Automatic Text Summarization: Past, Present and Future

Horacio Saggion, Thierry Poibeau

Chapter 1
Automatic Text Summarization: Past, Present and Future

Horacio Saggion and Thierry Poibeau

Abstract Automatic text summarization, the computer-based production of condensed versions of documents, is an important technology for the information society. Without summaries it would be practically impossible for human beings to get access to the ever growing mass of information available online. Although research in text summarization is over fifty years old, some efforts are still needed given the insufficient quality of automatic summaries and the number of interesting summarization topics being proposed in different contexts by end users ("domain-specific summaries", "opinion-oriented summaries", "update summaries", etc.). This paper gives a short overview of summarization methods and evaluation.

1.1 Introduction

Automatic Text Summarization, the reduction of a text to its essential content, is a very complex problem which, in spite of the progress in the area thus far, poses many challenges to the scientific community. It is also a relevant application in todays information society given the exponential growth of textual information online and the need to promptly assess the contents of text collections.

Research in automatic summarization started to attract the attention of the scientific community in the late fifties [35], when there was a particular interest in the automation of summarization for the production of abstracts of technical documentation. The interest in the area declined for a few years until Artificial Intelligence started to show interest in the topic [11].

Horacio Saggion
Department of Information and Communication Technologies, Universitat Pompeu Fabra, C/Tanger 122, Barcelona 08018, Spain e-mail: horacio.saggion@upf.edu

Thierry Poibeau
Laboratoire LaTTiCe-CNRS, École Normale Supérieure and Université Sorbonne-Nouvelle, 1 rue Maurice Arnoux, 92120 Montrouge, France e-mail: thierry.poibeau@ens.fr

It has long been assumed that summarization presupposes to understand the input text, which means that an explicit (semantic) representation of the text must be calculated so as to be able to identify its essential content. Therefore, text summarization became an interesting application to test the understanding capabilities of artificial systems. However, the interest for this approach to text summarization decreased rapidly given the complexity of the task, and text understanding became an open research area in itself.

There was a resurge of interest in summarization in the nineties with the organization of a number of relevant scientific events [72] and a peak of interest from the year 2000 with the development of evaluation programs such as the Document Understanding Conferences (DUC) [48] and the Text Analysis Conferences (TAC) [50] in the United States. These frameworks will be further detailed in section 4 dedicated to evaluation.

Two fundamental questions in text summarization are; *i)* how to select the essential content of a document, and *ii)* how to express the selected content in a condensed manner [70, 72]. Text summarization research has many times concentrated more on the product: the summary, and less on the cognitive basis of text understanding and production that underly human summarization. Some of the limitations of current systems would benefit from a better understanding of the cognitive basis of the task. However, formalizing the content of open domain documents is still a research issue, so most systems are only based on a selection of sentences from the set of original documents.

Where the transformation of the original text content into a summary is concerned, there are two main types of summaries: an *extractive* summary, a set of sentences from the input document and an *abstractive* summary (i.e., an abstract), a summary in which some of its material is not present in the input document [36]. Summaries can also be classified into *indicative* or *informative* depending on their intended purpose to *alert* or to *inform* respectively.

Early research in summarization concentrated on summarization of single documents (i.e., *single document summarization*), however in the current Web context, many approaches focus on summarization of multiple, related documents (i.e., *multi-document summarization*). Most summarization algorithms today aim at the generation of extracts given the difficulties associated to the automatic generation of well formed texts in arbitrary domains. It is generally accepted that there are a number of factors that determine the content to select from the source document and the type of output to produce [71], for example factors such as the audience (e.g., expert vs non-expert reader) certainly influences the material to select.

This short introduction overviews some classic works on content selection and realization, summarization tools and resources, and summarization evaluation. For the interested reader, there are a number of papers and books that provide extensive overviews of text summarization systems and methods [36, 23, 33].

1.2 Overview of Summarization Methods

Most summarization today is based on a sentence-extraction paradigm where a list of features believed to indicate the relevance of a sentence in a document or set of documents [36] is used as text interpretation mechanism. The approaches to sentence selection can be driven by statistical information or based on some informed summarization theory, which takes into account linguistic and semantic information.

1.2.1 Superficial Techniques

Basic statistical approaches to content selection have relied on the use of frequency computation to identify relevant document keywords [35]. The basic mechanism used is to measure the frequency of each word in a document and to adjust the frequency of words with additional corpus evidence such as the inverse document frequency of the word in a general document collection. This helps boost the score of words which are not too frequent in a general corpus and moderate the score of otherwise too frequent general words. These methods assume that the relevance of a given concept in a text is proportional to the number of times the concept is "mentioned" in the document, assuming that each distinct word corresponds to a different concept. However, counting concept occurrences in text is not a trivial task given the presence of synonymy (e.g., "dog" and "puppy" could refer to the concept dog) and coreferential expressions (e.g., "Obama" and "the President" could refer to the same individual) which contribute to text cohesion. Once keywords are identified in the document, sentences containing those keywords could be selected using various sentence scoring and ranking mechanisms (e.g., the relevance of a sentence could be proportional to the number of keywords it contains). More sophisticated techniques than using simple word counts also exists: for example topic signatures were proposed in [32] as a way to model relevant keywords in particular domains and as interpretation and sentence relevance determination mechanism.

The position of sentences in text [12, 30] is also deemed indicative of sentence relevance. For example in news stories, the first or leading paragraph usually contains the main information about the event reported in the news, while the rest of the text gives details as well as background information about the event, therefore selecting sentences from the beginning of the text could be a reasonable strategy. However, in scientific texts the introduction usually gives background information, while the main developments are reported in the conclusions, so the position strategy needs to be adapted to the text type to be summarized.

