NTCIR 17 Conference: Proceedings of the 17th NTCIR Conference on Evaluation of Information Access Technologies, December 12-15, 2023, Tokyo, Japan
DOI: https://doi.org/10.20736/0002001293

ditlab at the NTCIR-17 Transfer Task

Yuuki Tachioka
Denso IT Laboratory
Japan
tachioka.yuki@core.d-itlab.co.jp

ABSTRACT

The ditlab team participated in the Transfer task composed of dense retrieval and dense reranking subtasks. We trained sentence-BERT by using a Japanese version of mMARCO dataset and commonly used for both subtasks. We compared three types of models that were trained according to three types of losses: softmax, triplet, multiple negatives ranking losses. The results show that the multiple negatives ranking loss was the best for both subtasks. In addition, system fusions significantly improved the performance especially for the retrieval task.

KEYWORDS

dense retrieval, dense reranking, sentence-BERT, jMARCO

TEAM NAME

ditlab

SUBTASKS

Dense First Stage Retrieval
Dense Reranking

1 INTRODUCTION

The ditlab team participated in the Dense First Stage Retrieval and Dense Reranking tasks of the NTCIR-17 Transfer task [3]. To improve the performance of sparse information retrieval such as BM25 [8], dense retrieval has attracted many attentions. We used sentence-BERT [7] for reranking instead of monoBERT [6] used in the baseline provided by the task organizers. We trained sentence-BERT by using a Japanese version of mMARCO dataset (jMARCO) [1]. Similarity calculation based on sentence-BERT was commonly used for both subtasks.

2 LOSSES FOR TRAINING

We trained sentence-BERT according to three types of losses because which type of losses is the most effective is unclear for this task.

2.1 Softmax loss

This is a loss that was used in the original sentence-BERT paper [7]. For using this loss, negative samples are required. Training data consist of N_p positive sentence pairs (a_i, p_i) and N_n negative sentence pairs $(a_{i'}, n_{i'})$, where i is the index of positive pairs $i = 1, 2, ..., N_p$ and i' is the index of negative pairs $i' = 1, 2, ..., N_n$. If training data do not contain negative samples, a sentence n_i is sampled from p_j $(j \neq i)$ to make negative samples $(a_i, n_i = p_j)$. A softmax classifier is added on top of two transformers to classify two sentences are relevant or not. For (a_i, p_i), the output is relevant and for $(a_{i'}, n_{i'})$ its output is irrelevant.

2.2 Triplet loss

This loss minimizes the distance between anchor and positive and maximizes the distance between anchor and negative, simultaneously [9]. For using this loss, positive sample and negative sample corresponding to the same anchor are required. For each anchor a_i, if training data do not contain negative samples, a sentence n_i is sampled from p_j $(j \neq i)$ to make negative samples $(a_i, n_i = p_j)$, before triplet loss

$$\mathcal{L} = \sum_i \left(||a_i - p_i|| - ||a_i - n_i|| + \theta \right). \quad (1)$$

is optimized where θ is the margin.

2.3 Multiple Negatives Ranking loss

This loss is suitable to train embeddings for retrieval setups and only uses anchor-positive sentence pairs (a_i, p_i) where all pairs other than positive pairs (a_i, p_j) $(j = 1, 2, ..., N_p$ and $j \neq i)$ are negative pairs. For each a_i, we can obtain one positive example (a_i, p_i) and $N_p - 1$ negative examples (a_i, p_j). This loss aims to minimize the negative log-likehood for softmax normalized scores by sampling a negative example randomly in every batch [2].

3 EXPERIMENTS

3.1 Experimental setup

Transfer dataset composed of training set, which is NTCIR-1 [4] and evaluation set, which is NTCIR-2 [5]. The train set consists of over 330K documents with 83 search topics, while the eval set consists of 735K documents with 49 topics. The baseline codes, which are based on monoBERT [6] reranker, were provided by the task organizers[1].

The original test collections provide relevance scores at three levels from 0 to 2. Score 0 means that queries and documents are irrelevant, score 1 means that they are partially relevant, and score 2 means that they are relevant. We convert these scores into binary labels (0: irrelevant and 1: relevant) by ignoring the difference between relevant and partially relevant. The baseline monoBERT reranker was fine-tuned[2] on the sequential classification task, i.e., softmax loss, starting from a publicly available Japanese BERT model[3]. The baseline ranker used the probabilities of the label 1 (relevant) as the new document scores, and reranked the top 1000 documents based on those scores. Our sentence-BERT models were fine-tuned on both Transfer training set and jMARCO based on the three types of losses starting from a publicly available Japanese sentence-BERT model[4] and the cosine similarity of embedding vectors was used as the new document scores.

