

Pretrained Transformers for Text Ranking

BERT and Beyond

Synthesis Lectures on Human Language Technologies

Editor

Graeme Hirst, *University of Toronto*

Synthesis Lectures on Human Language Technologies is edited by Graeme Hirst of the University of Toronto. The series consists of 50- to 150-page monographs on topics relating to natural language processing, computational linguistics, information retrieval, and spoken language understanding. Emphasis is on important new techniques, on new applications, and on topics that combine two or more HLT subfields.

Pretrained Transformers for Text Ranking: BERT and Beyond

Jimmy Lin, Rodrigo Nogueira, and Andrew Yates
2021

Automated Essay Scoring

Beata Beigman Klebanov and Nitin Madnani
2021

Explainable Natural Language Processing

Anders Søgaard
2021

Finite-State Text Processing

Kyle Gorman and Richard Sproat
2021

Semantic Relations Between Nominals, Second Edition

Vivi Nastase, Stan Szpakowicz, Preslav Nakov, and Diarmuid Ó Séaghdha
2021

Embeddings in Natural Language Processing: Theory and Advances in Vector Representations of Meaning

Mohammad Taher Pilehvar and Jose Camacho-Collados
2020

Conversational AI: Dialogue Systems, Conversational Agents, and Chatbots

Michael McTear
2020

Natural Language Processing for Social Media, Third Edition

Anna Atefeh Farzindar and Diana Inkpen
2020

Statistical Significance Testing for Natural Language Processing

Rotem Dror, Lotem Peled, Segev Shlomov, and Roi Reichart
2020

Deep Learning Approaches to Text Production

Shashi Narayan and Claire Gardent
2020

Linguistic Fundamentals for Natural Language Processing II: 100 Essentials from Semantics and Pragmatics

Emily M. Bender and Alex Lascarides
2019

Cross-Lingual Word Embeddings

Anders Søgaard, Ivan Vulić, Sebastian Ruder, Manaal Faruqui
2019

Bayesian Analysis in Natural Language Processing, Second Edition

Shay Cohen
2019

Argumentation Mining

Manfred Stede and Jodi Schneider
2018

Quality Estimation for Machine Translation

Lucia Specia, Carolina Scarton, and Gustavo Henrique Paetzold
2018

Natural Language Processing for Social Media, Second Edition

Atefeh Farzindar and Diana Inkpen
2017

Automatic Text Simplification

Horacio Saggion
2017

Neural Network Methods for Natural Language Processing

Yoav Goldberg
2017

Syntax-based Statistical Machine Translation

Philip Williams, Rico Sennrich, Matt Post, and Philipp Koehn
2016

Domain-Sensitive Temporal Tagging

Jannik Strötgen and Michael Gertz
2016

Linked Lexical Knowledge Bases: Foundations and Applications

Iryna Gurevych, Judith Eckle-Kohler, and Michael Matuschek
2016

Bayesian Analysis in Natural Language Processing

Shay Cohen
2016

Metaphor: A Computational Perspective

Tony Veale, Ekaterina Shutova, and Beata Beigman Klebanov
2016

Grammatical Inference for Computational Linguistics

Jeffrey Heinz, Colin de la Higuera, and Menno van Zaanen
2015

Automatic Detection of Verbal Deception

Eileen Fitzpatrick, Joan Bachenko, and Tommaso Fornaciari
2015

Natural Language Processing for Social Media

Atefeh Farzindar and Diana Inkpen
2015

Semantic Similarity from Natural Language and Ontology Analysis

Sébastien Harispe, Sylvie Ranwez, Stefan Janaqi, and Jacky Montmain
2015

Learning to Rank for Information Retrieval and Natural Language Processing, Second Edition

