PA Ph&Tech at SemEval-2022 Task 11: NER Task with Ensemble Embedding from Reinforcement Learning

Qizhi LIN, Xiaopeng WANG, Xiandi JIANG, Benqi WANG, Qifeng XIAO
Ping An Puhui Enterprise Management Co. Ltd, Shanghai, China
linqizhi090@lu.com, wangxiaopeng069@lu.com
Changyong HOU, Yixuan QIAO, Jun WANG, Peng JIANG
Ping An Healthcare Technology, Beijing, China
wangjun916@pingan.com.cn

Abstract

From pretrained contextual embedding to document-level embedding, the selection and construction of embedding have drawn more and more attention in the NER domain in recent research. This paper aims to discuss the performance of ensemble embeddings on complex NER tasks. Enlightened by Wang's methodology, we try to replicate the dominating power of ensemble models with reinforcement learning optimizor on plain NER tasks to complex ones. Based on the composition of semeval dataset, the performance of the applied model is tested on lower-context, QA, and search query scenarios together with its zero-shot learning ability. Results show that with abundant training data, the model can achieve similar performance on lower-context cases compared to plain NER cases, but can barely transfer the performance to other scenarios in the test phase.

1 Introduction

Named Entity Recognition (NER) as a typical topic in the NLP field, has displayed numerous outstanding outcomes in recent years, especially benefiting from development of pretrained models. However, there are still challenging aspects that remain to be tackled, such as short texts (low-context), emerging entities, and complex entities. In this task, corpus from low-context, QA and search query scenarios are collected to represent different complex NER tasks (Meng et al., 2021). Wang et al. (2021) proposed an automated concatenation model on plain NER tasks with benchmark dataset like CONLL03, which automatically generates optimized concatenation of stack embeddings with reinforcement learning strategies for structure prediction tasks like NER. A similar methodology is applied on the given dataset to test if the challenges mentioned above can be properly addressed.

2 Related Work

The development of sequence tagging models can be concluded mainly into two separate parts: encoder and decoder. Ever since the BiLSTM-CRF model (Ma and Hovy, 2016) was brought up, it has been widely used as the decoder end of NER models, while researchers shifted increasing attention to the structure of encoder. Various categories of embeddings have been raised, including character embeddings, non-contextual embedding like word2vec (Mikolov et al., 2013) and pretrained contextualized embeddings like ELMo (Peters et al., 2018) and Flair (Akbik et al., 2018). With the emergence of Transformers, the performance of large pretrained contextualized models (Devlin et al., 2019)have sweeped the leaderboard in lots of NLP tasks. In terms of language, Multilingual BERT(M-BERT) (Pires et al., 2019) demonstrates excellent representation of multilingual embeddings, which was later surpassed by XLM-R (Conneau et al., 2020), a more powerful and comprehensive multilingual model. To better allocate these embedding methods, many recent researchers have tried different methods like ensemble, weighting and concatenation. Automated concatenation (Wang et al., 2021) is a superior method that automatically generates optimized concentration of stack embeddings with reinforcement learning strategies. It uses result accuracy as reward for the controller to decide which embeddings to drop.

3 Data

As described by Fetahu et al. (2021), the dataset (Malmasi et al., 2022a) of this semeval task (Malmasi et al., 2022b) mainly consists of three sources: Low-Context Wikipedia, MS-MARCO Question (Bajaj et al., 2018) and ORCAS Search Query (Craswell et al., 2020). Sentences from Wikipedia are parsed and linked pages are resolved to their respective Wikidata entities, to create a corpus of 1.4

Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 1444 - 1447
July 14-15, 2022 ©2022 Association for Computational Linguistics

million low-context sentences with annotated entities. Part of them form the train and dev dataset. By templating questions in MS-MARCO QnA dataset and 10 million Bing user queries from the ORCAS dataset, and slotting with random entities based on frequency, 17, 868 questions and 471, 746 queries are generated to form the test set together with the rest of Low-context sentences.

Several data processing and augmenting is conducted before training to cater for the applied model. Since document-level embeddings have been proved to be effective for performance improvement in NLP tasks, we adopt Yamada et al. (2020)'s method and extract features by sending the adjacent sentences in corpus together as a document to pretrained models. The number of sentences that each document contains is defined by grid search and eventually set to 30. Although the adjacent sentences are not sentimental related as a document, the document level representation still enriches the context and greatly resolves the low-context situation.

Another novelty of the semeval dataset is the unbalance between train and test set, in terms of quantity and content. To solve the quantity unbalancing, we manage to augment the train set by slotting the entities in sentences and altering with another one from the same type. In case of overfitting, we eventually augment the train set to 3 times of its original size.

