York University at TREC 2022: Deep Learning Track

Yizheng Huang and Jimmy Huang
Information Retrieval and Knowledge Management Research Lab
York University, Toronto, Canada
{hyz, jhuang}@yorku.ca

Abstract

The present study outlines the involvement of the YorkU group in the TREC 2022 Deep Learning Track. This year, the investigation into the fusion of BM25 and the deep learning model, which was initiated in the previous year, is pursued further. The findings from last year's experiments indicate that while the deep learning model was superior for most queries, BM25 demonstrated better performance for particular queries. In our contribution, the queries were classified: BM25 was utilized directly as the final ranking result for queries suited to it, whereas the results of the deep learning model were employed for queries incompatible with BM25. The experimental results indicate that this integrated approach yields improved results.

Keywords

BM25, Deep Learning, Passage Ranking, Relation Extraction

1 Introduction

Recent developments in deep learning-based dense passage retrieval [1] have exhibited remarkable superiority over traditional retrieval techniques such as TF-IDF and BM25 [2] on established question-answering and information retrieval (IR) datasets. These dense models are trained using annotated datasets and several of them have been demonstrated to outperform BM25 with as few as 1000 supervised examples trained using BERT [3, 4] as a pre-trained model with fine-tuning, indicating high potential for practical applications and suggesting a possibility of substitution for traditional retrieval methods. In a prior study [5], we posited that the dense retrieval model has not yet attained sufficient capability to entirely replace traditional methods. And we explored some of the fundamental limitations that still afflict dense retrievers.

The present study builds upon the research from last year, with a focus on investigating methods for the enhanced integration of BM25 and deep learning models. In the prior study, it was discovered that while the deep learning model outperformed the traditional retrieval model for most queries, BM25 demonstrated superior performance for specific queries, such as those involving proper names. As demonstrated in Table 1 of the prior study, BM25 was observed to outperform the BERT model for queries containing special entities, such as names of individuals, places, organizations, and specialized terms. These entities, which are frequently encountered in real-life queries, exhibit similar characteristics, such as a limited number of synonyms, specificity to a particular entity, and a relatively fixed position in the query. These features make BM25 particularly well-suited for handling these queries, while the deep learning model, which is based on word embeddings, may treat similar words as synonyms and therefore retrieve irrelevant passages. Hence, this study endeavors to investigate the possibility of combining BM25 and the deep learning model, such that BM25 results are utilized for specific queries and the results of the deep learning model are employed for other queries.

2 Related Work

Deep Learning in IR Prior to the emergence of dense retrieval models, traditional retrieval techniques, such as TF-IDF and BM25, were widely utilized in information retrieval systems [6, 7, 8, 9, 10]. These methods are characterized by relying on mathematical models to describe the retrieval process and using weighted term matching between queries and passages to determine similarity. Unlike dense models,

Table 1: BM25 was observed to outperform the BERT model for queries containing special entities from YorkU's results of the TREC 2021 Deep Learning Track with the evaluation of NDCG@10.

Query ID	Question	BM25	YorkU21a
168329	does light intensity or concentration of carbon dioxide have a higher rate of photosynthesis	0.7179	0.6937
190623	for what is david w. taylor known	0.5812	0.2962
508292	symptoms of neuroma	0.5000	0.3708
1128632	is levothyroxine likely to cause weight loss or weight gain	0.9009	0.8582

traditional retrieval models do not require training on labeled datasets. Although traditional retrieval models excel at lexical matching, they are deficient in capturing synonymy and semantic relationships.

In contrast, dense models utilize pre-trained language models, such as BERT, to compute similarity through embeddings learned from labeled datasets. These models typically employ two encoders - one for the query and one for the passage - and fine-tune the downstream tasks based on pre-trained models. Both queries and passages are represented as word embeddings, and the top passages with the highest similarity scores to the query are returned and ranked accordingly.

Despite their remarkable performance within their training domain, the effectiveness of dense retrievers in generalizing to new domain remains a challenge. Thakur et al. [11] introduced a zero-shot benchmark named BEIR, which demonstrated that dense retrieval models did not perform as well as BM25 in the majority of their datasets. Lewis et al. [12] discovered that the model had a tendency to memorize the training data, resulting from the substantial overlap between the training and test sets and the proclivity of deep learning models to overfit the training data for optimal performance. Chen et al. [13] devised the AmbER to assess entity disambiguation proficiency, and they found that the models performed significantly worse on rare entities than on common entities. Sciavolino's findings [14] were in line with this, indicating that the performance of dense retrieval models requires improvement in terms of generalizability, particularly when it comes to rare entities. Their study concluded that the integration of BM25 and deep learning models is a viable option.

