OPDAI at SemEval-2022 Task 11: A hybrid approach for Chinese NER using outside Wikipedia knowledge

Ze Chen, Kangxu Wang, Jiewen Zheng
Zijian Cai, Jiarong He and **Jin Gao**
Interactive Entertainment Group of Netease Inc., Guangzhou, China
{jackchen, wangkangxu, zhengjiewen}@corp.netease.com
{caizijian01, gzhejiarong, jgao}@corp.netease.com

Abstract

This article describes the OPDAI submission to SemEval-2022 Task 11 on Chinese complex NER. First, we explore the performance of model-based approaches and their ensemble, finding that fine-tuning the pre-trained Chinese RoBERTa-wwm model with word semantic representation and contextual gazetteer representation performs best among single models. However, the model-based approach performs poorly on test data because of low-context and unseen-entity cases. Then, we extend our system into two stages: (1) generating entity candidates by using neural model, soft-templates and Wikipedia lexicon. (2) predicting the final entity results within a feature-based rank model. For the evaluation, our best submission achieves an F_1 score of 0.7954 and attains the third-best score in the Chinese sub-track.

1 Introduction

Named Entity Recognition (NER)(Yadav and Bethard, 2019) aims to detecting the boundaries of named entities and recognizing their categories(e.g., person or location). It plays an important role in many downstream tasks, such as information extraction and question answering.

SemEval-2022 Task 11: Multilingual Complex Named Entity Recognition(MultiCoNER) (Malmasi et al., 2022b) is a shared task which encourages participants to develop NER system to detect semantically ambiguous and complex entities in short and low-context settings for 11 languages. Participants can build a NER model that works only for one language or for all the languages. And an additional track with code-mixed data are offered in this task. Different from ordinary NER, this task focuses on complex and unseen entities. Complex entities, like the titles of creative works(movie/book/software names) are harder to recognize. Additional test sets on questions and short search queries are offered in the test phase, which contains large proportion of unseen entities.

Our main interest is to build a NER system which can process complex entities and adapt to other domains in practical scenarios in Chinese language. This paper describes our two-stage hybrid approach for Chinese NER. A description of datasets provided in this shared task and additional datasets adopted in our system is given in Section 3. The implementation details of our system are listed in Section 4. We first experiment with model-based methods and integrate word semantic feature and gazetteer feature with neural model to improve inference performance. Further, we extend our model system to a two-stage prediction system: entity candidates generation and entity confidence ranking. Confidence ranking is used to pick out high-confidence entities from candidates. With the help of Wikipedia lexicon and soft templates for generating entity candidates, our hybrid system shows good performance and great domain adaption capability in the final evaluation phase.

2 Related Work

Our work is mainly related to the pre-trained language models and some specific strategies for Chinese NER task.

2.1 Pre-trained language models

Transformer-based Language Models e.g BERT (Devlin et al., 2018) have demonstrated that rich, unsupervised pre-training is an integral part of many natural language processing system. RoBERTa(Liu et al., 2019) is a BERT-based model with better performance which bring by different training strategies including training the model longer; bigger batches over more data; removing the next sentence prediction objective; training on longer sequences; and dynamically changing the masking pattern applied to the training data. RoBERTa-wwm (Cui et al., 2021) is a pre-trained language model which modifies the masked language model (MLM) task as a language correction

Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 1488 - 1493

manner and mitigates the discrepancy of the pre-training and fine-tuning stage. It could give significant gains in most various Chinese NER task (Yin et al., 2021).

2.2 Chinese NER

Compared with NER in English, Chinese NER is more difficult since there are no explicit word boundaries in sentences. Word-level information cannot be well modeled. Some approaches resort to performing Chinese NER directly at character-level (Sui et al., 2019; Ding et al., 2019) and some others perform word segmentation first (He and Sun, 2017). However, incorrect word segmentation will result in propagation errors in entity detection, and purely char-based approach will miss the word information. Pre-trained language models, such as BERT, can generate contextual embedding which can outperform other character or word-based approaches (Hu et al., 2020). More importantly, BPE subword segmentation method is employed by BERT-based models and word-level information is not explicitly modeled. Consequently, some researches introduce lexicon information into neural models which results in significant improvement (Ma et al., 2019; Liu et al., 2021). In this task, we implement similar strategies to integrate discrete lexicon and neural representation.

