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/10.20736/0002001319

Overview of the NTCIR-17 Transfer Task

Hideo Joho
University of Tsukuba
hideo@slis.tsukuba.ac.jp

Atsushi Keyaki
Hitotsubashi University
a.keyaki@r.hit-u.ac.jp

Yuki Oba
University of Tsukuba
s2230178@u.tsukuba.ac.jp

ABSTRACT

This paper provides an overview of the NTCIR-17 Transfer task, a pilot task that aims to bring together researchers from Information Retrieval, Machine Learning, and Natural Language Processing to develop a suite of technology for transferring resources generated for one purpose to another in the context of dense retrieval on Japanese texts. Two subtasks were proposed for this round: the Dense First Stage Retrieval subtask and the Dense Reranking subtask. We received 29 runs for the First Stage Retrieval and 25 runs for the Reranking subtask from three research groups. The evaluation results of these runs are presented and discussed in this paper.

KEYWORDS

dense retrieval, ad-hoc retrieval, reranking, test collection

SUBTASKS

Dense First Stage Retrieval
Dense Reranking

1 INTRODUCTION

One of the traditional retrieval models is called a vector space model [22]. This intuitive model is designed to represent both a query and documents in a common multidimensional vector space where the weight of indexed terms is used as its value. The relevance of documents can then be determined by the similarity between the query and document vectors. A typical implementation of the vector space model used *sparse* vectors.

More recently, researchers have proposed representing words using a fixed size of *dense* vectors, which are now widely known as word embeddings (e.g., [23]). With appropriate training, word embeddings enable us to compute semantic relationships between words in ways that were difficult with sparse vectors. Since then, a number of approaches have been proposed for word embeddings and their applications. Applied to Information Retrieval, retrieval models designed to use some form of word embeddings are called *dense* retrieval models, while traditional models (e.g., vector space model, BM25) are now referred to as *sparse* retrieval models.

Although there are many promising aspects of dense retrieval models, building effective dense models is expensive. Therefore, building on existing models and datasets is a common and important approach to the development of dense retrieval models. The *Resource Transfer Based Dense Retrieval (Transfer)*[1] task was hosted at the 17th NTCIR [3] as a pilot task to address this technical challenge. The focus of the first round of the Transfer task was on Japanese documents since these are largely unexplored settings in the literature.

The rest of the paper is structured as follows: Section 2 presents the test collection prepared for the Transfer task. Section 3 introduces two subtasks. Section 4 provides a technical overview of the teams who participated in the Transfer task. Section 5 offers meta-analyses of the performance of submitted runs. Finally, Section 6 summarises the work and discusses future directions.

2 TEST COLLECTION

The Transfer task used the ad-hoc retrieval test collections developed at NTCIR-1 [1] and NTCIR-2 [2] as the training (`train`) set and evaluation (`eval`) set, respectively. We chose these test collections for multiple reasons. First, the domain of the document collections is academic publications, which differ from web pages where most recent language resources are generated. This allows participants to benchmark their technologies from domain transformation perspectives. Second, it is known that the relevance judgments of these test collections are much deeper and more thorough than more recent ones. This enables us to evaluate the performance of proposed techniques with a higher level of confidence than collections with shallow judgments.

The `train` set (i.e., NTCIR-1) consists of over 330K documents with 83 search topics, while the `eval` set (i.e., NTCIR-2) consists of 735K documents with 49 topics. The documents in the training set are the titles and abstracts of academic conference papers (1988-1997), while those in the evaluation set are the titles and abstracts of academic conference papers (1997-1999) and grant reports (1988-1997). Note that the document collection of the evaluation set includes the documents of the training set, although the topics and relevance judgments are independent of each other. See the overview papers [1, 2] for details on the development of these test collections.

The original test collections provide graded relevance scores with A as Relevant, B as Partially Relevant, and C as Not Relevant. In this task, we converted them into numeric scores of 2, 1, and 0, respectively, for training. We used a binary score for evaluation. No additional relevance assessments were performed on submitted runs.

The task organisers provided a GitHub repository[2] which included Jupyter notebooks to assist task participants in accessing these test collections using the `ir_datasets` library [4] in a local setting. Participants were instructed not to access the queries of the `eval` set until the development of their system was completed and frozen.

