Statistical Approaches Toward Automatic Title Generation

Rong Jin
Language Technologies Institute
School of Computer Science
Carnegie Mellon University

A title is a compact representation that can help people capture a document's main idea without having to read through the entire document. Automatic title generation is a difficult natural language processing problem. It requires both the understanding of the essential content of a document and the knowledge of creating a headline that actually reflects the content with a few words. Therefore, the task of automatic title generation involves both natural language understanding and natural language synthesis.

Previous statistical approaches to title generation are based on the paper by Witbrock and Mittal (1999), where the process of title generation is divided into a phase of selecting title words for a document and a phase of organizing title words into a human readable sequence. In the work of Witbrock and Mittal and follow-ups, the phase of title word ordering is accomplished using an n-gram statistical language model and the phase of title word selection is realized by a Naïve Bayes method. In this thesis, we examine and compare seven different statistical methods for title word selection, including a nearest neighbor approach, K nearest neighbor approach, a decision tree approach, a statistical translation approach, a reverse information retrieval approach, a Naïve Bayes approach with a limited vocabulary, and a Naïve Bayes approach with a full vocabulary. In general, methods that are able to take into account all the words in the test document work better than methods that only consider a subset of document words.

The other dimension of this thesis is on the study of new model for title generation. A general probabilistic framework is proposed for the statistical approaches toward title generation, where previous works on title generation can be treated as special case of this framework. Furthermore, a new probabilistic model for title generation is derived from the general framework, which is able to overcome the problems with the previous statistical model on both the phase of title word selection and the phase of title word ordering. In the new probabilistic model, an intermediate state named 'information source' is proposed so that both the title and the document are created from this state. Unlike the previous work on title generation where titles are created directly from documents, in this new model, we will first infer the possible 'information sources' for a document and generate a title from those potential 'information sources'. Empirical studies over two different datasets have shown that the proposed model is able to outperform the previous model for title generation substantially.

Finally, we extend the title generation model to other related fields such as information retrieval and text categorization. For information retrieval, the basic idea is to treat a query as a 'title' and the problem of finding documents relevant to the query can be viewed as the problem of finding documents that fit in with the 'title'. Empirical studies over information retrieval have indicated that approaches based on the title generation model appear to work well for certain types of data.

The full version of the thesis is available on line http://www.cse.msu.edu/~rongjin/.