3 Data

We experiment using the datasets shared by MultiCoNER(Malmasi et al., 2022a) on Chinese monolingual track, which consist of 15300 training sentences, 800 validation sentences and 151661 test sentences. Entities are labeled using BIO scheme, and six entity types are involved: person(PER), location(LOC), group(GRP), corporation(CORP), product(PROD) and creative work(CW).

In order to make our model better adapt to ambiguous semantics and insufficient context, we built an entity lexicon from Wikipedia data. We parsed a Wikidata dump and mapped all the entities to our NER taxonomy following the rule from Table 1. We extracted about 3.8 million entities for Chinese language which were mapped to the entity types. Wikipedia entities with multiple categories can be mapped to different entity types in MultiCoNER. Moreover, pre-trained static word embeddings (Song et al., 2018), which provides 200-dimension vector representations for over 12 million Chinese words and phrases pre-trained on large-scale high-quality data, are adopted in this work for integrating the word-level information.

MultiCoNER Entity Types	Wikidata Entity Types
PER	human
	fictional human
GRP	music organization
	sports organization
	newspaper
	educational organization
	cultural institution
CORP	business
	enterprise
CW	creative work
LOC	location
PROD	product

Table 1: The entity types mapping between Wikipedia and MultiCoNER

4 Methodology

Our system classifies NER task into two stages: entity candidates generation and entity confidence ranking. Based on BERT-based model ensembles, soft-template methods and Wikipeida lexicon, we can first generate entity candidates. And then, handcrafted features for each entity candidate are extracted. Finally, a machine learning based rank model is used for entity confidence rank, the entities whose confidence score is above the threshold are regarded as the final predictions. Figure 1 gives an overview of our hybrid approach.

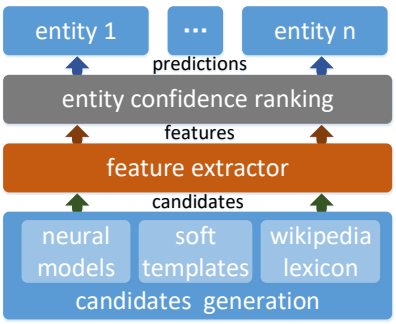


Figure 1: Overview of our hybrid approach

4.1 Entity Candidates Generation

Three approaches are adopted for entity candidates generation: model-based, template-based and lexicon-based. BERT-based model ensembles can achieve named entities recognition, however, its recall performance significantly decreases when dealing with low-context or new-entity cases. To alleviate this effect, soft templates and Wikipedia lexicon is integrated for candidates recall.

4.1.1 Model based approach

Figure 2 gives a glimpse of our model architecture, which consists of three layers: *encoding layer*, *aggregation layer* and *inference layer*. Pre-trained language models are adopted for sentence encoding which can grasp contextual information. However, sentences in Chinese are not naturally segmented, resulting in difficulties in Chinese NER task. Therefore, word semantic representation and contextual gazetteer representation are used in *aggregation* module for incorporating word lexicon information and boundary information into character representations. For further improvement, model ensemble methods are tried for prediction.

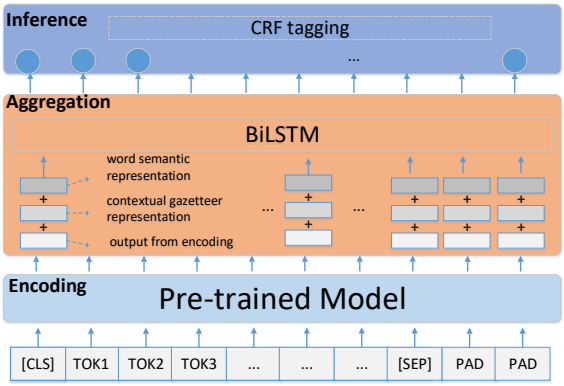


Figure 2: Model Architecture

Encoding: This layer is meant for sequence modeling to capture contextual semantic representation. BERT-based pre-trained models, such as BERT, RoBERTa and RoBERTa-wwm, which have been shown to capture implicit syntactic and semantic knowledge, are tried here.