Another superficial, easy-to-implement method consists in measuring the relevance of sentences by comparing them with the title of the document to be summarized, a sentence containing title words could be regarded as relevant and the more title words a sentence has the more relevant the sentence would be [12]. Yet another method looks at the presence of very specific cue-words or expressions in sentences

[12, 51] such as ”in conclusion”, ”to summarize”, or ”the objective” which could be regarded as pointing to relevant information out-of-context.

These now classical features are usually applied following an Edmundsonian empirical approach [12] or a machine learning framework. In the Edmundsonian approach a kind of linear combination of sentence features to score sentences is usded. In a machine learning approach [26], sentence extraction is seen as sentence classification where sentences in the document have to be classified as either “summary sentence” or “non summary sentence”. Superficial methods such as frequency, position, title, and cue-words are incorporated in one way or another in text summarization systems, even though the summaries they produce are far from those a human would produce.

Superficial techniques are also those based on graph representations [43, 14] which are so popular nowadays. These approaches compute sentence similarity values and use graph algorithms (e.g., random walks or PageRank) to weight word or sentence relevance to make decisions about sentence centrality. Graph-based approaches were also explored earlier for single document summarization of encyclopædic articles [69].

1.2.2 Knowledge-based Approaches

Knowledge-rich approaches either extend basic methods by the incorporation of sophisticated, yet general lexical resources or apply discourse organization theories in generic or specific contexts. Few approaches bring domain knowledge to the summarization enterprise. One of the best known artificial intelligence approaches to summarization was DeJong’s FRUMP system [11] which was based on text understanding and generation mechanisms. More specifically, the system was supposed to map chunks of text to rich conceptual structures such as sketchy scripts, which contained information about the key elements of specific story types (e.g., earthquakes, demonstrations). A main drawback of early artificial intelligence approaches was that they relied on manual, intensive coding of world knowledge which made such approaches impractical.

The situation is nowadays being corrected to some extent with studies into knowledge induction from text corpora. For example, in [28] clustering is applied to generate templates for specific entity types (actors, companies, etc.) and patterns are automatically produced that describe the information in the templates. In [8] narrative schemas are induced from corpora using coreference relations between participants in texts. Participants’ roles are identified and typical verb sequences in stereotypical situations are extracted. Induction results of these approaches have still to be incorporated and tested in summarization systems.

Lexical cohesion has long been considered a key component in assessing content relevance in text summarization. In [5] cohesive links between sentences were identified, based on the use of a thesaurus, and the resulting linked structure was used to select sentences. In [2], lexical chains that connect related words were created based

on WordNet relations [15], then sentences were selected depending on which chains sentences' words belong to.

A text is a complex linguistic unit, therefore many works rely on discourse structure or text organization theories for text interpretation and "sound" sentence selection. For example, in scientific domains, it is well known that texts follow a more or less predefined semantic or rhetorical organization [74]. Various works have identified categories of information such as "purpose", "method", "results", "conclusions" which make up the structure of scientific/technical documents and which are also present in research abstracts. In [29], a full model of summaries of empirical research was developed with 36 information types organized in a three-level hierarchy. The top of the hierarchy contains types of information which are very frequent in abstracts while the other two levels contain less frequent types of information. Automatic identification of some information types in textual input is essential to produce good quality abstracts. In [75], a Naïve Bayes machine learning algorithm was used to identify semantic information such as "Aim", "Background", "Contrast", etc. The method is similar to [26] although the problem is more challenging. In [64] a semantic dictionary and manually developed syntactic/conceptual patterns were used to identify sentences, instantiate templates, and generate indicative-informative abstracts.

In restricted domains, Information Extraction [16], the mapping of textual input into predefined template structures, can be used to extract information and then generate a summary. The approach has been applied in [52] where the objective was to produce short indicative informative abstracts (i.e., non-extractive summaries) and in [56] that uses templates instantiated from different articles referring to the same event to generate a coherent multi-document summary.

General discourse organization theories have also been used in summarization. Rhetorical Structure Theory (RST) [39] establishes that a text can be represented as a tree structure (i.e., rhetorical tree) linking text spans by using a set of predefined discourse relations. Because the rhetorical relations link text spans with different informational status (i.e., nucleus and satellite), text elements could be selected depending on their role in a relation. In [47] the rhetorical tree has been used to select sentences by pruning the tree while in [40] text units were promoted based on their status. However, analyzing discourse organization is still an open research issue and these kinds of techniques still need to be more robust in order to be more widely used.

1.2.3 Non-extractive Methods

In text summarization research, most of the attention has been paid to the problem of selecting information, whilst the problem of generating a new cohesive and coherent text has somehow received less attention. Non-extractive summarization in the current context refers to problems such as sentence compression, headline generation, cut-and-paste summarization, and sentence regeneration.

In headline generation [77] the objective is to produce a short title for a news story, this problem has been addressed following the statistical machine translation paradigm, where words have been selected from the document and then have been combined using a language model. In [78], two ways of generating headlines have been proposed, one based on the pruning of a syntactic tree which is supposed to generate well-formed headlines, a second one based on a Hidden Markov Model which is more robust.

