

CMB AI Lab at SemEval-2022 Task 11: A Two-Stage Approach for Complex Named Entity Recognition via Span Boundary Detection and Span Classification

Keyu PU, Hongyi LIU, Yixiao YANG, Jiangzhou JI, Wenyi LV and Yaohan HE

CMB AI Lab

{pukeyu, lhy24, yangyixiao}@cmbchina.com
{jesse, lvwymail, heyh18}@cmbchina.com

Abstract

This paper presents a solution for the SemEval-2022 Task 11 Multilingual Complex Named Entity Recognition. What is challenging in this task is detecting semantically ambiguous and complex entities in short and low-context settings. Our team (CMB AI Lab) propose a two-stage method to recognize the named entities: first, a model based on biaffine layer is built to predict span boundaries, and then a span classification model based on pooling layer is built to predict semantic tags of the spans. The basic pre-trained models we choose are XLM-RoBERTa and mT5. The evaluation result of our approach achieves an F1 score of 84.62 on sub-task 13, which ranks the third on the learder board.

1 Introduction

Named entity recognition (NER)(Tjong Kim Sang and De Meulder, 2003) is a fundamental task in natural language processing, aiming at identifying the spans of texts that refer to entities. NER is widely applied to information extraction and data mining(Lin et al., 2019)(Cao et al., 2019), which is greatly challenging in practical and open domain settings. However, the previous research has not paid much attention on processing complex and ambiguous named entities.

SemEval 2022 task 11 (Malmasi et al., 2022b) containing a total of 13 sub-tasks is a complex NER task which focuses on detecting semantically ambiguous and complex entities in short and low-context settings (Meng et al., 2021). For the purpose of testing the domain adaption capability of the participating models, the task not only set 11 base sub-tasks: English, Spanish, Dutch, Russian, Turkish, Korean, Farsi, German, Chinese, Hindi and Bangla, but also set two additional testing sets on questions and short queries: Multilingual, and code-mixed (Fetahu et al., 2021). We conduct a two-stage method to deal with the code-mixed sub-task, which achieves an F1 score of 84.62.

This paper is structured as follows. The related work of NER is briefly introduced in Section 2. The data for training and testing the model is presented in Section 3. The details of the two-stage method is described in Section 4. The experimental results of our method are exhibited in Section 5. Section 6 summarizes this paper.

2 Related Work

In the NLP field, the NER task is usually considered as a sequence labeling problem (Liu et al., 2018) (Lin et al., 2019) (Cao et al., 2019). With well-designed features, CRF-based models have achieved the leading performance (Lafferty et al., 2001) (Finkel et al., 2005) (Liu et al., 2011). Recently, neural network models have been exploited for feature representations (Chen and Manning, 2014). Moreover, contextualized word representations such as ELMo (Peters et al., 2018) and BERT (Devlin et al., 2019) have also achieved great success. As for NER, the end-to-end bi-directional LSTM + CRF model (Lample et al., 2016) (Yang et al., 2018) is one representative architecture. These models are only capable of recognizing regular named entities.

In e-commerce search domain, a common scenario is code-mixed queries, with query terms composed of multiple languages(Bhargava et al., 2016). The application for code-mixed Web queries still remains challenging(Gupta et al., 2014). A recent work proposed an NER hybrid approach for code-mixed queries, consisting of a gazetteer and tree based identifier(Bhargava et al., 2016). Another work leverages linguistic features to train a conditional random field (CRF) model, where the output is further processed using multi-lingual gazetteer lists(Gupta et al., 2016). In Seme Val 2022 task 11 code-mixed sub-task, we use a two-stage NER approach and get the 3rd place in the competition.

3 Data

In this task, we use the official raw data (Malmasi et al., 2022a) to train and test our model. Each line of texts in the data belongs to a sample, the languages involved are: English, Spanish, Dutch, Russian, Turkish, Korean, Farsi, German, Chinese, Hindi, and Bangla, some of which are also provided with code-mixed data as an additional sub-task. Entity types include Person, Location, Group, Corporation, Product, and Creative Work. The participants have to use their systems to accurately detect the entities and submit the predictions for the mixed languages task.