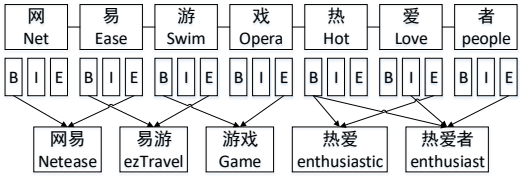
Aggregation: This layer focuses on incorporating word-level information and boundary-dependent features. We use a BiLSTM neural architecture to integrate encoding output with word semantic representation (Ma et al., 2019) and contextual gazetteer representation (Fetahu et al., 2021). An example is presented in Figure 3 for feature constructions. For each character c_i in the input sentence $s = c_1, c_2, ..., c_n$, three word sets(B/I/E) are constructed by:

$$B(c_i) = \{w_{i,k}, \forall w_{i,k} \in eL, i < k \leq n\}$$

$$I(c_i) = \{w_{j,k}, \forall w_{i,k} \in eL, 1 \leq j < i < k \leq n\}$$

$$E(c_i) = \{w_{j,k}, \forall w_{i,k} \in eL, 1 \leq j < i\}$$

Here, eL represents the Wikipedia entity lexicon. $w_{i,k}$ stands for the span that begins with c_i and ends with c_k. The average word embeddings of each word set are concatenated as a 600-dimension vector(200 for each word set) is regarded as the final word semantic representation. Contextual gazetteer representation for each character is a 13-dimension binary vector, which is introduced by Meng et al. (2021) .



(a) word semantic representation

	网 Net	易 Ease	游 Swim	戏 Opera	热 Hot	爱 Love	者 people
B-CORP	1	1	0	0	0	0	0
I-CORP	0	1	1	0	0	0	0
B-PER	0	0	0	0	0	0	0
I-PER	0	0	0	0	0	0	0
...	0	0	0	0	0	0	0
O	0	0	0	0	0	0	0

(b) contextual gazetteer representation

Figure 3: Features introduced in Aggregation Layer

Inference: Sequential conditional random field(CRF), which can capture the dependency between successive labels, is used in the inference layer for final prediction.

Ensemble: Majority voting method is applied to integrate different model results. In detail, for N different models, if more than N/2 models consider a predicted span belongs to the same entity category, the span is used for final prediction.

4.1.2 Lexicon based approach

An entity lexicon is built from wikidata with category mapping rule listed in Table 1. And then Aho-Corasick algorithm [1] is applied for span extraction. When predicting, word span in that Wikipedia lexicon is extracted as an entity candidate, if that word span maps to multiple entity categories, multiple candidates are generated.

4.1.3 Soft-templates based approach

Soft-templates are mined from training data by a simple statistical strategy. More specifically, we first replace entities in sentences with entity category placeholders, and those replaced frequently occurring sentences are regarded as soft-templates without manually labeled. To make improvements, templates differ only in placeholders are removed.

[1] https://pyahocorasick.readthedocs.io/

In the evaluation phase, additional soft-templates are mined from test data. Different from the procedures on training data, pseudo entity labels are generated from model prediction results.

4.2 Entity Confidence Ranking

The following paragraphs describe how we select high confident ones from entity candidates for the final predictions. We build a machine learning model with hand-crafted features to calculate the confidence score of each entity candidate. The hand-crafted features consist of 6-dimension lexical features and 16-dimension statistical features. An example of features extracted for an entity candidate in a sentence is given in figure 4.

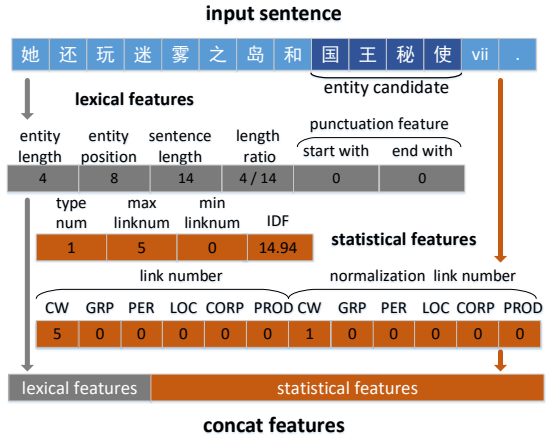


Figure 4: Handcraft Features

A. Lexical Features

- **entity length:** the length of the entity candidate.
- **entity position:** the start position of first character of entity candidate in the input sentence.
- **sentence length:** the length of the input sentence.
- **length ratio:** length ratio between entity candidate and the sentence.
- **punctuation feature:** a two-dimension binary feature indicating whether there is a punctuation at the beginning or end of the entity candidate.

