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 Séagdha
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

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

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