Efficient Randomized Algorithms for Text Summarization

Ahmed Mohamed and Sanguthevar Rajasekaran

Department of Computer Science & Engineering,
University of Connecticut, Storrs, CT 06268
{amohamed, rajasek}@engr.uconn.edu

Abstract. Text summarization is an important problem since it has numerous applications. This problem has been extensively studied and many approaches have been pro-posed in the literature for its solution. One such interesting approach is that of posing summarization as an optimization problem and using genetic algorithms to solve this optimization problem. In this paper we present elegant randomized algorithms for summarization based on sampling. Our experimental results show that our algorithms yield better accuracy than genetic algorithms while significantly saving on time. We have employed data from Document Understanding Conference 2002 and 2004 (DUC-2002, DUC-2004) in our experiments.

1 Introduction

Document summarization has been the focus of many researchers for the last decade, due to the increase in on-line information and the need to find the most important information in a (set of) document(s). There are different approaches to generate summaries depending on the task the summarization is required for. Summarization approaches usually fall into 3 categories (Mani and Maybury, 1999):

- *Surface-level* approaches tend to represent information in terms of shallow features, which are then selectively combined together to yield a salience function used to extract information;
- *Entity-level* approaches build an internal representation for text, modeling text entities and their relationships. These approaches tend to represent patterns of connectivity in the text (e.g., graph topology to help determine what is salient);
- *Discourse-level* approaches model the global structure of the text, and its relation to communicative goals.

Some approaches mix between two or more of the features of the above mentioned approaches, and the approaches discussed in this paper fall in that category, since they involve both surface and entity levels' features.

© A. Gelbukh (Editor)
Advances in Natural Language Proccesing
Research in Computing Science 18, 2006, pp. 195-200