Automatic Text Summarization

Chapter · September 2014

DOI: 10.1002/9781119004752.ch6

Juan-Manuel Torres-Moreno
Université d´Avignon et des Pays du Vaucluse

COGNITIVE SCIENCE AND KNOWLEDGE MANAGEMENT SERIES

Automatic Text Summarization

Juan-Manuel Torres-Moreno

iSTE

WILEY

Automatic Text Summarization

Series Editor
Jean-Charles Pomerol

Automatic Text Summarization

Juan-Manuel Torres-Moreno

iSTE

WILEY

First published 2014 in Great Britain and the United States by ISTE Ltd and John Wiley & Sons, Inc.

ISTE Ltd
27-37 St George's Road
London SW19 4EU
UK

www.iste.co.uk

John Wiley & Sons, Inc.
111 River Street
Hoboken, NJ 07030
USA

www.wiley.com

Library of Congress Control Number: 2014947781

British Library Cataloguing-in-Publication Data
A CIP record for this book is available from the British Library
ISBN 978-1-84821-668-6

Contents

Foreword by A. Zamora and R. Salvador

Foreword

The need to identify important information

Throughout history, the sheer amount of printed information and the scarce availability of time to read it have always been two major obstacles in the search for knowledge. Famous novelist W. Somerset Maugham wrote "It was customary for someone who read a book to skip paragraphs, pages or even whole sections of the book. But, to obviate paragraphs or pages without suffering losses was a dangerous and very difficult thing to do unless one had a natural gift for it or a peculiar facility for an on-going recognition of interesting things as well as for bypassing invalid or uninteresting matters."[1] Somerset Maugham called this the art of skipping pages, and even himself had an offer by a North American publisher to re-edit old books in abbreviated form. The publisher wanted him to omit everything except the argument, main ideas and personages created by the author.

[1] *Ten Novels and Their Authors* by W. Somerset Maugham.

The problem of information storage

In the November 1961 issue of the *Library Journal*, Malcolm M. Ferguson, Reference Librarian at the Massachusetts Institute of Technology, wrote that the December 12, 1960, issue of *Time* magazine included a statement that, in his opinion, "may provoke discussion and perplexity". The article reported that Richard P. Feynman, Professor of Physics, California Institute of Technology (who would later receive a Nobel prize), had predicted that an explosion of information and storage would soon occur on planet Earth and argued that it would be convenient to reduce the amount and size of present information to be able to store all the world's basic knowledge in the equivalent of a pocket-sized pamphlet. Feynman went on to offer a prize to anyone reducing the information of one page of a book to one twenty-five-thousandth of the linear scale of the original, in a manner that it could be read with an electron microscope.

One year after Ferguson's article, Hal Drapper from the University of California published a satirical article called "MS FND IN A LBRY" in the December 1961 issue of *The Magazine of Fantasy & Science Fiction*. Drapper poked fun at the idea of trying to cope with Feynman's predicted information explosion problem by compressing data to microscopic levels to help store the information and by the development of indexes of indexes in order to retrieve it.

The information explosion was and still is a real problem, but the exponential growth in the capacity of new electronic processors overcomes the barrier imposed by old paper archives. Electronic book readers, such as Amazon's Kindle, can now store hundreds of books in a device the size of a paperback book. Encoding information has even been taken to the molecular level, such as the synthetic organism created through genetic engineering at the J. Craig Venter Institute which used nucleotides in the organism's DNA to encode a message containing the names of the authors and contributors. And the message would replicate when the organism multiplies.

Automatic size reduction

Since the dawn of the computer age, various attempts have been made to automatically shrink the size of the documents into a human-readable format. Drapper suggested one experimental method which consisted of reducing the cumbersome alphabet to mainly consonantal elements (thus: thr cmbrsm alfbt ws rdsd t mnl cnsntl elmnts) but this was done to facilitate quick reading, and only incidentally would cut down the mass of documents and books to solve the information explosion. More sophisticated methods attempted to identify, select and extract important information through statistical analysis by correlating words from the title to passages in the text, and by analyzing the position in which sentences occurred in the document trying to assign importance to sentences by their positions within the text. We (Antonio Zamora and Ricardo Salvador) worked at Chemical Abstract Service (CAS), where abstracting and indexing performed manually was our daily job[2]. Realizing that it was difficult to recognize what was important in a document, we developed a computer program that started by trying to discover what was not important, such as clichés, empty phases, repetitive expressions, tables and grammatical subterfuges that were not essential for understanding the article. Our technique to eliminate non-significant, unessential, unsubstantial, trivial, useless, duplicated and obvious sentences from the whole text reduced the articles to the salient and interesting points of the document.