Hang Li
2014

Ontology-Based Interpretation of Natural Language

Philipp Cimiano, Christina Unger, and John McCrae
2014

Automated Grammatical Error Detection for Language Learners, Second Edition

Claudia Leacock, Martin Chodorow, Michael Gamon, and Joel Tetreault
2014

Web Corpus Construction

Roland Schäfer and Felix Bildhauer
2013

Recognizing Textual Entailment: Models and Applications

Ido Dagan, Dan Roth, Mark Sammons, and Fabio Massimo Zanzotto
2013

Linguistic Fundamentals for Natural Language Processing: 100 Essentials from Morphology and Syntax

Emily M. Bender
2013

Semi-Supervised Learning and Domain Adaptation in Natural Language Processing

Anders Søgaard
2013

Semantic Relations Between Nominals

Vivi Nastase, Preslav Nakov, Diarmuid Ó Séaghdha, and Stan Szpakowicz
2013

Computational Modeling of Narrative

Inderjeet Mani
2012

Natural Language Processing for Historical Texts

Michael Piotrowski
2012

Sentiment Analysis and Opinion Mining

Bing Liu
2012

Discourse Processing

Manfred Stede
2011

Bitext Alignment

Jörg Tiedemann
2011

Linguistic Structure Prediction

Noah A. Smith
2011

Learning to Rank for Information Retrieval and Natural Language Processing

Hang Li
2011

Computational Modeling of Human Language Acquisition

Afra Alishahi
2010

Introduction to Arabic Natural Language Processing

Nizar Y. Habash
2010

Cross-Language Information Retrieval

Jian-Yun Nie
2010

Automated Grammatical Error Detection for Language Learners

Claudia Leacock, Martin Chodorow, Michael Gamon, and Joel Tetreault
2010

Data-Intensive Text Processing with MapReduce

Jimmy Lin and Chris Dyer
2010

Semantic Role Labeling

Martha Palmer, Daniel Gildea, and Nianwen Xue
2010

Spoken Dialogue Systems

Kristiina Jokinen and Michael McTear
2009

Introduction to Chinese Natural Language Processing

Kam-Fai Wong, Wenjie Li, Ruifeng Xu, and Zheng-sheng Zhang
2009

Introduction to Linguistic Annotation and Text Analytics

Graham Wilcock
2009

Dependency Parsing

Sandra Kübler, Ryan McDonald, and Joakim Nivre
2009

Statistical Language Models for Information Retrieval

ChengXiang Zhai
2008

© Springer Nature Switzerland AG 2022
Reprint of original edition © Morgan & Claypool 2022

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means—electronic, mechanical, photocopy, recording, or any other except for brief quotations in printed reviews, without the prior permission of the publisher.

Pretrained Transformers for Text Ranking: BERT and Beyond
Jimmy Lin, Rodrigo Nogueira, and Andrew Yates

ISBN: 978-3-031-01053-8 paperback
ISBN: 978-3-031-0 181-7 PDF
ISBN: 978-3-031-00192-5 hardcover

DOI 10.1007/978-3-031-02181-7

A Publication in the Springer series
SYNTHESIS LECTURES ON HUMAN LANGUAGE TECHNOLOGIES

Lecture #53
Series Editor: Graeme Hirst, *University of Toronto*
Series ISSN
Print 1947-4040 Electronic 1947-4059

Pretrained Transformers for Text Ranking

BERT and Beyond

Jimmy Lin
University of Waterloo

Rodrigo Nogueira
University of Waterloo

Andrew Yates
University of Amsterdam and Max Planck Institute for Informatics

SYNTHESIS LECTURES ON HUMAN LANGUAGE TECHNOLOGIES #53

ABSTRACT

The goal of text ranking is to generate an ordered list of texts retrieved from a corpus in response to a query. Although the most common formulation of text ranking is search, instances of the task can also be found in many natural language processing (NLP) applications. This book provides an overview of text ranking with neural network architectures known as transformers, of which BERT (Bidirectional Encoder Representations from Transformers) is the best-known example. The combination of transformers and self-supervised pretraining has been responsible for a paradigm shift in NLP, information retrieval (IR), and beyond.

