Pretrained Transformers for Text Ranking: BERT and Beyond

Andrew Yates,[1] **Rodrigo Nogueira,**[2] and **Jimmy Lin**[2]
[1] Max Planck Institute for Informatics
[2] David R. Cheriton School of Computer Science, University of Waterloo

1 Overview

The goal of text ranking is to generate an ordered list of texts retrieved from a corpus in response to a query for a particular task. Although the most common formulation of text ranking is search, instances of the task can also be found in many text processing applications. This tutorial provides an overview of text ranking with neural network architectures known as transformers, of which BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019) is the best-known example. These models produce high quality results across many domains, tasks, and settings.

This tutorial, which is based on the preprint (Lin et al., 2020a) of a forthcoming book to be published by Morgan and & Claypool under the Synthesis Lectures on Human Language Technologies series, provides an overview of existing work as a single point of entry for practitioners who wish to deploy transformers for text ranking in real-world applications and researchers who wish to pursue work in this area. We cover a wide range of techniques, grouped into two categories: transformer models that perform reranking in multi-stage ranking architectures and learned dense representations that perform ranking directly.

2 Multi-Stage Ranking Architectures

The most straightforward application of transformers to text ranking is to convert the task into a text classification problem, and then sort the texts to be ranked based on the probability that each item belongs to the relevant class. The first application of BERT to text ranking, by Nogueira and Cho (2019), used BERT in exactly this manner. This *relevance classification* approach is usually deployed in a module that reranks candidate texts from an initial keyword search engine.

One key limitation of BERT is its inability to handle long input sequences and hence difficulty in ranking texts beyond a certain length (e.g., "full-length" documents such as news articles). This limitation is addressed by a number of models (Nogueira and Cho, 2019; Akkalyoncu Yilmaz et al., 2019; Dai and Callan, 2019b; MacAvaney et al., 2019; Wu et al., 2020; Li et al., 2020), and a simple retrieve-then-rerank approach can be elaborated into a multi-stage architecture with reranker pipelines (Nogueira et al., 2019a; Matsubara et al., 2020; Soldaini and Moschitti, 2020) that balance effectiveness and efficiency. On top of multi-stage ranking architectures, researchers have proposed additional innovations, including query expansion (Zheng et al., 2020), document expansion (Nogueira et al., 2019b; Nogueira and Lin, 2019) and term importance prediction (Dai and Callan, 2019a, 2020).

A natural question that arises is, "What's beyond BERT?" We describe efforts to build ranking models that are faster (i.e., lower inference latency), that are better (i.e., higher ranking effectiveness), or that manifest interesting tradeoffs between effectiveness and efficiency. These include ranking models that leverage BERT variants (Li et al., 2020), exploit knowledge distillation to train more compact student models (Gao et al., 2020a), and other transformer architectures, including ground-up redesign efforts (Hofstätter et al., 2020b; Mitra et al., 2020) and adapting pretrained sequence-to-sequence models (Nogueira et al., 2020; dos Santos et al., 2020). These discussions set up a natural transition to ranking based on dense representations, the other main category of approaches we cover.

3 Learned Dense Representations

Arguably, the single biggest benefit brought about by modern deep learning techniques to text ranking is the move away from sparse signals, mostly

Proceedings of NAACL-HLT 2021: Tutorials, pages 1–4
June 6–11, 2021. ©2021 Association for Computational Linguistics

limited to exact matches, to dense representations that are able to capture semantic matches to better model relevance. The potential of continuous dense representations for natural language analysis was first demonstrated nearly a decade ago with word embeddings on word analogy tasks (Mikolov et al., 2013). As soon as researchers tried to build representations for any larger spans of text: phrases, sentences, paragraphs, and documents, the same issues that arise in text ranking come into focus. In fact, ranking with dense representations predates BERT by many years (Huang et al., 2013; De Boom et al., 1999; Mitra et al., 2016; Henderson et al., 2017; Wu et al., 2018; Zamani et al., 2018).

In the context of transformers, the general setup of ranking with dense representations involves learning transformer-based encoders that convert queries and texts into dense, fixed-size vectors. In the simplest approach, ranking becomes the problem of approximate nearest neighbor (ANN) search based on some simple metric such as cosine similarity (Lee et al., 2019; Xiong et al., 2020; Lu et al., 2020; Reimers and Gurevych, 2019; MacAvaney et al., 2020; Gao et al., 2020b; Karpukhin et al., 2020; Zhan et al., 2020; Qu et al., 2020; Hofstätter et al., 2020a; Lin et al., 2020b). However, recognizing that accurate ranking cannot be captured via simple metrics, researchers have explored using more complex machinery to compare dense representations (Humeau et al., 2020; Khattab and Zaharia, 2020). Here, as with multi-stage ranking architectures, limitations on text length and effectiveness–efficiency tradeoffs are important considerations. It becomes increasingly difficult to accurately capture the semantics of longer texts with fixed-sized representations, and increasingly complex comparison architectures increase latency and may necessitate reranking designs.

