Towards Interactive Summarization of Large Document Collections

Benjamin Hättasch
TU Darmstadt, Germany
benjamin.haettasch@cs.tu-darmstadt.de

ABSTRACT

We present a new system for custom summarizations of large text corpora at interactive speed. The task of producing textual summaries is an important step to understand collections of topic-related documents and has many real-world applications in journalism, medicine, and many more. Our system consists of a sampling component that ranks and selects sentences from a given corpus and uses an integer-linear program (ILP) to produce the summary. Both components are called multiple times to improve the quality of the summarization iteratively. The human is brought into the loop to gather feedback in every iteration about which aspects of the intermediate summaries satisfy their individual information needs. That way, our system can provide a similar quality level as an ILP-approach working on the full corpus but with a constant runtime independent of the corpus size.

1 INTRODUCTION

Users like journalists or lawyers confronted with a large collection of unknown documents need to find the overall relation and event structure of those documents. An important step for this understanding process is to produce a concise textual summary that captures the information most relevant to a user's aims (e.g. degree of details or covered topics). While many automatic text summarization approaches have been suggested, there exist only a few that produce different summaries targeted at the individual user. One of those is the system recently proposed by P.V.S. and Meyer [1]. A major limiting factor however is that their system does not scale for large corpus sizes since the runtime of their approach which uses an ILP solver at its core grows exponentially with the amount of sentences and may take hours for each iteration. This hinders the user from performing an adequate amount of feedback rounds to get a suitable level of quality and customization.

Therefore, in our work we build upon their system but introduce a ranking based sampling component. That results in a constant computation time of each iteration depending on the sample size instead of the corpus size. With this new approximate summarization model we can guarantee interactive speeds even for large text collections to keep the user engaged in the process. The original system consists of a web-based interface that allows the user to provide feedback and a backend which computes the summaries using an ILP. The user requests a summary and can then annotate the concepts of the summary i.e. mark them as important or unimportant for her current goal. This process will be repeated iteratively until the user is satisfied with the quality.

In order to get a better overview of how the system works, we recommend the readers to watch this video: http://vimeo.com/257601765

DESIRES 2018, August 2018, Bertinoro, Italy

2 OVERVIEW

The main idea of this work is to enable the original system to achieve interactive response time on arbitrary large document collections with a similar quality of the resulting summary. A study [2] has shown that even small delays (more than 500 ms) significantly decrease a user's activity level, dataset coverage, and insight discovery rate, hence one should aim for lower runtimes.

Instead of looking at the complete document collection in every iteration, our approach only considers a sample from the documents per iteration and thus trades performance for quality of the summary. For creating the sample, two important factors thus play a role: The first factor is the *sample size* (i.e., the number of sentences in the sample), which determines the runtime of the summarization method; the second factor is the sampling procedure, that determines *which sentences* are part of the sample.

For deciding about the sample size, we need to be able to estimate the runtime for solving the ILP which mainly depends on its complexity (in number of constraints). In order to do so, we devised a cost function that maps the number of constraints to an estimated runtime. We use this function to derive the maximum sample size k such that the runtime stays below a chosen interactivity threshold.

For deciding which sentences should be contained in the sample, we developed a novel heuristic called *information density* that is computed for each sentence. It ranks the sentences by the weight density of concepts in it normalized by the sentence length. We then only select the top-k sentences based on this heuristic. The intuition is that sentences with a higher information density (containing more concepts rated as important) are more relevant to the user. With this sampling strategy, we are already able to archive a similar quality as the original system at a fraction of the runtime.

Our future work includes developing more advanced sampling strategies that can further improve the quality and increase the amount of feedback on different concepts. One direction would be to devise stratified sampling strategies using additional importance measures for (groups of) sentences. Furthermore, in addition to the current oracle-based approach for evaluation that gives feedback according to reference summaries we plan a user study.

REFERENCES

[1] P. V. S. Avinesh and C. M. Meyer. Joint optimization of user-desired content in multi-document summaries by learning from user feedback. In *ACL*, pages 1353–1363. ACL, 2017.
[2] Z. Liu and J. Heer. The effects of interactive latency on exploratory visual analysis. *IEEE transactions on visualization and computer graphics*, 20:2122–2131, 2014.

This work has been supported by the German Research Foundation as part of the Research Training Group AIPHES under grant No. GRK 1994/1.