Probing large language models for implicit learning of aphasia types

Thomas Hikaru Clark[1], Edward Gibson[1], Roger Levy[1]
[1]Massachusetts Institute of Technology, Department of Brain and Cognitive Sciences

Individuals with different subtypes of aphasia exhibit different patterns of impairment, with considerable individual variation [1]. We present an evaluation of large language models fine-tuned on transcriptions of utterances from individuals with aphasia of various subtypes and control participants, and probe the models for implicit learning of cues relating to aphasia type. Since acoustic cues and patient information are absent, this evaluation targets the degree of information shared between aphasia classification and the linguistic content of speech, as processed by artificial neural networks. **Data.** We use the English subset of AphasiaBank [2] and create separate training, validation, and testing partitions, such that each partition contains responses from a balanced number of individuals with Broca's, Wernicke's, and Conduction aphasia, as well as controls. Each partition came from disjoint picture description tasks in AphasiaBank. **Models.** We perform fine-tuning on the GPT-2 language model [3], training a separate model for each of the 4 categories of Broca, Wernicke, Conduction, and Control. Multiple learning rates were used, with a value of 5e-5 chosen based on validation set performance. **Evaluation.** We use Bayes' Rule to estimate $P(c \mid u)$, i.e. the probability of a given category *c*, given an utterance *u*, as described in Equation 1. Each test utterance is evaluated by all 4 fine-tuned models, yielding a probability for each test utterance under each model, corresponding to $P(u \mid c)$. Following Equation 1, probabilities are normalized to sum to 1 across the four models (e.g. if each model assigns the same probability to an utterance, then its score would be 0.25 under each model). Scores are averaged across utterances. If models implicitly learn features relevant for identifying aphasia type (in a way that generalizes beyond the specific conversational task), then the probability assigned to an utterance should be highest when evaluated by the model that matches its category. **Results.** Results are shown in Figure 1 as a pseudo-confusion matrix. Each column corresponds to a fine-tuned model trained on utterances from a specific subtype. Each row corresponds to held-out test data from a specific subtype. All diagonal values are above 0.25, indicating above chance performance, though for Broca and Conduction aphasia, utterances are on average assigned comparable probability by the Wernicke model as by their category-matched model. Control utterances are assigned low probability by every non-Control model, and the Control model assigns low probability to non-Control utterances. The results indicate that language models are able to implicitly learn some general linguistic features that are useful for aphasia classification, while there still exists confusability between different aphasia subtypes. Further research can expand the analysis to additional subtypes of aphasia and add interpretability to the results, in terms of which specific features are being learned by models during fine-tuning.

References:
[1] Rohde, A., Worrall, L., Godecke, E., O'Halloran, R., Farrell, A., & Massey, M. (2018). Diagnosis of aphasia in stroke populations: A systematic review of language tests. PLOS ONE, 13(3), e0194143. https://doi.org/10.1371/journal.pone.0194143
[2] MacWhinney, B., Fromm, D., Forbes, M., & Holland, A. (2011). AphasiaBank: Methods for Studying Discourse. Aphasiology, 25(11), 1286–1307. https://doi.org/10.1080/02687038.2011.589893
[3] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners. 24.

$$P(c \mid u) = \frac{P(u \mid c) \cdot P(c)}{\sum_{c'} P(u \mid c') \cdot P(c')}$$

c : category

u : utterance

Eq. 1: Bayes' Rule for calculating the probability $P(c \mid u)$ of a given category *c*, given an utterance *u*. We estimate $P(u \mid c)$ using a language model fine-tuned on utterances from category *c*, and assume a uniform prior $P(c)$, as the categories are balanced in the test dataset. With this assumption, $P(c \mid u)$ simplifies to $P(u \mid c)$ normalized by $\Sigma_{c'} P(u \mid c')$.

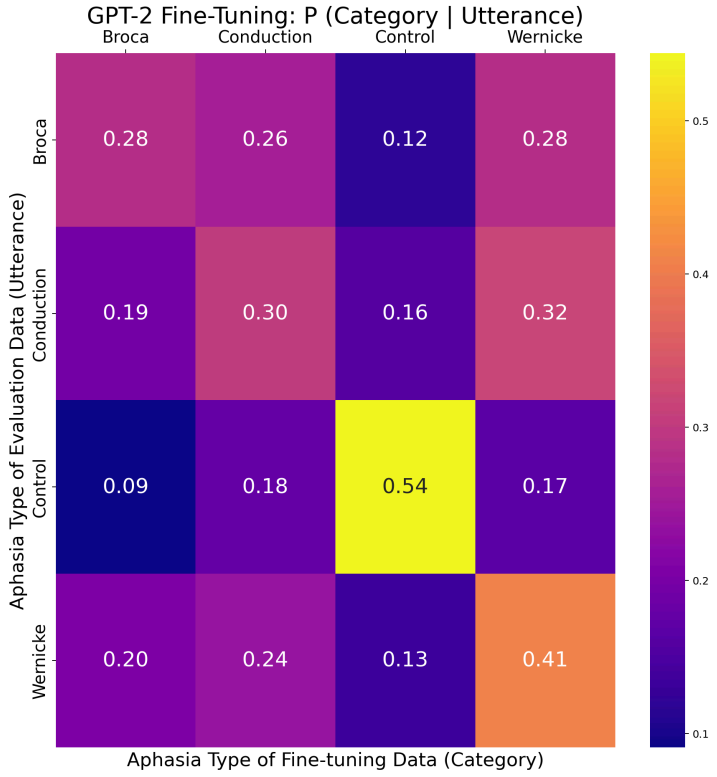

Fig. 1: Pseudo-confusion matrix for 3 types of aphasia and controls. Each column corresponds to a fine-tuned model trained on utterances from a specific subtype. Each row corresponds to held-out test data from a specific subtype. If fine-tuned language models implicitly learn general linguistic features relevant to classifying aphasia subtypes, then the highest value in each row should be on the diagonal. Note: each row may not sum exactly to 1 due to averaging across utterances.