Book Review

Pretrained Transformers for Text Ranking: BERT and Beyond

Jimmy Lin, Rodrigo Nogueira, and Andrew Yates
(University of Waterloo, University of Campinas, University of Amsterdam)

Morgan & Claypool (Synthesis Lectures on Human Language Technologies, edited by Graeme Hirst, volume 53), 2021, 325 pp; ISBN paperback: 9781636392288; ISBN ebook: 9781636392295; doi:10.2200/S01123ED1V01Y202108HLT053

Reviewed by
Suzan Verberne
Leiden Institute of Advanced Computer Science, Leiden University, the Netherlands

Text ranking takes a central place in Information Retrieval (IR), with Web search as its best-known application. More generally, text ranking models are applicable to any Natural Language Processing (NLP) task in which relevance of information plays a role, from filtering and recommendation applications to question answering and semantic similarity comparisons. Since the rise of BERT in 2019, Transformer models have become the most used and studied architectures in both NLP and IR, and they have been applied to basically any task in our research fields—including text ranking.

In a fast-changing research context, it can be challenging to keep lecture materials up to date. Lecturers in NLP are grateful for Dan Jurafsky and James Martin for yearly updating the 3rd edition of their textbook, making *Speech and Language Processing* the most comprehensive, modern textbook for NLP. The IR field is less fortunate, still relying on older textbooks, extended with a collection of recent materials that address neural models. The textbook *Pretrained Transformers for Text Ranking: BERT and Beyond* by Jimmy Lin, Rodrigo Nogueira, and Andrew Yates is a great effort to collect the recent developments in the use of Transformers for text ranking.

The introduction of the book is well-scoped with clear guidance for the reader about topics that are out of scope (such as user aspects). This is followed by an excellent history section, stating for example:

We might take for granted today the idea that automatically extracted terms from a document can serve as descriptors or index terms for describing the contents of those documents, but this was an important conceptual leap in the development of information retrieval. (p. 11)

Chapter 2 is a complete yet compact overview of the IR field and the context of Transformer-based ranking models, with a substantial section devoted to text collections and a good introduction to keyword search. This includes a discussion of variants of BM25 leading to different results, and a mention of pseudo-relevance feedback. Also, this chapter pays attention to terminology differences between fields, clearly aligning terms such as performance versus effectiveness.

https://doi.org/10.1162/coli_r_00468

Chapter 3 is extensive. It starts with an overview of BERT. This is an excellent introduction to the topic, not as detailed as Jurafsky and Martin's but sufficiently specific for students to understand it on a conceptual level. I will definitely include parts of these chapters in my lecture materials for master's students. The subsequent sections, in which the behavior of BERT for ranking is further analyzed quantitatively and qualitatively with even some non-published experiments, are very extensive, and perhaps too extensive—and unnecessary—for a textbook. Section 3.3 provides a clear explanation of the challenge of ranking full documents instead of passages. Interestingly, the whole of Chapter 3 describes the history of Transformer-based ranking, but since the history is so recent, it is all still relevant. To my taste, however, the summaries of individual papers with result tables are in places too detailed. I do appreciate that the authors spoke to the authors of some of these papers, sometimes even leading to a better explanation of the original methods. Sections 3.4 and 3.5 have good discussions of effectiveness–efficiency trade-offs, which is an important topic in text ranking because search engines cannot permit themselves long query latencies. Section 3.5 has an interesting discussion of sequence-to-sequence approaches to classification and ranking tasks.

In Chapter 4, applications of Transformer methods for document and query expansion are discussed. This is a topic that gets less attention in the literature than Transformer-based rankers, but it is important to address because it can be a very effective and, importantly, efficient strategy. Moreover, it gives interesting insights in what aspects of document content are the most important in relevance ranking. Especially the discussion of doc2query-T5 is insightful in this respect. The later sections of Chapter 4 are again very detailed and report experimental results from a number of papers; I would personally prefer to read more intuitions, and see examples and visual explanations of some of the most important methods instead.

Chapter 5 addresses representation learning for ranking, and in Section 5.1 the authors clearly put this task in the context of other NLP tasks, with a very relevant conclusion: A wide range of tasks can be represented as a concatenation of two texts and then fed into a BERT model, and this appears to be a successful paradigm in many contexts. In the words of the authors:

When faced with these myriad tasks, a natural question would be: Do these distinctions matter? With BERT, the answer is, likely not. (p. 201)

Section 5.2 relates text ranking with dense vectors to nearest neighbor search; indicating how efficient this approach is at query time because the documents are encoded at index time. After some history in Section 5.3, Section 5.4 provides a good explanation of bi-encoder architectures for retrieval (starting with SentenceBERT). Personally, I think the distinction between cross-encoders and bi-encoders could have been explained earlier in the book. Although cross-encoders reach higher retrieval effectiveness, bi-encoders are much more efficient at inference time and therefore too important to be described this late in the book. ColBERT, explained clearly in 5.5.2, is a key example of a model that has the efficiency of bi-encoders but approaches the effectiveness of cross-encoders.

Chapter 6 makes explicit that some related (potentially relevant) topics were omitted from the book on purpose, in particular, question answering, summarization, and conversational (interactive) search. Then it extensively goes into open research questions, which I interpret as an invitation and encouragement to the community for topics to dive into. One topic that I think should have received more attention is the application of Transformer models in domains other than the general domain, in particular,

legal and medical. For these domains, more ranking benchmarks have become available in recent years. We have seen that for some of those benchmarks—especially in query-by-document retrieval like case law retrieval or prior art retrieval, it is challenging to beat the lexical retrieval baseline (BM25), because there is less training data than in the common benchmarks, and queries and documents are much longer.

A challenge with writing a textbook about recent research topics is that it becomes outdated the moment it is published. I hope the authors will publish regular updates as science progresses. Overall, I think this book is a valuable resource that I have already recommended to my Ph.D. students, and that definitely provides me with materials for my master's courses. It is also relevant for NLP researchers dealing with text ranking problems—basically any task in which information relevance plays a role. This textbook could be the ideal cross-over from IR knowledge to the NLP field, using our common friend BERT as the bridge.

Suzan Verberne is an Associate Professor at the Leiden Institute of Advanced Computer Science at Leiden University. She is group leader of Text Mining and Retrieval. She obtained her Ph.D. in 2010 on the topic of Question Answering and has since then been working on the edge between NLP and IR. Her current work centers around interactive information access for specific domains. She is highly active in the NLP and IR communities, holding chairing positions in the large world-wide conferences. Suzan's e-mail address is `s.verberne@liacs.leidenuniv.nl`.