The model should be able to handle monolingual data from any of the languages and code-mixed cases. Tracks 2-12 are monolingual tracks, including English, Spanish, Dutch, Russian, Turkish, Korean, Farsi, German, Chinese, Hindi and Bengali. Participants are required to train a model for only one language. Track 13 is Code-mixed(Fetahu et al., 2021). This test data contains code-mixed samples. These samples include tokens from any of the 11 mentioned languages in the shared task. We participated in the Chinese monolingual track.

Dealing with complex and ambiguous entities in practical and open domain environments is a challenging NLP task(Meng et al., 2021),however, which has not received sufficient attention from the research community. The challenges of this task are mainly in three facts: 1)The complex entities problem, in which complex noun phrases, verbs, infinitives, or complete sentences, and other entities lack proper nouns, making it difficult to identify them(Ashwini and Choi, 2014).2)The ambiguous entities problem, in which some words are entities in some fields but not in others, especially in search, ASR (Automatic Speech Recognition), and other fields.3)The emerging entities problem, such as books, songs and movies, new works are released every week, and the problem in these fields is that the entities are growing faster, and it is more difficult to identify these newly growing entities.

For the complex entities problem, we use adversarial training and Child-Tuning to increase the representation capability at the parameter level, preventing the model from overfitting due to complex entities, so that the model does not simply remember the complex entities. For the ambiguous entities problem, we use a context-independent data augmentation strategy to replace entities in the data for semantic augmentation, so that the model can reduce the context dependency.For the emerging entities problem, we propose a progressive domain-adaptive pre-training mechanism to improve the performance of the model, so that the preceding methods yielded considerable results, with the F1 value in the Chinese monolingual track reaching 0.797.

Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 1518 - 1523
July 14-15, 2022 ©2022 Association for Computational Linguistics

2 Related Work

2.1 Named Entity Recognition

Named Entity Recognition is a fundamental problem in natural language processing, and it's a key component of many Natural Language Processing(NLP) activities like information extraction, question-answer systems, syntactic analysis, and machine translation. In general, the aim of named entity identification is to detect three major categories (entity, time, and number) of named entities in the text, as well as seven minor categories (person, institution, location, time, date, currency, and percentage).

NER usually consists of two parts: (1) entity boundary identification and (2) entity category determination. Entity boundaries are easier to identify in English because named entities have more visible indicators (the first letter of each word in the entity should be capitalized), and the work concentrates on establishing entity categories. In contrast to the entity category labeling subtask, the Chinese named entity identification task is more complex, and identifying entity boundaries is more challenging.

2.2 Complex and Ambiguous NER

For complex entities, such as titles of creative works (movie/book/song/software titles), not simple nouns and more difficult to identify. They can be any language component, such as a gerund ("Dial M for Murder"), and they don't appear to be traditional entities (names of people, places, organizations). Because of the syntactic ambiguity, identifying them based on their context is difficult. Finally, these entities increase at a quicker rate than traditional categories.

In datasets like CoNLL03/OntoNots, neural network models (e.g., Transformers) have gotten high scores(Devlin and Ming-Wei Chang, 2019). However these scores are driven by the usage of well-formed news texts, the existence of "easier" entities (e.g., names), and memorization due to entity overlap between training and test sets(Augenstein et al., 2017). These models perform significantly worse on complex/unseen entities. The failure of the NER system to recognize complicated items is responsible for a huge portion of their errors(Hanselowski et al., 2018).

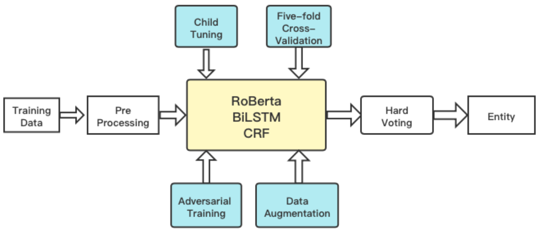

Figure 1: Overall system structure.

3 Our Method

3.1 Overall Approach

The structure of the base model we use is Roberta-wwm-ext-large+BiLSTM+CRF. The sequence samples are pre-trained to obtain their embedding representation, which is then contextually encoded by the BiLSTM(Xu et al., 2017; Ma and Hovy, 2016; Lample et al., 2016) layer, and the contextual encoding is decoded by the CRF(Lafferty et al.; Sutton et al., 2012) to obtain the final annotation result.

In addition, based on the structure of the base model, we first use a back-translation approach to introduce more domain-relevant training data. Related studies have shown that domain-adapted pre-training can improve the performance of the model for the corresponding domain-specific tasks, both with low and high resources. Second, data augmentation is used to extract entities in other languages translated into Chinese and match different contexts to obtain new data, and allow the model to do supervised training using the new data. Finally, we used five-fold cross-training, adversarial training, and child-tuning to improve the performance of the model. Our overall system structure is shown in Figure 1.

