Cut and Paste Based Text Summarization

Hongyan Jing and Kathleen R. McKeown
Department of Computer Science
Columbia University
New York, NY 10027, USA
hjing, kathy@cs.columbia.edu

Abstract

We present a cut and paste based text summarizer, which uses operations derived from an analysis of human written abstracts. The summarizer edits extracted sentences, using reduction to remove inessential phrases and combination to merge resulting phrases together as coherent sentences. Our work includes a statistically based sentence decomposition program that identifies where the phrases of a summary originate in the original document, producing an aligned corpus of summaries and articles which we used to develop the summarizer.

1 Introduction

There is a big gap between the summaries produced by current automatic summarizers and the abstracts written by human professionals. Certainly one factor contributing to this gap is that automatic systems can not always correctly identify the important topics of an article. Another factor, however, which has received little attention, is that automatic summarizers have poor text generation techniques. Most automatic summarizers rely on extracting key sentences or paragraphs from an article to produce a summary. Since the extracted sentences are disconnected in the original article, when they are strung together, the resulting summary can be inconcise, incoherent, and sometimes even misleading.

We present a cut and paste based text summarization technique, aimed at reducing the gap between automatically generated summaries and human-written abstracts. Rather than focusing on how to identify key sentences, as do other researchers, we study how to generate the text of a summary once key sentences have been extracted.

The main idea of cut and paste summarization is to reuse the text in an article to generate the summary. However, instead of simply extracting sentences as current summarizers do, the cut and paste system will "smooth" the extracted sentences by editing them. Such edits mainly involve cutting phrases and pasting them together in novel ways.

The key features of this work are:

(1) The identification of cutting and pasting operations. We identified six operations that can be used alone or together to transform extracted sentences into sentences in human-written abstracts. The operations were identified based on manual and automatic comparison of human-written abstracts and the original articles. Examples include sentence reduction, sentence combination, syntactic transformation, and lexical paraphrasing.

(2) Development of an automatic system to perform cut and paste operations. Two operations - sentence reduction and sentence combination - are most effective in transforming extracted sentences into summary sentences that are as concise and coherent as in human-written abstracts. We implemented a sentence reduction module that removes extraneous phrases from extracted sentences, and a sentence combination module that merges the extracted sentences or the reduced forms resulting from sentence reduction. Our sentence reduction model determines what to cut based on multiple sources of information, including syntactic knowledge, context, and statistics learned from corpus analysis. It improves the conciseness of extracted sentences, making them concise and on target. Our sentence combination module implements combination rules that were identified by observing examples written by human professionals. It improves the coherence of extracted sentences.

(3) Decomposing human-written summary sentences. The cut and paste technique we propose here is a new computational model which we based on analysis of human-written abstracts. To do this analysis, we developed an automatic system that can match a phrase in a human-written abstract to the corresponding phrase in the article, identifying its most likely location. This decomposition program allows us to analyze the construction of sentences in a human-written abstract. Its results have been used to train and test the sentence reduction and sentence combination module.

In Section 2, we discuss the cut and paste technique in general, from both a professional and computational perspective. We also describe the six cut and paste operations. In Section 3, we describe the