4 Looking Ahead

Learned dense representations complement sparse (bag-of-words) term-based representations central to keyword search techniques that have dominated the landscape for more than half a century. Together, hybrid multi-stage approaches (e.g., combining both ranking and reranking) present a promising future direction.

Despite the excitement in directly ranking with dense learned representations, we anticipate that reranking transformers will remain important in the future. For one, results from dense retrieval can usually be reranked to achieve even higher effectiveness. At a high level, there are three current approaches: *apply* existing transformer models with minimal modifications, *adapt* existing transformer models, perhaps adding additional architectural elements, and *redesign* transformer-based architectures from scratch. Which approach will prove to be most effective? The jury's still out.

Related, in NLP we see that the GPT family (Brown et al., 2020) continues to push the frontier of larger models, more compute, and more data. For text ranking, is the simple answer to build bigger models? Probably not, since ranking has important differences with many traditional NLP tasks. But if not, what are the evolving roles of zero-shot learning, distant supervision, transfer learning, domain adaptation, data augmentation, and task-specific fine-tuning? This remains an interesting open research question.

While there are aspects of text ranking with pretrained transformers that are well understood, many promising directions await further exploration. Looking ahead, we anticipate many more exciting developments!

5 Presenter Bios

Andrew Yates is a Senior Researcher at the Max Planck Institute for Informatics, where he heads a research group working in areas of information retrieval and natural language processing. Yates received his Ph.D. in Computer Science from Georgetown University in 2016.

Rodrigo Nogueira is a post-doctoral researcher at the University of Waterloo, an adjunct professor at UNICAMP, Brazil, and a senior research scientist at NeuralMind, Brazil. Nogueira received his Ph.D. from New York University in 2019.

Jimmy Lin holds the David R. Cheriton Chair in the David R. Cheriton School of Computer Science at the University of Waterloo. Prior to 2015, he was a faculty at the University of Maryland, College Park. Lin received his Ph.D. in Electrical Engineering and Computer Science from the Massachusetts Institute of Technology in 2004.

Acknowledgments

This work was supported in part by the Canada First Research Excellence Fund and the Natural Sciences and Engineering Research Council (NSERC) of Canada.

References

Zeynep Akkalyoncu Yilmaz, Wei Yang, Haotian Zhang, and Jimmy Lin. 2019. Cross-domain modeling of sentence-level evidence for document retrieval. 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 3490–3496, Hong Kong, China.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *arXiv:2005.14165*.

Zhuyun Dai and Jamie Callan. 2019a. Context-aware sentence/passage term importance estimation for first stage retrieval. *arXiv:1910.10687*.

Zhuyun Dai and Jamie Callan. 2019b. Deeper text understanding for IR with contextual neural language modeling. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2019)*, pages 985–988, Paris, France.

Zhuyun Dai and Jamie Callan. 2020. Context-aware document term weighting for ad-hoc search. In *Proceedings of The Web Conference 2020*, WWW '20, page 1897–1907.

Cedric De Boom, Steven Van Canneyt, Thomas Demeester, and Bart Dhoedt. 1999. Representation learning for very short texts using weighted word embedding aggregation. *Pattern Recognition Letters*, 80(C):150–156.

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.

Luyu Gao, Zhuyun Dai, and Jamie Callan. 2020a. Understanding bert rankers under distillation. In *Proceedings of the 2020 ACM SIGIR International Conference on Theory of Information Retrieval (ICTIR 2020)*, pages 149–152.

Luyu Gao, Zhuyun Dai, Zhen Fan, and Jamie Callan. 2020b. Complementing lexical retrieval with semantic residual embedding. *arXiv:2004.13969*.

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*.

Sebastian Hofstätter, Sophia Althammer, Michael Schröder, Mete Sertkan, and Allan Hanbury. 2020a. Improving efficient neural ranking models with cross-architecture knowledge distillation. *arXiv:2010.02666*.

Sebastian Hofstätter, Markus Zlabinger, and Allan Hanbury. 2020b. Interpretable & time-budget-constrained contextualization for re-ranking. In *Proceedings of the 24th European Conference on Artificial Intelligence (ECAI 2020)*, Santiago de Compostela, Spain.

Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry Heck. 2013. Learning deep structured semantic models for web search using clickthrough data. In *Proceedings of 22nd International Conference on Information and Knowledge Management (CIKM 2013)*, pages 2333–2338, San Francisco, California.

Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2020. Poly-encoders: Architectures and pre-training strategies for fast and accurate multi-sentence scoring. In *Proceedings of the 8th International Conference on Learning Representations (ICLR 2020)*.

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.

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 2020)*, pages 39–48.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6086–6096, Florence, Italy.

Canjia Li, Andrew Yates, Sean MacAvaney, Ben He, and Yingfei Sun. 2020. PARADE: Passage representation aggregation for document reranking. *arXiv:2008.09093*.

Jimmy Lin, Rodrigo Nogueira, and Andrew Yates. 2020a. Pretrained transformers for text ranking: BERT and beyond. *arXiv:2010.06467*.

Sheng-Chieh Lin, Jheng-Hong Yang, and Jimmy Lin. 2020b. Distilling dense representations for ranking using tightly-coupled teachers. *arXiv:2010.11386*.

Wenhao Lu, Jian Jiao, and Ruofei Zhang. 2020. Twin-BERT: Distilling knowledge to twin-structured bert models for efficient retrieval. *arXiv:2002.06275*.

Sean MacAvaney, Franco Maria Nardini, Raffaele Perego, Nicola Tonellotto, Nazli Goharian, and Ophir Frieder. 2020. Expansion via prediction of importance with contextualization. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2020)*, page 1573–1576.

Sean MacAvaney, Andrew Yates, Arman Cohan, and Nazli Goharian. 2019. CEDR: Contextualized embeddings for document ranking. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2019)*, pages 1101–1104, Paris, France.

Yoshitomo Matsubara, Thuy Vu, and Alessandro Moschitti. 2020. Reranking for efficient transformer-based answer selection. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2020)*, page 1577–1580.

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 26 (NIPS 2013)*, pages 3111–3119, Lake Tahoe, California.

Bhaskar Mitra, Sebastian Hofstatter, Hamed Zamani, and Nick Craswell. 2020. Conformer-kernel with query term independence for document retrieval. *arXiv:2007.10434*.

Bhaskar Mitra, Eric Nalisnick, Nick Craswell, and Rich Caruana. 2016. A dual embedding space model for document ranking. *arXiv:1602.01137v1*.

Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with BERT. *arXiv:1901.04085*.

Rodrigo Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. 2020. Document ranking with a pre-trained sequence-to-sequence model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 708–718.

Rodrigo Nogueira and Jimmy Lin. 2019. From doc2query to docTTTTTquery.

Rodrigo Nogueira, Wei Yang, Kyunghyun Cho, and Jimmy Lin. 2019a. Multi-stage document ranking with BERT. In *arXiv:1910.14424*.

Rodrigo Nogueira, Wei Yang, Jimmy Lin, and Kyunghyun Cho. 2019b. Document expansion by query prediction. In *arXiv:1904.08375*.

Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Xin Zhao, Daxiang Dong, Hua Wu, and Haifeng Wang. 2020. RocketQA: An optimized training approach to dense passage retrieval for open-domain question answering. *arXiv:2010.08191*.

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.

Cicero Nogueira dos Santos, Xiaofei Ma, Ramesh Nallapati, Zhiheng Huang, and Bing Xiang. 2020. Beyond [CLS] through ranking by generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1722–1727.

Luca Soldaini and Alessandro Moschitti. 2020. The cascade transformer: an application for efficient answer sentence selection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5697–5708.

Ledell Wu, Adam Fisch, Sumit Chopra, Keith Adams, Antoine Bordes, and Jason Weston. 2018. StarSpace: Embed all the things! In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI 2018)*.

Zhijing Wu, Jiaxin Mao, Yiqun Liu, Jingtao Zhan, Yukun Zheng, Min Zhang, and Shaoping Ma. 2020. Leveraging passage-level cumulative gain for document ranking. In *Proceedings of The Web Conference 2020*, WWW '20, page 2421–2431.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. 2020. Approximate nearest neighbor negative contrastive learning for dense text retrieval. *arXiv:2007.00808*.

Hamed Zamani, Mostafa Dehghani, W. Bruce Croft, Erik Learned-Miller, and Jaap Kamps. 2018. From neural re-ranking to neural ranking: Learning a sparse representation for inverted indexing. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM 2018)*, pages 497–506, Torino, Italy.

Jingtao Zhan, Jiaxin Mao, Yiqun Liu, Min Zhang, and Shaoping Ma. 2020. RepBERT: Contextualized text embeddings for first-stage retrieval. *arXiv:2006.15498*.

Zhi Zheng, Kai Hui, Ben He, Xianpei Han, Le Sun, and Andrew Yates. 2020. BERT-QE: Contextualized Query Expansion for Document Re-ranking. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4718–4728.