LexTrim: A Lexical Cohesion based Approach to Parse-and-Trim Style Headline Generation

Ruichao Wang, Nicola Stokes, William Doran, Eamonn Newman,
John Dunnion, Joe Carthy

Intelligent Information Retrieval Group, Department of Computer Science,
University College Dublin, Ireland.
{rachel, nicola.stokes, william.doran, eamonn.newman,
john.dunnion, joe.carthy}@ucd.ie

Abstract. In this paper we compare two parse-and-trim style headline generation systems. The Topiary system uses a statistical learning approach to finding topic labels for headlines, while our approach, the LexTrim system, identifies key summary words by analysing the lexical cohesion structure of a text. The performance of these systems is evaluated using the ROUGE evaluation suite on the DUC 2004 news stories collection.

1 Introduction

A headline is a very short summary (usually less than 10 words) describing the essential message of a piece of text. Like other types of summaries, news story headlines are used to help a reader to quickly identify information that is of interest to them in a presentation format such as a newspaper or a website. Although newspaper articles have already been assigned headlines, there are other types of news text sources, such as transcripts of radio and television broadcasts, where this type of summary information is missing. In 2003 the Document Understanding Conference (DUC) added the headline generation task to their annual summarisation evaluation. This task was also included in the 2004 evaluation plan where summary quality was automatically judged using a set of n-gram word overlap metrics called ROUGE [1]. The best performing system at this workshop was the Topiary approach [2] which generated headlines by combining a set of topic descriptors generated from the DUC 2004 corpus with a compressed version of the lead sentence, e.g. (Topic Descriptors) *BIN_LADEN EMBASSY BOMBING:* (Compressed Lead Sentence) *FBI agents this week began questioning relatives of the victims.*

Topiary-style summaries perform well in the ROUGE evaluation for a number of reasons. Firstly, summarisation researchers have observed that the lead sentence of a news story is in itself often an adequate summary of the text. However, it has also been observed that additional important information about a topic may be spread across other sentences in the text. The success of the Topiary-style summaries at DUC 2004 can be attributed to fact that this technique takes both of these observations into consideration when generating titles.