Jordan Frimpter JEF180001
Henry Kim HTK180000
CS 4395.001 Human Language Technology
Dr. Mazidi

ACL Paper Summary: GPT-D

Full Paper Title
GPT-D: Inducing Dementia-related Linguistic Anomalies by Deliberate Degradation of Artificial Neural Language Models
https://aclanthology.org/2022.acl-long.131/

The code for the project is publicly available on GitHub:
https://github.com/linguisticanomalies/hammer-nets.

Problems Addressed by the Paper
This paper covers the work of natural language processing and biomedical researchers in leveraging the GPT-2 transformer machine learning model and a degraded version called GPT-D to detect linguistic anomalies associated with dementia in text as well as generate new text that exhibits these irregularities. The authors claim this approach reaches state-of-the-art performance on detection of dementia-related anomalies that was previously achieved by supervised machine learning models. Furthermore, the authors claim their approach can generalize better to classifying text from conversational domains as compared to previous work where the models were trained on text collected from specific tasks.

Authors
The authors of the paper are Changye Li, David Knopman, Weizhe Xu, Trevor Cohen, and Serguei Pakhomov. Changye Li, the first author, is a Ph.D. student of the Institute for Health Informatics at the University of Minnesota who is studying NLP applications with biomedical data. She currently has 20 citations on Google Scholar, and her Ph.D. advisor is Dr. Serguei Pakhomov of the Department of Pharmaceutical Care and Health Systems at the University of Minnesota. Dr. Pakhomov has 6941 citations on Google Scholar, the most of any of the authors, and researches NLP on medical records as well as the effects of diseases and medications on speech production. Dr. Knopman is a clinical neurologist who studies cognitive disorders like dementia. Weizhe Xu is a Ph.D. student from the Department of Biomedical and Health Informatics at the University of Washington. Dr. Cohen is an adjunct professor of psychology at the Department of Biomedical and Health Informatics, University of Washington with 3922 citations on Google Scholar for his research on distributional semantics.

Summary
Alzheimer's disease is a neurodegenerative disorder characterized by degradation of cognitive ability. The disease affects over 50 million diagnosed people and many more who are still undiagnosed. Diagnosis is typically performed with cognitive exams, brain imaging, and behavioral observations of the patient. This paper addresses the problem of diagnosing

Alzheimer's disease through the use of natural language processing. A text analysis technique that can detect speech irregularities associated with dementia could be used as a screening tool for identifying people that may have Alzheimer's disease. The authors seek to develop such a technique that can generalize to conversation topics outside of a structured task.

The methods used in this project of course build upon previous research in machine learning, specifically the Transformer model, previous models used for detecting speech indicative of Alzheimer's disease, and the datasets of speech collected from people with dementia. The GPT-D model created by the authors of this paper is a variation of GPT-2. In the 2019 paper "Language Models are Unsupervised Multitask Learners" by Radford et al., GPT-2 was used to demonstrate unsupervised learning is a viable alternative to supervised learning in learning on task-based datasets. GPT-2 is a Transformer, which is a model that was proposed by Vaswani et al. in the paper "Attention is All You Need" in 2017. Transformers are deep learning models that use an attention mechanism to process the positional relevance of tokens in an input, and they can do so in a way that is more parallelized than a recurrent neural network.

When training and testing their models, the authors used data from DementiaBank (DB), the AD Recognition through Spontaneous Speech (ADReSS) Challenge, and the Carolinas Conversation Collection (CCC). The DB dataset consists of transcripts from healthy participants and participants with dementia completing tasks designed to detect speech patterns associated with Alzheimer's disease, and ADReSS is a balanced subset of DB. On the other hand, CCC consists of transcripts of non-task-based interviews with healthy people and people diagnosed with dementia. In the 2020 paper "To BERT or Not To BERT: Comparing Speech and Language-based Approaches for Alzheimer's Disease Detection," Balagopalan et al. achieved an 83.3% accuracy on classifying the AD vs non-AD samples in the ADReSS test set using a Bidirectional Encoder Representations from Transformers (BERT) model. This experiment showed that a fine-tuned BERT classification model could even outperform a learning model that used expert-defined features on the AD text classification task.

While BERT's encoder-based structure makes it good at tasks for understanding language like classification, it is not well-suited to generating text like GPT-2. Rather than retrain an entirely new model on a whole new corpus of text from people with AD, the authors of this paper take the pretrained GPT-2 and create a degraded version called GPT-D by impairing the attention mechanism of the model. This approach is meant to compensate for the small amount of data available from people with AD compared to the large corpus available from cognitively healthy people. Several impairment schemes were tried including impairments to the embedding layer and several different combinations of masking to the attention mechanism layer. The authors found the best performance for the classification task was the Cumulative impairment pattern where 50% of attention heads in the first 9 layers of the attention mechanism were masked as zeroes.

To classify whether text was generated by someone with AD or not, the ratio of model's perplexity (PPL) was used. PPL is a measure of how well a piece of text fits a language model, so it indicates the likelihood that text was generated by a model. A text sample from someone

with AD should have lower perplexity with respect to the model trained on data from people with AD. Using a cumulative impairment scheme on GPT-D and comparing perplexities of GPT-D and GPT-2 on the ADReSS test set, the authors achieved a classification accuracy of 85%. This exceeds the 83.3% average achieved by Balgopalan et al., so the authors argue this approach is a viable alternative to fine-tuning a BERT model for the task.

The authors further argue that their approach based on comparing perplexities generalizes better than the previous approach using BERT. They found that impairment patterns on GPT-D that worked well for the ADReSS or DB datasets could achieve accuracies close to or above 70% on the conversational CCC dataset. In contrast the authors argue that when fine-tuning BERT on ADReSS and DB, BERT does not generalize well to being applied to the CCC data.

The other advantage of GPT-D is that it can be used to generate new text that has anomalies associated with AD. By prompting GPT-D with Bird et al.'s synthetic image descriptions, the authors caused GPT-D to generate new text which used more repetitions (measured in type-to-token ratio) and words with higher lexical frequency compared to GPT-2. This suggests the degraded model mimics speech patterns associated with AD. In examining salience visualizations of sample text generated by GPT-2 and GPT-D, the authors found that GPT-D gave more evenly distributed weights to the words when using them to predict other words. The authors state this indicates "the model is uncertain with respect to what it should consider as important" as compared to the original GPT-2, which gave non-stopwords more importance. Although informal, this finding that AD speech can be simulated by impairing the attention mechanism of a GPT model may give neurologists some insight into how attention mechanisms work in the brain or are impaired by dementia.

References

Li, Changye, David Knopman, Weizhe Xu, Trevor Cohen, and Serguei Pakhomov. "GPT-D: Inducing dementia-related linguistic anomalies by deliberate degradation of artificial neural language models." *arXiv preprint arXiv:2203.13397* (2022).
Code: https://github.com/LinguisticAnomalies/hammer-nets

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog,* 1(8):9

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc

Aparna Balagopalan, Benjamin Eyre, Frank Rudzicz, and Jekaterina Novikova. 2020. To bert or not to bert: Comparing speech and language-based approaches for alzheimer's disease detection. *Proc. Interspeech 2020*, pages 2167–2171.

Helen Bird, Matthew A Lambon Ralph, Karalyn Patterson, and John R Hodges. 2000. The rise and fall of frequency and imageability: Noun and verb production in semantic dementia. *Brain and language*, 73(1):17–49