KTH Computer Science
and Communication

Resource Lean and Portable Automatic Text Summarization

MARTIN HASSEL

Doctoral Thesis
Stockholm, Sweden 2007

Akademisk avhandling som med tillstånd av Kungl Tekniska högskolan framlägges till offentlig granskning för avläggande av filosofie doktorsexamen i datalogi måndagen den 11 juni 2007 klockan 13.15 i Sal E2, Lindstedtsvägen 3, KTH Campus Valhallavägen, Kungl Tekniska högskolan, Stockholm.

TRITA-CSC-A 2007:9 • ISSN-1653-5723 • ISRN-KTH/CSC/A--07/09--SE • ISBN-978-91-7178-704-0

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-rendundant form. Apart from the major languages of the world there are a lot of languages for which large bodies of data aimed at language technology research to a high degree are lacking. There might also not be resources available to develop such bodies of data, since it is usually time consuming and requires substantial manual labor, hence being expensive. Nevertheless, there will still be a need for automatic text summarization for these languages in order to subdue this constantly increasing amount of electronically produced text.

This thesis thus sets the focus on automatic summarization of text and the evaluation of summaries 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 or evaluation of summaries and, preferably, created as part of natural literary processes. Moreover, the summarization systems should be able to be easily assembled using only a small set of basic language processing tools, again, not specifically aimed at summarization/evaluation. The summarization system should thus be near language independent as to be quickly ported between different natural languages.

The research put forth in this thesis mainly concerns three computerized systems, one for near language independent summarization – The HolSum summarizer; one for the collection of large-scale corpora – The KTH News Corpus; and one for summarization evaluation – The KTH eXtract Corpus. These three systems represent three different aspects of transferring the proposed summarization method to a new language.

One aspect is the actual summarization method and how it relates to the highly irregular nature of human language and to the difference in traits among language groups. This aspect is discussed in detail in Chapter 3. This chapter also presents the notion of “holistic summarization”, an approach to self-evaluative summarization that weighs the fitness of the summary as a whole, by semantically comparing it to the text being summarized, before presenting it to the user. This approach is embodied as the text summarizer HolSum, which is presented in this chapter and evaluated in Paper 5.

A second aspect is the collection of large-scale corpora for languages where few or none such exist. This type of corpora is on the one hand needed for building the language model used by HolSum when comparing summaries on semantic grounds, on the other hand a large enough set of (written) language use is needed to guarantee the randomly selected subcorpus used for evaluation to be representative. This topic briefly touched upon in Chapter 4, and detailed in Paper 1.

The third aspect is, of course, the evaluation of the proposed summarization method on a new language. This aspect is investigated in Chapter 4. Evaluations of HolSum have been run on English as well as on Swedish, using both well established data and evaluation schemes (English) as well as with corpora gathered “in the wild” (Swedish). During the development of the latter corpora, which is discussed in Paper 4, evaluations of a traditional sentence ranking text summarizer, SweSum, have also been run. These can be found in Paper 2 and 3.

This thesis thus contributes a novel approach to highly portable automatic text summarization, coupled with methods for building the needed corpora, both for training and evaluation on the new language.

TRITA-CSC-A 2007:9 • ISSN-1653-5723 • ISRN-KTH/CSC/A--07/09--SE • ISBN-978-91-7178-704-0