3 TASKS

NTCIR-17 Transfer task consisted of two subtasks: Dense First Stage Retrieval and Dense Reranking. Participants were allowed to submit up to ten runs for each of the subtasks.

[1] https://github.com/ntcirtransfer/transfer1/discussions/1

[2] https://github.com/ntcirtransfer/transfer1

3.1 Dense First Stage Retrieval subtask

This subtask is essentialy an ad-hoc retrieval task. Participants were asked to use the `title` field of the original topic files as the input both in the training and evaluation sets. A sample query was *feature dimensionality reduction* (`qid:0005` in `train` set). The output was the top 1,000 document IDs.

We used nDCG [5] (@1000) as the evaluation metric with a binary relevance judgement.

3.2 Dense Reranking subtask

This subtask is designed to develop second-stage retrieval techniques in a multi-stage retrieval framework. More specifically, we asked participants to rerank the top 1,000 documents that were retrieved by BM25 model in the same way as the first stage subtask. Therefore, the input was the query and the top 1,000 document IDs, and the output was the 100 reranked document IDs. Not all topics had 1,000 documents in the initial ranking.

The top 1000 documents were retrieved by PyTerrier (v 0.9.2) [6] where both queries and documents were tokenised by SudachiPy (v 0.5.4) [7] with its core dictionary and `SplitMode.A`.

We used nDCG@20 and MRR [8] as the evaluation metrics with a binary relevance judgement.

4 PARTICIATED SYSTEMS

This section describes an overview of participated systems. Please refer to individual participant papers for the details of their implementations.

4.1 Dense First Stage Retrieval subtask

4.1.1 ditlab. The team from ditlab [11] submitted ten runs for the first subtask, employing a sentence-BERT framework (`sonoisa/sentence-bert-base-ja-mean-tokens-v2`[3]) which was enhanced through the application of various loss functions including softmax loss [19], triplet loss [20], and multiple negatives ranking loss. Additionally, they presented runs that combined the results from the previously mentioned methods with those from BM25. For fine-tuning their models, they utilised not only the training set but also a Japanese version of the mMARCO dataset [17] (called *jMARCO* in this paper).

4.1.2 KANDUH. The KANDUH team [12] submitted six runs for the first subtask, by fine-tuning a DeBERTa model (`ku-nlp/deberta-v2-base-japanese`[4]), which is a decoding-enhanced BERT with disentangled attention [21]. Fine-tuning was done by the train set, jMARCO, or both, through a bi-encoder and cross-encoder methods. The ranking was based on the similarity between query and document embeddings. The team also examined the effectiveness of the embeddings provided by Azure and OpenAI.

4.1.3 KASYS. The KASYS team [13] submitted a total of nine runs for the first subtask, utilising three dense retrieval models: Contriever [14], ColBERT [15], and SPLADE [16]. Out of these nine runs, seven were based on various combinations of the dense retrieval models, fine-tuned using datasets including MS MARCO, mMARCO [17], and our own training set (NTCIR-1). The remaining two runs employed a fused ranking approach, integrating models such as ColBERT with BM25 and ColBERT with SPLADE.

4.1.4 Organiser. The organiser team presented six baseline runs utilising the ColBERT and DPR models [18]. The runs labeled `ColBERT_J_` and `DPR_J_` were implemented using a Japanese BERT model (`cl-tohoku/bert-large-japanese-v2`[5]), which was trained on a Japanese adaptation of the mMARCO dataset [17]. Conversely, the `ColBERT_X_25_` and `DPR_X_25_` runs employed a multilingual RoBERTa model (`xlm-roberta-large`[6]), trained with the original MS MARCO dataset. The final set of runs, `ColBERT_X_TT_` and `DPR_X_TT_`, also used the `xlm-roberta-large` model, but these were trained on the Japanese version of the mMARCO dataset [17].

For implementation, the team relied on ColBERT v1[7] and Tevatron[8] frameworks, respectively.

4.2 Dense Reranking subtask

4.2.1 ditlab. The ditlab team entered ten runs for the second subtask, utilising sentence-BERT models in a fashion akin to their approach for the first subtask. They computed the new document score by assessing the cosine similarity between the embedding vectors.

