Chapter 2
Portable Text Summarization

Martin Hassel
Stockholm University, Sweden

Hercules Dalianis
Stockholm University, Sweden

ABSTRACT

Today, with digitally stored information available in abundance, even for many minor languages, this information must by some means be filtered and extracted in order to avoid drowning in it. Automatic summarization is one such technique, where a computer summarizes a longer text to a shorter non-redundant form. The development of advanced summarization systems also for smaller languages may unfortunately prove too costly. Nevertheless, there will still be a need for summarization tools for these languages in order to curb the immense flow of digital information. This chapter sets the focus on automatic summarization of text using as few direct human resources as possible, resulting in what can be perceived as an intermediary system. Furthermore, it presents the notion of taking a holistic view of the generation of summaries.

INTRODUCTION

Text summarization is the process of creating a summary of one or more texts. This summary may serve several purposes. One might, for example, want to get an overview of a document set in order to choose what documents one needs to read in full. Another plausible scenario would be getting the gist of a constant news flow, without having to wade through inherently redundant articles run by several news agencies, in order to find what might differ in reports from different parties. With digitally stored information available in abundance and in a myriad of forms, even for many minor languages, it has now become near impossible to manually search, sift and choose which information one should incorporate. Instead this information must by some means be filtered and extracted in order to avoid drowning in it. Automatic summarization is one such technique.

The title of this chapter sets the focus on summarization of text, automatically carried out by a computer program using methods more or less directly transferable from one language to an-
Portable Text Summarization

other. This accomplished by using as few human resources as possible. The resources that are used should to as high extent as possible be already existing, not specifically aimed at summarization and, preferably, created as part of natural literary processes. Moreover, the summarization system should be able to be easily assembled using only a small set of basic language processing tools, again, not specifically aimed at summarization. The summarization system should thus be near language independent as to be quickly ported between different natural languages. The motivation for this is as simple as intuitive. Apart from the major languages of the world, there simply are a lot of languages for which large bodies of data aimed at language technology research, let alone research in automatic text summarization, are lacking. There might also not be resources available to develop such bodies of data, since it is usually time-consuming and hence expensive. Nevertheless, there will still be a need for sufficiently efficacious automatic text summarization for these languages, acting as intermediate states, in order to subdue this constantly increasing amount of electronically produced text.

The application areas for automatic text summarization are extensive. As the amount of information on the Internet grows abundantly, it is difficult to select relevant information. Information is published simultaneously on many media channels in different versions, for instance, a paper newspaper, web newspaper, SMS news flash, mobile radio newscast, and a spoken newspaper for the visually impaired. Also, these may today be accessed by a myriad of display devices, sporting a wide range of presentation capacity. Customization of information for different channels and formats is an immense editing job that notably involves shortening of original texts. Automatic text summarization can automate this work completely, or at least assist in the process by producing a draft summary. Also, documents can be made accessible in other languages by first summarizing them before translation, which in many cases would be sufficient to establish the relevance of a foreign language document, and hence save human translators work since they need not translate every document manually. Automatic text summarization can also be used to summarize a text before an automatic speech synthesizer reads it, thus reducing the time needed to absorb the key facts in a document. In particular, automatic text summarization can be used to prepare information for use in small mobile devices, such as a PDA, which may need considerable reduction of content.

BACKGROUND

Summarization approaches are often divided into two main groups, text extraction and text abstraction. Text abstraction is in many aspects similar to what humans abstractors do when writing an abstract, even though professional abstractors often utilize surface-level information such as headings, key phrases and position in the text as well as the overall organization of the text into more or less genre specific sections (Liddy 1991, Endres-Niggemeyer et al. 1995, Cremmins 1996). The parsing and interpretation of text is a venerable research area that has been investigated for many years. In this area we have a wide spectrum of techniques and methods ranging from word by word parsing to rhetorical discourse parsing as well as more statistical methods, or a mixture of all. Also the generation of text is a vigorous research field with techniques ranging from canned text and template filling to more advanced systems with discourse planners and surface realizers.

Text extraction, on the other hand, simply re-uses a subset of the original text, thus preserving the original wording and structure of the source text. Sometimes the extracted fragments are post-edited, for example by deleting subordinate clauses or joining incomplete clauses to form complete clauses (Jing and McKeown 2000, Jing 2000). Most of the research in the field of automatic text summarization has to the nature been that of text