Relationship Extraction Relationship extraction techniques can be broadly categorized into traditional and neural network-based methods. Traditional methods encompass manual, unsupervised, and supervised approaches. The manual method [15] involves incorporating linguistic knowledge to construct a linguistic model based on words, syntax, or semantics. The model is then utilized to match preprocessed sentences and establish the corresponding linguistic relation. In contrast, the unsupervised approach [16] seeks to determine semantic relations by extracting entities and their contexts and grouping them based on similarities in contextual information. The supervised approach views relationship extraction as a classification problem and employs data-based features [17], derived from entity context semantics or syntax, to train classifiers for each related entity. In testing, the classifier can recognize the relation of a new entity if its features are similar. However, this method requires high-quality features. An alternative approach, proposed by Mooney et al. [18], is to design a kernel function for classification.

Recent advancements in relation extraction techniques have led to the widespread adoption of neural network-based methods. Liu et al. [19] were the pioneering researchers to apply the Convolutional Neural Network (CNN) model to relation extraction, transforming sentences into word embeddings through the use of a synonym dictionary and other lexical features. The output of the model is the relationship classification probability between entities. Xu et al. [20] sought to enhance the semantic aspect of the method by using the shortest dependency path (SDP) and incorporating the central part of the sentence as input, while removing irrelevant words for improved accuracy. Zhang et al. [21] posited that relation extraction necessitates comprehensive and continuous information from all words in the sentence and utilized bi-directional long short-term memory networks (BLSTM) for sentence-level representation and feature improvement. Zhou et al. [22] proposed the use of an attention mechanism for BLSTM to extract essential features from the data, without relying on external resources. However, for a complete representation of the sentence, additional knowledge or resources, such as knowledge graphs, may be required. Ji et al. [23] introduced the concept of relational vectors from knowledge graphs to represent the features of relationships. Qin et al. [24] focused on improving the quality of the dataset by using reinforcement learning to filter mislabeled sentences and reduce data noise, thereby forming a new high-confidence training dataset that enhances the performance of the trained model.

3 Our Methods

Our approach is a simplistic yet effective technique in conjunction with relation extraction to determine the suitability of a query for retrieval through BM25. The ranking produced by the deep learning model is then replaced with the ranking generated by BM25 for these queries, thereby improving retrieval performance. Details will be presented in the following Section 3.1 and 3.2. During this year's participation, we presented two outcomes: the dense retrieval YorkU22a and the non-dense retrieval YorkU22b.

Table 2: Different performance of BM25 based on different queries from YorkU's results of the TREC 2021 Deep Learning Track with the evaluation of NDCG@10.

Query ID	Question	BM25	YorkU21a
1129560	accounting definition of building improvements	0.3274	0.6423
168329	does light intensity or concentration of carbon dioxide have a higher rate of photosynthesis	0.7179	0.6937
225975	how does my baby get submitted for medicaid after birth	0.0000	0.4885

3.1 Relation Extraction

The results of last year's experiments indicate that BM25 is not as effective for short queries as it is for queries containing special entities. As demonstrated in Table 2, the query "1129560" ("accounting definition of building improvements") is a short query, but BM25's performance is not good. Conversely, the query "168329" ("does light intensity or concentration of carbon dioxide have a higher rate of photosynthesis") is a longer, semantically complex query that would typically be better suited to deep learning models, however, BM25 yields superior results. Despite the lack of semantic relationship between the words in the query, BM25 is still able to identify relevant documents by the presence of several special entities, such as "light intensity", "carbon dioxide", and "photosynthesis". Particularly, the entity "photosynthesis" is a term with limited synonyms, and BM25 tends to perform well in the presence of such entities.

On the other hand, the query "225975" ("how does my baby get submitted for medicaid after birth") yields an 0.00 NDCG score in the top-10 hits for BM25 due to the absence of special entities. Therefore, we refer that entities such as "photosynthesis", "david w. taylor", "neuroma", and "levothyroxine" as strong entities, and BM25 tends to perform better in the query with strong entities.

Based on this analysis, the goal of relation extraction is set as the identification of strong entities. These entities are classified, including names of individuals, locations, organizations, and specialized terms such as medical terms. In other words, the strong entities are the core keywords in the sentence.

We adopt the YAKE model [25], which is a lightweight unsupervised automatic keyword extraction method that relies on extracted statistical text features to select the most relevant keywords in the text. Also, we utilize Gensim and Rake-NLTK (Rapid Automatic Keyword Extraction algorithm with the NLTK toolkit) to identify strong entites. If their results match with the YAKE model, the extracted entites are proved to be correct. Otherwise, the entity is discarded.