4.2.2 KANDUH. The KANDUH team submited six runs for the second subtask in a similar manner to their first subtask.

4.2.3 KASYS. The KASYS team submitted four runs for the second subtask by reranking the BM25 top 1000 documents using `KASYS-FIRST-1` to `KASYS-FIRST-5` as a reranker, resulting in the generation of runs `KASYS-SECOND-1` to `KASYS-SECOND-5`, respectively.

4.2.4 Organiser. The organising team offered a baseline run (`ORG-SECOND-1`) that utilized a monoBERT reranker [9, 10]. This reranker was developed through the fine-tuning of the Japanese BERT Model (`cl-tohoku/bert-japanese`), specifically for a sequential classification task. The model was trained with inputs structured as "`[CLS]`Query`[SEP]`Document`[SEP]`" and used relevance scores as labels. In this setup, graded labels were converted into binary scores (e.g., Scores of `2` and `1` became `1`, while Score `0` remained `0`). This input format was derived from the qrels of the `train` set, which contained over 260K samples.

For the evaluation set, the fine-tuned classifier was then used on inputs with the same structure to infer labels. The likelihood of a document being relevant (label `1`) was used as its new score. Utilizing this approach, the top 1000 documents for each topic were reranked based on these probabilities.

5 META ANALYSES

We have received 29 runs for Dense First Stage Retrieval and 25 runs for Dense Reranking. In addition, the organisers provided six runs and one run for the two subtasks, respectively. This section presents the meta analysis of these submitted runs.

[3]https://huggingface.co/sonoisa/sentence-bert-base-ja-mean-tokens-v2
[4]https://huggingface.co/ku-nlp/deberta-v2-base-japanese
[5]https://huggingface.co/cl-tohoku/cl-tohoku/bert-large-japanese-v2
[6]https://huggingface.co/xlm-roberta-large
[7]https://github.com/stanford-futuredata/ColBERT
[8]https://github.com/texttron/tevatron

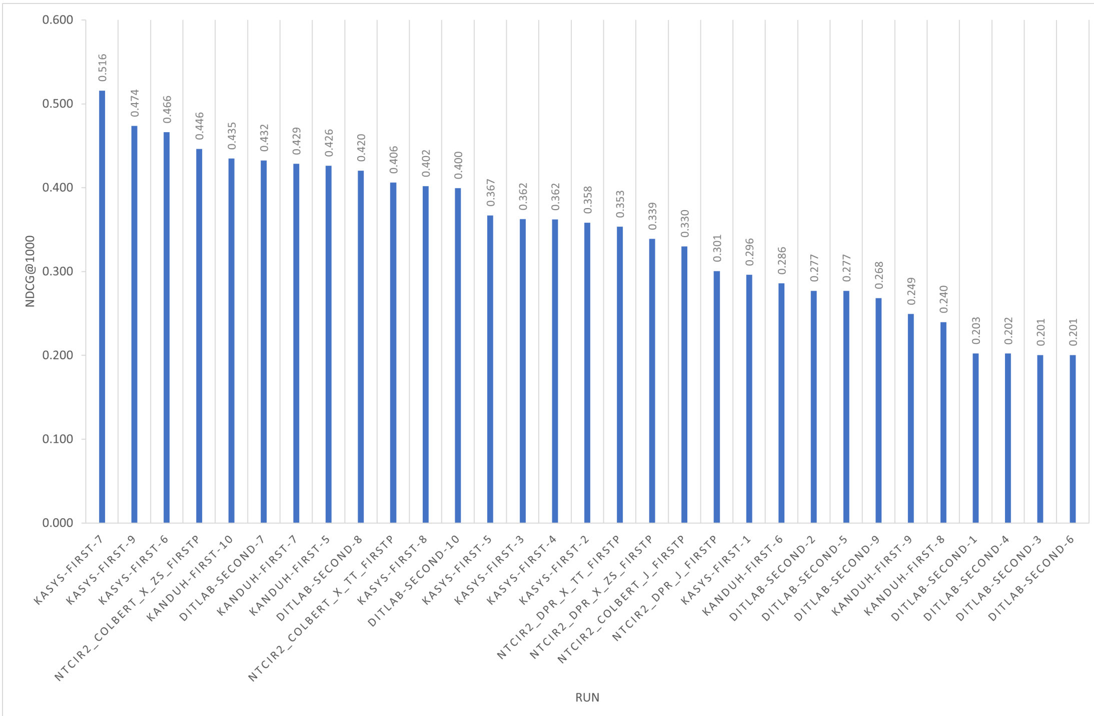

Figure 1: Performance of Dense First Stage Retrieval subtask (nDCG@1000)