This book provides a synthesis of existing work as a single point of entry for practitioners who wish to gain a better understanding of how to apply transformers to text ranking problems and researchers who wish to pursue work in this area. It covers a wide range of modern techniques, grouped into two high-level categories: transformer models that perform reranking in multi-stage architectures and dense retrieval techniques that perform ranking directly. Two themes pervade the book: techniques for handling long documents, beyond typical sentence-by-sentence processing in NLP, and techniques for addressing the tradeoff between effectiveness (i.e., result quality) and efficiency (e.g., query latency, model and index size). Although transformer architectures and pretraining techniques are recent innovations, many aspects of how they are applied to text ranking are relatively well understood and represent mature techniques. However, there remain many open research questions, and thus in addition to laying out the foundations of pretrained transformers for text ranking, this book also attempts to prognosticate where the field is heading.

KEYWORDS

pretrained transformers, BERT, neural information retrieval, effectiveness–efficiency tradeoffs, reranking, multi-stage ranking, dense retrieval, representational learning, document expansion, query expansion

Contents

	Preface	xv
	Acknowledgments	xvii
1	Introduction	1
1.1	Text Ranking Problems	4
1.2	A Brief History	10
1.2.1	The Beginnings of Text Ranking	11
1.2.2	The Challenges of Exact Match	12
1.2.3	The Rise of Learning to Rank	16
1.2.4	The Advent of Deep Learning	18
1.2.5	The Arrival of BERT	22
1.3	Roadmap, Assumptions, and Omissions	23
2	Setting the Stage	25
2.1	Texts	25
2.2	Information Needs	27
2.3	Relevance	30
2.4	Relevance Judgments	33
2.5	Ranking Metrics	34
2.6	Community Evaluations and Reusable Test Collections	40
2.7	Descriptions of Common Test Collections	46
2.8	Keyword Search	56
2.9	Notes on Parlance	59
3	Multi-Stage Architectures for Reranking	63
3.1	A High-Level Overview of BERT	64
3.2	Simple Relevance Classification: monoBERT	71
3.2.1	Basic Design of monoBERT	73
3.2.2	Exploring monoBERT	78
3.2.3	Investigating How BERT Works	86

3.2.4	Nuances of Training BERT	89
3.3	From Passage to Document Ranking	94
3.3.1	Document Ranking with Sentences: Birch	96
3.3.2	Passage Score Aggregation: BERT–MaxP and Variants	103
3.3.3	Leveraging Contextual Embeddings: CEDR	110
3.3.4	Passage Representation Aggregation: PARADE	118
3.3.5	Alternatives for Tackling Long Texts	124
3.4	From Single-Stage to Multi-Stage Rerankers	127
3.4.1	Reranking Pairs of Texts	130
3.4.2	Reranking Lists of Texts	135
3.4.3	Efficient Multi-Stage Rerankers: Cascade Transformers	137
3.5	Beyond BERT	141
3.5.1	Knowledge Distillation	142
3.5.2	Ranking with Transformers: TK, TKL, CK	146
3.5.3	Ranking with Sequence-to-Sequence Models: monoT5	152
3.5.4	Ranking with Sequence-to-Sequence Models: Query Likelihood ..	159
3.6	Concluding Thoughts	161
4	Refining Query and Document Representations	163
4.1	Query and Document Expansion: General Remarks	164
4.2	Pseudo-Relevance Feedback with Contextualized Embeddings: CEQE ...	167
4.3	Document Expansion via Query Prediction: doc2query	172
4.4	Term Reweighting as Regression: DeepCT	178
4.5	Term Reweighting with Weak Supervision: HDCT	183
4.6	Combining Term Expansion with Term Weighting: DeepImpact	186
4.7	Expansion of Query and Document Representations	189
4.8	Concluding Thoughts	192
5	Learned Dense Representations for Ranking	195
5.1	Task Formulation	196
5.2	Nearest Neighbor Search	202
5.3	Pre-BERT Text Representations for Ranking	204
5.4	Simple Transformer Bi-Encoders for Ranking	206
5.4.1	Basic Bi-Encoder Design: Sentence-BERT	208
5.4.2	Bi-Encoders for Dense Retrieval: DPR and ANCE	211
5.4.3	Bi-Encoders for Dense Retrieval: Additional Variations	219

