Query-focused Summarization: Summarization is Easy, Retrieval is Hard

Tal Baumel
Ben-Gurion University

Raphael Choen
Ben-Gurion University

Jumana Nassour
Ben-Gurion University

Michael Elhadad
Ben-Gurion University

Abstract

The task of Query-focused Summarization (QFS) consists of producing a summary of a documents cluster that answers a specific query. As in generic summarization (GS), QFS must select content which is *central* and *non-redundant*, but in contrast, QFS must also be *responsive* to the query. Despite this dissimilarity, it has been observed that algorithms for QFS and GS obtain very similar results. We investigate this surprising fact.

We compare the DUC 2005 dataset with a new dataset, the Query Chain Focused Summarization (QCFS) dataset where document clusters are still centered around a main topic, but contain more variation than in DUC 2005. We compare the behavior of representative baseline sentence extraction summarization algorithms on these 2 datasets.

We model QFS as a 2-stage process: retrieve content most relevant to the query from the document cluster; and extract a concise non-redundant subset of central sentences from the relevant content. We compare different retrieval models and find that the quality of the retrieval model dominates: given a strong passage retrieval method, QFS is improved; but if the input document cluster is varied enough, baseline summarization algorithms do not succeed in jointly retrieving relevant content and summarizing it. The importance of the retrieval component was hidden in previous research because standard datasets were massively relevant to queries. We find that a combination of a strong relevance-based passage retrieval model combined with a simple KLSum summarization component produces strong QFS results.

Reference

Badrinath, Rama and Venkatasubramaniyan, Suresh and Veni Madhavan. 2011. Improving query focused summarization using look-ahead strategy, *ECIR 2011*.

Bendersky and Kurland. 2008. Utilizing Passage-based Language Models for Document Retrieval. *ECIR 2008, LNCS 4956*, 162-174.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3 (2003): 993-1022.

Celikyilmaz, Asli, and Dilek Hakkani-Tur. 2010. A hybrid hierarchical model for multi-document summarization. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics*, ACL 2010.

Dang H.T.. 2005. Overview of DUC 2005. *National Institute of Standards and Technology* (NIST), http://duc.nist.gov/pubs.html#2005.

Daumé III, Hal, and Daniel Marcu. 2006. Bayesian query-focused summarization. *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*. *ACL 2006*.

Gupta, Nenkova and Jurafsky. 2007. Measuring Importance and Query Relevance in Topic-focused Multi-document Summarization, *ACL 2007*.

Haghighi and Vanderwende. 2009. Exploring content models for multi-document summarization. *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. ACL 2009.

Dogan, Islamaj, et al. 2009. Understanding PubMed User Search Behavior through Log Analysis. Database: *The Journal of Biological Databases & Curation* 2009.

Jagadeesh, J., Prasad Pingali, and Vasudeva Varma. 2005. A relevance-based language modelling approach to DUC 2005. *Proceedings of Document Understanding Conferences (along with HLT-EMNLP 2005)*, Vancouver, Canada.

Jiwei Li, Sujian Li. 2013. A Novel Feature-based Bayesian Model for Query Focused Multi-document Summarization. *Transactions of the Association for Computational Linguistics*, 1 (2013) 89–98.

Kurland, Oren, and Lillian Lee. 2010. PageRank without Hyperlinks: Structural Reranking using Links Induced by Language Models. *ACM Transactions on Information Systems (TOIS)* 28.4 (2010): 18.

Lavrenko, Victor, and W. Bruce Croft. 2001. Relevance based language models. *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2001*.

Liu, Xiaoyong, and W. Bruce Croft. 2002. Passage retrieval based on language models. *Proceedings of the eleventh international conference on Information and knowledge management. ACM, 2002*.

Marchionini, G. 2006. "Exploratory search: from finding to understanding." *Commun. ACM* 49(4): 41-46.

Otterbacher J., Erkan G., and Radev D.. 2009. Biased LexRank: Passage retrieval using random walks with question-based priors. *Information Processing & Management* 45.1 (2009): 42-54.

Ponte, Jay M., and W. Bruce Croft. 1998. A Language Modelling Approach to Information Retrieval. *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 1998.

Vanderwende, Suzuki, Brockett and Nankova. 2007. Beyond SumBasic: Task-Focused Summarization with Sentence Simplification and Lexical Expansion, *Information Processing and Management, Special Issue on Summarization*, 43(6).

White R.W and. Roth R.A., 2009. *Exploratory Search: Beyond the Query–Response Paradigm*, Morgan Claypool.

Wei, Xing, and W. Bruce Croft. 2006. LDA-based Document Models for Ad-hoc Retrieval." *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2006.

Yi X. and Allan J. 2009. A comparative study of utilizing topic models for information retrieval. *Advances in Information Retrieval*. Springer Berlin Heidelberg, 2009. 29-41.