DANGNT-SGU at SemEval-2022 Task 11: Using Pre-trained Language Model for Complex Named Entity Recognition

Dang Tuan Nguyen, Huy Khac Nguyen Huynh
Saigon University, Ho Chi Minh City, Vietnam
{dangnt, hnkhuy}@sgu.edu.vn

Abstract

This paper describes a system that we built to participate in the SemEval 2022 Task 11: MultiCoNER Multilingual Complex Named Entity Recognition, specifically the track Mono-lingual in English. To construct this system, we used Pre-trained Language Models (PLMs). Especially, the Pre-trained Model base on BERT is applied for the task of recognizing named entities by fine-tuning method. We performed the evaluation on two test datasets of the shared task: the Practice Phase and the Evaluation Phase of the competition.

1 Introduction

In natural language processing, Named Entity Recognition (NER) is traditionally a fundamental task that aims to recognize groups of words used to identify people, organizations, places, times, etc. in documents. SemEval-2022 Task 11 focuses on detecting semantically complex and ambiguous entities (Malmasi et al., 2022b). This is a challenging NLP task that has not sufficiently received attention from the research community. In practice, available named entity recognition models find it hard to recognize complex named entities, such as titles of works (*movie/book/song/software titles*) that are not simple nouns (Malmasi et al., 2022b).

Type	Description
Complex entities	Not all entities are proper noun -Noun phrases: *Eternal Sunshine of the Spotless Mind* -Gerunds: *Saving private Ryan* -Full clauses: *Mr. Smith goes to Washington…*
Ambiguous entities and Contexts	Not always entities *-Inside out (movie),* *-Bonanza (game)* *-On the Beach (movie)…*
Emerging Entities	New books/songs/movies released daily, weekly…

Table 1: SemEval-2022 Task 11 illustrates some types of complex and ambiguous entities

To enable the recognition of such entities, we believe that it is necessary to make full use of Pre-trained Language Models (PLMs) to increase efficiency as well as reduce training time, which is also the method we applied in this study.

2 Related Work

Nowadays, there are many NER works and different approaches to solving this problem, such as LSTM (Long Short-Term Memory) (Nut Limsopatham and Nigel Collier, 2016), Transformers (Ashish Vaswani et al., 2017). In particular, the BERT model (Jacob Devlin et al., 2019). Many relevant studies also approach the direction of combining deep learning models and adjusting model parameters to improve the efficiency of named entity recognition in the text (Zheng Yuan et al., 2021). However, entities with complex and ambiguous names appear continuously in increasing numbers. This has posed emerging challenges for the natural language processing field. Many research show that clarifies that named substance acknowledgement is particularly troublesome in circumstances with a low context or in scenarios where the named substances are especially complex (Meng et al., 2021). The misrecognition of entities with complex and ambiguous names significantly affects the performance of natural language processing systems (Andreas Hanselowski et al., 2018). In this paper, we will use PLMs to find ways to improve accuracy in such cases.

PLMs are Pre-trained Language Models with large datasets to be utilized in specific

Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 1483 - 1487
July 14-15, 2022 ©2022 Association for Computational Linguistics

tasks. In particular, BERT has drawn great attention and it is the favorite research direction of the NLP community after achieving state-of-the-art on 11 NLP tasks (Jacob Devlin et al., 2019). Based on BERT, there have been many research directions from the NLP research community: distilBERT (Victor Sanh et al., 2019), RoBERTa (Yinhan Liu et al., 2019), and so on.

3 Data

We will use the training and evaluation **English** dataset provided at SemEval 2022 Task 11: MultiCoNER Multilingual Complex Named Entity Recognition to conduct the research with 15300 training examples and 800 validation examples (Malmasi et al,. 2022a).

Language	Train	Dev
BN-Bangla	15,300	800
DE-German	15,300	800
EN-English	15,300	800
ES-Spanish	15,300	800
FA-Farsi	15,300	800
HI-Hindi	15,300	800
KO-Korean	15,300	800
NL-Dutch	15,300	800
RU-Russian	15,300	800
TR-Turkish	15,300	800
ZH-Chinese	15,300	800
Total	168,300	8,800

Table 2: Lists the sizes of the datasets

The data will be formatted according to CoNLL (Malmasi et al., 2022a)

Text	Format
kingdom	B_CW
hospital	I-CW
,	O
lewiston	B-LOC
from	O
stephen	B-PER
king	I-PER
of	O
the	O
same	O
name	O

Table 3: Data Format use in Task 11: Multi-CoNER Multilingual Complex Named Entity Recognition

For the missions of the competition, we will focus on 6 types of entities:

- PER: *Person*
- LOC: *Location*
- GRP: *Group*
- CORP: *Corporation*
- PROD: *Product*
- CW: *Creative Work*

In total, there are 13 tags: B-PER, I-PER, B-LOC, I-LOC, B-GRP, I-GRP, B-CORP, I-CORP, B-PROD, I- PROD, B-CW, I-CW, and O with B is the beginning of a named entity, I is all the words inside an entity and all words outside an entity (Malmasi et al., 2022b)

4 Methodology

In this paper, we use the Transformer library of HuggingFace (Wolf et al.,2020). Currently, the HuggingFace library is considered the most powerful and widely accepted Pytorch interface to deal with BERT. In addition to the support for a wide range of pre-trained language models, the library also includes pre-built modifications of BERT tailored to specific tasks. Then we tested several models based on BERT downloaded from Huggingface as BERT, RoBERTa, and XML-RoBERTa with different versions.

