Automatic Annotation of Corpora for Text Summarisation: A Comparative Study

Constantin Orăsan

Research Group in Computational Linguistics
School of Humanities, Languages and Social Sciences
University of Wolverhampton
Stafford St., Wolverhampton, WV1 1SB, UK,
Email: C.Orasan@wlv.ac.uk,
WWW home page: http://www.wlv.ac.uk/~in6093/

Abstract. This paper presents two methods which automatically produce annotated corpora for text summarisation on the basis of human produced abstracts. Both methods identify a set of sentences from the document which conveys the information in the human produced abstract best. The first method relies on a greedy algorithm, whilst the second one uses a genetic algorithm. The methods allow to specify the number of sentences to be annotated, which constitutes an advantage over the existing methods. Comparison between the two approaches investigated here revealed that the genetic algorithm is appropriate in cases where the number of sentences to be annotated is less than the number of sentences in an ideal gold standard with no length restrictions, whereas the greedy algorithm should be used in other cases.

1 Introduction

Annotated corpora are essential for most branches in computational linguistics, including automatic summarisation. Within computational linguistics, annotated corpora are normally considered a gold standard, and are used to train machine learning algorithms and evaluate the performance of automatic summarisation methods. In order to be used for these purposes, the annotation usually indicates the importance of each sentence. In this paper, the term gold standard is used only to refer to sets of sentences marked as important, and not to the whole annotated document. This approach was taken in order to facilitate the explanation to follow, and it does not prevent the methods presented in this paper being used to produce gold standards where the important sentences are annotated within the document.

The decision as to whether a sentence is important enough to be annotated can be taken either by humans or by programs. When humans are employed in the process, producing such corpora becomes time consuming and expensive. Methods which automatically build annotated corpora are cheap, but have some drawbacks. Section 2 presents brief details about the existing methods for producing such corpora for text summarisation.