3.2 Model Structure

Our basic model structure is shown in Figure 2. The sequence samples get their embedding representation through the pre-training model. Then BiLSTM is connected to the embedding representation for context encoding, and CRF is used to decode the context representation. Finally the annotation result is obtained.

The Bert model, among them, uses Roberta-wwm-ext-large, a joint publication of Harbin Institute of Technology and Pengcheng Lab that uses a full-word Mask scheme in the pre-training phase(Cui et al., 2020). If part of a complete word WordPiece subword is masked, other parts of the

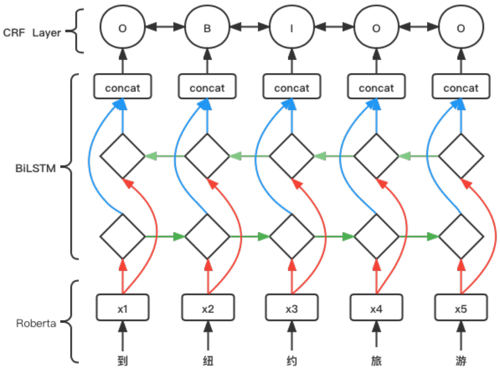

Figure 2: Our basic model structure.

same word will be masked as well, cancels the Next Sentence Prediction, and uses the training model with max len=512.

The underlying model structure is of the form Roberta-wwm-ext-large+BiLSTM+CRF.

3.3 Training Method

Five-fold Cross-voting: We use five-fold cross-validation to divide the training set into five different datasets, and the inconsistencies of entity labeling in each dataset are various. We fix the same model structure, train five models on five training sets, and integrate their prediction results on the same test set by hard voting.

Adversarial Training: To obtain a model with better robustness, we use adversarial training to improve the stability of the model.Referring to the FGM(Miyato et al., 2016) adversarial training mechanism, we directly impose a small disturbance on the embedding representation of the model and assume the embedding representation of the input text sequence $[v_1, v_2, \ldots, v_T]$ as x. Then the small disturbance r_{adv} applied each time is:

$$r_{adv} = \epsilon \cdot g/\|g\|_2 \tag{1}$$

$$g = \nabla_x L(\theta, x, y) \tag{2}$$

The meaning of the formulas is to move the input one step further in the direction of rising loss, which will make the model loss rise in the fastest direction, thus forming an attack. In contrast, the model needs to find more robust parameters in the optimization process to deal with attacks against samples.

Among them, applying a small disturbance to the embedding characterization simulates the natural error of the dataset in the labeling to a certain extent. It encourages the model to find more robust parameters during the training process to weaken the impact of aleatoric uncertainty. Then the model's embedding representation will be optimized together with the model. Adversarial training will make the model more tolerant of changes brought about by model parameter fluctuations, thereby decreasing the impact of epistemic uncertainty.

Child-Tuning: In the Fine-tuning process， there is a mismatch between the "high number of parameters" of the large-scale pre-trained model and the "limited number of labeled samples" in the Fine-tuning phase. Child-Tuning proposes, like regular Fine-tuning, using the entire model's parameters to encode the input samples in the forward direction, but without adjusting the huge number of parameters when updating the parameters in the backward direction, i.e. using only a portion of the Child Network. Child-Tuning can be divided into two stages:

- Confirmation of Child Network is found in the pre-trained model, and 0-1Mask of Gradients corresponding to Weights is generated;
- After the gradient is calculated by backward propagation, only the parameters in Child Network are updated, while the other parameters remain unchanged.

Among the above steps,Step 2 is the simplest of the above steps to change the parameters. It's done with a gradient mask, which means that after computing the gradient of each parameter position, it's multiplied by a 0-1 matrix gradient mask, with the positions belonging to the Child Network parameters corresponding to 1 and those not belonging to 0, and the parameters are updated.

The key to this method is to identify the Child Network mentioned in the preceding steps, one of which is the task-independent algorithm Child-Tuning F, whose main advantage is that it is simple and effective; in the Fine-tune process, it only needs to get a Gradients Mask by sampling from the Bernoulli distribution in each update iteration, which is equivalent to randomly discarding part of the gradients when the network parameters are updated.

$$w_{t+1} = w_t - \eta \frac{\partial \zeta(w_t)}{\partial w_t} \odot M_t \tag{3}$$

$$M_t \sim Bernoulli(P_F) \tag{4}$$

Another is the task-related algorithm Child-Tuning-D, which overcomes the disadvantage that Child-Tuning-F treats different downstream tasks with the same policy and treats different model parameters equally.Child-Tuning uses the Fisher Information Matrix (FIM)(Tu et al., 2016) to estimate the importance of each parameter for the downstream task, and, in line with previous work, approximates the diagonal matrix of FIM to calculate the importance score of each parameter relative to the downstream task (i.e., assuming that the parameters are independent of each other), and then selects the parameter with the highest score as the Child-Network.

