CIP at TREC 2021 Deep Learning Track

Xuanang Chen[1,2], Ben He[1,2], Le Sun[2], and Yingfei Sun[1]

[1] University of Chinese Academy of Sciences, Beijing, China
[2] Chinese Information Processing Laboratory, Institute of Software, Chinese Academy of Sciences, Beijing, China
chenxuanang19@mails.ucas.ac.cn
{benhe, yfsun}@ucas.ac.cn, sunle@iscas.ac.cn

Abstract. This paper describes the CIP participation in the TREC 2021 Deep Learning Track. Akin to our previous participation, we adopt the passage-level BERT re-ranker in the re-ranking subtask of the document ranking task. Besides, we utilize the MS MARCO v1 passage dataset and both the MS MARCO v2 passage and document datasets to generate hopefully sufficient training data, and BERT re-ranker is fine-tuned on these three kinds of training data one by one. Meanwhile, we adopt pairwise hinge loss rather than pointwise cross-entropy loss this year for model training, to boost the ranking effectiveness.

1 Introduction

The CIP participation in the TREC 2021 Deep Learning (DL) track focuses on the re-ranking subtask of the document ranking task. In the TREC 2021 DL track, a new, larger and cleaner corpus MS MARCO v2 is released, and it unifies the passage and document datasets. Based on our participation [1, 2] in the TREC 2019 and 2020 DL tracks, we remain to use the (passage-level) BERT re-ranker [3] as our main neural ranking model, but we adopt the pairwise hinge loss [4] rather than the cross-entropy loss for model training this year. Besides, we still use the MS MARCO v1 passage dataset as an important part of our training data as it provides large-scale ready-made training triples (a query, a positive passage, and a negative passage). Before being fine-tuned on the MS MARCO v2 document dataset, the BERT re-ranker has been trained on the passage ranking datasets (MS MARCO v1 and v2) as in [2]. And, *K-Max-AvgP* [2] still acts as an effective option for the document score aggregation under the MS MARCO v2 corpus.

2 Method

The passage-level BERT re-ranker has been widely-used in the document ranking task [5–7], and it produces a relevance score for every passage in a document and then use these scores of passages to aggregate the relevance score of this document, e.g. *MaxP* [5], *K-Max-AvgP* [2]. The input format of passage-level

BERT re-ranker is [CLS] [query] [SEP] [passage] [SEP], and the final pooled hidden vector of the [CLS] token is fed into a single layer feed-forward network to obtain the probability (score) of the passage being relevant. We use the pairwise hinge loss to fine-tune the BERT re-ranker, because it can produce better re-ranking effectiveness than the cross-entropy loss in our experiments. Akin to [2], we adopt a passage filter step after splitting documents and preserve top-ranked passages in a relevant document to construct the training data in the MS MARCO v2 document dataset. But we preserve all passages for the validation and test set, to completely judge the relevance of query-document pairs.

3 Experiments

3.1 Data

Training data. Although the new MS MARCO v2 corpus is released, we still use the passage dataset in the MS MARCO v1 corpus. Thus, in our experiments, totally three datasets are used for model training:

- **Passage v1:** In the MS MARCO v1 passage dataset, we sample about 4.2 M training triples from the provided training triple ids as our training data.
- **Passage v2:** In the MS MARCO v2 passage dataset, as a lack of ready-made training triples, we sample negative passages from BM25 top list (namely, the top-500 to top-1000 ranked passages) for positive passages in the qrels file, and finally we get about 1.4 M training triples.
- **Document v2:** In the MS MARCO v2 document dataset, we use the official provided top-100 file of train queries to generate the training data. Specifically, for a train query, the judged positive documents and the negative documents in top-100 list are split into up to 32 overlapping passages (100 whole words and 50 overlapping words), and the title (up to 16 words) and headings (up to 32 words) are added to the beginning of every passage if they are available. Then, for every passage in the positive documents, a negative passage is sampled from the passage pool of negative documents. Finally, a BERT re-ranker trained on Passage v1 training data is used to score all passages in the positive documents, and only five top-ranked passages are retained along with their sampled negative passages. Thus, here we finally get about 1.6 M training triples.

Validation data. When BERT re-ranker is trained with Passage v1, we use 43 and 54 test queries in TREC 2019 and 2020 DL track, and re-rank BM25 top-1000 candidates. When BERT re-ranker is trained with Passage v2, we use official Dev set (3,903 queries) and official top-100 list for model validation. When BERT re-ranker is trained with Document v2, we use the provided two official validation sets (namely, 43 and 45 test queries from TREC 2019 and 2020 DL track) and the BM25 top-100 documents for model selection. If two kinds validation sets are used, we use the average metric of them. Besides, all top documents are split into overlapping passages (100 whole words and 50

overlapping words), which are all retained without any filter step. As for the test set, the official top-100 candidates are re-ranked and are processed in the same way as the validation set.

Table 1. The summary of submitted runs. Aggregation refers the way to get the score of a document according to the scores of its split passages during inference, *MaxP* means the max score of passages and *4-Max-AvgP* means the average score of the top-4 ranked passages.

