PART 1

Foundations

COPYRIGHTED MATERIAL

1

Why Summarize Texts?

In the 1780s, Joseph Joubert [1] was already tormented by his ambition to summarize texts and condense sentences. Though he did not know it, he was a visionary of the field of automatic text summarization, which was born some two and a half centuries later with the arrival of the Internet and the subsequent surge in the number of documents. Despite this surge, the number of documents which have been annotated (with Standard Generalized Markup Language (SGML), Extensible Markup Language (XML) or their dialects) remains small compared to unstructured text documents. As a result, this huge volume of documents quickly accumulates to even larger quantities. As a result, text documents are often analyzed in a perfunctory and very superficial way. In addition, different types of documents, such as administrative notes, technical reports, medical documents and legal and scientific texts, etc., have very different writing standards. Automatic text analysis tasks and text mining [2] [BER 04, FEL 07, MIN 02] as exploration, information extraction (IE), categorization and classification, among others, are therefore becoming increasingly difficult to implement [MAN 99b].

1.1. The need for automatic summarization

The expression "too much information kills information" is as relevant today as it has ever been. The fact that the Internet exists in multiple languages does nothing but increase the aforementioned

1. http://en.wikipedia.org/wiki/Joseph_Joubert

2. "Text mining, also referred to as text data mining, roughly equivalent to text analytics, refers to the process of deriving high-quality information from text. 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).

difficulties regarding document analysis. Automatic text summarization helps us to efficiently process the ever-growing volume of information, which humans are simply incapable of handling. To be efficient, it is essential that the storage of documents is linked to their distribution. 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]. However, unfortunately this is not always possible.

Summaries written by the authors of online documents are not always available: they either do not exist or have been written by somebody else. In fact, summaries can either be written by the document author, professional summarizers [3] or a third party. Minel *et al.* [MIN 01] have questioned why we are not happy with the summaries written by professional summarizers. According to the authors there are a number of reasons: "[...] because the cost of production of a summary by a professional is very high. [...] Finally, the reliability of this kind of summary is very controversial". Knowing how to write documents does not always equate with knowing how to write *correct* summaries. This is even more true when the source document(s) relate to a specialized domain.

Why summarize texts? There are several valid reasons in favor of the – automatic – summarization of documents. Here are just a few [ARC 13]:

1) Summaries reduce reading time.

2) When researching documents, summaries make the selection process easier.

3) Automatic summarization improves the effectiveness of indexing.

4) Automatic summarization algorithms are less biased than human summarizers.

3. The term "summarizer" will henceforth be used to refer to an agent (either a human or an artificial system) whose role is to condense one or several documents into a summary.

5) Personalized summaries are useful in question-answering systems as they provide personalized information.

6) Using automatic or semi-automatic summarization systems enables commercial abstract services to increase the number of texts they are able to process.

In addition to the above, the *American National Standards Institute*[4] (ANSI) [ANS 79] states that "a well prepared abstract enables readers to identify the basic content of a document quickly and accurately, to determine its relevance to their interests, and thus to decide whether they need to read the document in its entirety". Indeed, in 2002 the SUMMAC report supports this assertion by demonstrating that "summaries as short as 17% of full text length sped up decision-making by almost a factor of 2 with no statistically significant degradation in accuracy" [MAN 02].

1.2. Definitions of text summarization

The literature provides various definitions of text summarization. In 1979, the ANSI provided a concise definition [ANS 79]:

DEFINITION 1.1.– *[An abstract] is an abbreviated, accurate representation of the contents of a document, preferably prepared by its author(s) for publication with it. Such abstracts are useful in access publications and machine-readable databases.*

According to van Dijk [DIJ 80]:

DEFINITION 1.2.– *The primary function of abstracts is to indicate and predict the structure and content of the text.*

According to Cleveland [CLE 83]:

DEFINITION 1.3.– *An abstract summarizes the essential contents of a particular knowledge record, and it is a true surrogate of the document.*

