Digit·HVM revista digital d'humanitats

Automatic Text Summarization

Salvador Climent Roca

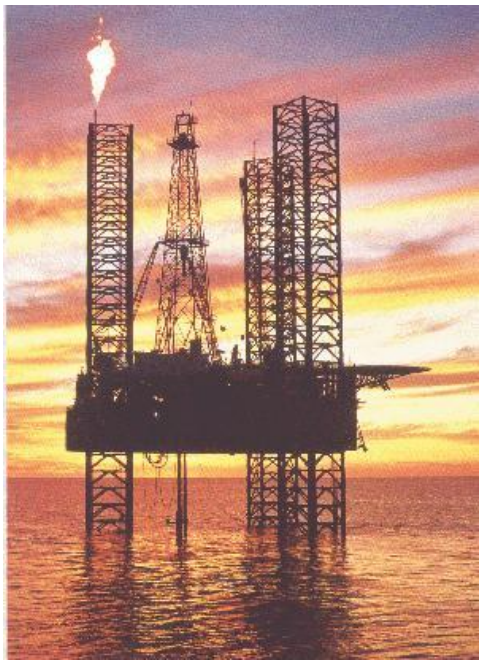
Professor of Humane Studies and Philology Studies at UOC
scliment@campus.uoc.es

Abstract: The coming of the WWW and of the Internet as a telecommunications network has changed the concept of what is considered information: the person who is best informed is not the one with the most information but the one with the best means for obtaining and assimilating (consuming) exactly the information acquired. This situation is proving to be a great stimulus for research into and the development of applications in the field of technology for recovering and extracting information. In this context, automated document summary systems are a new step forward towards optimising the treatment of documentation in digital formats and for tailoring it to the needs of users. This article outlines the main lines of research into the creation of automated summaries and its relationships with other areas of linguistic engineering.

1. Preface

It is no use talking here again about the so—called information overload problem. Retrieving, managing and digesting information are hard and crucial tasks in our society; we therefore call for systems and applications able to help us in such heavy duties: given the billions of pages the Web comprises, *please help me to select just that precise information which I am interested in*.

Well, we have Internet search engines. If it is the case that we can handle them appropriately, soon we will be able to reduce the billion—page problem to some hundreds of them. This is not heaven but it is certainly better than nothing. Moreover, to the extent that present or future search engines make use of a suitable system for classifying documents by relevance, we will progressively approach our aim —to track down the few texts which eventually, yes, we really need to read.



To come to terms with that final sorting out, there is something which would prove to be extremely useful: someone or something to give us an idea about the content of those dozens of documents, i.e., to give us a grasp of the gist of the text.

A wide range of research on language technologies focuses on such and related problems, many of them involving to a greater or smaller extent some kind of human language processing inasmuch as most of the information flowing around the world, as anyone can notice, happens to be codified in a natural language.

Information Retrieval, Filtering, Classification and Extraction, either in their monolingual or multilingual dimensions, are key fields of research in linguistic technologies. In this environment, automatic summarization of documents may probably be the finishing touch. Our favorite search engine will be one which would allow us to make our request (in our language) in a way like this: *Search for documents in any language talking about that matter, choose the most relevant ones, sort them according to those criteria and give me back a fifty—word summary of each one (in my language)*.

2. Analogies

Automatic text summarizing shares, or can share, properties with both Information Retrieval and Information Extraction. In fact, the two main lines of research in summarization map such technologies.

We mean by Information Retrieval the task of, e.g., Internet search engines, that is, among a (huge) amount of documents, retrieving those corresponding to certain criteria —usually, a set of key words. On the other hand, Information Extraction is about processing one or more documents in order to extract from them some information which is relevant to us and then generating a brand new text (or another structure, as for instance a template) containing exactly that precise relevant information. Both basic lines of research on automated text summarizing are analogous to these.

Summaries by extraction, i.e. *extracts*, process one (or several) documents, seen as a collection of sentences. Among such a collection they retrieve and give back those considered most relevant —or else those responding to certain criteria. In this case, the summary is a subset of the set of sentences of the original text.