B. Statistical Features

These features are extracted from wiki pages, which can capture entity type information from the link relationships among wiki entities. Two entities are thought to have a link relationship if there is any sitelink in one page which can link to the other. For each entity candidate e_i, the associated wiki pages $P(e_i) = \{p_1, p_2, ..., p_n\}$ are those who have the same label name or alias name to e_i. The number of links which redirect to an entity whose type is $etype_j$ in page p_t is represented as $linknum(p_t, etype_j)$. $etype_j, j = 0, ..., 5$ is one of the six entity types(e.g. PER,LOC,etc.). Therefore, the total number of links for e_i are calculated as $linknum(e_i, etype_j) = \sum_{t=1}^{n} linknum(p_t, etype_j)$.

- **IDF**: inverse document frequency of an entity candidate is calculated from wiki pages.
- **link number**: $\{linknum(e_i, etype_j), j = 0, 1, ..., 5\}$, a six-dimension vector, indicating the number of links in the associated wiki pages of six entity types separately.
- **normalization link number**: normalization value of link number based on entity types, $\{\frac{linknum(e_i, etype_j)}{\sum_j linknum(e_i, etype_j)}, j = 0, 1, ..., 5\}$
- **maximum link number**: maximum value of link number, $\max_{0 \le j \le 5} linknum(e_i, etype_j)$
- **minimum link number**: minimum value of link number, $\min_{0 \le j \le 5} linknum(e_i, etype_j)$
- **number of link types**: the number of entity types where $linknum(e_i, etype_j) > 0$

Then, we adopt LightGBM (Ke et al., 2017) as the multi-label classifier, which is a gradient boosting tree model and performs well on unbalanced classification tasks. For an entity candidate, the confidence on its entity type is calculated by this classifier. If the confidence score is greater than the threshold, e_i is used for the final prediction.

5 Experiments and Results

5.1 Experiment Setup

Our implementation is based on a powerful NLP framework Flair(Akbik et al., 2019), and the Transformers library by HuggingFace(Wolf et al., 2019) for the pre-trained models and corresponding tokenizers.

We first experiment on development dataset with different encoding and aggregation strategies, and we later do an ensemble of these models. During training, the data is processed by batches of size 32 and the maximum length of each sentence is set to 256. In all experiments, we use AdamW optimizer with learning rate set to 2e-5 and train our models for a maximum of 30 epochs. Then, we implement LightGBM for entity confidence ranking with learning rate set to 5e-2, maximum number of leaves set to 50, max-depth set to 6. For the evaluation phase, we mix train and development data and split it randomly for 10-fold cross validation.

Model	Dev P	Dev R	Dev F1
BERT	0.824	0.840	0.832
RoBERTa	0.850	0.837	0.842
RoBERTa-wwm	0.851	0.846	0.847
RoBERTa-wwm-large (I)	0.856	0.860	0.858
+ WE (II)	0.854	**0.863**	0.859
+ GZ (III)	**0.887**	0.842	0.859
+ WE + GZ(IV)	0.867	0.858	**0.861**
Ensemble (V)	**0.913**	0.853	**0.875**

Table 2: Best results achieved by each model on dev dataset, WE and GZ are the shorthand of word semantic representation and gazetteer representation separately

Methods	Dev P	Dev R	Dev F1
Ensemble(V)	0.913	0.853	0.875
Lexicon + LightGBM	0.842	0.779	0.810
Lexicon + V + LightGBM	**0.904**	**0.881**	**0.890**

Table 3: Hybrid results on dev dataset

5.2 Neural Model Results

Table 2 shows the performance of different models and their ensemble approach. The experimental results show that RoBERTa with whole word mask(RoBERTa-wwm) can outperform others in this Chinese language task. The first four rows show the performance of pre-trained Chinese language models. Model I to IV represent different aggregation strategies: no feature introduced, word semantic feature(WE for short) only, gazetteer feature(GZ for short) only, combination of word semantic representation and gazetteer feature representation. We find that the pre-trained model with WE introduced can achieve higher recall and with GZ introduced can achieve higher precision. The last row gives us the results of an ensemble(V) of model I to IV. Model ensemble results in an improvement of about 0.6% on macro-F1 score, indicating that predictions of model I to IV have good complementarity.