4. http://www.ansi.org

Nevertheless, it is important to understand that these definitions describe summaries produced by people. Definitions of automatic summarization are considerably less ambitious. For instance, automatic text summarization is defined in the Oxford English dictionary [5] as:

DEFINITION 1.4.– *The creation of a shortened version of a text by a computer program. The product of this procedure still contains the most important points of the original text.*

Automatically generated summaries do not need to be stored in databases (unlike the ANSI summaries) as they are generated online in accordance with users' needs. After all, the main objective of automatic summarization is to provide readers with information and give him or her exclusive access to the source literature [MOE 00, MAN 01]. Nevertheless, summarizing text documents is a process of *compression*, which involves the loss of information. The process is automatic when it is carried out by an algorithm or a computer system. But what information should be included in the summary that is provided to the user? [SPÄ 93, SPÄ 97]. An intuitive answer would be that the generated summary must contain the most important and representative content from the source document. But how can we measure the representativeness and the significance of the information? This is one of the key questions automatic text summarization algorithms are trying to answer.

Karen Spärck-Jones and Tetsuya Sakai [SAK 01] defined 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.*

According to Radev *et al.* [RAD 00]:

DEFINITION 1.6.– *Text Summarization (TS) is the process of identifying salient concepts in text narrative, conceptualizing the*

5. http://www.oed.com.

relationships that exist among them and generating concise representations of the input text that preserve the gist of its content.

In 2002, Radev *et al.* [RAD 02a] introduced the concept of multidocument summarization and the length of the summary in their definition:

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 significantly less than that.*

According to Horacio Saggion and Guy Lapalme [SAG 02b], in terms of function, a summary is:

DEFINITION 1.8.– *A condensed version of a source document having a recognizable genre and a very specific purpose: to give the reader an exact and concise idea of the contents of the source.*

The ratio between the length of the summary and the length of the source document is calculated by the compression rate τ:

$$\tau = \frac{|\text{Summary}|}{|\text{Source}|} \qquad [1.1]$$

where $|\bullet|$ indicates the length of the document in characters, words or sentences. τ can be expressed as a percentage.

So what is the optimal value for the compression rate? The ANSI recommends that summaries be no longer than 250 words [ANS 79]. [BOR 75] indicate that a rate of $\tau = 10\%$ is desirable for summaries. In contrast, [GOL 99] maintain that the length of the summary is not connected to the length of the source text. Finally, [RAD 02a] and [HOV 05] specify that the length of the summary must be less than half of that of the source document. In fact, [LIN 99]'s study shows that the best performances of automatic summarization systems are found with a compression rate of τ = 15 to 30% of the length of the source document.

We are now going to introduce a definition 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 significant amount of relevant information from the source text. Its compression rate τ is less than a third of the length of the original document.*

Generating summaries demands that the summarizer (both human and algorithm) make an effort to select, reformulate and create a coherent text containing the most informative segments of a document. The notion of segments of information is left purposefully vague. Summaries can be guided by a particular profile, topic or query, as is the case for multidocument summarization [MAN 99a, MOR 99, MAN 01. Finally, coherence, cohesion as well as the temporality of the information presented must also be respected.

Many different types of document summarizations exist. There are two main reasons for this: first, there are many different types and sources of documents and, second, people have different uses for summaries and are, thus, not looking for the same type of document summarization. In any case, there is a general consensus that the process of summarizing a document requires a person to engage significant cognitive effort. To understand this process, it is interesting to take a look at how people produce summaries. What are the precise steps professional summarizers follow in their work? In his book, "The Art of Abstracting", Edward Cremmins [CRE 96] shows us that professional summarizers take 8 to 12 minutes to summarize a standard scientific article. Clearly, it requires considerably less time to summarize a text than to understand it. Other studies have backed up Cremmins' findings. [MIN 01] indicates that a professional summarizer – when he or she is a specialist in the field – can

6. *A summary is a text that is produced from one or more texts, that contains a significant portion of the information in the original text(s) and that is no longer than half of the original text(s).*

summarize a text of 10 or so pages in 10 or so minutes, though when the text is outside his or her area of expertise, it takes him or her around an hour to complete. In his studies about professional human summarizers [CRE 93, CRE 96], Cremmins identifies two phases in the summarization process: *local analysis* (the content of a sentence) and *global analysis* (content which is connected through several sentences).