3.2 Dense Retrieval

In this year's full-passage ranking task, we submitted two runs: YorkU22a and YorkU22b. The YorkU22b is a non-dense retrieval and serves as the first stage of the dense retrieval, YorkU22a. Since this year's work builds on last year's efforts, we adopt the same dense retrieval architecture. The Sentence-BERT (SBERT) [26] model was used as the pre-training model, and the Bi-Encoder was applied to generate the first ranking result of the dense retrieval, i.e., YorkU22b. After performing relation extraction on the query, we identify queries containing strong entities, and compute their BM25 ranking results through Anserini [27]. The results of the strong entity queries in YorkU22b were then substituted with their corresponding BM25 results, forming a new first-stage ranking. Finally, YorkU22a was obtained by using the Cross-Encoder, with the scope of documents limited to those retrieved in the previous ranking.

4 Results

Our experiment runs are denoted as: Anserini_BM25, YorkU22a, YorkU22a−BM25, YorkU22a+BM25, YorkU22b, YorkU22b+BM25. Table 3 presents the detailed descriptions of these runs are as follows.

Table 3: Runs Description

Runs	**Description**
Anserini_BM25	The BM25 baseline.
YorkU22b	The first ranking obtained from SBERT.
YorkU22b+BM25	The first ranking combined with BM25.
YorkU22a	The re-ranking based on the first ranking combined with BM25.
YorkU22a−BM25	The re-ranking based on the first ranking without combining BM25.
YorkU22a+BM25	The optimal result combines re-ranking and BM25.

Table 4: Results with Different Runs.

Runs	MAP@100	P@10	NDCG@10	NDCG@100
Anserini_BM25	0.0325	0.1421	0.2692	0.2133
YorkU22b	0.1130	0.3947	0.5076	0.3408
YorkU22b+BM25	0.1162	0.4066	0.5181	0.3471
YorkU22a−BM25	0.1989	0.5288	0.6003	0.4587
YorkU22a	0.2003	0.5316	0.6089	0.4610
YorkU22a+BM25	0.2011	0.5303	0.6144	0.4628

The experimental results of all runs are presented in Table 4. It can be observed that the traditional BM25 model, serving as the baseline, exhibits significantly lower performance compared to the other models. However, in the first-stage ranking, replacing the results of the deep learning model for the strong entity queries with their BM25 results leads to improved performance. The re-ranking results of the dense retrieval that underwent this method also showed improvement. Furthermore, the possibility of combining the re-ranked results with BM25 was explored, as depicted in Table 5. It is noteworthy that even after re-ranking, YorkU22a still demonstrates inferior results compared to BM25, and directly replacing them with BM25 results improves retrieval performance. This is reflected in the result of the experiment, represented by YorkU22a+BM25. This suggests that the deep learning model fails to fully capture all the information obtained from BM25 in the first ranking, and there remains room for optimization in the results after replacing the queries with strong entities.

Table 5: The queries that BM25 outperform YorkU22a at the evaluation of NDCG@10.

Query ID	Question	BM25	YorkU22a
2003157	how to cook frozen ham steak on nuwave oven	0.3089	0.2415
2006211	what does auslan interpreted performance mean	0.4203	0.2393

5 Conclusion and Future Work

It is a widely acknowledged fact that deep learning-based information retrieval models outperform the traditional BM25. However, BM25 remains a prevalent information retrieval algorithm due to its mathematical explanation of the retrieval process to a certain extent, which is lacking in deep learning models. On the other hand, training deep learning models is resource-intensive and requires large amounts of training data, often unavailable in fields such as biology and medicine. It is feasible to utilize a small-scale pre-trained model combined with traditional retrieval methods to improve performance. The approach proposed in this paper attempts to address this issue by employing relation extraction to identify strong entity queries and combining the results of the deep learning model with the BM25 results of such queries to achieve better results. In the future, we consider using other sentence features, like positional context, to improve retrieval performance.

Acknowledgement

This research is supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada and the York Research Chairs (YRC) program.

References

[1] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 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, November 2020. Association for Computational Linguistics.

[2] Stephen E. Robertson, Steve Walker, Micheline Hancock-Beaulieu, Mike Gatford, and A. Payne. Okapi at TREC-4. In Donna K. Harman, editor, *Proceedings of The Fourth Text REtrieval Conference, TREC 1995, Gaithersburg, Maryland, USA, November 1-3, 1995*, volume 500-236 of *NIST Special Publication*. National Institute of Standards and Technology (NIST), 1995.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019.

[4] Md. Tahmid Rahman Laskar, Jimmy Xiangji Huang, and Enamul Hoque. Contextualized embeddings based transformer encoder for sentence similarity modeling in answer selection task. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 5505–5514. European Language Resources Association, 2020.

[5] Yizheng Huang and Jimmy Huang. York University at TREC 2021: Deep Learning Track. In *Proceedings of The 30th Text REtrieval Conference, TREC 2021*, 2021.

[6] Ben He, Jimmy Xiangji Huang, and Xiaofeng Zhou. Modeling term proximity for probabilistic information retrieval models. *Inf. Sci.*, 181(14):3017–3031, 2011.

