A Scalable MMR Approach to Sentence Scoring for Multi-Document Update Summarization

Florian Boudin [♮] and **Marc El-Bèze** [♮]
[♮] Laboratoire Informatique d'Avignon
339 chemin des Meinajaries, BP1228,
84911 Avignon Cedex 9, France.
florian.boudin@univ-avignon.fr
marc.elbeze@univ-avignon.fr

Juan-Manuel Torres-Moreno [♮,♭]
[♭] École Polytechnique de Montréal
CP 6079 Succ. Centre Ville H3C 3A7
Montréal (Québec), Canada.
juan-manuel.torres@univ-avignon.fr

Abstract

We present SMMR, a scalable sentence scoring method for query-oriented update summarization. Sentences are scored thanks to a criterion combining query relevance and dissimilarity with already read documents (history). As the amount of data in history increases, non-redundancy is prioritized over query-relevance. We show that SMMR achieves promising results on the DUC 2007 update corpus.

1 Introduction

Extensive experiments on query-oriented multi-document summarization have been carried out over the past few years. Most of the strategies to produce summaries are based on an extraction method, which identifies salient textual segments, most often sentences, in documents. Sentences containing the most salient concepts are selected, ordered and assembled according to their relevance to produce summaries (also called extracts) (Mani and Maybury, 1999).

Recently emerged from the Document Understanding Conference (DUC) 2007[1], update summarization attempts to enhance summarization when more information about knowledge acquired by the user is available. It asks the following question: has the user already read documents on the topic? In the case of a positive answer, producing an extract focusing on only new facts is of interest. In this way, an important issue is introduced: redundancy with previously read documents (history) has to be removed from the extract.

A natural way to go about update summarization would be extracting temporal tags (dates, elapsed times, temporal expressions...) (Mani and Wilson, 2000) or to automatically construct the timeline from documents (Swan and Allan, 2000). These temporal marks could be used to focus extracts on the most recently written facts. However, most recently written facts are not necessarily new facts. Machine Reading (MR) was used by (Hickl et al., 2007) to construct knowledge representations from clusters of documents. Sentences containing "new" information (i.e. that could not be inferred by any previously considered document) are selected to generate summary. However, this highly efficient approach (best system in DUC 2007 update) requires large linguistic resources. (Witte et al., 2007) propose a rule-based system based on fuzzy coreference cluster graphs. Again, this approach requires to manually write the sentence ranking scheme. Several strategies remaining on post-processing redundancy removal techniques have been suggested. Extracts constructed from history were used by (Boudin and Torres-Moreno, 2007) to minimize history's redundancy. (Lin et al., 2007) have proposed a modified Maximal Marginal Relevance (MMR) (Carbonell and Goldstein, 1998) re-ranker during sentence selection, constructing the summary by incrementally re-ranking sentences.

In this paper, we propose a scalable sentence scoring method for update summarization derived from MMR. Motivated by the need for relevant novelty, candidate sentences are selected according to a combined criterion of query relevance and dissimilarity with previously read sentences. The rest of the paper is organized as follows. Section 2

[1] Document Understanding Conferences are conducted since 2000 by the National Institute of Standards and Technology (NIST), http://www-nlpir.nist.gov

Coling 2008: Companion volume – Posters and Demonstrations, pages 23–26
Manchester, August 2008