Often both phases of the analysis are linked to the summarizer's prior knowledge of a topic; therefore, it is difficult to create an algorithm for this process [END 95a, END 95b, END 98]. Rath, Resnick and Savage published surprising results about professional summarizers [RAT 61]. Four professional summarizers produced summaries and there was a 25% overlap rate [7] between the summary and the source document; after a six month interval the overlap rate fell to 55% for a summary produced by the same summarizer. What made the summarizer change his or her sentence choice over these six months?

This question cannot be answered directly, but hypotheses about the variability and plasticity of language can be put forward. In Poibeau's words [POI 11]: "there is a number of ways of saying the same thing, or, at least, slightly different things, which will nevertheless be interpreted in an almost identical fashion [...] I can [...] express the same idea in two different ways after an interval of just one minute". This underlying plasticity is what makes languages so rich, yet it is simultaneously the reason they are so difficult to process automatically.

In simple terms, the human production of summaries involves two phases: understanding the source text and writing a concise and shortened version of it. Both understanding the text and the production of summaries require linguistic and extra-linguistic skills and knowledge on the part of the summarizer. An approximation of this process is illustrated in Figure 1.1.

7. The number of identical sentences between two documents.

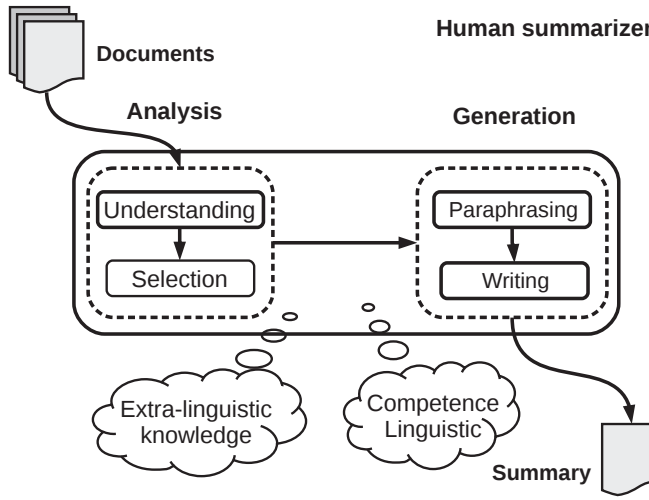

Figure 1.1. *Humans' production of summaries*

Is it worth modeling and trying to replicate the abstracting process of humans? Probably not. The mechanisms of this highly complex process are not very well understood and, what is more, the level of technology it requires is not currently available. Fortunately, other studies suggest that the results of the abstracting process can be approximated by algorithms. What is even more interesting is that automatic summarization (and the overall quality of content), though a long way off summaries produced by people, already do have exploitable properties. Better still, some algorithms can quickly and efficiently process large volumes of documents. Although people produce better summaries (in terms of readability, content, form and conciseness), automatic summarization is accessible, can be quickly distributed and costs very little. Automatic text summarization is an interesting alternative – rather than a replacement – to manual summarization [GEN 09b].

1.3. Categorizing automatic summaries

Summaries have two main functions: direct functions and indirect functions. The direct functions provide an overview of the document

(essential information), update the user (update summary), eliminate language barriers (multi- and cross-lingual summary) or facilitate information retrieval (IR). The indirect functions enable documents to be classified and indexed, keywords to be extracted and so on.

Summaries can be categorized according to different sets of criteria, such as their function, type and document source, etc. [MOE 00, MAN 01]:

– According to their function [VAS 63, EDM 69, HOV 05]:

- *indicative summary*: an indicative summary provides information about the topics discussed in the source document. It resembles a table of contents;

- *informative summary*: an informative summary aims to reflect the content of the source text, possibly explaining the arguments. It is a shorted version of the document.

