Lexical Cohesion Based Topic Modeling for Summarization

Gonenc Ercan and Ilyas Cicekli

Dept. of Computer Engineering
Bilkent University, Ankara, Turkey
ercangu@cs.bilkent.edu.tr, ilyas@cs.bilkent.edu.tr

Abstract. In this paper, we attack the problem of forming extracts for text summarization. Forming extracts involves selecting the most representative and significant sentences from the text. Our method takes advantage of the lexical cohesion structure in the text in order to evaluate significance of sentences. Lexical chains have been used in summarization research to analyze the lexical cohesion structure and represent topics in a text. Our algorithm represents topics by sets of co-located lexical chains to take advantage of more lexical cohesion clues. Our algorithm segments the text with respect to each topic and finds the most important topic segments. Our summarization algorithm has achieved better results, compared to some other lexical chain based algorithms.

Keywords: text summarization, lexical cohesion, lexical chains.

1 Introduction

Summary is the condensed representation of a document's content. For this reason, they are low cost indicators of relevance. Summaries could be used in different applications both as informative tools for humans and as similarity functions for information retrieval applications. Summaries could be displayed in search results as an informative tool for the user. The user can measure the relevance of a document that he gets as a result of a search on Internet by just looking its summary. In order to measure similarities between documents, their summaries can be used instead of whole documents, and indexing algorithms can index their summaries instead of whole documents.

Depending on its content, summaries can be categorized into two groups: *extract* and *abstract*. If a summary is formed of sentences that appear in the original text, it is called as an *extract*. A summarization system targeting extracts should evaluate each sentence for its importance. Abstracts are the summaries that are formed from paraphrased or generated sentences. Building abstracts has additional challenges.

Different clues can be exploited to evaluate the importance of sentences. There are extractive summarization systems that take advantage of surface level features like word repetition, position in text, cue phrases and similar features that are easy to compute. Ideally, a summarization system should perform full understanding, which is very difficult and only domain dependant solutions are currently available.

Some summarization algorithms including ours, rely on more sophisticated clues that require deeper analyses of the text. A meaningful text is not a random sequence of