Transformer language models: architectures, word sense disambiguation, and targeted syntactic testing

9.19(0) Fall 2023, Instructor: Roger Levy

8 November 2023

Controlled syntactic testing of neural language models

In this class we will develop some controlled syntactic tests of an autoregressive language model, GPT–2. (The same principles would apply to more recent "large" language models like GPT–4 or LLaMa, but those are either not open or too large to easily run on Colab.) The Colab notebook for this exercise is here:

https://colab.research.google.com/drive/1OQ0UgQVCMK9kwRweGFcmMj8JpW7hQB_7?usp=sharing

Examples we will look at include:

- Testing the model's capabilities in subject–verb agreement;
- Looking at how well the model continues the beginnings of multiply center-embedded sentences;
- Investigating whether the model "garden paths" in human-like ways in the face of incremental syntactic ambiguity.