– According to the number of documents for summarization:

- single document: a summary of one document;

- multidocument: a summary of a not necessarily heterogeneous group of documents often about a specific topic.

– 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.);

- *literary*; a summary of narrative documents, literary texts, etc;

- *encyclopedic*: a summary of encyclopedic documents, such as Wikipedia;

- *social networks*: a summary of blogs and very short documents (such as tweets).

– According to type:

- *extract*: an assembly of fragments from the original document;

- *abstract*: summarizing by reformulating: a summary produced by rewriting and/or paraphrasing the text;

- *sentence compression*: a summary containing the same number of sentences as the original text, but with a reduced sentence length.

– According to the type of summarizer:

- *author summary*: a summary written by the author of the document which reflects 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 field, 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. Therefore, update summaries only show important new information and avoid repeating information.

– According to the target audience:

- *without a profile*: a summary which is independent of the needs and profile of the user; the summary is based uniquely on information from the source documents;

- *based on a user profile*: summaries targeted at users interested in a specialized domain (chemistry, international politics, sports, the economy, etc.).

State-of-the-art systems are able to generate indicative and informative summaries. Informative summaries are considerably harder to produce than indicative summaries as they require that the information contained in the source text be properly understood, generalized, organized and synthesized. However, they can be

8. *A Generic summary provides the author's point of view, while a Query-based summary focuses on material of interest to the user: http://www.cs.cmu.edu/ref/mlim/chapter3.html.*

substituted at the source, unlike indicative summaries which merely enable an idea of the content of the texts to be produced.

[GEL 99] defined another type of summary: topic selection summary. This type of summary is supposed to generate a succinct report centered around the topics – rather than the ideas – discussed in a document. For instance, one document may talk about science, technology and transport whereas another may discuss politics, war and so on.

Currently, the scientific community is conducting research into the production of multidocument (mono- or multilingual) summaries guided by a query, update summaries and summaries of specialized domains. International campaigns run by the *National Institute of Standards and Technology* (NIST)[9], *Document Understanding Conferences* (DUC) held from 2001 to 2007 and *Text Analysis Conferences* (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 corpora according to tasks and evaluating participating systems.

Figure 1.2 shows a basic approximation of the process of generating automatic text summaries. The source text (single or multidocument) is processed by a summarization system according to a query and certain parameters, such as the compression rate, the desired type of summary and the desired function. The system produces an extract, an abstract or a summary via compression. The summarization module in the center of the figure will be discussed in more detail in the following chapter.

1.4. Applications of automatic text summarization

There are countless applications of automatic text summarization. Here are just a few:

– increasing the performance of traditional IR and IE systems (a summarization system coupled with Question-Answering (QA) system);

9. http://www.nist.gov.

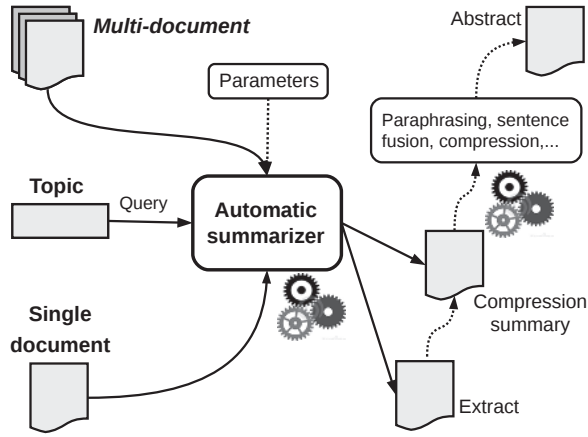

Figure 1.2. *Simplified abstracting process. The source documents and the topic go in. The parameters are the compression rate τ, the type, etc. Extracts, abstracts or sentence compression summaries come out*

– 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.;

– meeting summarization;

– biographical extracts [MCK 01, DAU 02, HAR 03];

10. "RSS (Rich Site Summary); originally RDF Site Summary; often dubbed Really Simple Syndication, uses a family of standard web feed formats to publish frequently updated information: blog entries, news headlines, audio and video. An RSS document (called "feed", "web feed" or "channel") includes full or summarized text, and metadata, like publishing date and author's name" (source: Wikipedia: http://en.wikipedia.org/wiki/RSS).