Together with headline generation, sentence compression has received considerable attention, the objective here is for example to reduce sentence size by eliminating components which might be unnecessary. In [22], compression was carried out by removing sentence components using manually developed rules operating on syntactic trees which consider syntactic restrictions (e.g., obligatory verb arguments) and contextual information. In [25], a noisy-channel model which learns based on an aligned corpus how to translate uncompressed to compressed sentences. The method only dealt with removal of elements and used a very small corpus of compressed sentences, it was therefore extended in [76]. In professional abstracting settings [13], it was observed that in order to produce abstracts, textual fragments are usually combined to create new sentences and sometimes new linguistic material is included. Simulation of text abstracting operations have been implemented using rule-based and machine learning approaches in [21, 61, 62].

1.2.4 Multi-document Summarization

A multi-document summary is a brief representation of the essential contents of a set of related documents. The relation between the documents can be of various types, for example documents can be related because they are about the same entity, or because they discuss the same topic, or because they are about the same event or the same event type.

Fundamental problems when dealing with multi-source input in summarization are the detection and reduction of redundancy as well as the identification of contradictory information. In [56], the multi-document summarization problem is studied in the context of multi-source information extraction in specific domains. Here templates instantiated from various documents are merged using specific operators aiming at detecting identical or contradictory information.

In an information retrieval context, where multi-document summaries are required for a set of documents retrieved from a search engine in response to the query, the Maximal Marginal Relevance (MMR) method [7] can be applied. The method scores text passages (e.g., paragraphs) iteratively taking into consideration the relevance of each passage to a user query and the redundancy of the passage with respect to summary content already selected. In the case of generic summarization, computing similarity between sentences and the centroid of the documents to summarize has resulted in competitive summarization solutions [55, 63]. Do-

main specific multi-document summarization techniques are explored in a chapter on summarizing images in this volume (cf. chapter **??**).

Sentence ordering is also an issue for multi-document summarization. In single-document summarization it is assumed that presenting the information in the order this information appears in the input document would generally produce an acceptable summary. By contrast, in multi-document summarization particular attention has to be paid to how sentences extracted from multiple sources are going to be presented. Various techniques exist for dealing with sentence ordering, for example if sentences are timestamped by publication date, then they could be presented in chronological order. However this is not always possible because recognizing the date of a reported event is not trivial and not all documents contain a publication date.

Sentence ordering can also be conceived as to represent the different topics to be addressed in the summary. For example, a clustering algorithm can be used to identify topics in the set of input documents and discover in what order the topics are presented in the input documents [3], this in turn could be used to present sentences in an order similar to that observed in the input set. A probabilistic approach to sentence ordering seeks to estimate the likelihood of a sequence of sentences. It tries to find a locally optimal order by learning ordering constraints for pairs of sentences [27]. An entity-grid model [1] tries to represent coherent texts by modelling entity roles (e.g., subject, object) in consecutive sentences, the model is able to discriminate coherent and incoherent texts. Advanced techniques improving on these methods are presented in this volume **??**.

1.2.5 Multilingual Summarization

While most summarization research so far has been carried out for the English language probably because of the availability of data and evaluation resources, summarization in languages other than English is not rare.

Activities to promote summarization in Japanese have been undertaken in the Text Summarization Challenges [46], a series of evaluations of text summarization systems with tasks such as single document summarization and topic focused multi-document summarization. Summarization in a cross-lingual environment was studied in the 2001 Johns Hopkins research workshop [65] where evaluation resources and summarization algorithms for English and Chinese were developed [54]. The 2005 Multilingual Summarization Evaluation concentrated on summarization from mixed input in Arabic and English, where the challenge was to generate output from automatic translations [58].

Various research projects have produced multilingual summarization technology based on features already used for the English language: the SUMMARIST research project has produced a summarizer available for Korean and Spanish, the SweSum summarizer works for Scandinavian languages and the MUSE system is a language independent summarizer which has been tested for Arabic and Hebrew. The 2011

Text Analysis Conference in its text summarization evaluation program has included a pilot task on Multilingual Summarization which objective is to evaluate the application of language independent summarization algorithms. In the multilingual task ten clusters with ten documents (in Arabic, Czech, English, French, Greek, Hebrew, and Hindi) were produced for evaluation, and each system had to produce summaries in at least 4 different languages. Moreover, various researchers argue their summarizers to be language independent [42, 59, 24].

1.3 Summarization Resources

1.3.1 Text Summarization Tools

From a commercial point of view, there are a number of software products offering summarization capabilities (e.g., Microsoft Autosummarize option) as well as a number of companies offering standalone summarization applications (e.g., Copernic, Pertinence). From the research point of view, a few tool-kits exist, probably the best known text summarization software for research available today is MEAD [54] a set of Perl components for summarization of English as well as other languages such as Chinese. MEAD implements classical sentence relevance features such as position and term frequency, but also implements multi-document summarization features such as centroid. MEAD has been used many times for comparison purposes in text summarization research. SUMMA [60] is another available tool which relies on the GATE framework [41] for document processing and text representation. It is based mainly on statistical approaches such as frequency computation, cue-word identification, position, title, centroid, etc. It also exploits the vector space model to represent documents, sets of documents, sentences, and other textual units to be able to compute text relevance based on similarity measures between text units. SUMMA is implemented in Java and used within GATE as a plug-in or as a Java library for standalone applications. Both MEAD and SUMMA systems need to be adjusted for optimal performance therefore requiring training data to set up parameters.

1.3.2 Text Summarization Datasets

Data is fundamental in any scientific activity, and it is of paramount relevance in text summarization. Without data it will be impossible to formulate working hypotheses, verify them, and adjust system parameters. Over the past few years, various datasets have been produced for the study of text summarization, among them datasets pertaining to the various evaluation frameworks we mention in Section 1.4 (mainly The Document Understanding Conference and the Text Analysis Conference).

