Query-Based Summarization: A survey

Mariana Damova[1], Ivan Koychev[2]*

[1] Ontotext, Sofia, Bulgaria
mariana.damova@ontotext.com
[2] Faculty of Mathematics and Informatics, University of Sofia, Bulgaria
koychev@fmi.uni-sofia.bg

Abstract. This paper presents a survey of recent extractive query-based summarization techniques. We explore approaches for single document and multi-document summarization. Knowledge-based and machine learning methods for choosing the most relevant sentences from documents with respect to a given query are considered. Further, we expose tailored summarization techniques for particular domains like medical texts. The most recent developments in the field are presented with opinion summarization of blog entries.

1 Introduction

This survey is motivated by the idea of making e-books more intelligent [10], in par-ticular enabling them to "answer" users' queries. To find the needed information in books users usually do not want to spend a long time searching, browsing or skim-ming them. They will be happy to have a "guru" nearby that can provide them with the right answer almost simultaneously. For this purpose we had a close look at the area of automated text summarization. Recently, with the increasing of information available online, those approaches have been developed very extensively. In the realm of automatic summarization different kinds of summarization have been attempted. Along with [8] we distinguish between the following types of summaries according to specific criteria.

Summary construction methods:

- *abstractive* summaries produce generated text from the important parts of the documents;
- *extractive* summaries identify important sections of the text and use them in the summary as they are.

Number of Sources for the summary

- *single document* summaries represent a single document.
- *multi-document* summaries are produced from multiple documents and they have to deal with three major problems:
  - recognizing and coping with *redundancy*;
  - identifying *important differences* among documents;
  - ensuring *summary coherence*.

* Also associated with Institute of Mathematics and Informatics - BAS

Summary trigger:

- *generic summaries* present in concise manner the main topics of a given text;
- *query-based summaries* are constructed as an answer to an information need ex-pressed by a user's query, where:
  - *indicative* summaries point to information of the document, which helps the user to decide whether the document should be read or not;
  - *informative* summaries provide all the relevant information to represent the original document.

The technology of summarization benefits from an intensive development in the last years. The DUC conferences and competitions contribute to this evolution. How-ever, not many surveys of the field have been produced. Das et al [5] presents a tho-rough overview of the field, which starts with historical information from 50 years ago when the field of summarization was shaped. It investigates approaches in the realm of single document and multi-document summarization and pay special atten-tion to the evaluation techniques used to rank summarization systems.

This paper presents a survey of *extractive query-based* summarization techniques with approaches for single document and multi-document summarization based on knowledge-based or machine learning methods for choosing the most relevant sen-tences from the documents with respect to a given query. The selected applications are viewed as different use cases of summarization systems.

2 General Purpose Query-Based Summarization Approaches

2.1 Approaches based on Document Graphs

Jagadeesh et. al. presents an extractive multi-document summarization method, that represents the documents as graphs [7]. The document graph is produced from a plain text document, by first tokenizing, then parsing it into NPs. The relations are generated following heuristic rules. A centric graph is produced from all source documents and guides the summarizer in its search for candidate sentences to be added to the output summary. The query-based summarization is done in three ways:

a. The centric graph of the documents is compared with the concepts in the query;
b. The graph of the document and a graph of the query are generated and the simi-larity between each sentence and the query are measured, the best sentences or-dered chronologically according to their appearance in the input documents pro-duce the summary;
c. A query modification technique is used by including the graph of a selected sen-tence to the query graph.

The best results come from summarizer (b).

The method in [1] shows how answers to questions can be improved by

extracting more information about the topic with summarization techniques, based on text analysis for query-based single document extracts. The RST (Rhetorical Structure Theory) is used to create a graph representation of the document - a weighted graph in which each node represents a sentence and the weight of an edge represents the distance between two sentences. If a sentence is relevant to an answer, a second sentence is evaluated as relevant too, based on the weight of the path between the two sentences. The approach is of two steps. First the relations between sentences are defined in a discourse graph. Then, a graph search algorithm is used to extract the most salient sentences from the graph for the summary. The sentences with the cheapest path from the entry point are selected.

2.2 Approaches using linguistics

The approach in [4] is based on HMM (Hidden Markov Model) for sentence selec-tion within a document and a question answering algorithm for generation of a multi-document summary. The developed system CLASSY makes use of linguistics, pat-terns with lexical cues for sentence and phrase elimination. Typographic cues like title paragraph and other specific paragraphs are used to detect the topic description and obtain question-answering capability. In a separate pre-processing step a named entity identifier ran on all document sets, generates lists of entities for the categories of location, person, date, organization, and evaluates each topic description looking for keywords. After all linguistic processing, and query terms generated, HMM model is used to score the individual sentences classifying them as summary and non-summary sentences.

The approach in [10] is a multi-document summarizer that uses query-interpretation to analyze the given user profile and topic narrative for document clus-ters before creating the summary. It is based on basic elements, a head-modifier rela-tion triple representation of document content which is created by using a parser to produce a syntactic parse tree and a set of ‘cutting rules’ to extract just the valid basic elements from the tree. Scores are assigned to the sentences based on their basic elements, and then standard filtering and redundancy removal techniques are applied before generating the summaries which consists in outputting the topmost sentences until the required sentence limit is reached.

2.3 Machine-learning approaches

In the approach of [6] information retrieval techniques are combined with summarization techniques in producing the summary extracts. This approach incorporates a new notion of sentence importance independent of query into the final scoring. The sentences are scored using a set of features from all sentences, normalized in a maximum score and the final score of a sentence is calculated using a weighted linear combination of the individual feature values. The top scoring sentences are selected for the summary until the summary length reaches the desired limit. A new feature - Information Measure - captures the sentence importance based on the distribution of its constituent words in the domain corpus. The formula consists of two parts:

a. a query dependent ranking of a document/sentence;
b. the explicit notion of importance or prior of a document/sentence.