By the late 1970s, we could manufacture indicative abstracts for a fraction of a dollar. These abstracts contained 60–70% of the same sentences chosen by professional abstractors. Some professional abstractors began to worry that they could lose their jobs to a machine. However, primary journals started providing abstracts prepared by the authors themselves; so there was no demand for automatic abstracting. The Internet has changed all that. Many news feeds with unabridged text have become available that can overwhelm anyone looking for

[2] See sections 1.5 and 3.1.4.

information. Yes, presently there is a real need for automatic abstracting.

The future

Today's smart phones have more computational power than many mainframe computers of the 20th Century. Speech recognition and automatic translation have evolved from being experimental curiosities to tools that we use every day from Google. We are at the threshold of artificial intelligence. IBM's Watson program won a one-million dollar Jeopardy contest against the two best human champions. Cloud-based computing has removed all the constraints of memory size from our portable devices. It is now possible to access extensive knowledge bases with simple protocols. The ease with which dictionaries can be accessed allows us to use synonyms in our contextual searches so that "bird flu" will retrieve "avian influenza". Great advances in automatic abstracting have been made during the last 40 years. It is quite possible that in the next quarter of a century there will be computer programs with enough cognition to answer questions such as "What were the important points of this article?". This is exactly what automatic abstracting strives to accomplish. For specific tasks, the behavior of these new programs will be indistinguishable from that of humans. These expectations are not only dreams about a distant future ... we may actually live to see them become reality.

This book by Juan-Manuel Torres-Moreno presents the approaches that have been used in the past for automatic text summarization and describes the new algorithms and techniques of state-of-the-art programs.

Antonio ZAMORA
Ricardo SALVADOR
August 2014

Foreword by H. Saggion

Automatic Text Summarization

Juan-Manual Torres-Moreno

Text summarization, the reduction of a text to its essential content, is a task that requires linguistic competence, world knowledge, and intelligence. Automatic text summarization, the production of summaries by computers is therefore a very difficult task. One may wonder whether machines would ever be able to produce summaries which are indistinguishable from human summaries, a kind of Turing test and a motivation to advance the state of the art in natural language processing. Text summarization algorithms have many times ignored the cognitive processes and the knowledge that go into text understanding and which are essential to properly summarize.

In *Automatic Text Summarization*, Juan-Manuel Torres- Moreno offers a comprehensive overview of methods and techniques used in automatic text summarization research from the first attempts to the most recent trends in the field (e.g. opinion and tweet summarization).

Torres-Moreno makes an excellent job of covering various summarization problems, starting from the motivations behind this interesting subject, he takes the reader on a long research journey that spans over 50 years. The book is organized into more or less traditional topics: single and multi-document summarization, domain-specific summarization, multi-lingual and cross-lingual summarization. Systems, algorithms and methods are explained in detail often with illustrations, assessment of their performances, and limitations. Torres-Moreno pays particular attention to intrinsic summarization evaluation metrics that are based on vocabulary comparison and to international evaluation programs in text summarization such as the Document Understanding Conference and the Text Analysis Conference. Part of the book is dedicated to text abstracting, the ultimate goal of text summarization research, consisting of the production of text summaries which are not a mere copy of sentences and words from the input text.

While various books exist on this subject, Torres-Moreno's covers interesting system and research ideas rarely cited in the literature. This book is a very valuable source of information offering in my view the most complete account of automatic text summarization research up to date. Because of its detailed content, clarity of exposition, and inquisitive style, this work will become a very valuable resource for teachers and researchers alike. I hope the readers will learn from this book and enjoy it as much as I have.