SummBank is a dataset of parallel English and Chinese documents containing, in addition to source documents, multi-document summaries for sets of related stories (i.e., 40 clusters), relative utility judgements for sets of sentences (e.g., an indication of how valuable the sentence is for a summary), and automatic summaries [66]; it has been used in large scale evaluation experiments [57]. The CAST corpus [18] contains a set of documents where each sentence and sentence fragment has been annotated as either essential or non essential; this kind of dataset could be of help for developing sentence selection and sentence reduction algorithms. The Ziff-Davis corpus [17] which has been partially used for experiments in cut-and-paste summarization [20] and sentence reduction [25], contains newspaper articles and their human written abstracts. The abstracts were automatically sentence aligned to their source documents therefore being of value for studying non-extractive summarization.

1.4 Evaluation

A good summary must be easy to read and give a good overview of the content of the source text. Since summaries tend to be more and more oriented towards specific needs, it is necessary to tune existing evaluation methods accordingly. Unfortunately these needs do not give a clear basis for evaluation and the definition of what is a good summary remains to a large extent an open question.

Therefore, the evaluation of human or automatic summaries is known to be a difficult task. It is difficult for humans, which means the automation of the task is even more challenging and hard to assess. However, because of the importance of the research effort in automatic summarization, a series of proposals have been made to partially or fully automate the evaluation [73, 38]. It is also useful to note that in most cases automatic evaluations already correlate positively with human evaluations.

1.4.1 Evaluating Automatically Produced Summaries

A series of evaluation campaigns have been organized since the late 1990 in the US, which provided a forum for evaluation and discussion. These campaigns are essentially SUMMAC (1996-1998) [37], DUC (the Document Understanding Conference, 2000-2007) [49] and more recently TAC (the Text Analysis Conference, 2008-) [45]. Evaluation in these conferences is based on human as well as automatic scoring of the summaries proposed by participants. Therefore, these conferences have played a major role in the design of evaluation measures; they also play a role in the meta-evaluation of scoring methods since it is possible to check to what extend the scores obtained automatically correlate with human judgements.

Broadly, one can say that there are three main difficulties in the automatic evaluation of summaries: *i*) it is necessary to determine what are the most important pieces of information that should be kept from the initial text; *ii*) evaluators must be able to automatically recognize these pieces of information in the candidate summary since this information can be expressed using various expressions; *iii*) lastly, the readability (including grammaticality and coherence) of the summary should be evaluated.

In this section we mainly refer to evaluation methods that can be applied to extractive summaries. These summaries are made of extracts from the original text, which means it is possible to evaluate their quality by comparing their content (i.e. sequences of words) to reference summaries. If this comparison is not possible (for example in the case of abstractive summaries), then manual methods are the only reliable solution within the current state of the art.

Even for extractive summaries, evaluation methods range from purely manual approaches to purely automatic ones, and there are of course a lot of possibilities in between. Manual approaches refer to methods where a human evaluates a candidate summary from different points of view, for example coverage, grammaticality or style; this kind of evaluation is necessary but is known to be highly subjective. Automatic approaches compare segments of texts from the candidate summary with one or several reference abstracts; this approach is easy to reproduce but cannot be applied when the system uses reformulation techniques. Mixed approaches allow one to manually analyse and score the most important pieces of information and rank the candidate summaries according to these (the most important pieces of information must be contained in the candidate summary, independently of their linguistic formulation).

A lot of diverse approaches for summary evaluation have been proposed in the last two decades. This prevents us from being comprehensive. We will instead focus on three methods that have been widely used during recent evaluation campaigns (esp. during the last Text Analysis Conferences organized by NIST [10]): ROUGE (a fully automatic method), PYRAMID (a mixed method) and a series of indicators resulting from a manual evaluation. Lastly, we will consider a recent body of research aiming at evaluating summaries with no human reference.

1.4.2 Manual Methods

The most obvious and simple way to evaluate a summary is to have assessors evaluating its quality. For example for DUC, the judges had to evaluate the coverage of the summary, which means they had to give a global score assessing to what extent the candidate summary covers the text given as input. In more recent frameworks, and especially in TAC, query-oriented summaries have to be produced: judges then have to evaluate the responsiveness of the summary, that is to say to what extent a given summary answers the query given in input.

Manual evaluation can also provide some indicators to assess the quality and readability of a text. A good summary is supposed to be:

- syntactically accurate;
- semantically coherent;
- logically organized
- without redundancy.

These different points are too complex to be fully automatically calculated, especially semantic coherence and logical organization. In order to get a reliable evaluation of these different aspects, it is necessary to get human judgements. For the DUC [49] and TAC [10] campaigns, human experts had to give different scores to each candidate summary, using the following indicators:

- grammaticality;
- non redundancy;
- focus (integration of the most important pieces of information of the original text);
- structure and coherence.

Experts had to give a grade between 0 (void) and 10 (perfect) for each of these indicators. For TAC 2009, reference summaries written by human experts got an average grade of 8.8/10 (TAC did not provide the figure for each of the criteria in isolation). Thus, this grade can be seen as the upper bound score reachable by candidate summaries.

1.4.3 Automatic Methods Based on a Comparison with a Manual Reference

Since the early 2000s, a series of measures have been proposed to automate the evaluation of summaries. Most of these measures are based on a direct comparison of the produced and the reference summaries [67, 57]. We detail here the two most popular measures: Rouge and Pyramid.