In a data file, samples are separated by blank lines. Each data instance is tokenized and each line contains a single token in the first column with the associated label in the last (4th) column. The second and third columns are underscores (`_`) to separate the tokens and the labels. The entities are labeled with the BIO scheme, which means that the token tagged O is not a part of the entity, the token tagged B-X is the first token of an X type entity, and the remaining tokens of the entity are tagged as I-X.

When the amount of training data is insufficient or unevenly distributed, data augmentation can quickly expand the corpus to avoid overfitting. At the same time, data augmentation can also improve the robustness of the model, preventing the performance of the model from being greatly reduced once the data only changes slightly. We build a dictionary with all entities of the same type to randomly replace the entities in each sample, and translate the replaced entities into other languages to expand the dataset, which is similar to autoencoders in the computer vision. However, translation between different languages relies on a large number of parallel corpuses, and requires training first.

4 Methodology

The two-stage method we use in this task includes two separated models to recognize the named entities: one for predicting the boundaries of the spans, and the other for predicting the semantic tags of the spans. The processing flow of our approach is depicted in Figure 1.

4.1 Text Encoders

The models we employed are both trained based on XLM-RoBERTa_{LARGE} and mT5_{LARGE}.

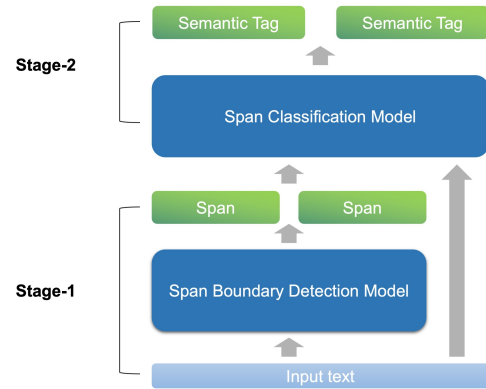


Figure 1: The processing flow of our approach

XLM-RoBERTa (Conneau et al., 2020): XLM-RoBERTa_{LARGE} is pre-trained on 2.5TB of filtered common crawl data containing 100 languages, which consists of 24 transformer layers, 16 self-attention heads per layer, and a hidden size of 1024. In order to deal with a large number of common words in natural language corpus, BPE (byte pair encoding), a coding schema mixed by the character level and the word level representation, is utilized to process the text data.

mT5 (Xue et al., 2021): mT5_{LARGE} is a multilingual pre-trained text-to-text transformer which is pretrained on the common crawl-based dataset corpus, covering 101 languages. We only use the encoder of the mT5_{LARGE} consisting of 24 transformer layers.

4.2 Boundary Detection of Spans

The first stage of our method is to extract the phrase. As shown in Figure 2, we built a boundary detection model by connecting the last hidden states of the pre-trained model to the biaffine layer (Yu et al., 2020) to obtain the span boundaries, which can also be regarded as a named entity recognition (NER) model that only recognize one single category.

The output of the biaffine model is a span boundary matrix as illustrated in Figure 3. All pairs of start-end tokens have corresponding scores indicating whether they are the spans we need. Figure 4 shows an example of a span boundary matrix: the reason why *Jackie Ma* and *the louvre museum* are the entities in this sentence is that they are the two pairs of start-end tokens.

To better detect the span’s boundary, we build a dictionary from the training data. Words that appear in training data more than twice are selected and then splice together with the original sentence