5.3 Hybrid Approach Results

Table 3 shows the results of the hybrid approach on dev dataset. We can find that the recall value of model ensemble(V) is far less than precision from the first row. By integrating lexicon-based and ensemble model-based results for entity candidates generation, and adopting LightGBM for entity confidence ranking, the recall value improves more than 3% and the macro-F1 increases by 1.4%.

The results of hybrid approaches on test dataset are given in Table 4. We can find that with lexicon-based entity candidates generation, the recall improves a lot. By integrating lexicon, model V with a LightGBM confidence ranking model, the F1 score increases more than 10%. To make further improvements, soft templates are used for additional candidates generation, which helps gain an improvement of 0.9% on F1 socre.

Methods	Test P	Test R	Test F1
Ensemble(V)	0.710	0.673	0.678
Lexicon + LightGBM	0.484	0.858	0.359
Lexicon + V + LightGBM	0.786	0.806	0.786
+ soft-template(Hybrid)	**0.805**	**0.794**	**0.795**

Table 4: Results of different approaches on test

Methods	LOWNER F1	Orcas F1	MSQ F1
Ensemble(V)	**0.854**	0.582	0.683
Hybrid	0.852	**0.747**	**0.822**

Table 5: Results of different domains on test

To further analyze the differences and respective advantages of different approaches, detail results of different domains are listed in Table 5. For the ensemble model V, when compared to LOWNER results, the results for MSQ and ORCAS are worse. This large gap shows that the existing model approach cannot generalize well.

6 Conclusion

In this paper, we introduce a hybrid approach for Chinese NER, which contains two stages: entity candidates generation and confidence ranking. We find that integrating word semantic representation and gazetteer representation can improve the performance of neural model-based approach. In particular, our ensemble model of different aggregation strategies performs better. However, due to the limitation of training data, the performance of neural model-based approach drops sharply in other new domains.

Considering that there are new entities and other domain sentences in test sets, we can use Wikipeida lexicon and soft templates to help recall unseen entities, and an entity confidence ranking model is built which results in significant improvement on test sets. For future work, semi-supervised approaches or data augmentation methods could be explored to alleviate the limitation of training data.

References

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. Flair: An easy-to-use framework for state-of-the-art nlp. In *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. Pre-training with whole word masking for chinese bert. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3504–3514.

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*.

Ruixue Ding, Pengjun Xie, Xiaoyan Zhang, Wei Lu, Linlin Li, and Luo Si. 2019. A neural multi-digraph model for chinese ner with gazetteers. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1462–1467.

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.

Hangfeng He and Xu Sun. 2017. A unified model for cross-domain and semi-supervised named entity recognition in chinese social media. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Yuting Hu, Suzan Verberne, D Scott, N Bel, and C Zong. 2020. Named entity recognition for chinese biomedical patents. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 627–637. International Committee on Computational Linguistics.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30.

Wei Liu, Xiyan Fu, Yue Zhang, and Wenming Xiao. 2021. Lexicon enhanced chinese sequence labeling using bert adapter. *arXiv preprint arXiv:2105.07148*.

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 preprint arXiv:1907.11692*.

Ruotian Ma, Minlong Peng, Qi Zhang, and Xuanjing Huang. 2019. Simplify the usage of lexicon in chinese ner. *arXiv preprint arXiv:1908.05969*.

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.

Yan Song, Shuming Shi, Jing Li, and Haisong Zhang. 2018. Directional skip-gram: Explicitly distinguishing left and right context for word embeddings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 175–180.

Dianbo Sui, Yubo Chen, Kang Liu, Jun Zhao, and Shengping Liu. 2019. Leverage lexical knowledge for chinese named entity recognition via collaborative graph network. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3830–3840.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Vikas Yadav and Steven Bethard. 2019. A survey on recent advances in named entity recognition from deep learning models. *arXiv preprint arXiv:1910.11470*.

Xunwei Yin, Shuang Zheng, and Quanmin Wang. 2021. Fine-grained chinese named entity recognition based on roberta-wwm-bilstm-crf model. In *2021 6th International Conference on Image, Vision and Computing (ICIVC)*, pages 408–413.