Summaries by abstraction, i.e. *abstracts*, make use of rather more complex linguistic technology techniques, as long as the output is not a simple number of sentences retrieved as *found* in the input text but instead a brand new document which has been generated by processing the information contained there. Such a strategy makes use of a wide set of techniques in natural language processing which up to

now are far from being state—of—the—art. For that reason the most significant advances in the field relate to extracts. Therefore we are going to focus on them in the rest of the paper.

When talking about a (relatively) new and promising technology such as this it is important to state a caution for non—experts —one should avoid false expectations, as used to happen in the dawn of Machine Translation and to a certain extent is still the case with other technologies, such as Speech Dialogue—to—Machine Systems. One should not expect ready—to—use perfect systems able to replace humans in their work. Instead, one should expect some kind of realistic applications which could either help people or release them from a number of boring and recurrent tasks.

To put it another way, following the comparison above, in the same way that Machine Translation cannot provide ready—to—publish documents but drafts which always need human post—editing, what Text Summarizers are going to provide are extracts which will be useful to get the gist of documents, and perhaps also useful to classify them in some topic structure, and eventually helpful for deciding whether or not we shall go and read the full original text.

3. Typologies

When one hears talking about automatic summarizers for the first time, a couple of questions usually come to mind: (1) *"So how do these programs do it? Do they do it the same way WE would?"* And (2) *"What about those automatic summaries? Are they good, really? Are they as good as those a person can do?"*

For the moment, we had better postpone the first question, but it is convenient to put forward that, as usually happens in Linguistic Engineering (and in general, in Artificial Intelligence), automatic systems seldom behave following the same strategies as humans, among other reasons because, in fact, we still do not have a very clear idea of what kind of strategies are they. Nor do computing systems have at their disposal the same kind of information that we humans have. In Artificial Intelligence, we mean by intelligent system one that emulates *intelligent* behaviour. In other words, the one that carries on a task that, if carried out by a human, we could say that he or she has made use of a certain degree of intelligence.

But let us face the second question. *Those automatic summaries… are they good, really?* As usual, the first answer is another question —what is a good summary, exactly? Well, we can say, that depends on several factors. Especially on what kind of things we expect to find in the summary.

Maybe we are generically interested to know what the author is saying there —or maybe what we want to know is what the author says *about some particular issue*. We are talking about the difference between "what's this article about?" and "what does this article have to say about the Tokyo Stock Exchange?"

Alternatively, maybe the fact is that we already have previous information about the issue and all we want to know is what is new about it. In that case we don't want to be told the whole tale — a summary containing the novel facts will be enough.

The kind of summary we might want can also be related to the quantity and quality of the information we want to get. It may be sufficient to us to know more or less what the topic of the document is for classification purposes — or maybe we need a new document condensing its content. In the former case, perhaps we want to have that indicative summary as a coherent text, or perhaps a simple list of fragments or keywords will be enough —or maybe we prefer just a topic index. In the latter case, depending on the amount of documents and time we have available, we might want fairly short extracts — fifty or one hundred words, for instance. Or maybe we have time so we can handle longer summaries.

Going further, maybe we want a system good enough to give us a single summary of a set of documents. For instance, *tell me everything the papers say today about Napster*.

Such examples correspond to some of the different typologies and current research lines in the field: indicative summary (to be used in document classification and filtering) vs. informative summary (to be read instead of or as an advance of the complete document). Generic summary vs. query—based (answering a query defining the reader's interests). Generic summary vs. update (offering just novel facts and neglecting information which is already known). Multi—document summary (condensing several texts treating the same point). Summaries with different levels of compression of the input document. And so on. Any one of these genres needs different techniques of human language processing.

Besides, we should notice that there are many different types of documents to be summarized – newspaper news, articles, e—mails, scientific papers, web sites, brochures, lists... even books, maybe.

Moreover, in every case the level of marking up of the text may also differ, from documents in plain text to highly codified XML pages, passing through HTML or any other mark—up language. Genre and level of structure of the input text are extremely important variables to pay attention to before facing the task, to the extent that most of the techniques which may prove to be useful in one case, may be useless in the other.

So that, summing up, what makes a good summary? Well, the first thing we have to know is: a good summary *of what and to get what*.