– automatic extraction and generation of titles [BAN 00];

– domain-specific summarization (domains of medicine, chemistry, law, etc.) [FAR 05, BOU 08c];

– opinion summarization [11] [LIU 09], etc.

In addition to these applications, the scientific challenge of replicating a difficult cognitive task, namely the human abstracting process, must also be mentioned. In fact, research aims to make a machine understand a text written in natural language. This would enable, at least theoretically, real – relevant, correct, coherent and fluid – abstracts of source documents to be produced.

1.5. About automatic text summarization

The discipline of automatic text summarization originated with Hans Peter Luhn's research in the 1950s. Already worried about the ever-growing volume of information available at the time, Luhn built a summarizer for scientific and technical documents in the field of chemistry. After this pioneering research, several founding works were written including Edmundson and Wyllys' research in 1961 and 1969 [EDM 61, EDM 69], Rush *et al.'s* work in 1971–1975 [RUS 71, POL 75] and Gerald Francis DeJong's studies in 1982 [DEJ 82]. In 1993, research took off once again with work by Spärck-Jones [12] [SPÄ 93] and Julian Kupiec *et al.* [KUP 95]. This research helped to respark an interest in automatic summarization.

Though incomplete, Figure 1.3 shows the landmark works, books and conferences in the field of automatic summarization. This work will look at these accomplishments in more detail. Below, in italics, is a list of outstanding works in the domain of automatic text summarization:

– "Summarizing Information" by Endres-Niggemeyer [END 98];

11. See TAC 2008 *Opinion Summarization Task*: http://www.nist.gov/tac/2008/summarization/op.summ.08.guidelines.html.

12. Karen Spärck-Jones (born August 26, 1935 in Huddersfield, Yorkshire, died April 4, 2007) was a British scientist specializing in NLP. She was one of the pioneers of IR.

– "Advances in Automatic Text Summarization" by Mani and Mayburi [MAN 99a];

– "Automatic Indexing and Abstracting of Document Texts" by Moens [MOE 00];

– "Automatic Summarization (Natural Language Processing)" by Mani [MAN 01];

– "Automatic Summarization" by Nenkova and McKeown [NEN 11].

TIPSTER and SUMMAC's campaigns in 1997 and 1998 [MAN 02] as well as the NII Testbeds and Community for Information access Research (NTCIR) workshop held in Japan in 2001 and 2002 enabled automatically produced summaries to be evaluated on a large scale (see section 8.4). In 2005, the *Multilingual Summarization Evaluation* (MSE) [13] conducted an international evaluation of Arabic and English summaries. NIST conferences, DUC (2001–2007), TAC from 2008 and CLEF-INEX from 2011 have taken over this role (see section 8.6).

Automatic text summarization is currently the subject of intensive research, particularly, though not exclusively, in Natural Language Processing (NLP). In fact, automatic summarization has benefited from the expertise of several related fields of research (see Figure 1.4) [BRA 95, IPM 95, MAY 95, MCK 95b, SPÄ 95]. Among these fields of research, computer science (understood in a broad and transversal sense), artificial intelligence, and more specifically its symbolic and cognitive methods, have helped summarization [DEJ 79, DEJ 82, ALT 90, BAR 99]. Cognitive sciences have studied the abstracting process [FAY 85, LEM 05]. [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]. Finally, [CRE 93, END 95a, END 95b,

13. http://en.wikipedia.org/wiki/DARPA_TIDES_program.

CRE 96, END 98]'s studies have clarified the techniques employed by professional summarizers when generating summaries.