$$F^(i)(w) = \frac{1}{|D|}\sum_{i=1}^{|D|}(\frac{\partial logp(y_i|x_j;w)}{\partial w^(i))^2} \qquad (5)$$

3.4 Data Augmentation

Since the amount of data in the test set is 10 times the amount of data in the training set, the amount of data in the training set is obviously insufficient, so we use data augmentation to get more data. Data augmentation techniques are already standard in the image field, and data augmentation is achieved by techniques such as flipping, rotating, mirroring, and Gaussian white noise on images. However, in the field of NLP, there are four ways of data augmentation: synonym substitution, random insertion, random swapping, and random deletion. In this paper, we use random swapping, i.e., random replacement of entities in a sentence with other entities of the same type. We extract entities in English, German, and Dutch translated into Chinese as replacement entities.

4 Experiment

4.1 Dataset Introduction

SemEval 2022 Task 11: MultiCoNER Multilingual Complex Named Entity Recog- nition,Track 9 - Chinese provide 15,300 training data, 800 validation sets, and at least 150,000 final test data(Malmasi et al., 2022a). Six types of entities are included: people, places, organizations, products, companies, and creative works.The sentence and character statistics of the training set, development set, and test set are shown in Table 1.

4.2 Evaluation Metrics

This task takes strict macro F1 as the evaluation metric. The macro F1 evaluation metric looks at each entity category equally compared to micro F1, and for the strict F1 evaluation metric, it is considered correct only when both entity boundaries and entity types agree with the standard answer.

Dataset	Type	Train	Dev	Test
Chinese	Sentence	15.3K	0.8K	153K
	Char	382.1K	20K	1835K

Table 1: Statistics of dataset.

4.3 Pre-Processing

Text Expansion: Inevitably there are many combinations of number strings, English and Chinese in Chinese datasets, which are usually done in Chinese ner based on a single character. To maintain uniformity, we expand the English and number strings into a single English letter and a single number.

Labeling Scheme: The annotation scheme is a method for marking character sequences in sequence annotation tasks, by which the type and location of entities in a sentence can be uniquely determined. Moreover, the annotation scheme affects the named entity recognition performance. We use the BIOES annotation scheme which has better performance compared to the BIO annotation scheme.

4.4 Model Parameters

The basic model structure is RoBERTa-wwm-ext-large+BiLSTM-CRF. The batch size is set to 64, the BERT learning rate is set to 1e-5, the BiLSTM+CRF learning rate is set to 1e-3, the training epoch is set to 50, AdamW is used as the optimizer, and a dropout of 0.3 is used.

For the dev set, we used the basic model of RoBERTa-wwm-ext-large+BiLSTM+CRF with two training methods: Child-Tuning and adversarial training.

For the test set, the training data from 10 additional languages were first translated into Chinese using the back-translation approach to do continue pre-training, and then a further pre-training model with 7 epochs and 15 epochs was obtained by changing the training duration.

For the model after continued pre-training, we used data augmentation to extract entities from other languages and translate them into Chinese, and divided the Chinese training set of 15,300 data into five parts, each containing 3,600 training data,

Type	CW	PER	LOC	GRP	CORP	PROD
Num	8732	14851	7610	7250	5220	4354

Table 2: Number of entities for data augmentation.

Model	Precision	Recal	F1
roberta-large	0.8810	0.8510	0.8648
Roberta+fgm	0.8840	0.8600	0.8710
Roberta+child-tuning-F	0.8775	0.8553	0.8656
Roberta+child-tuning-D	0.8722	0.8629	0.8668
Roberta-large+child-tuning-F+fgm	0.8900	0.8424	0.8629
Roberta-large+child-tuning-D+fgm	0.8929	0.8491	0.8686

Table 3: Results on the Chinese dev set.

	(1)	(2)	(3)	(4)	(5)	(6)
7epoch-micro	+	-	+	+	-	+
15epoch-micro	+	-	+	-	+	+
7epoch-macro	+	+	-	-	-	+
15epoch-macro	+	+	-	-	-	+
Data Augmentation	+	+	+	+	+	-
F1	0.7844	0.7835	0.7823	0.7792	0.7812	0.7970

Table 4: Results on the Chinese test set.