But we were also posing the following question: those automatic summaries, are they as good as if a man or a woman did them? In this case another question applies as a first answer: *as if they were done by... whom?*

One of the usual and necessary tasks to be carried out in any research is evaluation of the results. In the case of automatic text summarization, evaluation is usually done by comparing automatic summaries against some kind of reference—summary built up by humans. To achieve that one usually asks a set of referees to summarize (or get *extracts* from) a test corpus of documents.

So, it has been proved that the very obtainment of a single reference—summary is not a trivial task at all. Given the same document, rarely do the referees agree on what information is relevant to be selected in a summary. What's more, in some evaluation tests the same referees have been asked to summarize the same documents letting a lapse of several weeks between one test and the other. In such cases, referees turned out to agree with themselves in a range of about only 55%.

This drawback is usually solved using statistical methods, so managing to reach some kind of average summary to be used as a benchmark. This seems to be a fairly reasonable way of evaluating automatic summarization systems. However, the above comments help us to notice that a question such as *are automatic summaries as good as if humans did them?* cannot usually have a straightforward answer.

4. Techniques

These are the problems. What about solutions? What techniques are used to summarize automatically?

They are usually classified in three families: based on the *surface* (no linguistic analysis is performed); based on *entities* named in the text (there is some kind of lexical acknowledgement and classification); and based on *discourse structure* (some kind of structural, usually linguistic, processing of the document is required).

Commercial products usually make use of *surface* techniques. They simply handle text as strings of symbols and they compute them in order to select some as a representation of the document. One classical method is selection of statistically frequent terms in the document. E.g. those sentences (in fact, we should better talk of strings beginning with a capital letter and ending with a full stop, more or less) containing more of the most frequent terms (*strings*) will be selected as a summary of the document.

Another group of methods is based on position: position in the text, in the paragraph, in depth or embedding of the section, etc. The name of the game is selecting those fragments occupying the most promising positions. Typically, to summarize newspaper news a good bet might be to select the first paragraph(s). In scientific papers, maybe one should turn to extract, among others, the *abstract* itself, conclusions and bibliography.

Other methods gain profit from outstanding parts of the text: titles, subtitles, and*leads*... It is supposed that sentences containing words of the title are better candidates to summarize the whole document. Other methods compute sentence relevance standing on the presence of some kind

of *bonus* or *penalty* words —such as “important”, “in summary”, and so on.

A second family of methods is that based on entities. In this case, there is a higher degree of linguistic analysis. I say ‘higher’ and not ‘some’ because, although it sometimes can’t be detected, *surface* methods usually do require some kind of linguistic technology. For instance, segmenting a text in sentences is not a trivial matter of capitals and full stops since one can find capitals and full stops not signaling the boundaries of a sentence (e.g., think about ‘e.g.’), so some kind of linguistic knowledge to decide in which kind of cases which kind of symbols bound a sentence or not has to be applied.

Methods based on entities are grounded on techniques allowing us to acknowledge linguistic units among the mass of alphanumeric symbol strings. Which string between blanks is a noun, which one is a verb... To do this, lemmatizers, morphological parsers and part—of—speech taggers are needed since, on the one hand, one same string might belong to different parts of speech (e.g. ‘bomb’, noun or verb), and, on the other, different strings might be instances of the same part of speech (e.g. ‘bomb’ and ‘bombed’). Depending on the complexity and accuracy of the method, more sophisticated syntactic or semantic parsers may be needed.

These kinds of methods may be able to pick up entities in the text and detect different kinds of relationships existing between them: identity of forms or lemmas, semantic relations (a cat, a motorbike and a bus are all types of vehicles), topic relations (a referee, a goalkeeper and a free—kick are all terms used in football), etc.

Standing on this base, it is possible to build up a representation of the connectivity inside the text, in terms of graphs, in a way that the system could decide which parts (sentences) are especially relevant, so they become candidates to belong to the extract.

In order to implement this kind of methods it is specially important to have available lexical databases (as EuroWordNet), named entity recognizers (e.g. to detect proper names, a kind of words which by their own nature are not listed in dictionaries or lexical repositories) and anaphora resolution systems (e.g. to acknowledge that some pronoun refers to some noun previously mentioned in the text).

Finally, simple methods based on structure can for instance take advantage of the hypertextual scaffolding of an HTML page. More complex methods using linguistic technology resources and techniques such as those mentioned above and others might build a rhetoric structure of the document, allowing its most relevant fragments to be detected. Such techniques are based on detection of discourse markers such as connectors or adverbs. These methods still belong to the field of pure basic research.

