CoDiS Lab

ISDS

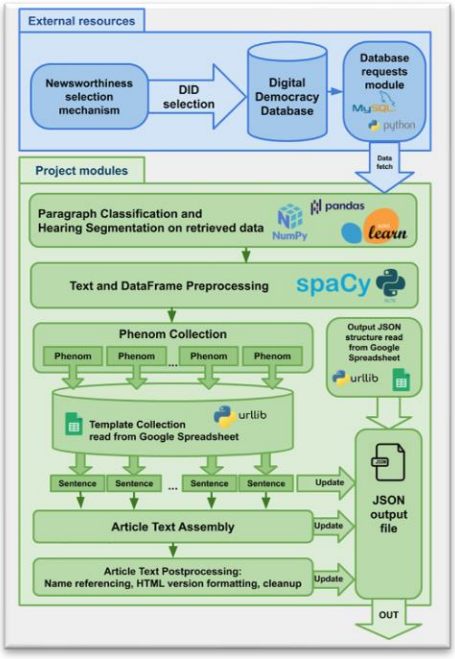
Master Theses

Automatic Summary Generation from Legislative Proceedings

Student: Anastasiia Klimashevskaia
Supervisor: Christian Gütl
Co-Supervisor: Foaad Khosmood (California Polytechnic State University)

Computer scientists have been trying to tackle the task of transcript summarization for decades, introducing different techniques and solutions, broadening the experience in both extractive and abstractive summarization. However, the field of transcript text summarization appears to be less researched and fairly new. The methods of summarization for articles or other well-structured, grammatically correct texts are quite often not applicable in such a case at all or yield poor results. Moreover, transcripts with several speechmakers and various narratives require taking the speakers into consideration and keeping track of the discourse. Lastly, a lot of the summaries produced with some of those techniques just tend to sound "robotic", especially the extractive summaries, where a coherent flow of sentences with smooth transitions between paragraphs is quite often missing.

This thesis suggests a novel approach to summarization of legislative proceedings transcripts using so-called "phenom"-capturing technique in an attempt to solve some of the aforementioned issues. A phenom is a specific pattern appearing in the text that is deemed to be worth extracting and presenting in the summary. It can be a long back-and-forth discussion between two people, a pull-quote of interest, an emotionally charged claim or a mention of a well-known person, organization or other entity. Those features tend to appear in certain parts of the text more often, thus a classification of text fragments has to be performed first to split the text in certain chunks bearing different functions in the transcript. Luckily, legislative meetings are mainly quite consistent and well-structured in this sense, with the organizers trying to stick to the agenda. After the parts of the text are classified and split into sections, the phenom extraction is performed, collecting facts to be filled into text templates crafted for each phenom. In the end, those generated sentences and paragraphs can be put together in the summary article and presented to the reader.



Findings and lessons learned revealed that the whole system is built in a flexible way so the phenoms that the consumer is not interested in can be easily left out or, if need be, other phenoms can be added and incorporated. The evaluation user study has shown that the phenom system concept and the fusion of extractive and abstractive approaches have proven to be a viable option of producing factually and grammatically correct summarization articles with some room for improvement. Certain steps of the system can use more sophisticated mechanisms discovering other approaches to boost the intermediate results such as paragraph classification or automated neural-net-based template generation instead of a bank of hand-written ones.