[7] Xiangji Huang and Qinmin Hu. A bayesian learning approach to promoting diversity in ranking for biomedical information retrieval. In James Allan, Javed A. Aslam, Mark Sanderson, ChengXiang Zhai, and Justin Zobel, editors, *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2009, Boston, MA, USA, July 19-23, 2009*, pages 307–314. ACM, 2009.

[8] Xiangji Huang, Ming Zhong, and Luo Si. York University at TREC 2005: Genomics track. In Ellen M. Voorhees and Lori P. Buckland, editors, *Proceedings of the Fourteenth Text REtrieval Conference, TREC 2005, Gaithersburg, Maryland, USA, November 15-18, 2005*, volume 500-266 of *NIST Special Publication*. National Institute of Standards and Technology (NIST), 2005.

[9] Xiangji Huang, Fuchun Peng, Dale Schuurmans, Nick Cercone, and Stephen E. Robertson. Applying machine learning to text segmentation for information retrieval. *Inf. Retr.*, 6(3-4):333–362, 2003.

[10] Jiashu Zhao, Jimmy Xiangji Huang, and Ben He. CRTER: using cross terms to enhance probabilistic information retrieval. In Wei-Ying Ma, Jian-Yun Nie, Ricardo Baeza-Yates, Tat-Seng Chua, and W. Bruce Croft, editors, *Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2011, Beijing, China, July 25-29, 2011*, pages 155–164. ACM, 2011.

[11] Nandan Thakur, Nils Reimers, Andreas Ruckl'e, Abhishek Srivastava, and Iryna Gurevych. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models. *Neurips Datasets And Benchmarks*, 2021.

[12] Patrick Lewis, Pontus Stenetorp, and Sebastian Riedel. Question and answer test-train overlap in open-domain question answering datasets. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1000–1008, Online, April 2021. Association for Computational Linguistics.

[13] Anthony Chen, Pallavi Gudipati, Shayne Longpre, Xiao Ling, and Sameer Singh. Evaluating entity disambiguation and the role of popularity in retrieval-based NLP. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4472–4485, Online, August 2021. Association for Computational Linguistics.

[14] Christopher Sciavolino, Zexuan Zhong, Jinhyuk Lee, and Danqi Chen. Simple entity-centric questions challenge dense retrievers. In *2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021*, pages 6138–6148. Association for Computational Linguistics (ACL), 2021.

[15] Marti A Hearst. Automatic acquisition of hyponyms from large text corpora. In *COLING 1992 Volume 2: The 14th International Conference on Computational Linguistics*, 1992.

[16] Takaaki Hasegawa, Satoshi Sekine, and Ralph Grishman. Discovering relations among named entities from large corpora. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 415–422, 2004.

[17] Bryan Rink and Sanda Harabagiu. Utd: Classifying semantic relations by combining lexical and semantic resources. In *Proceedings of the 5th international workshop on semantic evaluation*, pages 256–259, 2010.

[18] Raymond Mooney and Razvan Bunescu. Subsequence kernels for relation extraction. *Advances in neural information processing systems*, 18, 2005.

[19] ChunYang Liu, WenBo Sun, WenHan Chao, and Wanxiang Che. Convolution neural network for relation extraction. In *Proceedings of the Advanced Data Mining and Applications: 9th International Conference, ADMA 2013, Hangzhou, China, December 14-16, 2013, Proceedings, Part II 9*, pages 231–242. Springer, 2013.

[20] Kun Xu, Yansong Feng, Songfang Huang, and Dongyan Zhao. Semantic relation classification via convolutional neural networks with simple negative sampling. *Conference On Empirical Methods In Natural Language Processing*, 2015.

[21] Shu Zhang, Dequan Zheng, Xinchen Hu, and Ming Yang. Bidirectional long short-term memory networks for relation classification. In *Proceedings of the 29th Pacific Asia conference on language, information and computation*, pages 73–78, 2015.

[22] Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short papers)*, pages 207–212, 2016.

[23] Guoliang Ji, Kang Liu, Shizhu He, and Jun Zhao. Distant supervision for relation extraction with sentence-level attention and entity descriptions. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.

[24] Pengda Qin, Weiran Xu, and William Yang Wang. Robust distant supervision relation extraction via deep reinforcement learning. *Annual Meeting Of The Association For Computational Linguistics*, 2018.

[25] Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Jorge, Célia Nunes, and Adam Jatowt. Yake! keyword extraction from single documents using multiple local features. *Information Sciences*, 509:257–289, 2020.

[26] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *Conference On Empirical Methods In Natural Language Processing*, 2019.

[27] Peilin Yang, Hui Fang, and Jimmy Lin. Anserini: Enabling the use of lucene for information retrieval research. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1253–1256, 2017.