Horacio SAGGION
Research Professor at Universitat Pompeu Fabra
Barcelona, August 2014

Notation

The main notations used in this book are the following:

V	the set of words in a text
$\|V\| = N$	the cardinality of the set V (N words)
$s \in D$	an element s belonging to the set D
$A \backslash B$	the complement of sets A and B
	elements of A without the elements of B
λ	a linear combination parameter, $0 \leq \lambda \leq 1$
P	the precision of a system
R	the recall of a system
$\\|\boldsymbol{x}\\|$	the norm of the vector $\boldsymbol{x}$: $\\|\boldsymbol{x}\\| = \sum_i x_i^2$
$C(\bullet)$	the occurrence number of an event
C_w^s	the occurrences of the word w in the sentence s
$\mathcal{C}$	independent and identically distributed (i.i.d.) set of observations: $\mathcal{C} = \{x^{(1)}, x^{(2)}, \ldots, x^{(n)}\}$
$\mathcal{LL}(\mathcal{C})$	the log-probability of observations: $\mathcal{LL}(\mathcal{C}) = \log(\mathcal{L}(\mathcal{C}))$
ρ	the number of sentences of a document
$\vec{D} = (s_1, s_2, \ldots, s_\rho)$	a document of ρ sentences (VSM model)
$\vec{Q}$	a query of N terms, $Q = (q_1, q_2, \ldots, q_N)$
Sum	a summary of a document

Π	a set of n summaries $\Pi = \{\mathrm{Sum}_1, \mathrm{Sum}_2, \dots, \mathrm{Sum}_n\}$
ε	an empty word
w	a word over an alphabet Σ
$\|w\|$	the length of the word w ($\|\varepsilon\| = 0$)
s	a sentence composed of N words
s_μ	a sentence $\mu \subset \{1 \dots \rho\}$
$\vec{s_\mu}$	a sentence vector $\mu \subset \{1 \dots \rho\}$ (VSM model)
$\|s\| = \|w_1, w_2, \dots, w_N\|$	the length of a sentence of N words w_j
w_j	a word of s in positions $j \subset \{1 \dots \|s\|\}$
$\omega(s)$	the normalized weight of the sentence s, $0 \leq \omega(s) \leq 1$
$\omega_{i,j}$	the weight of the word j in the sentence i
σ	a discourse segment $\in s$
$p_e(\sigma)$	the probability of eliminating a discourse segment σ
$adj[s]$	adjacency matrix of the sentence s (VSM model)
$S[i, j]$	sentences × terms matrix (VSM model)
$\mathrm{sim}(x, y)$	function that quantifies the similarity between 2 objects
τ	compression rate (characters, words or sentences)
ϵ	error of tolerance for convergence
$\mathcal{D}_{JS\|KL}(A\|\|B)$	Jensen–Shannon (JS) and Kullback–Leibler (KL) divergence between two probability distributions A and B

Introduction

Tormented by the cursed ambition always to put a
whole book in a page, a whole page in a sentence,
and this sentence in a word. I am speaking of myself[1]
Joseph Joubert (1754–1824), *Pensées, essais et maximes.*
Gallica http://www.bnf.fr

The need to summarize texts

Textual information in the form of digital documents quickly accumulates to huge amounts of data. Most of this large volume of documents is unstructured: it is unrestricted text and has not been organized into traditional databases. Processing documents is therefore a perfunctory task, mostly due to the lack of standards. Consequently, it has become extremely difficult to implement automatic text analysis tasks. Automatic text summarization (ATS), by condensing the text while maintaining relevant information, can help to process this ever-increasing, difficult to handle, mass of information.

Summaries are the most obvious way of reducing the length of a document. In books, abstracts and tables of content are different ways

[1] *"S'il est un homme tourmenté par la maudite ambition de mettre tout un livre dans une page, toute une page dans une phrase, et cette phrase dans un mot, c'est moi".*

of representing the condensed form of the document. But what exactly is a text summary? The literature provides several definitions. One definition states that the summary of a document is a reduced, though precise, representation of the text which seeks to render the exact idea of its contents. Its principal objective is to give information about and provide privileged access to the source documents. Summarization is automatic when it is generated by software or an algorithm. ATS is *a process of compression with loss of information*, unlike conventional text compression methods and software, such as those of the gzip family[2]. Information which has been discarded during the summarization process is not considered representative or relevant. In fact, determining the relevance of information included in documents is one of the major challenges of automatic summarization.