4.1 BASE-BERT

BERT is a transformers model pretrained on a large corpus of English data in a self-supervised fashion (Jacob Devlin et al., 2019). We use the most basic version of BERT and are downloaded from Huggingface with two versions which are **bert-base-uncased** and **bert-large-uncased.**

4.2 RoBERTa

BERT is trained simultaneously with 2 tasks called *Masked LM* (to predict the missing word in a sentence) and *Next Sentence Prediction* (NSP-to predict the next sentence in the current sentence). Meanwhile, to enhance the training process, instead of using the next sentence prediction mechanism from BERT, **RoBERTa** uses a dynamic masking technique, thereby the mask tokens will be changed during the process. The use of a larger batch size shows better performance in training (Yinhan Liu et al., 2019), we applied on **roberta-base** version from Hugging face.

4.3 XML-RoBERTa

XLM-RoBERTa model pre-trained on 2.5TB of filtered CommonCrawl data containing 100 languages. It was introduced in (Conneau, Alexis et al., 2019). The **xlm-roberta-large** version we chose to use.

4.4 Experimental Setup

Resources are required to get started; as a large neural network will have to be trained, we use Google Colab Pro to make the best use of the GPU and TPU resources from this environment.

Regarding the hyper-parameter, we have inspected many different values around the "*work well across all tasks*" value recommendations (according to Jacob Devlin et al., 2019) and this is the parameter set that gives the best results: *Batch size: 64, Learning rate: 2e-5, Number of epochs: 10, Maximum sequence length: 128.* Besides, with the desire to increase the recognition efficiency, we try to add a Linear layer and a softmax layer on top to recognize the entity, however, it has no effect on the results.

We import BERT's word tokenizer, which is used to convert text into tokens corresponding to BERT's vocabulary set. BERT requires specifically formatted inputs. For each encoded input sentence, we need to generate *input ids*, *segment mask, attention mask, label*

When the input data is correctly formatted, there comes the fine-tuning stage of the BERT models. For this task, we start with the modification of the pre-trained BERT model to provide outputs for the named entity recognition task, and then we continue to train the model on the dataset until the whole model is correspondent to the task. This is a prominent advantage of using Huggingface Pytorch library which already contains a set of interfaces designed for many NLP tasks.

To evaluate the model, we will use the **Macro Precision, Recall, and F1-score** indexes. The Precision ratio of the number of positive points correctly predicted by the model to the total number of points predicted by the model is Positive. The Recall is the ratio of the number of positive points the model correctly predicted to the total number of points that are actually Positive. F1-score is the harmonic mean of precision and recall (assuming these two quantities are different from 0) (Powers, David, 2008)

$$macro - Precision = \frac{tp}{tp + fp}$$

$$macro - Recall = \frac{tp}{tp + fp}$$

$$macro\ - F1 = \frac{2 * Precision * Recall}{Precision + Recall}$$

5 Results

This section describes the evaluation results of our system with PLMs based on BERT. These results are calculated using the script provided by SemEval 2022 Task 11: MultiCoNER Multilingual Complex Named Entity Recognition (Malmasi et al., 2022b)

Model	**PER**			**LOC**			**CW**			**GRP**			**CORP**			**PROD**		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
base-bert-uncased	0.924	0.959	0.941	0.823	0.893	0.857	0.667	0.761	0.711	0.827	0.879	0.852	0.860	0.798	0.828	0.658	**0.864**	**0.747**
bert-large-uncased	**0.959**	0.966	0.962	0.852	**0.910**	**0.880**	0.743	0.790	0.766	0.848	**0.937**	0.890	**0.904**	**0.829**	**0.865**	**0.706**	0.735	0.720
roberta-large	0.906	0.897	0.901	0**.863**	0.889	0.876	0.674	0.716	0.694	0.793	0.847	0.819	0.875	0.798	0.835	0.687	0.776	0.728
xml-roberta-large	0.956	**0.976**	**0.966**	0.857	0.893	0.874	**0.772**	**0.830**	**0.800**	**0.874**	0.911	**0.892**	0.859	0.819	0.838	0.671	0.789	0.725

Table 4: Summary of the scores of all the models tested with 6 types of entities in this paper use dataset on the SemEval 2022 Task 11 (Malmasi et al., 2022a)

Based on the results of table 4, the two models bert-large-uncased and xml-roberta-large are slightly better than the other two models in each entity type. In which xml-roberta-large gives the best results in 3 entity types: PER, CW, GRP, while bert-large-uncased is slightly better in two entity types LOC, CORP and interestingly base-bert-uncased gave the best results in the PROD type.