1.4.3.1 ROUGE

The ROUGE measures (*Recall-Oriented Understudy for Gisty Evaluation*) have been introduced by [31]. These measures are based on the comparison of n-grams (i.e. a sequence of n elements) between the candidate summary (the summary to be evaluated) and one or several reference summaries[1]. Most of the time, several reference summaries are used for comparison, which allows more flexibility and a fairer

[1] Rouge was inspired by BLEU, a measure used in the evaluation of machine translation also based on the comparison of n-grams [53].

evaluation. There are several variants of ROUGE and we present the most widely used of them below.

- **ROUGE-*n***: This measure is based on a simple comparison of n-grams (most of the time a sequence of 2 or 3 elements, more rarely 4). A series of n-grams (hence series of sequences of n consecutive words) is extracted from the reference summaries and the candidate summary. The score is the ratio between the number of common n-grams (between the candidate summary and the reference summary) and the number of n-grams extracted from the reference summary only.
- **ROUGE-*L***: In order to overcome some of the shortcomings of ROUGE-*n*, more precisely the fact that the measure may be based on too small sequences of text, ROUGE-L takes into account the longest common sequence between two sequences of text divided by the the length of one of the text. Even if this method is more flexible than the previous one, it still suffers from the fact that all n-grams have to be continuous.
- **ROUGE-SU**: Skip-bi-gram and uni-gram ROUGE takes into account bigrams as well as unigrams. However, the bi-grams, instead of being just continuous sequences of words, allows insertions of words between their first and last element. The maximal distance between the two elements of the bi-gram corresponds to a parameter (n) of the measure (often, the measure is instantiated with $n = 4$). During TAC 2008, it has been shown that ROUGE-SUn was the most correlated measure with human judgements.

ROUGE has been very useful for comparing different summaries based on extractive methods, but the use of n-grams only is a strong limitation since it requires an exact match of different units of text. More recently, other evaluation measures have been developed to better capture the semantics of texts.

1.4.3.2 Pyramid

As we have seen, ROUGE measures are based on the discovery of perfect matches between some sequences of the candidate summary and some of the reference summaries. These methods are thus inefficient if the candidate summary has been produced using reformulation techniques. PYRAMID [44] is supposed to overcome some of these issues.

First, the most important pieces of information to be included in reference summaries are extracted[2]. This can be done by a group of experts or directly by analyzing reference summaries. These pieces of information are called SCU (*Summarization Content Units*). They are then weighted: if a SCU only appears in one of the reference summaries it will get a low score, but if it is included in all the reference summaries, it will get a high score. The analogy with a pyramid comes here: the result of this process is a lot of SCU with a low score as the basis of the pyramid, and a few SCU with a high score at its top.

[2] ([19] also proposed an approach of this kind with the notion of *Basic Units*

A list of linguistic expressions is associated with each SCU. It is then possible to map sequences of text with SCU. The last step is then obvious: the evaluation identifies all the SCU contained in the candidate summary and calculates a score for the candidate summary, based on the number and the weight of the SCU it contains.

This evaluation is more precise than ROUGE but requires some manual work to identify the SCU, associate linguistic expressions to them and calculate the weights necessary for the evaluation. However, according to TAC 2008, this method seems to better correlate with human judgements than ROUGE probably because it takes into account some of the semantics of the text.

1.4.4 Automatic Evaluation with no Manual Reference

Even if the previous automatic methods have proven useful for evaluation, they require that reference summaries be available. Several authors have observed that this is not always the case and reference summaries are costly to produce when they are not directly available. Moreover, even when reference summaries are available, the definition and weighting of SCU remains a difficult and time-consuming task. On the other hand, the input text used to produce the summary is of course always available and contains valuable information to evaluate the summaries derived from it.

The fact that information from the original text can be used to directly evaluate summaries with no manual reference was first explored in [34]. The authors proposed to use four classes of easily computable features that are supposed to capture aspects of the input text: input and output text size (coverage is generally proportional to the length of the summary), information-theoretic properties of the distribution of words in the input, presence of descriptive (topic) words and similarity between the documents in multi-document inputs. Then they directly compared a set of candidate summaries using these features and the Jensen-Shannon measure. The authors showed on different tasks (query-focused and update summaries, using different test sets from different evaluation campaigns) that their approach correlates favourably with Pyramid.

Finally, [68] showed that this method is effective in most cases but not always: they specifically pointed out that the method does not perform so well for biographical information or for the summarization of opinions. They propose a new method, mainly based on n-grams, skip bi-grams and the Jensen-Shannon measure. Further, their approach provides interesting results on different summarization tasks for different languages.

1.5 Conclusion and Perspectives

This introductory chapter provided a quick insight into recent trends in automatic summarization methods. However, despite more than 50 years of research, there is still room for improvement.

The most effective and versatile methods used so far in automatic summarization rely on extractive methods: they aim at selecting the most relevant sentences from the collection of original documents in order to produce a condensed text rendering important pieces of information. As we have seen, these kinds of methods are far from being ideal: in multi-document summarization, the selection of sentences from different documents leads to redundancy, which in turn must be eliminated. Moreover, most of the time only a part of a sentence is relevant, but extracting only sub-sentences is still far from being operational. Lastly, extracting sentences from different documents may produce an inconsistent and/or hard-to-read summary.

These limitations suggest a number of desirable improvements. We detail here three very active research trends, some of them being illustrated in the papers in this volume.