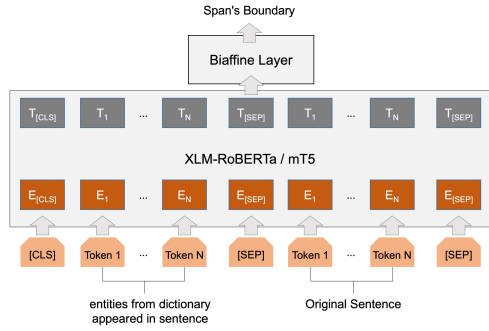


Figure 2: Span boundary detection model

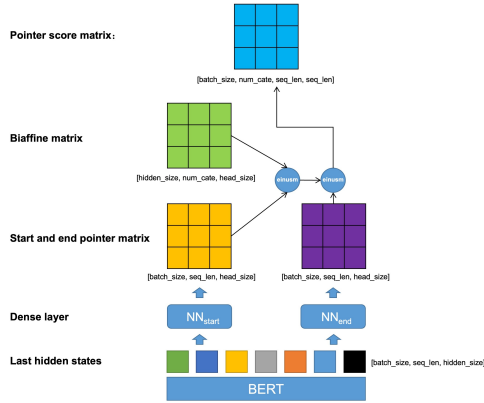


Figure 3: Biaffine model: using start-end pointer to explore all spans

into the input text fragment, as shown below:

[CLS] *word selected 1* [SEP] *word selected 2*
[SEP] ... [SEP] *sentence* [SEP].

4.3 Span Classification

The second stage of our method is to classify the spans obtained in the first stage. We built a classification model to determine which semantic tag the span belongs to. As shown in Figure 5, a full connection layer is connected to the pooling layer of the pretrained model to output the score for each category, and then is activated by a softmax function, with the cross entropy set as the loss function of the classification model.

It should be noted that the text fragment to be classified is constructed by the phrase extracted from the first model (boundary detection) and the original sentence. Similar to the sentence pair classification, a sample sentence before BPE applied appears as below:

[CLS] *phrase extracted* [SEP] *sentence* [SEP].

		end							
		Jackie	Ma	has	been	to	the	louvre	museum
start	Jackie	0	1	0	0	0	0	0	0
	Ma		0	0	0	0	0	0	0
	has			0	0	0	0	0	0
	been				0	0	0	0	0
	to					0	0	0	0
	the						0	0	1
	louvre							0	0
museum									0

Figure 4: Span boundary matrix: scores of all start-end token-paire for detecting boundary of span

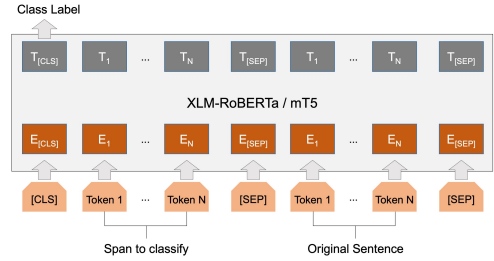


Figure 5: Span classification model

4.4 Training Procedures

In the first stage (span boundary detection), the Adam optimizer with a learning rate of 6×10^{-6} is employed, the batch size is set to 12, and the model is trained for 30 epochs. As for the second stage (span classification), the learning rate of the Adam optimizer is set to 8×10^{-6} , and the batch size and the number of epoch amounts are 16 and 40 respectively. For each stage, we use both the 10-fold cross-validation.

During the training phase of the span classification model, FGM adversarial learning is applied to improve the robustness of model: the samples are mixed with some fairly small disturbances that might lead to misclassification, and the neural network is then adapted to the disturbances to be robust to the adversarial samples.

4.5 Ensemble Model

The ensemble of deep learning models has a great improvement on the test dataset. We ensemble the predictions of span boundary detection models by voting strategy to get the best span boundary. Besides, the predictions of span classification models

are also combined to get the final predictions. Each model is trained on different dataset augmented and based on different text encoders (i.e., XLM-RoBERTa and mT5). What is most conspicuous, however, is that the strategy of the two-stage model performs better than the traditional ner model in this task.

5 Results

A two-stage method is employed to complete the code-mixed language sub-task. Based on two pre-trained models (XLM-RoBERTa and mT5), we adopted a variety of optimization schemes, such as: biaffine network structure, two-stage entity prediction, adding distantly supervised dictionary and adversarial training, all of which have achieved a certain improvement, according to the evaluation results shown in Table 1. Lastly, we voted on all prediction results in terms of ensemble learning idea to get the final submission file.