5.1 Dense First Stage Retrieval subtask

The result of submitted runs for the first subtask is presented in Figure 1. As can be seen, the top five runs include all participating teams, suggesting that all teams developed compentitive systems for the dense first staga retrieval. Among those, however, `KASYS-FIRST-7` was better than other systems. `KASYS-FIRST-7` was a fusion of multilingual version of ColBERT models (ColBERT-X) and BM25. Although top performing systems tend to use BM25 rankings one way or another, ColBERT-X based models were generally performing well in our datasets.

5.2 Dense Reranking subtask

The result of submitted runs for the second subtask is presented in Figure 2 (MRR@100) and 3 (nDCG@20). As for MRR metric, 14 runs outperformed the organiser's baseline run, and the top six runs include all participating teams within a close score range (0.7117 to .7449). The best performing run was `KASYS-SECOND-3` which was based on a Contriever fine-tuned on the original MS MARCO and the train set with random negatives. `DITLAB-SECOND-9` was also performing well.

As for nDCT@20, 12 runs outperformed the baseline run, and the top five runs include all participating teams. The best performing run was `KANDUH-SECOND-7` which was a cross-encoder version of the DeBERT model trained on jMARCO followed by the train set. `KASYS-SECOND-5`, a Contriever fine-tuned on MS MARCO and the train set with hard negatives, peformed well too.

These results suggests an advantage of contrastive learning methods adapted by Contriever in the precision-oriented metrics. However, the DeBERT model can perform well with appropriate data for fine-turning.

6 CONCLUSIONS

This marked the first NTCIR Transfer Task, with participants addressing the challenges of adapting existing resources to the Japanese language, covering academic texts and web content originally in English. This initial phase establishes a baseline for future research on dense retrieval models in these particular scenarios.

Upcoming research will closely examine the effects of transitioning between languages and domains on performance. The quality of translation and its varying impact on different search topics also presents a significant area for investigation. The ultimate objective is to design and assess a new dense retrieval model that is shaped by the outcomes of this task.

ACKNOWLEDGMENTS

The authors thank task participants for their contributions. We also thank the NTCIR project at NII for providing the platform of our research.

REFERENCES

[1] 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, August 30 - September 1, 1999, pp.11-44.

