Terms Derived from Frequent Sequences for Extractive Text Summarization[*]

Yulia Ledeneva,[1] Alexander Gelbukh,[1] René Arnulfo García-Hernández [2]

[1] Natural Language and Text Processing Laboratory,
Center for Computing Research, National Polytechnic Institute, DF 07738, Mexico
yledeneva@yahoo.com, www.Gelbukh.com

[2] Instituto Tecnologico de Toluca, Mexico
renearnulfo@hotmail.com

Abstract. Automatic text summarization helps the user to quickly understand large volumes of information. We present a language- and domain-independent statistical-based method for single-document extractive summarization, i.e., to produce a text summary by extracting some sentences from the given text. We show experimentally that words that are parts of bigrams that repeat more than once in the text are good terms to describe the text's contents, and so are also so-called maximal frequent sentences. We also show that the frequency of the term as term weight gives good results (while we only count the occurrences of a term in repeating bigrams).

1 Introduction

A summary of a document is a (much) shorter text that conveys the most important information from the source document. There are a number of scenarios where automatic construction of such summaries is useful. For example, an information retrieval system could present an automatically built summary in its list of retrieval results, for the user to quickly decide which documents are interesting and worth opening for a closer look—this is what Google models to some degree with the snippets shown in its search results. Other examples include automatic construction of summaries of news articles or email messages to be sent to mobile devices as SMS; summarization of information for government officials, businessmen, researches, etc., and summarization of web pages to be shown on the screen of a mobile device, among many others.

The text summarization tasks can be classified into single-document and multi-document summarization. In single-document summarization, the summary of only one document is to be built, while in multi-document summarization the summary of a whole collection of documents (such as all today's news or all search results for

[*] Work done under partial support of Mexican Government (CONACyT, SNI, SIP-IPN, COTEPABE-IPN, COFAA-IPN). The authors thank Rada Mihalcea for useful discussion.