It is important to note that in all methods based on extraction, sooner or later one has to face the problem of text reposition. It is clear that when creating a text using fragments of a previous original, reference chains and, in general, text cohesion, are easily lost. A good system based on extraction should be able to compare and analyze both the original text and the resulting extract to further insert lexical material that re—make textual cohesion.

5. References and products

To know about the state of the art in automatic text summarization we should pay attention to both commercial systems and research.

Among the former we have Inxight (Xerox), Copernic and the well-known Microsoft Word summarizer. A fairly complete list of systems can be found in the Mitre Orgsite. With respect to systems in progress which can be tested online, SweSum andExtractor (the technology powering Copernic Summarizer) are very interesting ones.

In the field of research, relevant sites where to collect information are those of theUniversity of Ottawa, Dragomir Radev (researcher of the Columbia University) and, above all, the large bibliography maintained by Radev itself in the site of the University of Michigan. In Spain, we are building a prototype of a summarizer for newspaper news in the framework of the Hermes project.

Last, it must be noted that the programmes of a wider scope are actions prompted by the The Defense Advanced Research Projects Agency (DARPA) of the USA, such as TIDES (a program for detection, extraction and summarization of multilingual information) and the Document Understanding Conference (DUC) which proposes a series of large-scale experiments for the evaluation of techniques and systems for text understanding and summarization.

6. Appendix

6.1. 100 - word extract, by Microsoft Word

This is the 100-word extract of the paper generated by Microsoft Word

> Automatic text summarizing shares, or can share, properties with both Information Retrieval and Information Extraction.
> Summaries by extraction, i.e. extracts, process one (or several) documents, seen as a collection of sentences.
> Those automatic summaries… are they good, really?
> Multi-document summary (condensing several texts treating the same point). Summaries with different levels of compression of the input document.
> In the case of automatic text summarization, evaluation is usually done by comparing automatic summaries against some kind of reference-summary built up by humans

6.2. 300 - word extract, by SweSum

This is the 300-word extract of the paper generated by SweSum

> In this environment, automatic summarization of documents may probably be the finishing touch. Automatic text summarizing shares, or can share, properties with both Information Retrieval and Information Extraction. Both basic lines of research on automated text summarizing are analogous to these. In this case, the summary is a subset of the set of sentences of the original text. Those automatic summaries… are they good, really? Multi-document summary (condensing several texts treating the same point). In the case of automatic text summarization, evaluation is usually done by comparing automatic summaries against some kind of reference-summary built up by humans. This seems to be a fairly reasonable way of evaluating automatic summarization systems. Other methods gain profit from outstanding parts of the text: titles, subtitles, and leads... These kinds of methods may be able to pick up entities in the text and detect different kinds of relationships existing between them: identity of forms or lemmas, semantic relations (a cat, a motorbike and a bus are all types of vehicles), topic relations (a referee, a goalkeeper and a free-kick are all terms used in football), etc. Standing on this base, it is possible to build up a representation of the connectivity inside the text, in terms of graphs, in a way that the system could decide which parts (sentences) are especially relevant, so they become candidates to belong to the extract. It is clear that when creating a text using fragments of a previous original, reference chains and, in general, text cohesion, are easily lost. To know about the state of the art in automatic text summarization we should pay attention to both commercial systems and research.

Related links:

- DARPA Information Technology Office: Translingual Information, Detection, Extraction and Summarization (TIDES): http://www.darpa.mil/ito/research/tides/
- The Text Summarization Project. Universitat of Ottawa: http://www.site.uottawa.ca/tanka/ts.html

Digit·HVM revista digital d'humanitats

- A Bibliography of Research in Text Summarization:
  http://www.si.umich.edu/~radev/summarization/large-bib.doc
- SweSum:
  http://www.nada.kth.se/~xmartin/swesum/index-eng.html
- Extractor:
  http://extractor.iit.nrc.ca/on_line_demo.html

[Published on: september 2001]

Recommended citation:

CLIMENT ROCA, Salvador (2001). “Automatic text summarization”. *Digithum*, issue 3 [article online].
DOI: http://dx.doi.org/10.7238/d.v0i3.598