Natural Language Processing for Law and Social Science, Spring 2022

Course Project Proposal

Gökberk Özsoy <goezsoy@student.ethz.ch>

Topic: 93- Build a congressman speech generator with embedded speaker ID's. Analyze geometry of resulting speaker embeddings.

The topic is from the suggested list shared by our TA, and approval of Professor Ash is received. It is a conditioned text generation problem. Specifically, the model will generate speech as if it was addressed by the congressman given as input. The dataset is U.S. Congressional Record Corpus, for the years 1995-2020, with the speaker metadata, which is provided by Professor Ash.

Text generation is the natural outcome of the language modeling task, where the state-of-the-art performance is obtained by transformer based models such as GPT. Common way for conditioned text generation via these models is providing a natural text as a prompt. This is as opposed to our task's condition which will be speaker embeddings. (Yet, giving the speaker name as a prompt is an option to try.) Thus, this project requires being creative on crafting a design for obtaining categorical speaker embeddings and fusing these embeddings with the generation pipeline. Initially, a pre-trained GPT-Neo will be used as the generative model. Further analysis on the geometry of resulting speaker embeddings will also be conducted.