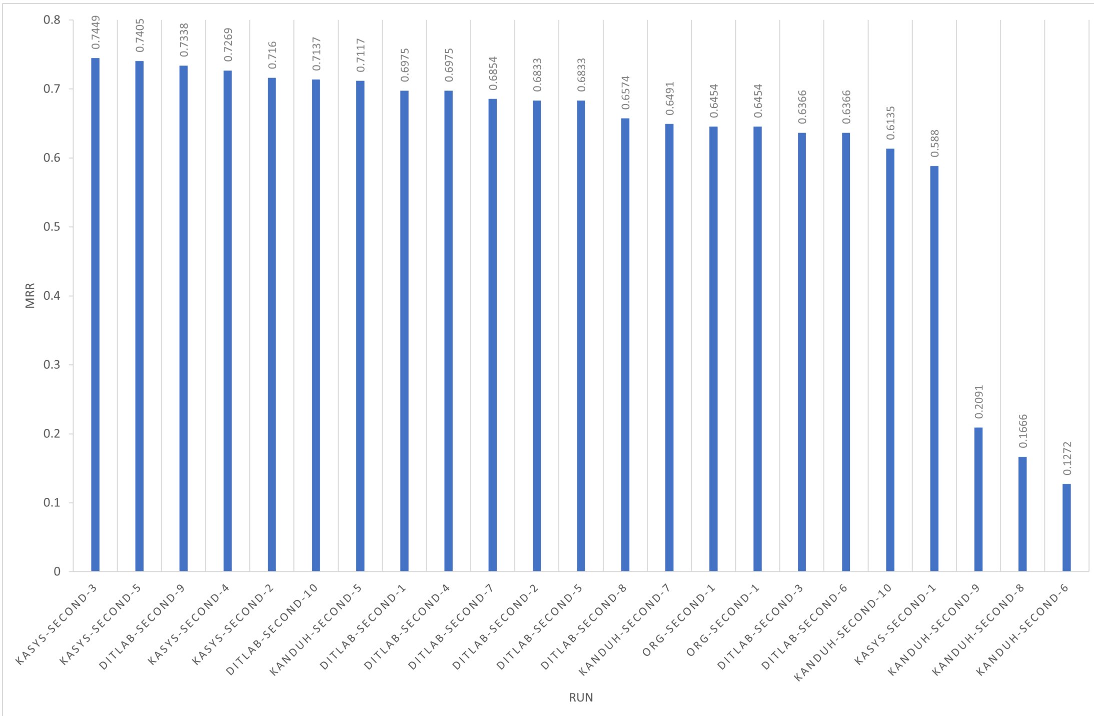

Figure 2: Performance of Dense Reranking subtask (MRR@100)

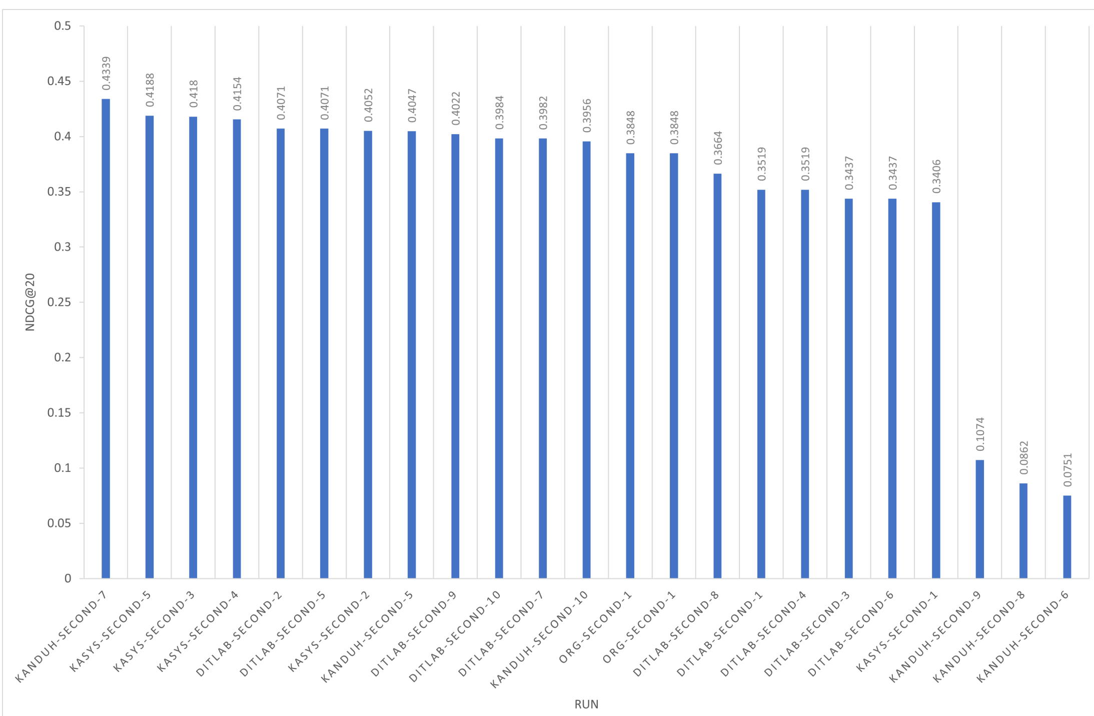

Figure 3: Performance of Dense Reranking subtask (nDCG@20)

[2] Noriko Kando, Kazuko Kuriyama, 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, May 2000- March 2001.

[3] Takehiro Yamamoto and Zhicheng Dou. (2023). Overview of the NTCIR-17. In Proceedings of the 17th NTCIR Conference on Evaluation of Information Access Technologies, December 12-15, 2023, Tokyo, Japan. https://doi.org/10.20736/0002001332

[4] Sean MacAvaney, Andrew Yates, Sergey Feldman, Doug Downey, Arman Cohan, and Nazli Goharian. (2021). Simplified Data Wrangling with ir_datasets. In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21). Association for Computing Machinery, New York, NY, USA, 2429–2436. https://doi.org/10.1145/3404835.3463254

[5] Kalervo Järvelin and Jaana Kekäläinen. (2002). Cumulated gain-based evaluation of IR techniques. ACM Trans. Inf. Syst. 20, 4 (October 2002), 422–446. https://doi.org/10.1145/582415.582418

[6] Craig Macdonald and Nicola Tonellotto (2020) Declarative Experimentation inInformation Retrieval using PyTerrier. In: Proceedings of ICTIR 2020.