5.5	Enhanced Transformer Bi-Encoders for Ranking	223
5.5.1	Multiple Text Representations: Poly-encoders and ME-BERT	224
5.5.2	Per-Token Representations and Late Interactions: ColBERT	227
5.6	Knowledge Distillation for Transformer Bi-Encoders	230
5.7	Concluding Thoughts	234
6	Future Directions and Conclusions	239
6.1	Notable Content Omissions	239
6.2	Open Research Questions	241
6.3	Final Thoughts	253
	Bibliography	255
	Authors' Biographies	307

Preface

BERT, unveiled by Google in October 2018, represents the culmination of a series of developments in NLP that date back many years. Together with its siblings, cousins, and intellectual descendants, these pretrained transformer models have, without exaggeration, brought about a paradigm shift in technologies that process, analyze, and otherwise manipulate human language text (and beyond). Pretrained transformers are responsible for substantial quality improvements in a range of downstream tasks, including natural language inference, syntactic and semantic analysis, text classification, question answering, summarization, and many more...and yes, text ranking.

The goal of text ranking is to generate an ordered list of texts retrieved from a corpus in response to a user's query. The most common formulation of text ranking is search, but many other NLP tasks involve aspects of text ranking as well. Despite being a relatively recent innovation, the foundations of how to apply BERT and other pretrained transformers to tackle text ranking are already quite sturdy. In many cases, improvements in effectiveness are substantial, robust, and have been widely replicated in many application scenarios. Building on these empirical foundations, this book provides a synthesis of existing work. We imagine these pages to be useful for practitioners interested in real-world solutions as well as researchers who wish to pursue work in this area.

This book covers a wide range of techniques for text ranking with pretrained transformers, grouped into two categories: transformer models that perform reranking in multi-stage architectures and dense retrieval techniques that perform ranking directly. Examples in the first category include approaches based on relevance classification, evidence aggregation from multiple segments of text, and query expansion methods. The second category involves using transformers to learn dense vector representations of texts, where ranking is formulated as comparisons between query and document vectors that take advantage of nearest neighbor search.

Two themes pervade our book: techniques for handling long documents, beyond typical sentence-by-sentence processing in NLP, and techniques for addressing the tradeoffs between effectiveness (i.e., result quality) and efficiency (e.g., query latency, model and index size). Much effort has been devoted to developing ranking models that address the mismatch between document lengths and the length limitations of input to transformers. The computational costs of inference with transformers have led to the development of alternatives and variants that aim for different tradeoffs, both within multi-stage architectures as well as with dense learned representations.

While many aspects of how to apply pretrained transformers to text ranking are well understood and can be considered mature techniques, there remain many open research questions

and unresolved issues. Thus, in addition to laying out the foundations of pretrained transformers for text ranking, this book also attempts to prognosticate where the field is heading.

It is quite remarkable that BERT is only about three years old. Taking a step back and reflecting, the field has seen an incredible amount of progress in a short amount of time. We expect even more exciting things to come!

Jimmy Lin, Rodrigo Nogueira, and Andrew Yates
August 2021

Acknowledgments

This book was supported in part by the Canada First Research Excellence Fund and the Natural Sciences and Engineering Research Council (NSERC) of Canada. We would like to thank the TPU Research Cloud and Compute Canada for resources used to obtain new results in this work.

In addition, we'd like to thank the following people for comments on earlier drafts of this book: Chris Buckley, Danqi Chen, Maura Grossman, Sebastian Hofstätter, Kenton Lee, Sheng-Chieh Lin, Xueguang Ma, Bhaskar Mitra, Jheng-Hong Yang, Scott Yih, and Ellen Voorhees. Special thanks goes out to two anonymous reviewers for their insightful comments and helpful feedback.

Jimmy Lin, Rodrigo Nogueira, and Andrew Yates
August 2021