- In order to overcome the redundancy problem, researchers are actively working on a better representation of the text content and, more interestingly, are now trying to provide summaries tailored towards specific user needs. Evaluation tasks proposed in TAC reflect this trend since recent evaluations concerned, among other things, the production of opinion-based summaries and the production of update summaries (where only new information should be selected), see for example [10]. In this new context, even if the Edmundsonian approach (based on the recognition of cues and key phrases from the set of original documents) is still widely used, the integration of new methods and new modules has proven necessary (e.g., the integration of an opinion mining module for opinion-based summarization [6]), leading to a more fine grained representation of the original set of documents.
- As explained above, sentence compression is another very active domain of research. Rather than just selecting relevant sentences from original documents, sentence compression aims at keeping only the part of the sentence that is really meaningful and of interest for the abstract. Most of the time, compression rules are defined manually [22] even if recent experiments tried to automatically learn these rules from a set of examples [9]. The improvement brought by this method in readability and summarization quality still needs to be assessed: most approaches in sentence compression require an accurate analysis of the input sentence in order to provide reliable results, which is not always possible given the state of the art in parsing. So, sentence compression is a promising source of improvement but its application still needs to be validated
- One of the main drawback of extractive methods is often the lack of readability of the text produced. In most systems, sentence ordering is based on simple heuristics (e.g. location of the sentence in the original documents) that are not enough to produce a coherent text. Recent research aims at finding new methods for pro-

ducing more coherent texts. For example [4] suggest that calculating the local coherence between candidate sentences may lead to more readable summaries. Developing a global document model may also help.

Finally, the evaluation of automatic summarization is still an open issue. Recent automatic methods (like Pyramid, see above) have proven more consistent that human-based evaluation. However, the lack of consensus between humans when evaluating summaries is a real problem. The development of more focused summaries may lead to a more consistent evaluation and to a better convergence between human and automatic evaluation methods, which is highly desirable. However, automatic summarization evaluation is still a very promising research area with many challenges ahead.

The tenet of this research goes a lot further than just evaluation, since evaluating a summary involves having an accurate description of the content of the documents. Hence, automatic summarization is not *just* another natural language application: it raises important issues related to artificial intelligence and cognitive science. Beyond practical applications, research in the domain may lead to a better understanding of human comprehension.

Acknowledgments

Horacio Saggion is grateful to a fellowship from Programa Ramón y Cajal, Ministerio de Ciencia e Innovación, Spain. Thierry Poibeau is supported by the "Empirical Fundations of Linguistics" labex, Sorbonne-Paris-Cité. We acknowledge the support from the editors of this volume.

References

1. Barzilay, R.: Modeling local coherence: An entity-based approach. In: In Proceedings of ACL 2005, pp. 141–148 (2005)
2. Barzilay, R., Elhadad, M.: Using Lexical Chains for Text Summarization. In: Proceedings of the ACL/EACL'97 Workshop on Intelligent Scalable Text Summarization, pp. 10–17. Madrid, Spain (1997)
3. Barzilay, R., Elhadad, N., Mckeown, K.R.: Inferring strategies for sentence ordering in multidocument news summarization. Journal of Artificial Intelligence Research **17**, 2002 (2002)
4. Barzilay, R., Lapata, M.: Modeling local coherence: An entity-based approach. Computational Linguistics **34**(1), 1–34 (2008)
5. Benbrahim, M., Ahmad, K.: Text Summarisation: the Role of Lexical Cohesion Analysis. The New Review of Document & Text Management pp. 321–335 (1995)
6. Bossard, A., Généreux, M., Poibeau, T.: Cbseas, a summarization system – integration of opinion mining techniques to summarize blogs. In: Proceedings of the 12th Meeting of the European Association for Computational Linguistics (system demonstration), EACL '09 (2009)
7. Carbonell, J.G., Goldstein, J.: The use of MMR, diversity-based reranking for reordering documents and producing summaries. In: Research and Development in Information Retrieval, pp. 335–336 (1998)

