論 文 の 内 容 の 要 旨

Abstract

論文題目　Reconstruction and Bayesian Nonparametrics Based Multi-Document Summarization

（再構成とノンパラメトリックベイズ手法に基づく複数文書要約）

氏　名　馬　騰飛

Document summarization aims to extract the most important information from a single document or a cluster of documents. It plays an increasingly important role with the exponential growth of web documents. Over the past half a century, there are various approaches proposed to solve the problem from many different perspectives, most of which directly selected summary sentences using sentence ranking or greedy selection approaches. Generally the quality of a summary should be determined by three properties: relevance, diversity and coverage. However, the sentence ranking methods and greedy selection approaches hardly simultaneously consider the three properties, and they could not provide a solution which selects best overall sentences. Therefore optimizing all three properties jointly with a global sentence selection procedure has been attractive.

In this thesis, we solve the summarization problem and unify all aims from a novel perspective. We assumed that original documents should be reconstructed from the best summary with least information loss. From this assumption, we first propose a reconstruction

based optimization framework for multi-document summarization. We brought in various information-theoretic measures and regarded the “minimum distortion” as the objective function. We defined three reconstruction models for optimization of the distortion measures, gaining state-of-the-art summarization results.

Moreover, we studied a new problem in summarization called summary length determination. Traditional summarization systems require users to pre-define a bounded length for summaries. However, how to find the proper summary length is quite a problem; and keeping all summaries restricted to the same length is not always a good choice. Following our reconstruction assumption, we developed a Bayesian nonparametric model to automatically determine the proper summary length. The model is demonstrated to own good summary qualities and to determine rational summary length. Finally, we consider the case that the real categories of documents are not known, and advanced the hybrid nested Dirichlet process to extend traditional Bayesian nonparametric topic analysis, which is a preprocessing step for document summarization. The topic analysis itself also provides visualization for abstractive summarization of the documents.