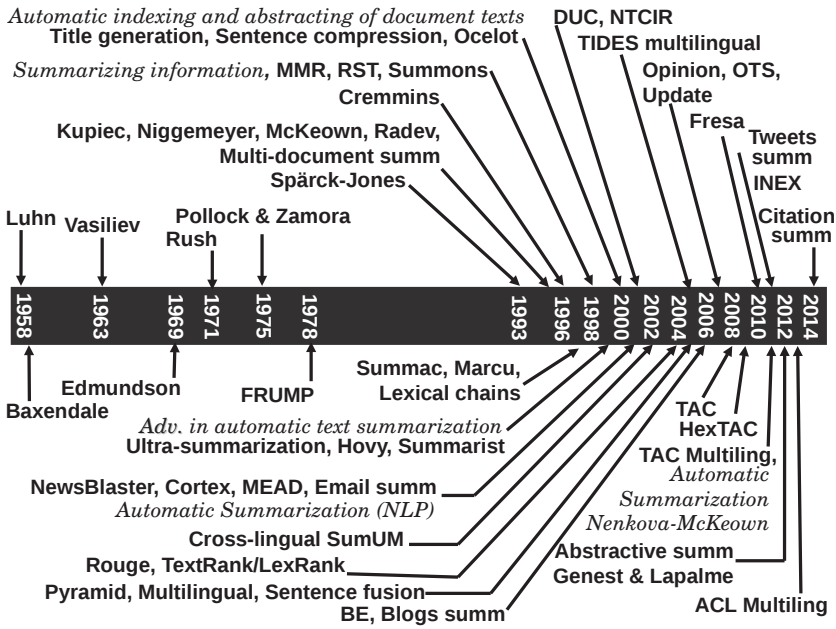

Figure 1.3. *Highlights of automatic text summarization*

One final important point must also be considered with regard to automatic summarization. Knowing how to construct efficient systems which produce summaries is not enough: it is also necessary to know how to evaluate the *quality* of the summaries generated. Therefore, coming up with a clear definition of quality in relation to automatic summarization is essential. In IR (such as web search engines), the performance of a system is evaluated by the relatively objective measures of precision, recall and F-measure. This enables the relevance ranking of documents retrieved by a system in relation to users' queries to be quantified.

However, in automatic text generation, sentence compression and automatic text summarization, what goes in and comes out of the

system are documents written in natural language. Therefore, it is difficult to evaluate the relevance of the results. In the concrete case of text summarization, how can we know if the summary produced by an algorithm is correct? How can we determine whether one algorithm produces higher quality summaries than another? Worse still, what does *correct* mean in this applied area of science? What is the ratio of quality and the length, content and form of summaries? In recent years, the scientific community has attempted to develop ways of evaluating quality, but to date no definitive answers have been found. This topic will be picked up in Chapter 8.

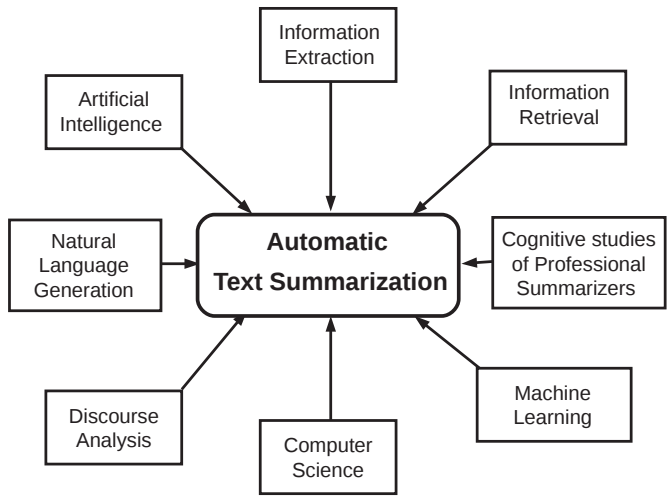

Figure 1.4. *Fields of research which have influenced the development of automatic text summarization*

In a little homage to J. Joubert, who was tormented by the ambition to condense a whole book into a page, we have created two summaries of this work. The first (see Figure 1.5) is an author summary of 154 words (compression rate of approximately 0.12 %). The second (Figure 1.6) summary has been generated by a system from (`untex`-ted) LaTeX sources of this book. The text as a whole (apart from the tables, figures and appendices) were automatically produced by the CORTEX

algorithm (see section 3.7). The summary was generated in less than 6 s and did not undergo any manual processing. [14]

