Meeting Notes—Summary LLM

Bas Donkers, Dennis Fok, Finn-Ole Höner

January 31, 2024

18-01-2024

- Status: First steps towards a summary LLM, basic computations seem to work.
- Admin: Setup of MS Teams group
- Technical Note: The `output` object is a tuple of length "positions", and each row of the contained matrix is one beam.

Dutch LLMs

I found some dutch language models on Huggingface, e.g. a GPT-2 version which I use for the illustration below (also available with more parameters, see corresponding paper). There are also other "dutch" versions of other language models, see this collection. One caveat might be that all these models are in some way fine-tuned/transfer learned from the English version of the model, so not "natively" trained on only dutch corpora. However, it seems to work in our current "brute force" optimization.See attachements for the targetHeeft het zin om appels en sinaasappelen te vergelijken?? (sic)

Speeding up computations

Currently, we are performing too many computations. We can make use of the fact that we know the target string. Essentially, we can compute each "next word" in parallel, instead of sequentially. This also circumvents the problem that we need to think about generation strategies, as we can essentially just make a forward pass through the LLM.

TO-DOs

- Finn speeds computations up
- Finn tries to write draft for grant proposal
- Reaching out to SKIM, slide deck, NDA