CHALMERS

Multi-Document Summarization and Semantic Relatedness

OLOF MOGREN

Thesis to be defended in public on **November 20**, **2015** at **10:00** in room **ML2, Hörsalsvägen 7B, Chalmers** for the Degree of Licentiate of Engineering. The defense will be conducted in English.

Discussion leader:
Tapani Raiko, Assistant Professor, Aalto University

The thesis is available at:

Chalmers University of Technology
412 96 Göteborg, Sweden
Phone: +46 (0)31-772 10 51

Multi-Document Summarization and Semantic Relatedness

OLOF MOGREN

Department of Computer Science and Engineering
Chalmers University of Technology

Abstract

Automatic summarization is the process of presenting the contents of written documents in a short, comprehensive fashion. Many approaches have been proposed for this problem, some of which extract content from the input documents (extractive methods), and others that generate the language in the summary based on some representation of the document contents (abstractive methods).

This thesis is concerned with extractive summarization in the multi-document setting, and we define the problem as choosing the most informative sentences from the input documents, while minimizing the redundancy in the summary. This definition calls for a way of measuring the similarity between sentences that captures as much as possible of the meaning. We present novel ways of measuring the similarity between sentences, based on neural word embeddings and sentiment analysis. We also show that combining multiple sentence similarity scores, by multiplicative aggregation, helps in the process of creating better extractive summaries.

We also discuss the use of information extraction for improving the quality of automatic summarization by providing ways of assessing the salience of information elements, as well as helping with the fluency of the output and providing the temporal dimension.

Furthermore, we present graph-based algorithms for clustering words by co-occurrence, and for summarizing short online user-reviews by computing bicliques. The biclique algorithm provides a fast, simple algorithm for summarization in many e-commerce settings.

Keywords: Multi-Document Summarization, Semantic Relatedness, Natural Language Processing, Neural Word Embeddings