With the XLM-RoBERTa + crf method as the baseline, and an end-to-end structure, we get an F1 score of 79.7. After using the biaffine network structure and two-stage optimization architecture instead, the F1 score improves to 80.9 and 81.3 respectively. In addition, the two-stage optimization architecture introducing supervised dictionary, adversarial training and data augmentation obtains F1 scores of 82.5, 83.1 and 82.7 respectively. Compared to the XLM-RoBERTa, the prediction results acquired based on the mT5 pre-trained model are improved by an average of 0.4 points. As a result, the final evaluation scores gained by voting is 84.62.

6 Conclusion

Aiming at the complex multilingual ambiguity and lack of context in this competition, we adopt a deep learning network model for entity extraction based on the biaffine attention mechanism, and carry out transfer learning based on different pre-trained models such as RoBERTa and mT5.

Through adversarial training, the robustness of the model is enhanced, and the two-stage training also improves the performance of the model in few-shot scenarios. Besides, a remote supervised dictionary is added to revise the results, and the entity dictionary for random replacement and multilingual machine translation is used for data augmentation. Usually for enhanced data, it is necessary to give a weight less than 1, which is different from

Comparison of different methods	
Method	F1(%)
Baseline XLM-RoBERTa+crf	79.7
Biaffine	80.9
Two-Stage	81.3
w/Dict	82.5
w/adv train	83.1
w/data augmentation	82.7
Baseline mT5+crf	80.2
Biaffine	81.2
Two-Stage	81.7
w/Dict	82.8
w/adv train	83.6
w/data augmentation	83.2
Ensemble strategy	84.62

Table 1: The code-mixed sub-task evaluation results

real data. Data augmentation can also alleviate the problem of data imbalance. Ultimately, the best result (F1 score of 84.62) is achieved via ensemble learning voting strategy.

References

- Rupal Bhargava, Bapiraju Vamsi Tadikonda, and Yashvardhan Sharma. 2016. Named Entity Recognition for Code Mixing in Indian Languages using Hybrid Approach. In *FIRE*.
- Yixin Cao, Zikun Hu, Tat-seng Chua, Zhiyuan Liu, and Heng Ji. 2019. Low-Resource Name Tagging Learned with Weakly Labeled Data. 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 261–270.
- Danqi Chen and Christopher D Manning. 2014. A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 740–750.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.
- 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, Volume 1 (Long and Short Papers)*, pages 4171–4186.
- 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.
- Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL’05)*, pages 363–370.
- Deepak Gupta, Shubham Tripathi, Asif Ekbal, and Pushpak Bhattacharyya. 2016. A Hybrid Approach for Entity Extraction in Code-Mixed Social Media Data. *Money*, 25(66).
- Parth Gupta, Kalika Bali, Rafael E Banchs, Monojit Choudhury, and Paolo Rosso. 2014. Query Expansion for Mixed-Script Information Retrieval. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pages 677–686.
- John Lafferty, Andrew McCallum, and Fernando C.N. Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the eighteenth international conference on machine learning, ICML, volume 1*, pages 282–289.
- 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.
- Ying Lin, Liyuan Liu, Heng Ji, Dong Yu, and Jiawei Han. 2019. Reliability-aware Dynamic Feature Composition for Name Tagging. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 165–174.
- Liyuan Liu, Jingbo Shang, Xiang Ren, Frank Fangzheng XU, Huan Gui, Jian Peng, and Jiawei Han. 2018. Empower Sequence Labeling with Task-Aware Neural Language Model. In *Thirty-Second AAAI Conference on Artificial Intelligence*.
- Xiaohua Liu, Shaodian Zhang, Furu Wei, and Ming Zhou. 2011. Recognizing Named Entities in Tweets. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 359–367.
- 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.
- Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237.
- Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.
- Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498.
- Jie Yang, Shuailong Liang, and Yue Zhang. 2018. Design Challenges and Misconceptions in Neural Sequence Labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3879–3889.
- Juntao Yu, Bernd Bohnet, and Massimo Poesio. 2020. Named Entity Recognition as Dependency Parsing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6470–6476.