A Survey on Automatic Text Summarization

Dipanjan Das André F.T. Martins

Language Technologies Institute
Carnegie Mellon University
{dipanjan, afm}@cs.cmu.edu

November 21, 2007

Abstract

The increasing availability of online information has necessitated intensive research in the area of automatic text summarization within the Natural Language Processing (NLP) community. Over the past half a century, the problem has been addressed from many different perspectives, in varying domains and using various paradigms. This survey intends to investigate some of the most relevant approaches both in the areas of single-document and multiple-document summarization, giving special emphasis to empirical methods and extractive techniques. Some promising approaches that concentrate on specific details of the summarization problem are also discussed. Special attention is devoted to automatic evaluation of summarization systems, as future research on summarization is strongly dependent on progress in this area.

1 Introduction

The subfield of summarization has been investigated by the NLP community for nearly the last half century. Radev et al. (2002) define a *summary* as "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". This simple definition captures three important aspects that characterize research on automatic summarization:

- Summaries may be produced from a *single document* or *multiple documents*,
- Summaries should preserve important information,
- Summaries should be short.

Even if we agree unanimously on these points, it seems from the literature that any attempt to provide a more elaborate definition for the task would result in disagreement within the community. In fact, many approaches differ on the manner of their problem formulations. We start by introducing some common terms in the