4 Methodology

After dealing with data, a common approach for better results is to fine-tune the transformer-based embeddings first, where sentences are sent to the model which is connected to a linear layer for tag prediction. Different language embeddings are selected for different tracks of the task. For English models, we fine-tune BERT-base, BERT-large, M-BERT, XLNET (Yang et al., 2019), Roberta (Liu et al., 2019), XLM-R separately and concatenate these embeddings with basic settings of other embeddings like ELMo, Flair and fastText (Bojanowski et al., 2017), for final training. For Dutch, Spanish and German models, we fine-tine M-BERT, XLM-R and BERT for each language respectively. The parameters for the fine-tuning process are 10 max epochs with batch size of 1 and learning rate of 5.0×10^{-06}.

The concatenated embeddings are used as inputs for a sequence tagging model with BiLSTM layers and CRF layer. The accuracy of the model is used as the reward for reinforcement learning to train the controller, which uses the policy gradient method to maximize the expected reward. We refer to Wang's search space algorithm for the design of reward function and gradient update. The parameters for the training process are 25 maximum episodes, 70 maximum epochs with batch size of 32 and learning rate of 5.0×10^{-06}.

	Dutch	Spanish	German
XLM-R+Fine-tune	0.8985	0.8502	0.8983
Final+Fine-tune	0.9030	0.8612	0.9106

Table 1: Dev Results

5 Results

Due to limitations in time and calculation resources, we only present the results on non-English language models. Table 1 shows the comparison between fine-tuning marco-F1 scores and final marco-F1 scores on dev set for each language in our experiment. It clearly shows that the automatic concatenation method is effective in improving the performance of fine-tuned embeddings. We can also find that XLM-R model plays a vital important role in the concatenated embeddings for non-English models.

Table 2 shows detailed final results on the test set for each language. All scores are calculated as F1 scores. As mentioned in the data section, apart from the performance on low-context NER tasks, the dataset also focuses on the zero-shot learning ability of models on QA and search query scenarios. It can be seen that with fine-tune and reinforcement learning training on low-context corpus, the model achieves high performance on the responsive part in test set, but fails to maintain the performance when making predictions on corpus from other sources.

6 Conclusion

We focus on the performance of automatic concatenated embedding model on semeval complex NER task, and draw the conclusion that with proper data processing methods, the model can learn excellent sequence tagging ability from low-context corpus and achieve outstanding performance on corresponsive part in test set, but cannot transfer such ability to QA and search query domains in test set. Meanwhile, document-level feature extraction and data augmenting by slotting and altering entities are

	Dutch			Spanish			German		
	LOWNER	MSQ-NER	ORCAS-NER	LOWNER	MSQ-NER	ORCAS-NER	LOWNER	MSQ-NER	ORCAS-NER
LOC	0.9222	0.7061	0.4618	0.8519	0.6700	0.4608	0.8230	0.6797	0.4051
PER	0.9432	0.8429	0.6993	0.9315	0.8210	0.6195	0.8837	0.8000	0.6663
PROD	0.7930	0.5456	0.6178	0.7170	0.4327	0.5721	0.7311	0.4307	0.5223
GRP	0.8716	0.3515	0.4148	0.8175	0.3771	0.4662	0.8085	0.4308	0.4462
CORP	0.8791	0.5062	0.5121	0.8567	0.4811	0.5008	0.7631	0.4937	0.4592
CW	0.8176	0.6466	0.4601	0.7578	0.5273	0.4084	0.7603	0.5122	0.4155
macro-F1	0.8711	0.5998	0.5276	0.8221	0.5515	0.5149	0.7950	0.5579	0.4858
overall macro-F1	0.7205			0.6966			0.6675		

Table 2: Cross-language Test Results

proved to be reproductably effective for common NER tasks.

References

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2018. Ms marco: A human generated machine reading comprehension dataset.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

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, Online. Association for Computational Linguistics.

Nick Craswell, Daniel Campos, Bhaskar Mitra, Emine Yilmaz, and Bodo Billerbeck. 2020. Orcas: 18 million clicked query-document pairs for analyzing search.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

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.

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, Berlin, Germany. Association for Computational Linguistics.

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.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

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, New Orleans, Louisiana. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Xinyu Wang, Yong Jiang, Nguyen Bach, Tao Wang, Zhongqiang Huang, Fei Huang, and Kewei Tu. 2021. Automated Concatenation of Embeddings for Structured Prediction. In *the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (**ACL-IJCNLP 2021**)*. Association for Computational Linguistics.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. LUKE: Deep contextualized entity representations with entity-aware self-attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6442–6454, Online. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *CoRR*, abs/1906.08237.