According to the results in the table 5, it can be observed that testing with the XML-RoBERTa model (Conneau, Alexis et al., 2019). with version xml-roberta-large gives good results compared to the rest of the models when it comes to average performance, followed by the bert-large-uncased model (Jacob Devlin et al., 2019).

Model	Precision	Recall	F1-score
base-bert-uncased	0.804	0.869	0.835
bert-large-uncased	**0.849**	0.876	**0.863**
roberta-large	0.812	0.832	0.822
xml-roberta-large	0.845	**0.882**	**0.863**

Table 5: Summary of the scores of all the models tested with the value of macro average performance in this paper on the SemEval 2022 Task 11 (Malmasi et al., 2022a)

However, in general, the results between the models are not much different, so to improve the efficiency of the NER model, in addition to using PLMs, it is necessary to learn more about other improvement methods to achieve positive results. When analyzing each measure, the BERT model with the bert-large-uncased version gave the highest Precision result (0.849), showing that the BERT model has the highest accuracy in entity labeling. Among the tested models, xml-roberta-large gives the best results in 2 measures: recall and F1-score. Therefore, to participate in the Evaluation phase of SemEval 2022 Task 11: MultiCoNER Multilingual Complex Named Entity Recognition (Malmasi et al., 2022b) we selected the model with PLM xml-roberta-large performed on the data set of this phase. with data size of 15300 training examples and 217.818 test examples (Malmasi et al., 2022a), the result is F1-score of 0.67, details as below table 6:

	Precision	Recall	F1-score
Type	**0.6651**	**0.677**	**0.6689**
LOC	0.6401	0.7405	0.6866
PER	**0.8233**	**0.8729**	**0.8474**
PROD	0.6092	0.6433	0.6258
GRP	0.6448	0.6449	0.6448
CW	**0.5508**	**0.5781**	**0.5641**
CORP	0.7222	0.5823	0.6448

Table 6: Results of our system evaluation in dataset validation of Evaluation Phase (Malmasi et al., 2022a)

In Evaluation Phase, the test dataset in this phase is 14 times larger in size than the training dataset (217,818 vs 15300 examples), which is a big challenge for this year's competition.

6 Conclusion

In this paper, we proudly introduced our semantically complex and ambiguous entity recognition system. We tested on different PLMs to evaluate the effectiveness of entity recognition. We trained each model with the dataset provided by the contest (Malmasi et al., 2022a): model BERT with two versions bert-base-uncased and bert-large-uncased, model RoBERTa with roberta-large version and model XML-RoBERTa with xml-roberta-large, and we also tried to fine-tune the hyperparameters to increase recognition efficiency, but there were no significant changes. The evaluation results in the Practice Phase are quite positive. Finally, based on the evaluation results, we used the xml-roberta-large model to participate in the Evaluation phase of the competition, but the results were not good.

In future work, we plan to continue testing on other PLMs. In addition, we will also extend the approach by applying deep learning techniques and theories of language to improve the accuracy in the task of recognizing possible entities, especially complex entities

References

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

Malmasi, Shervin and Fang, Anjie and Fetahu, Besnik and Kar, Sudipta and Rokhlenko, Oleg. 2022b. SemEval-2022 Task 11: Multilingual Complex Named Entity Recog- nition (MultiCoNER). In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022).* Association for Computational Lin- guistics.

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 Lan- guage Technologies*, pages 1499–1512.

Nut Limsopatham, Nigel Collier. 2016. Bidirectional LSTM for Named Entity Recognition in Twitter Messages. *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, pp: 145–152.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin. 2017. Attention Is All You Need. *arXiv:1706.03762 [cs.CL].*

Zheng Yuan, Yijia Liu, Chuanqi Tan, Songfang Huang, Fei Huang. 2021. Improving Biomedical Pretrained Language Models with Knowledge, *BioNLP 2021, arXiv:2104.10344 [cs.CL].*

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. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 103–108, Brussels, Belgium. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pretraining 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. 4171–4186.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. In *Proceedings of the 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing (EMC2) co-located with the Thirty-third Conference on Neural Information Processing Systems (NeurIPS 2019)*. 1–5.

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(2019)*

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pier- ric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020.Transform- ers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Conneau, Alexis & Khandelwal, Kartikay & Goyal, Naman & Chaudhary, Vishrav & Wenzek, Guillaume & Guzman, Francisco & Grave, Edouard & Ott, Myle & Zettlemoyer, Luke & Stoyanov, Veselin. 2020. Unsupervised Cross-lingual Representation Learning at Scale. *8440-8451. 10.18653/v1/2020.acl-main.747.*

Powers, David. (2008). Evaluation: From Precision, Recall and F-Factor to ROC, *Informedness, Markedness & Correlation. Mach. Learn. Technol*, 2.