[1] https://github.com/ntcirtransfer/transfer1
[2] This did not use jMARCO for training.
[3] cl-tohoku/bert-japanese3
[4] sonoisa/sentence-bert-base-ja-mean-tokens-v2

Table 1: Subtask1 results.

softmax	triplet	mnegrank	BM25	map	recip-rank	nDCG@5	nDCG@10	nDCG@15	nDCG@20	nDCG@1000
✓				0.0530	0.3250	0.2015	0.1893	0.1750	0.1627	0.2022
	✓			0.0453	0.3199	0.2017	0.1795	0.1620	0.1499	0.2005
		✓		**0.0797**	**0.4508**	**0.3019**	**0.2662**	**0.2412**	**0.2249**	**0.2771**
✓			✓	**0.1706**	**0.6570**	**0.4700**	**0.4284**	**0.4015**	**0.3826**	**0.4342**
✓	✓	✓		0.0797	0.4236	0.2771	0.2613	0.2406	0.2235	0.2684
	✓	✓	✓	0.1485	0.5764	0.4345	0.3775	0.3468	0.3221	0.4204
✓	✓	✓	✓	0.1318	0.5454	0.3878	0.3548	0.3246	0.3030	0.3995

Table 2: Subtask2 results.

softmax	triplet	mnegrank	map	recip-rank	nDCG@5	nDCG@10	nDCG@15	nDCG@20	nDCG@1000
✓			0.1520	**0.6975**	0.4855	0.4185	0.3744	0.3519	0.4079
	✓		0.1527	0.6366	0.4352	0.3948	0.3672	0.3437	0.4068
		✓	**0.1699**	0.6833	**0.5479**	**0.4707**	**0.4258**	**0.4071**	**0.4283**
✓	✓		0.1620	0.6574	0.4825	0.4437	0.3936	0.3664	0.4146
✓		✓	0.1718	**0.7338**	**0.5651**	**0.4887**	**0.4361**	**0.4022**	**0.4295**
	✓	✓	0.1738	0.6854	0.5188	0.4667	0.4237	0.3982	0.4275
✓	✓	✓	**0.1749**	0.7137	0.5321	0.4845	0.4271	0.3984	**0.4295**

3.2 Subtask1 (Dense retrieval)

Dense retrieval subtask is an ad-hoc retrieval task. The input was the title field of the original topic files and the output was the top (reranked) 1,000 document IDs. Table 1 shows the results of map, reciprocal rank, and nDCG@{5,10,15,20,1000}. The upper three rows show the result of a single model. The multiple negatives ranking loss was the best for all metrics. The lower four rows show the result of a system fusion. Three out of four fusions improved the performance, which demonstrates the effectiveness of system fusions and the results of BM25 are required to improve the performance. Among them[5], the fusion of softmax and triplet was the best.

3.3 Subtask2 (Dense reranking)

Dense reranking subtask is prepared for two-stage retrieval systems. For given the top 1,000 documents that were retrieved by BM25 model, dense models rerank them. The input was the query and the top 1,000 document IDs, and the output was the 100 reranked document IDs. Table 2 shows the results of map, reciprocal rank, and nDCG@{5,10,15,20,1000}. In this case, the performance differences between losses were much smaller than the retrieval task. This also shows the effectiveness of the multiple negatives ranking loss except reciprocal rank. The performance improvement by system fusions was smaller than the retrieval task but the system fusions improved the performance. In average, the fusion of softmax and multiple negatives ranking loss was the best.

4 CONCLUSION

We used sentence-BERT for similarity computation, which is applied to NTCIR-17 Transfer task. We compared three types of models that were trained according to three types of losses: softmax, triplet, multiple negatives ranking losses. The results show that the multiple negatives ranking loss was the best for both retrieval and reranking. In addition, system fusions significantly improved the performance especially for the retrieval task.

[5]We cannot validate all of the fusions because the maximum number of submitted runs is limited to 10.

REFERENCES

[1] Luiz Henrique Bonifacio, Israel Campiotti, Roberto de Alencar Lotufo, and Rodrigo Frassetto Nogueira. 2021. mMARCO: A Multilingual Version of MS MARCO Passage Ranking Dataset. *CoRR* abs/2108.13897 (2021). arXiv:2108.13897 https://arxiv.org/abs/2108.13897

[2] Matthew Henderson, Rami Al-Rfou, Brian Strope, Yun hsuan Sung, Laszlo Lukacs, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. 2017. Efficient Natural Language Response Suggestion for Smart Reply. arXiv:1705.00652 [cs.CL]

[3] Hideo Joho, Atsushi Keyaki, and Yuki Oba. 2023. Overview of the NTCIR-17 Transfer Task. In *Procceings of The 17th NTCIR Conference on Evaluation of Information Access Technologies.* https://doi.org/10.20736/0002001319

[4] Noriko Kando, Kazuko Kuriyama, Toshihiko Nozue, Koji Eguchi, Hiroyuki Kato, and Souichiro Hidaka. 1999. Overview of IR Tasks at the First NTCIR Workshop. In *Proceedings of the First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition.* 11–44.

[5] Noriko Kando, Kazuko Kuriyama, and Masaharu Yoshioka. 2001. Overview of Japanese and English Information Retrieval Tasks (JEIR) at the Second NTCIR Workshop. In *Proceedings of the Second NTCIR Workshop on Research in Chinese & Japanese Text Retrieval and Text Summarization.*

[6] Rodrigo Nogueira, Wei Yang, Kyunghyun Cho, and Jimmy Lin. 2019. Multi-Stage Document Ranking with BERT. arXiv:1910.14424 [cs.IR]

[7] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing.* Association for Computational Linguistics. https://arxiv.org/abs/1908.10084

[8] Stephen Robertson and Hugo Zaragoza. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Journal Foundations and Trends in Information Retrieval* 3, 4 (2009), 333–389.

[9] Kilian Q Weinberger and Lawrence K Saul. 2009. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research* 10 (2 2009), 207–244.