8. Chambers, N., Jurafsky, D.: Unsupervised learning of narrative schemas and their participants. In: ACL/AFNLP, pp. 602–610 (2009)
9. Cohn, T., Lapata, M.: Sentence compression as tree transduction. Journal of Artificial Intelligence Research (JAIR) **34**, 637–674 (2009)
10. Dang, H.T., Owczarzak, K.: Overview of the tac 2008 opinion question answering and summarization tasks. In: Proceedings of the TAC 2008 Workshop, vol. Notebook Papers and Results (2008)
11. DeJong, G.: An Overview of the FRUMP System. In: W. Lehnert, M. Ringle (eds.) Strategies for Natural Language Processing, pp. 149–176. Lawrence Erlbaum Associates, Publishers (1982)
12. Edmundson, H.: New Methods in Automatic Extracting. Journal of the Association for Computing Machinery **16**(2), 264–285 (1969)
13. Endres-Niggemeyer, B.: SimSum: an empirically founded simulation of summarizing. Information Processing & Management **36**, 659–682 (2000)
14. Erkan, G., Radev, D.: Lexrank: Graph-based lexical centrality as salience in text summarization. Journal of Artificial Intelligence Research (JAIR) **22**, 457–479 (2004)
15. Fellbaum, C. (ed.): WordNet: An Electronic Lexical Database. The MIT Press (1998)
16. Grishman, R.: Information extraction: techniques and challenges. In: M.T. Pazienza (ed.) Information Extraction. A multidisciplinary approach to an Emerging Information Technology, no. 1299 in Lecture Notes in Artificial Intelligence. Springer (1997)
17. Harman, D., Liberman, M.: Tipster complete. Tech. rep., University of Pennsylvania, USA (1993)
18. Hasler, L., Orăsan, C., Mitkov, R.: Building better corpora for summarisation. In: Proceedings of Corpus Linguistics, pp. 309–319. Lancaster, UK (2003)
19. Hovy, E., Lin, C.Y., Zhou, L., Fukumoto, J.: Automated summarization evaluation with basic elements. In: In Proceedings of the Fifth Conference on Language Resources and Evaluation (LREC) (2006)
20. Jing, H.: Using hidden markov modeling to decompose human-written summaries. Computational Linguistics **28**, 527–543 (2002)
21. Jing, H., McKeown, K.: The Decomposition of Human-Written Summary Sentences. In: M. Hearst, F. Gey, R. Tong (eds.) Proceedings of SIGIR'99 – 22nd International Conference on Research and Development in Information Retrieval, pp. 129–136. University of California, Berkeley (1999)
22. Jing, H., McKeown, K.: Cut and Paste Based Text Summarization. In: Proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics, pp. 178–185. Seattle, Washington, USA (2000)
23. Jones, K.S.: Automatic summarising: The state of the art. Inf. Process. Manage. **43**(6), 1449–1481 (2007)
24. Kabadjov, M.A., Atkinson, M., Steinberger, J., Steinberger, R., der Goot, E.V.: Newsgist: A multilingual statistical news summarizer. In: ECML/PKDD (3), pp. 591–594 (2010)
25. Knight, K., Marcu, D.: Statistics-based summarization - step one: Sentence compression. In: Proceedings of the 17th National Conference of the American Association for Artificial Intelligence. AAAI (2000)
26. Kupiec, J., Pedersen, J., Chen, F.: A Trainable Document Summarizer. In: Proc. of the 18th ACM-SIGIR Conference, pp. 68–73. Seattle, Washington, United States (1995)
27. Lapata, M.: Probabilistic Text Structuring: Experiments with Sentence Ordering. In: Proceedings of the 41st Meeting of the Association of Computational Linguistics, pp. 545–552. Sapporo, Japan (2003)
28. Li, P., Jiang, J., Wang, Y.: Generating Templates of Entity Summaries with an Entity-Aspect Model and Pattern Mining. In: Proceedings of ACL. ACL, Uppsala (2010)
29. Liddy, E.D.: The Discourse-Level Structure of Empirical Abstracts: An Exploratory Study. Information Processing & Management **27**(1), 55–81 (1991)
30. Lin, C., Hovy, E.: Identifying Topics by Position. In: Fifth Conference on Applied Natural Language Processing, pp. 283–290. Association for Computational Linguistics (1997)

31. Lin, C.Y.: Rouge: a package for automatic evaluation of summaries. In: Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004). Barcelona, Spain (2004)
32. Lin, C.Y., Hovy, E.: The automated acquisition of topic signatures for text summarization. In: Proceedings of the COLING Conference. Saarbrumlcken, Germany (2000)
33. Lloret, E., Palomar, M.: Text summarisation in progress: a literature review. Artificial Intelligence Review pp. 1–41 (2011)
34. Louis, A., Nenkova, A.: Automatically evaluating content selection in summarization without human models. In: Proceedings of EMNLP'09, pp. 306–314 (2009)
35. Luhn, H.P.: The Automatic Creation of Literature Abstracts. IBM Journal of Research Development **2**(2), 159–165 (1958)
36. Mani, I.: Automatic Text Summarization. John Benjamins Publishing Company (2001)
37. Mani, I., Klein, G., House, D., Hirschman, L., Firmin, T., Sundheim, B.: Summac: a text summarization evaluation. Natural Languae Engineering **8**, 43–68 (2002). DOI 10.1017/S1351324901002741. URL http://portal.acm.org/citation.cfm?id=973860.973864
38. Mani, I., Maybury, M.T.: Advances in Automatic Text Summarization. MIT Press (1999)
39. Mann, W., Thompson, S.: Rhetorical Structure Theory: towards a functional theory of text organization. Text **8**(3), 243–281 (1988)
40. Marcu, D.: From Discourse Structures to Text Summaries. In: The Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization, pp. 82–88. Madrid, Spain (1997)
41. Maynard, D., Tablan, V., Cunningham, H., Ursu, C., Saggion, H., Bontcheva, K., Wilks, Y.: Architectural Elements of Language Engineering Robustness. Journal of Natural Language Engineering – Special Issue on Robust Methods in Analysis of Natural Language Data **8**(2/3), 257–274 (2002)
42. Mihalcea, R.: Language independent extractive summarization. In: AAAI, pp. 1688–1689 (2005)
43. Mihalcea, R., Tarau, P.: TextRank: Bringing order into texts. In: Proceedings of EMNLP-04and the 2004 Conference on Empirical Methods in Natural Language Processing (2004)
44. Nenkova, A., Passonneau, R., McKeown, K.: The pyramid method: Incorporating human content selection variation in summarization evaluation. ACM Transactions on Speech and Language Processing **4**(2), 1–23 (2007)
45. NIST: Proceedings of the Text Analysis Conference. NIST, Gaithesburg (2008)
46. Okumura, M., Fukusima, T., Nanba, H., Hirao, T.: Text summarization challenge 2 text summarization evaluation at ntcir workshop 3. SIGIR Forum **38**(1), 29–38 (2004)
47. Ono, K., Sumita, K., Miike, S.: Abstract Generation Based on Rhetorical Structure Extraction. In: Proceedings of the International Conference on Computational Linguistics, pp. 344–348 (1994)
48. Over, P., Dang, H., Harman, D.: DUC in context. Inf. Process. Manage. **43**, 1506–1520 (2007)
49. Over, P., Dang, H., Harman, D.: Duc in context. Inf. Process. Manage. **43**, 1506–1520 (2007). DOI 10.1016/j.ipm.2007.01.019. URL http://portal.acm.org/citation.cfm?id=1284916.1285157
50. Owczarzak, K., Dang, H.: Overview of the tac 2010 summarization track. In: Proceedings of TAC 2010. NIST (2010)
51. Paice, C.D.: Constructing Literature Abstracts by Computer: Technics and Prospects. Information Processing & Management **26**(1), 171–186 (1990)
52. Paice, C.D., Oakes, M.P.: A Concept-Based Method for Automatic Abstracting. Tech. Rep. 27, Library and Information Commission (1999)
53. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL '02, pp. 311–318 (2002)
54. Radev, D., Allison, T., Blair-Goldensohn, S., Blitzer, J., Çelebi, A., Dimitrov, S., Drabek, E., Hakim, A., Lam, W., Liu, D., Otterbacher, J., Qi, H., Saggion, H., Teufel, S., Topper, M., Winkel, A., Zhang, Z.: MEAD — A platform for multidocument multilingual text summarization. In: Conference on Language Resources and Evaluation (LREC). Lisbon, Portugal (2004)