[7] Kazuma Takaoka and Sorami Hisamoto and Noriko Kawahara and Miho Sakamoto and Yoshitaka Uchida and Yuji Matsumoto. (2018) Sudachi: a Japanese Tokenizer for Business. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).

[8] Craswell, N. (2009). Mean Reciprocal Rank. In: LIU, L., ÖZSU, M.T. (eds) Encyclopedia of Database Systems. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-39940-9_488

[9] Rodrigo Nogueira, Wei Yang, Kyunghyun Cho, Jimmy Lin (2019) Multi-Stage Document Ranking with BERT. arXiv:1910.14424. https://doi.org/10.48550/arXiv.1910.14424

[10] Ronak Pradeep, Rodrigo Nogueira, Jimmy Lin. (2021). The Expando-Mono-Duo Design Pattern for Text Ranking with Pretrained Sequence-to-Sequence Models. arXiv:2101.05667. https://doi.org/10.48550/arXiv.2101.05667

[11] Yuuki Tachioka. (2023). ditlab at the NTCIR-17 Transfer Task. In Proceedings of the 17th NTCIR Conference on Evaluation of Information Access Technologies, December 12-15, 2023, Tokyo, Japan. https://doi.org/10.20736/0002001293

[12] Tomoya Hashiguchi, Ryota Mibayashi, Huu-Long Pham, Wakana Kuwata, Yuka Kawada, Yuya Tsuda, Takehiro Yamamoto and Hiroaki Ohshima. (2023). KANDUH at the NTCIR-17 Transfer Task. In Proceedings of the 17th NTCIR Conference on Evaluation of Information Access Technologies, December 12-15, 2023, Tokyo, Japan. https://doi.org/10.20736/0002001330

[13] Kenya Abe, Kota Usuha and Makoto P. Kato. (2023). KASYS at the NTCIR-17 Transfer Task. In Proceedings of the 17th NTCIR Conference on Evaluation of Information Access Technologies, December 12-15, 2023, Tokyo, Japan. https://doi.org/10.20736/0002001331

[14] Gautier Izacard and Mathilde Caron and Lucas Hosseini and Sebastian Riedel and Piotr Bojanowski and Armand Joulin and Edouard Grave (2022) Unsupervised Dense Information Retrieval with Contrastive Learning, arXiv.2112.09118.

[15] Omar Khattab and Matei Zaharia. (2020). ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. In Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20). Association for Computing Machinery, New York, NY, USA, 39–48. https://doi.org/10.1145/3397271.3401075

[16] Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. (2021). SPLADE: Sparse Lexical and Expansion Model for First Stage Ranking. In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21). Association for Computing Machinery, New York, NY, USA, 2288–2292. https://doi.org/10.1145/3404835.3463098

[17] Luiz Henrique Bonifacio, Israel Campiotti, Roberto de Alencar Lotufo, and Rodrigo Frassetto Nogueira. (2021). mMARCO: A Multilingual Version of MS MARCO Passage Ranking Dataset. arXiv:2108.13897

[18] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. (2020). Dense Passage Retrieval for Open-Domain Question Answering. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 6769–6781, Online. Association for Computational Linguistics.

[19] 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 and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

[20] 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.

[21] Pengcheng He, Xiaodong Liu, Jianfeng Gao, Weizhu Chen. (2021). DEBERTA: DECODING-ENHANCED BERT WITH DISENTANGLED ATTENTION. In Proceedings of the Ninth International Conference on Learning Representations. https://openreview.net/forum?id=XPZIaotutsD

[22] G. Salton, A. Wong, and C. S. Yang. (1975). A vector space model for automatic indexing. Communications of the ACM, 18, 11 (Nov. 1975), 613–620. https://doi.org/10.1145/361219.361220

[23] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. (2013). Distributed representations of words and phrases and their compositionality. In Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2 (NIPS'13). Curran Associates Inc., Red Hook, NY, USA, 3111–3119.