# Event-based Multi-document Summarization

Luís Carlos dos Santos Marujo
Language Technologies Institute, Carnegie Mellon University, USA
Instituto Superior Técnico, Universidade de Lisboa, Portugal
INESC-ID Lisbon, Portugal
*lmarujo@cs.cmu.edu*

**Abstract**

Daily amount of news reporting real-world events is growing exponentially. At the same time, Organizations are looking for information about current and past events that affects them, such as mergers and acquisitions of companies. The Organizations need to obtain event information in a fast and summarized form to make decisions. Event-based retrieval and summarization systems offer an efficient solution to this problem. Most summarization research work uses news stories. Although this type of documents is characterized by conveying information about events, almost all work concentrates on approaches that do not take into account this aspect.

The proposed multi-document summarization methods are based on the hierarchical combination of single-document summaries. We improved our multi-document summarization methods using event information. Our approach is based on a two-stage single-document method that extracts a collection of key phrases, which are then used in a centrality-as-relevance passage retrieval model. To adapt centrality-as-relevance single-document summarization for multi-document summarization that is able to use event information, we needed a good and adaptable baseline system. Because the key phrase extraction play a significant role in the summarization, we improved a state-of-the-art key phrase extraction toolkit using four additional sets of semantic features. The event detection method is based on Fuzzy Fingerprint, which is a supervised method trained on documents with annotated event tags. We explored three different ways to integrate event information, achieving state-of-the-art results in both single and multi-document summarization using filtering and event-based features. To cope with the possible usage of different terms to describe the same event, we explored distributed representations of text in the form of word embeddings, which contributed to improve the multi-document summarization results.

The automatic evaluation and user study performed show that these methods improve upon current state-of-the-art multi-document summarization systems on two mainstream evaluation datasets, DUC 2007 and TAC 2009. We show a relative improvement in ROUGE-1 scores of 16% for TAC 2009 and of 17% for DUC 2007. We have also obtained improvements in ROUGE-1 upon current state-of-the-art single-document summarization systems of between 32% in clean data and 19% in noisy data. These improvements derived from the inclusion of key phrases and event information. The extraction of key phrases was also refined with additional pre-processing steps and features, which lead to a relative improvement in NDCG scores of 9%. The introduction of Fuzzy Fingerprints for event detection enabled

the detection of all event types, while the best competitor, an SVM with enhanced features, only detects roughly 85% of the different types of events. This lead to a large increase in the G-Mean and variants results when using the Fuzzy Fingerprints method.

This doctoral work was carried under the Dual Ph.D. program in Languages and Information Technologies of Carnegie Mellon Portugal Program. The program took place at both the Language Technologies Institute at Carnegie Mellon University (CMU) and at the Department of Computer of Science and Engineering at Instituto Superior Técnico (IST). The doctoral work was performed under the supervision of University Professor Jaime Carbonell (CMU), Distinguished Career Professor Anatole Gershman (CMU), Assistant Professor David Martins de Matos (IST), and Assistant Professor Joo P. Neto (IST). Full Professor Ricardo Baeza-Yates (Yahoo!Research/Univ. Pompeu Fabra/Univ. de Chile), Research Professor Eduard Hovy (CMU), Associate Professor Ana Paiva (IST), and Full Professor Isabel Trancoso served as dissertation committee members.

**Available online at:** http://l2f.inesc-id.pt/ ldsm/dissertation-lmarujo.pdf