	魔	鬼	军	团	博	物	馆
True Table	B-CW	I-CW	I-CW	I-CW	O	O	O
Baseline	O	O	O	O	O	O	O
Other Method	B-GRP	I-GRP	I-GRP	I-GRP	O	O	O
Our Method	B-CW	I-CW	I-CW	I-CW	O	O	O
F1	0.7844	0.7835	0.7823	0.7792	0.7812	0.7970	

Table 5: Case Study.

and replaced the entities in each part with the translated entities, and added the development set to each data augmentation to get the new five-fold data. Since the total number of entities contained in the other 10 languages is large, we selected three languages with high Baseline scores, English, German and Dutch, to do the data augmentation. Since there are more entities appearing in the three languages, we need to do some filtering by downsampling method.We use the down-sampling method, retaining once for entities that appear once and twice for those that appear twice or more. The final number of entities obtained is shown in Table 2.

4.5 Experimental Results and Analysis

We have conducted a large number of experiments locally and the results are shown in Table 3.

The experimental results of the final test set are shown in Table 4. "+" represents the final results using the fifty-fold cross-validation results of the model to participate in hard voting, and "-" represents not using.

Looking at the results of the local experiments and the test set, we found the following two problems:

- **Why does the use of adversarial training and Child-Tuning in local experiments show a drop in scores?** According to our previous experience, adversarial training and Child-Tuning are effective for improving system performance. We believe this is because local experiments use the development set as the test set for validation, and the number of GRP tags in the Chinese development set is very small, resulting in a strong influence of GRP tags on the results and the score drop phenomenon.
- **Why does the score drop when using data augmentation in a test set?** After using the data augmentation method, the score showed a decreasing trend, which we believe is due to the overfitting phenomenon of adding too many entities, resulting in an imbalance between the number of entities and the number of sentences. Later, we'll experiment with changing the amount of data enhanced entities to see if we can improve the model's performance.

5 Case Study

Our model undergoes the above approach and the above challenges are effectively improved. It is able to identify ambiguous and complex entities in shorter contexts.The case study is shown in Table 5. As you can see, our method can easily predict entities like creative work compared to Baseline and other methods.

6 Conclusion and Future Work

To address the challenges of complex entities, ambiguous entities and emerging entities problem, we propose adversarial training and Child-Tuning training methods, context-independent data augmentation strategies, and a progressive domain-adaptive pre-training mechanism to improve the performance of the named entity recognition system.

In the future, we will focus our efforts on strategies and methods to enhance the use of data and hope to make a good progress.

Acknowledgements

This work is supported by the National Key Research and Development Program of China (No. 2020AAA0106400), the National Natural Science

Foundation of China (No.62176257，61976211). This work is also supported by the CCF-Tencent Open Research Fund and the Youth Innovation Promotion Association CAS.

References

Sandeep Ashwini and Jinho D Choi. 2014. Targetable named entity recognition in social media. *arXiv e-prints*, pages arXiv–1408.

Isabelle Augenstein, Leon Derczynski, and Kalina Bontcheva. 2017. Generalisation in named entity recognition: A quantitative analysis. *Computer Speech & Language*, 44:61–83.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for chinese natural language processing. *arXiv preprint arXiv:2004.13922*.

Jacob Devlin and Kristina Toutanova Ming-Wei Chang, Kenton Lee. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv*.

Besnik Fetahu, Anjie Fang, Oleg Rokhlenko, and Shervin Malmasi. 2021. Gazetteer Enhanced Named Entity Recognition for Code-Mixed Web Queries. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1677–1681.

Andreas Hanselowski, Hao Zhang, Zile Li, Daniil Sorokin, Benjamin Schiller, Claudia Schulz, and Iryna Gurevych. 2018. Ukp-athene: Multi-sentence textual entailment for claim verification. *arXiv preprint arXiv:1809.01479*.

John Lafferty, Andrew McCallum, and Fernando Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074.

Shervin Malmasi, Anjie Fang, Besnik Fetahu, Sudipta Kar, and Oleg Rokhlenko. 2022a. MultiCoNER: a Large-scale Multilingual dataset for Complex Named Entity Recognition.

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)*. 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.

Takeru Miyato, Andrew M Dai, and Ian Goodfellow. 2016. Adversarial training methods for semi-supervised text classification. *arXiv preprint arXiv:1605.07725*.

Charles Sutton, Andrew McCallum, et al. 2012. An introduction to conditional random fields. *Foundations and Trends® in Machine Learning*, 4(4):267–373.

Ming Tu, Visar Berisha, Martin Woolf, Jae-sun Seo, and Yu Cao. 2016. Ranking the parameters of deep neural networks using the fisher information. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2647–2651. IEEE.

Kai Xu, Zhanfan Zhou, Tianyong Hao, and Wenyin Liu. 2017. A bidirectional lstm and conditional random fields approach to medical named entity recognition. In *International Conference on Advanced Intelligent Systems and Informatics*, pages 355–365. Springer.