AUTOMATIC TEXT SUMMARIZATION

Automatic text summarization is a discipline of Natural Language Processing (NLP) which aims to condense text documents. Condensing the text signifies producing a shortened version of the source document which contains the main points of this document. The process involves the loss of information. The author will provide an overview of approaches adopted over time, from Luhn in 1958 to present. Several methods which have been proposed to resolve the issues of automatic summarization will be studied: single and multi-document summarization algorithms, cross- and multi-lingual summarization, specialized summarization, tweets and email summarization, automatic sentence compression and rhetorical analysis summarization. The open-ended and difficult issue of summary evaluation will also be discussed, making reference to both manual and automatic approaches. Several applications of automatic text summarization will also be shown. This work is aimed at people interested in the techniques and algorithms of NLP (students, researchers, linguists, computer scientists, mathematicians, engineers etc.).

Figure 1.5. *Author summary of this work (8 sentences, 154 words, compression rate* $\tau = 0.17\ \%$*)*

According to the GNU/Linux `wc` command, the source document contained 3,982 sentences and 71,143 words. The compression rate was fixed at seven sentences (307 words), representing a rate of approximately 0.27%.

14. Conducted on a machine with GNU/Linux v13.10, CPU Intel Core^{TM} i5-3317U$\times$4, 1.70GHz and 4 Gb of RAM.

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. The extract summarizer is composed of modules of analysis, extraction and summary generation. In comparison to single-summarization MDS, new difficulties arise with MDS: clustering similar documents, designing and implementing anti - redundancy measures, high compression rates meaning that sources are aggressively condensed, taking into account the temporality of texts and correctly resolving anaphoric references, among others. Formally, the problem of MDS can be expressed as follows: given a topic Q and a set of D relevant documents, the DUC task consists of generating a short, coherent and well - organized summary which answers the questions asked by the topic. The third NTCIR workshop centered on evaluating IE systems, Question - Answering systems and ATS systems. Users are asking for increasingly higher performing ATS systems which produce quality personalized summaries, are able to process large volumes of heterogeneous documents which over time have become multilingual and, moreover, that the task be completed quickly. There are multiple applications for ATS: news, RSS, email, web page and blog summarization; specialized document, meeting and report summarization; biographic extracts; opinion summarization; automatic title generation etc.

Figure 1.6. *Automatic summary of this work generated by the* CORTEX *system (7 sentences, 224 words, compression rate* $\tau = 0.17\ \%$*)*

An automatic evaluation of the quality of the summaries [15] gives the artificial summary a score of 0.00965 and the author summary a score of 0.00593. The reader, of course, has the final say and can judge for himself or herself the relevance of the summaries produced.

15. See FRESA automatic evaluation in section 8.9.2. The FRESA score falls between $[0, 1]$: 0 indicates a bad summary and 1 indicates that the summary is very similar to the source text.

1.6. Conclusion

Until we have found a solution to the difficult task of making a machine automatically understand text, automatic summarization will remain a – poor – approximation of human summarization. That said, the automatic text summarization methods available have come a long way and achieved significant goals. It is likely to take at least another 50 years to resolve all the issues concerning text summarization. So we can ask ourselves the following questions. How sophisticated should automatic summarization systems be? Is understanding fundamental to the generation of "real" summaries created by reformulating sentences (and not merely providing an extract)? Can statistical methods accomplish this task or is something else required? What should the role of linguistic tools and resources be in automatic text summarization systems? How can we break out of the extraction stage and start to generate real abstracts? How can we objectively measure the overall quality and sensitivity of the content of summaries and the automatic systems that produce them?

This work aims to show how the scientific community has attempted to tackle and find answers to these difficult questions. The answers found are both theoretical – formulating models – and practical – developing technological artifacts and real automatic text summarization systems which are applied to concrete tasks.