# From Extractive to Abstractive Summarization: A Journey

Parth Mehta · Prasenjit Majumder

# From Extractive to Abstractive Summarization: A Journey

Springer

Parth Mehta
Information Retrieval and Language Processing Lab
Dhirubhai Ambani Institute of Information and Communication Technology
Gandhinagar, Gujarat, India

Prasenjit Majumder
Information Retrieval and Language Processing Lab
Dhirubhai Ambani Institute of Information and Communication Technology
Gandhinagar, Gujarat, India

ISBN 978-981-13-8933-7 ISBN 978-981-13-8934-4 (eBook)
https://doi.org/10.1007/978-981-13-8934-4



This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

# Preface

Research in the field of text summarisation has primarily been dominated by investigations of various sentence extraction techniques with a significant focus towards news articles. In this book, we intend to look beyond generic sentence extraction and instead focus on domain-specific summarisation, methods for creating ensembles of multiple extractive summarisation techniques and using sentence compression as the first step towards abstractive summarisation.

We begin by proposing two new corpora, related to legal and scientific articles, for domain-specific summarisation. The first corpus is a collection of judgements delivered by the Supreme Court of India, with corresponding handwritten summaries written by legal experts. The second dataset is a collection of scientific articles related to the domain of computational linguistics and indexed in the ACL anthology. These two tasks highlight the challenges in domain-specific summarisation. The legal summaries are recall-oriented (trying to retain as much information as possible) and semi-extractive (reuses phrases and sentences from the original document). In comparison, the abstracts of ACL articles are more precision-oriented and abstractive. Both collections have a reasonable number of article-summary pairs, enabling us to use data-driven techniques. We annotate both these corpora through a completely automatic process described without requiring any manual intervention. Excluding newswire corpora, where the summaries are usually article headlines, the proposed collections are among the largest openly available collections for document summarisation. We propose a completely data-driven technique for sentence extraction for legal and scientific articles. In both legal and ACL corpora, the summaries have a predefined format. Hence, it is possible to identify *summary worthy* sentences depending on whether they contain certain key phrases. Our proposed approach uses attention-based neural network to automatically identify these key phrases from pseudo-labelled data, without requiring any annotation or handcrafted rules. The proposed model outperforms the existing baselines and state-of-the-art systems by a large margin.

Next, we explore methods for combining several sentence extraction techniques to build an ensemble. A large number of sentence extraction techniques exist, none of which guarantee better performance than the others. As a part of this book, we

explore whether it is possible to leverage this variance in performance for generating an ensemble of several extractive techniques. In the first model, we study the effect of using multiple sentence similarity scores, ranking algorithms and text representation techniques. We demonstrate that such variations can be used for improving rank aggregation, leading to better sentence rankings. Using several sentence similarity metrics, with any given ranking algorithm, always generates better abstracts. We also propose several content-based aggregation models. Given the variation in performance of extractive techniques across documents, the a priori knowledge about which technique would give the best result for a given document can drastically improve the result. In such case, an oracle ensemble system can be made which chose the best possible summary for a given document. In the proposed content-based aggregation models, we estimate the probability of a summary being good by looking at the amount of content it shares with other candidate summaries. We present a hypothesis that a good summary will necessarily share more information with another good summary, but not with a bad summary. We build upon this argument to construct several content-based aggregation techniques, achieving a substantial improvement in the ROUGE scores.

In the end, we propose another attention-based neural model for sentence compression. We use a novel context encoder, which helps the network to handle rare but informative terms better. We contrast our work to some sentence compression and abstractive techniques that have been proposed in the past few years. We present our arguments for and against these techniques and build a further road map for abstractive summarisation. In the end, we present the results on an end-to-end system which performs sentence extraction using stand-alone summarisation systems as well as their ensembles and then uses the sentence compression technique for generating the final abstractive summary. Overall, this book is aimed at providing alternates to generic extractive summarisation in the form of domain-specific as well as ensemble techniques, gradually leading to sentence compression and abstractive summarisation.

Gandhinagar, India

Parth Mehta
Prasenjit Majumder

# Contents

# About the Authors

**Dr. Parth Mehta** completed his M.Tech. in Machine Intelligence and his Ph.D. in Text Summarization at Dhirubhai Ambani Institute of ICT (DA-IICT), Gandhinagar, India. At the DA-IICT he was part of the Information Retrieval and Natural Language Processing Lab. He was also involved in the national project "Cross Lingual Information Access", funded by the Govt. of India, which focused on building a cross-lingual search engine for nine Indian languages.

Dr. Mehta has served as reviewer for the journals Information Processing and Management and Forum for Information Retrieval Evaluation. Apart from several journal and conference papers, he has also co-edited a book on text processing published by Springer.

**Prof. Prasenjit Majumder** is an Associate Professor at Dhirubhai Ambani Institute of ICT (DA-IICT), Gandhinagar and a Visiting Professor at the Indian Institute of Information Technology, Vadodara (IIIT-V). Professor Majumder completed his Ph.D. at Jadavpur University in 2008 and worked as a postdoctoral fellow at the University College Dublin, prior to joining the DA-IICT, where he currently heads the Information Retrieval and Language Processing Lab. His research interests lie at the intersection of Information Retrieval, Cognitive Science and Human Computing Interaction. He has headed several projects sponsored by the Govt. of India.

He is one of the pioneers of the Forum for Information Retrieval Evaluation (FIRE), which assesses research on Information Retrieval and related areas for South Asian languages. Since being founded in 2008, FIRE has grown to become a respected conference, drawing participants from across the globe. Professor Majumder has authored several journal and conference papers, and co-edited two special issues of Transactions in Information Systems (ACM). He has co-edited two books: 'Multi Lingual Information Access in South Asian Languages' and 'Text Processing,' both published by Springer.