RTVMONITOR

Topic-based text segmentation of generated transcripts
MSc research topic | 3 - 6 months

RTV Monitor

RTV Monitor makes audiovisual content accessible and searchable. With our Automatic Speech Recognition (ASR) system, we can generate transcripts of any content broadcasted on radio and television, as well as parliament streams and popular podcasts. Our ASR handles Dutch, Flemish, French and Austrian speech. We create extensive queries for our customers, enabling them to receive notifications when their company, brand, organisation, a person of interest or competitors are being mentioned in the media. In addition, we also curate tailored summaries. Currently, we are expanding our ASR system with enhanced algorithms and additional languages.

Problem description

News, as it appears in traditional articles, is easily sectioned. Each different article often has a different or new topic. Within audiovisual media (radio, tv, podcasts), the beginning and end of a specific topic are not so obvious. Therefore, the marking of the beginning and end of a different subject is now a manual and time-consuming task. Doing so requires our colleagues to listen to the original audio. Detecting topic boundaries can be very useful for downstream tasks, such as text summarization and classification.

Expectations

For this research, we would like you to develop a model able to conduct when one subject starts and ends and a new one begins. You will be working with a dataset of 100.000+ transcripts generated by our ASR system. In addition, the start/end of subjects will be hand-labeled. How you compose the research is up to you. You may think of comparing or combining different models, fine-tuning hyperparameters, using supervised or unsupervised approaches and so on.

Requirements

- Good understanding of NLP
- Proficiency in Python and Pytorch/TensorFlow
- Proficiency in the Dutch language

Articles

1. Arnold, S., Schneider R., Cudré-Mauroux P., Gers F.A., Löser A. (2019) SECTOR: A Neural Model for Coherent Topic Segmentation and Classification. Transactions of the Association for Computational Linguistics 2019; 7 169–184. DOI: https://doi.org/10.1162/tacl_a_00261
2. Koshorek, O., Cohen, A., Mor, N., Rotman, M., Berant, J. (2018). Text Segmentation as a Supervised Learning Task. NAACL. https://aclanthology.org/N18-2075.pdf

Additional information

RTV Monitor is part of the SoundAware Group, a collection of Dutch Media Monitoring companies focusing on monitoring of Music Airplay, Advertising, and Spoken Word (RTV Monitor).
For this research project you will be working together with our AI and engineering team, consisting of both internal and external staff, whereby you are expected to take the lead in the project.
For this assignment you can work from our office in Amsterdam as well as from home, while respecting the Covid-19 rules.

Contact information

Alex Terpstra
Executive Director
alex.terpstra@rtvmonitor.nl
06-53141293

Justo van der Werf
Computer Science Engineer
justo.van.der.werf@rtvmonitor.nl
06-13424378

Office address:
Johnny River building
Joan Muyskenweg 22
Amstel Business Park
Amsterdam

Websites:
www.rtvmonitor.nl
www.soundawaregroup.com