55. Radev, D.R., Jing, H., Budzikowska, M.: Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies. In: ANLP/NAACL Workshop on Summarization. Seattle, WA (2000)
56. Radev, D.R., McKeown, K.R.: Generating natural language summaries from multiple on-line sources. Computational Linguistics **24**(3), 469–500 (1998)
57. Radev, D.R., Teufel, S., Saggion, H., Lam, W., Blitzer, J., Qi, H., Çelebi, A., Liu, D., Drabek, E.: Evaluation challenges in large-scale document summarization. In: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1, ACL '03, pp. 375–382 (2003)
58. Saggion, H.: Multilingual Multidocument Summarization Tools and Evaluation. In: Proceedings of LREC 2006 (2006)
59. Saggion, H.: Experiments on semantic-based clustering for cross-document coreference. In: Proceedings of the Third Joint International Conference on Natural Language Processing, pp. 149–156. AFNLP, AFNLP, Hyderabad, India (2008)
60. Saggion, H.: SUMMA: A Robust and Adaptable Summarization Tool. Traitement Automatique des Langues **49**(2), 103–125 (2008)
61. Saggion, H.: A classification algorithm for predicting the structure of summaries. In: UCNLG+Sum '09: Proceedings of the 2009 Workshop on Language Generation and Summarisation, p. 31–38. Association for Computational Linguistics, Association for Computational Linguistics, Morristown, NJ, USA (2009)
62. Saggion, H.: Learning predicate insertion rules for document abstracting. In: CICLing, pp. 301–312 (2011)
63. Saggion, H., Gaizauskas, R.: Multi-document summarization by cluster/profile relevance and redundancy removal. In: Proceedings of the Document Understanding Conference 2004. NIST, Boston, USA (2004)
64. Saggion, H., Lapalme, G.: Generating Indicative-Informative Summaries with SumUM. Computational Linguistics **28**, 497–526 (2002)
65. Saggion, H., Radev, D., Teufel, S., Lam, W.: Meta-evaluation of Summaries in a Cross-lingual Environment using Content-based Metrics. In: Proceedings of COLING 2002, pp. 849–855. Taipei, Taiwan (2002)
66. Saggion, H., Radev, D., Teufel, S., Wai, L., Strassel, S.: Developing Infrastructure for the Evaluation of Single and Multi-document Summarization Systems in a Cross-lingual Environment. In: LREC 2002, pp. 747–754. Las Palmas, Gran Canaria, Spain (2002)
67. Saggion, H., Teufel, S., Radev, D., Lam, W.: Meta-evaluation of summaries in a cross-lingual environment using content-based metrics. In: Proceedings of the 19th international conference on Computational linguistics - Volume 1, COLING '02, pp. 1–7 (2002)
68. Saggion, H., Torres-Moreno, J.M., da Cunha, I., SanJuan, E., Velazquez-Morales, P.: Multilingual summarization evaluation without human models. In: In Proceedings of COLING (2010)
69. Salton, G., Allan, J., Singhal, A.: Automatic text decomposition and structuring. Information Processing & Management **32**(2), 127–138 (1996)
70. Sparck Jones, K.: What Might Be in a Summary? In: K. Knorz, Womser-Hacker (eds.) Information Retrieval 93: Von der Modellierung zur Anwendung (1993)
71. Sparck Jones, K.: Automatic Summarizing: Factors and Directions. In: I. Mani, M. Maybury (eds.) Advances in Automatic Text Summarization. MIT Press, Cambridge MA (1999)
72. Sparck Jones, K., Endres-Niggemeyer, B.: Automatic Summarizing. Information Processing & Management **31**(5), 625–630 (1995)
73. Spärck Jones, K., Galliers, J.R.: Evaluating natural language processing systems. Springer, Berlin (1996)
74. Swales, J.: Genre analysis: English in academic and research settings. Cambridge University Press (1990)
75. Teufel, S., Moens, M.: Argumentative classification of extracted sentences as a first step towards flexible abstracting. In: I. Mani, M. Maybury (eds.) Advances in Automatic Text Summarization, pp. 155–171. The MIT Press (1999)

76. Turner, J., Charniak, E.: Supervised and Unsupervised Learning for Sentence Compression. In: ACL (2005)
77. Witbrock, M.J., Mittal, V.O.: Ultra-summarization: A statistical approach to generating highly condensed non-extractive summaries. In: In SIGIR99, pp. 315–316 (1999)
78. Zajic, D., Dorr, B., Lin, J., Schwartz, R.: Multi-candidate reduction: Sentence compression as a tool for document summarization tasks. In: Information Processing and Management Special Issue on Summarization, p. 43 (2007)