Don-Thuan Le
CS 410: TIS Tech Review

I briefly describe the readings referenced for topic modeling and bioinformatics and further wonder how GPT-2 could affect the medical industry.

In the bioinformatics research paper [2], Liu et al breaks down PLSA and LDA with its current application to bioinformatics. This is significant because biological datasets are one of the areas that rapidly accumulate data and are directly used for research and studies for health. PLSA allows us to generate the probability of topics from documents and the probability of words from the topics. LDA is an extension of this that allows the user to add a bias to the topic model with a Dirichlet prior variable. By combining topic modeling with different data such as gene-sampling and protein models, researchers were able to identify key points. For example, the main task for gene sequence data was “to characterize a set of common genomic features shared by the same species.

GPT-2 [1] aims to use unsupervised learning to create a model that is robust to many different language tasks such as reading comprehension, question and answering, translations and summarizing. And from the (GPT-2 paper) you can see how the model compares to other models that are state of the art but trained specifically for each task. GPT-2 is trained on a massive amount of text with 1.5B parameters and, although not trained for a specific task, GPT-2 is able to compete with other models.

I read these articles separately while looking for a topic to review but decided to mash them together. This could be very naive to assume as I am not a health professional, but if the human body is able to be processed as text such as through genome and protein sequences, what would be the result if a model with 1.5B parameters were trained on current bioinformatic

datasets? Robust modeling of bioinformatics such as the way GPT-2 is created to complete many different tasks could potentially lead to a model that generalizes the whole human body.

Similar to how a general PLSA model finds topics over large amounts of text that are harder for humans to pinpoint, a model that compares to GPT-2 (and maybe even GPT-3) has the potential to find connections between diseases and genetic patterns throughout the human body. We have learned that general text is hard for computers to process because of ambiguity, context, and the way text is created for humans to process and not for computers. On the other hand, the human body is finite and exhibits context. For example, sickness creates bacteria and the human body fights these bacteria by creating antibodies. These elevated levels of antibodies can act as "context" for the model and therefore have less ambiguity than that of general natural language.

As referenced in the GPT-2 paper, one of the difficulties is being able to evaluate the model effectively because many text datasets overlap unknowingly. Some datasets even have repeat documents. Although bioinformatics data is personal and is required to have consent from patients to share, it is a constant stream of new data because it's continuously being collected in hospitals around the world. Where language is a barrier which currently limits text data to a specific language, bioinformatics can be collected without such limitations.

To summarize, my general curiosity stems from the possibility of applying bioinformatics data to a model as large as GPT-2 and what patterns could be found that we currently do not see in the human body. Being less ambiguous, and with more context, a model could better generalize the human body, but this also depends on how much of the human body can truly be represented as text.

References

Radford, Alec, et al. *Language Models Are Unsupervised Multitask Learners - Openai*. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Liu, Lin, et al. “An Overview of Topic Modeling and Its Current Applications in Bioinformatics.” *SpringerPlus*, Springer International Publishing, 20 Sept. 2016, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5028368/.