The summarization process

For human beings, summarizing documents to generate an adequate abstract is a cognitive process which requires that the text be understood. However, in a few weeks interval, the same person could write very different summaries. However, after an interval of several weeks, the same person can write very different summaries. This demonstrates, in part, the difficulty of automating the task. Generating a summary requires considerable cognitive effort from the summarizer (either a human being or an artificial system): different fragments of a text must be selected, reformulated and assembled according to their relevance. The coherence of the information included in the summary must also be taken into account. In any case, there is a general consensus that the process of summarizing documents is, for humans, a difficult cognitive task.

Fortunately, automatic summarization is an application requiring an extremely limited understanding of the text. Therefore, current systems of ATS have set out to replicate the results of the abstracting process and not the process itself, of which we still have a limited understanding.

[2] For more information, see http://www.gzip.org/.

Although great progress has been made in automatic summarization in recent years, there is still a great number of things to achieve.

From the user's perspective, people are not always looking for the same type of summary. There is also another type of user: automatic systems which use the results of a summarization system as the foundation for other tasks. Many different types and sources of documents exist (both textual and/or multimedia), such as legal, literary, scientific and technical documents, e-mails, tweets, videos, audio and images. As a result, there is no such thing as one type of summary. Sources and user expectations have prompted many applications to be created. Even for text documents, there is a large number of automatic summarization applications in existence (for people or machines):

– generic summarization;

– multi-document summarization;

– specialized document summarization: biomedical, legal texts, etc.;

– web page summarization;

– meeting, report, etc., summarization;

– biographical extracts;

– e-mail and e-mail thread summarization;

– news, rich site summary (RSS) and blog summarization;

– automatic extraction of titles;

– tweets summarization;

– opinion summarization;

– improving the performance of information retrieval systems, and so on.

Automatic text summarization

ATS became a discipline in 1958 following H.P. Luhn's research into scientific text summarization. Two or three important works [EDM 61, EDM 69, RUS 71] were completed before 1978, but they were followed by some 20 years of silence. In the early 1990s, however, the works of K. Spärck-Jones and J. Kupieck improved this landscape. Currently, ATS is the subject of intensive research in several fields, including natural language processing (NLP) and other related areas.

ATS has benefited from the expertise of a range of fields of research: information retrieval and information extraction, natural language generation, discourse studies, machine learning and technical studies used by professional summarizers. Answers have been found to several questions concerning ATS, but many more remain unsolved. Indeed, it appears that 50 years will not suffice to resolve all the issues concerning ATS. For instance, although generating a summary is a difficult task in itself, evaluating the quality of the summary is another matter altogether. How can we objectively determine that the summary of one text is better than another? Does a “perfect” summary exist for each document? What objective criteria should exist to evaluate the content and form of summaries? The community is yet to find answers to these questions.

About this book

Since 1971, roughly 10 books have been published about document summarization: half of these are concerned with automatic summarization. This book is aimed at people who are interested in automatic summarization algorithms: researchers, undergraduate and postgraduate students in NLP, PhD students, engineers, linguists, computer scientists, mathematicians and specialists in the digital humanities. Far from being exhaustive, this book aims to provide an introduction to ATS. It will therefore offer an overview of ATS theories and techniques; the readers will be able to increase their knowledge on the subject.

The book is divided into two parts, consisting of four chapters each.

– ☞ I) Foundations:
  - Chapter 1. Why Smmarize Texts?
  - Chapter 2. Automatic Text Summarization
  - Chapter 3. Single-Document Summarization
  - Chapter 4. Guided Multi-Document Summarization

– ☞ II) Emerging Systems:
  - Chapter 5. Multi- and Cross-Lingual Summarization
  - Chapter 6. Source and Domain-Specific Summarization
  - Chapter 7. Text Abstracting
  - Chapter 8. Evaluating Document Summaries

The conclusion and two appendices complete this book. The first appendix deals with NLP and information retrieval (IR) techniques, which is useful for an improved understanding of the rest of the book: text preprocessing, vector model and relevance measures. The second appendix contains several resources for ATS: software, evaluation systems and scientific conferences. A website providing readers with examples, software and resources accompanies this book: http://ats.talne.eu.

This book is first and foremost a pragmatic look at what is eminently an applied science. A coherent overview of the field will be given, though chronology will not always be respected.

Juan-Manuel TORRES-MORENO
Laboratoire Informatique d'Avignon
Université d'Avignon et des Pays de Vaucluse
France, August 2014