Run ID	Passage v1	Passage v2	Document v2	Aggregation
CIP_run1	✓		✓	*4-Max-AvgP*
CIP_run2	✓	✓	✓	*4-Max-AvgP*
CIP_run3	✓		✓	*MaxP*

Table 2. Evaluation results on TREC 2021 DL test queries in the document re-ranking subtask. Validation 1 (2) consists the 43 (45) test queries from TREC 2019 (2020) DL track, but under the MS MARCO v2 document corpus. The TREC 2021 DL test set consists of 57 queries. The best values are highlighted in boldface.

Run ID	Validation 1	Validation 2	TREC 2021 DL Test			
	NDCG@10	NDCG@10	MAP	NDCG@10	P@10	MRR
CIP_run1	**0.4003**	0.3675	0.2445	0.6755	0.8158	0.9505
CIP_run2	0.3914	**0.3732**	**0.2478**	**0.6783**	0.8140	0.9373
CIP_run3	0.3925	0.3718	0.2457	0.6668	**0.8175**	**0.9567**

3.2 Model

As for BERT re-ranker, we adopt the pre-trained BERT-Large model (bert-large-uncased) [8], and it is fine-tuned on above three kinds of training data as described in Section 3.1 with the training order of Passage v1, Passage v2 and Document v2. The query has up to 32 tokens, and the concatenation of query, passage, separator tokens has the maximum length of 256 tokens. For the document ranking task, we use *MaxP* or *K-Max-AvgP* aggregation method, and K is set as 4 in our submitted runs. However, for the passage ranking during model validation using Passage v1 and v2, there is no score aggregation. According to the used training data and the final aggregation way for test set in TREC 2021 DL track, we summarize our three submitted runs in Table 1.

We carry out our experiments on three TITAN RTX 24G GPUs with Mixed Precision Training [9]. We use Adam optimizer with a weight decay of 0.01, and the learning rate is set as 3e-6 for the whole fine-tuning procedure with batch size of 32. Besides, the BERT re-ranker is trained for 1 epoch using Passage v1 and Passage v2, and 2 epochs using Document v2. The margin in hinge ranking

loss is set as 1 in all training steps. We save a checkpoint per 5,000 training steps, and select the checkpoint according to the best NDCG@10 in Passage v1 and Document v2, and the best MRR@10 in Passage v2.

3.3 Results

The evaluation results of our submitted runs for document re-ranking subtask are shown in Table 2. For a more comprehensive comparison, we also present the validation results on test queries from both TREC 2019 and 2020 DL test sets. From the above results, we find that CIP_run2 outperforms other two runs on both Validation 2 and TREC 2021 DL test sets in terms of NDCG@10, and meanwhile CIP_run1 behaves better than other two runs on Validation 1 set in terms of NDCG@10. But CIP_run3 behaves better than other two runs on TREC 2021 DL test set in terms of P@10 and MRR. Thus, the submitted three runs seem comparable to each other.

4 Conclusions

In this paper, we describe the system based on BERT model for the document re-ranking subtask in TREC 2021 Deep Learning track. Our experiments demonstrate again that the BERT re-ranker fine-tuned on passage dataset can be transferred to the document ranking task effectively. Meanwhile, *K-Max-AvgP* aggregation method behaves better than *MaxP* when using the passage-level BERT re-ranker for document ranking task. In future work, we plan to investigate the uses of passage-document mapping released in the MS MARCO v2 corpus, like how to use the mapped passage information for document ranking.

References

1. Chen, X., Li, C., He, B., Sun, Y.: UCAS at TREC-2019 deep learning track. In: TREC. NIST Special Publication, vol. 1250. National Institute of Standards and Technology (NIST) (2019)
2. Chen, X., He, B., Sun, L., Sun, Y.: ICIP at TREC-2020 deep learning track. In: TREC. NIST Special Publication, vol. 1266. National Institute of Standards and Technology (NIST) (2020)
3. Nogueira, R., Cho, K.: Passage re-ranking with BERT. CoRR **abs/1901.04085** (2019)
4. Herbrich, R., Graepel, T., Obermayer, K., et al.: Large margin rank boundaries for ordinal regression. Advances in large margin classifiers **88**(2), 115–132 (2000)
5. Dai, Z., Callan, J.: Deeper text understanding for IR with contextual neural language modeling. In: SIGIR. pp. 985–988. ACM (2019)
6. Yang, W., Zhang, H., Lin, J.: Simple applications of BERT for ad hoc document retrieval. CoRR **abs/1903.10972** (2019)
7. Zhang, X., Yates, A., Lin, J.: Comparing score aggregation approaches for document retrieval with pretrained transformers. In: ECIR (2). Lecture Notes in Computer Science, vol. 12657, pp. 150–163. Springer (2021)

8. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: NAACL-HLT (1). pp. 4171–4186. Association for Computational Linguistics (2019)
9. Micikevicius, P., Narang, S., Alben, J., Diamos, G.F., Elsen, E., García, D., Ginsburg, B., Houston, M., Kuchaiev, O., Venkatesh, G., Wu, H.: Mixed precision training. In: ICLR (Poster). OpenReview.net (2018)