This allows query independent forms of evidence to be incorporated into the ranking process.

FastSum [9] is based on word-frequency features of clusters, documents and top-ics. Summary sentences are ranked by a regression Support Vector Machine. The me-thod involves sentence splitting, filtering candidate sentences and computing the word frequencies in the documents of a cluster, topic description and the topic title. All sentences in the topic cluster are ranked for summarizability. The topic contains a topic title and a topic description. The former is a list of key words or phrases describing the topic, and the later contains the query or queries. The features used are word-based and sentence-based. Word-based features are computed based on the probability of words for the different containers. Sentence-based features include the length and position of the sentence in the document. Because of adopting Least Angle Regression, a new approach for selecting features, FastSum can rely on a minimal set of features leading to fast processing times, e.g. 1250 news documents per 60 seconds.

3 Applications Tailored Systems

3.1 Subject domain ontology based approach

A query-based Medical Information Summarization System Using Ontology Know-ledge [2] proposes a technique using UMLS and ontology from National Library of Medicine. The summarization algorithm is term-based, and only terms defined in UMLS are recognized and processed. The summarization procedure is:

a. revising the query with UMLS ontology knowledge;
b. calculating distance of each sentence in the document to the finalized query;
c. calculating pair-wise distances among the candidate sentences, then dividing the candidate sentences into groups based on a threshold and selecting highest-ranked one from each group.

When it is determined which sentences will be included in the summary, three dif-ferent “scores” are generated and normalized with the length of the sentence.

3.2 Opinion Summarization

Systems that can identify trends in news streams will become more and more impor-tant with the time. This first report of automatic sentiment summarization [3] in the legal domain is based on processing a set of legal questions with a system consisting of a semi-automatic Web blog search module and FastSum. The input to the summa-rization task is some opinion-related questions about the target, and a set of docu-ments that contain answers to the questions. The output is a summary for each target that summarizes the answers to the questions. FastSum is modified for sentiment integration. A filter identifies sen-

tences that are unlikely to be in a good summary. Another filter deals with the sentiment of the sentence, using a sentiment tagger to de-termine the sentiment of the extracted sentences. It is a sentiment polarity tagger based on unigram term lookup using gazetteers of positive and negative polarity indi-cating terms based on the General Inquirer. The final summary is created from the ranked sentence list after a redundancy removal step. The relevant and substantive blog entries on legal topics of interest are harvested by blog search engines, e.g. blog-searchengine.com.

4 Conclusions

This paper presented an overview of a variety of query-based summarization ap-proaches implemented in different applications. Our goal was to present different use cases of query-based summarization. All described systems participate in the DUC competitions in the last couple of years and some of them score highly (top 4, top 7, top 6) like CLASSY and FastSum.

Acknowledgements. This research is supported by the SmartBook project, subsidized by the Bulgarian National Science Fund, under Grant D002-111 /15.12.2008.

References

1. Wauter Bosma (2005). Query-Based Summarization using Rhetorical Structure Theory. In: Ton van der Wouden, Michaela Poss, H. Reckman and C. Cremers, ed., 15th Meeting of CLIN, 2005
2. Ping Chen, Rakesh Verma (2006). A Query-based Medical Information Summarization System Using Ontology Knowledge. In Proceedings of the 19th IEEE Symposium on Computer-Based Medical Systems (CBMS'06) , 2006
3. Jack G. Conrad, Jochen L. Leidner, Frank Schilder, Ravi Kondadadi (2009). Query-based Opinion Summarization for Legal Blog Entries. In Proceedings of the 12th International Conference on Artificial Intelligence and Law (ICAIL 2009), ACM Press, Barcelona, Spain
4. John M. Conroy, Judith D. Schlesinger, Jade Goldstein Stewart (2005). CLASSY Query-Based Multi-Document Summarization. In DUC 05 Conference Proceedings, Boston, USA
5. Dipanjan Das, Andre F.T. Martins (2007). A Survey on Automatic Text Summarization. Literature Survey for the Language and Statistics II course at CMU, Pittsburg, USA, 2007
6. Jagadeesh J, Prasad Pingali, Vasudeva Varma (2007). Capturing Sentence Prior for Query-Based Multi-Document Summarization. In RIAO, http://dblp.uni-trier.de
7. Ahmed A. Mohamed, Sanguthevar Rajasekaran (2006). Query-Based Summarization Based on Document Graphs. In Proceedings of IEEE International Symposium on Signal Processing and Information Technology, pp.408-410, Vancouver, Canada, 2006
8. Dragomir Radev. Text summarization - Tutorial at ACM SIGIR Conference, Sheffield, UK, July 25, 2004 http://www.summarization.com/sigirtutorial2004.ppt
9. Frank Schilder, Ravikumar Kondadadi (2008). FastSum: Fast and accurate query-based multi-document summarization. In Proceedings of the 46th meeting of the Association for Computational Linguistics, Columbus, Ohio
10. Koychev I., Nikolov, R. and Dicheva D.: SmartBook: The New Generation e-Book, Proc. of BooksOnline'09 Workshop, in conjunction with ECDL 2009, Corfu, October 2, 2009
11. Liang Zhou, Chin-Yew, Eduard Hovy (2006). Summarizing Answers for Complicated Questions. In Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC-2006), Genoa, Italy