Extended Abstract

Introduction

This work is part of our efforts to produce automated tools for identification and fine-grained classification of paraphasias within discourse, the production of which is the hallmark characteristic of most people with aphasia (PWA). We address the initial step for that goal: automatically identifying paraphasias in transcripts of discourse.

Aims

We fine-tune a machine learning-based large language model (LLM) to automatically identify paraphasias in Cinderella story retellings. The downstream use-case of this model is for clinicians to more easily analyze paraphasias produced during discourse by being able to automatically identify candidate paraphasias quickly and accurately. We had two research objectives: 1) develop and demonstrate the utility of a classifier for automatically identifying paraphasias in discourse; 2) explore the impact of clinical characteristics on classifier performance.

Method

Data consisted of 353 Cinderella story retellings from 254 PWA from the English AphasiaBank database (MacWhinney et al., 2011). Demographic and clinical information are shown in Table 1.

Following our protocol in Salem et al. (2023), we defined paraphasias as word-level errors made to the lemma of content words (i.e., nouns, verbs, adjectives, adverbs) and excluded from analysis all other kinds of word-level errors (e.g., dysfluency, plurality). This left 3,107 paraphasias out of 93,842 total productions.

We used our pre-trained LLM BORT (Beyond Orthographically Restricted Transformers; Gale et al., 2023), designed for usage on text with a mix of orthographic and phonemic transcriptions. Using 10-fold cross validation to prevent overfitting, we fine-tuned BORT to classify each all tokens in its transcript as paraphasia or non-paraphasia. Examples are shown in Table 2.

After fine-tuning, we used Receiver Operating Characteristic (ROC) analysis to determine the optimal threshold for final classification, by jointly maximizing the true positive rate (sensitivity), and minimizing the false positive rate (1-specificity) from our model's predictions. We evaluated the performance of the final classifier by calculating sensitivity, specificity, positive predictive value (PPV), and accuracy.

We also calculated stratified metrics based on clinical characteristics of the participant: fluency, severity, and mean length of utterance in words (MLU). We tested whether differences in accuracy for each stratification were significant using two-sided z-tests for independent proportions.

Results

Figure 1 shows the ROC curve (AUC = 0.957) and optimal threshold (0.044), which achieved 0.867 sensitivity, 0.923 specificity, and 0.921 accuracy. Figure 2 shows a heat map illustrating prediction probability levels for each production in a sample transcript. Table 3

shows our model's performance metrics stratified by clinical characteristics. We achieved higher accuracy on transcripts from participants with fluent aphasia, less severe aphasia, and higher MLU. All differences in accuracy were significant according to the z-tests with $p < 0.001$.

Discussion

Due to the imbalanced nature of the data—out of 93,842 total productions only 3,107 were paraphasias—if a classifier predicted all productions were non-paraphasias, it would achieve 0.967 accuracy (with 1.0 specificity, 0.0 sensitivity). Thus, it is important to consider sensitivity to properly evaluate performance. We achieved high sensitivity (0.867), alongside high specificity (0.923), demonstrating high performance despite imbalanced data.

Our classifier identified 6,991 non-paraphasias as paraphasias (e.g., "mopping" in Figure 2), in addition to 2,694 correctly classified paraphasias, reflected in our low PPV of 0.278. However, for our use-case, we prioritized high sensitivity and capturing potential paraphasias, at the expense of an inflated false positive rate, since it is easier for clinicians to narrow down from potential options than to have to identify paraphasias initially.

Our model performed significantly better on transcripts from participants with fluent aphasia, less severe aphasia, and higher MLU. This higher performance came via higher specificity; sensitivity was higher in non-fluent, more severe, and lower MLU PWA. This dichotomy is likely due to a few factors. PWA with more severe aphasia had a higher proportion of paraphasias, leading to lower specificity. Additionally, the PWA with severe aphasia produced more neologisms than less severe PWA, and neologisms are easier for an automated system to identify as paraphasias than, e.g., semantic paraphasias, due to being transcribed phonemically. If accepted, we will present results stratified by paraphasia type.

This work is a successful proof-of-concept demonstrating the utility of developing a clinical tool for automatic identification of paraphasias produced during discourse. A limitation of this work is that it assumes the availability of fine-grained transcriptions; recent promising advances in clinical automatic speech recognition raise the possibility of a technical solution to this problem. These findings take us closer to automatic aphasic discourse analysis, opening up possibilities for novel applications beyond assessment (e.g., AAC).

References

1. Cho-Reyes, S., & Thompson, C. K. (2012). Verb and sentence production and comprehension in aphasia: Northwestern Assessment of Verbs and Sentences (NAVS). Aphasiology, 26(10), 1250– 1277. https://doi.org/10.1080/02687038.2012.693584
2. Gale, R., Salem, A., Fergadiotis, G., & Bedrick, S. (2023). Mixed Orthographic/Phonemic Language Modeling: Beyond Orthographically Restricted Transformers (BORT). *Proceedings of the 8th Workshop on Representation Learning for NLP (RepL4NLP 2023)*, 212–225. https://doi.org/10.18653/v1/2023.repl4nlp-1.18
3. Kaplan, E., Goodglass, H., & Weintraub, S. (Eds.). (2001). Boston naming test (2. ed). Lippincott, Williams & Wilkins.
4. Kertesz, A. (2012). Western Aphasia Battery—Revised [Data set]. American Psychological Association. https://doi.org/10.1037/t15168-000
5. MacWhinney, B. (2000). The CHILDES Project: Tools for Analyzing Talk (3rd ed.). Lawrence Erlbaum Associates.
6. Salem, A. C., Gale, R. C., Fleegle, M., Fergadiotis, G., & Bedrick, S. (2023). Automating Intended Target Identification for Paraphasias in Discourse Using a Large Language Model. *Journal of Speech, Language, and Hearing Research*, 1–18. https://doi.org/10.1044/2023_JSLHR-23-00121

Tables

Table 1

Demographic characteristics

Characteristic	**Value**
Age (years)	
M (*SD*)	61.48 (12.39)
Min - Max	25.60 - 90.72
Missing (*N*)	3
Gender	
M (*N*)	141
F (*N*)	113
Race	
White (*N*)	218
African American (*N*)	25
Asian (*N*)	2
Hispanic/Latino (*N*)	7
Native Hawaiian/ Pacific Islander (*N*)	1
Mixed (*N*)	1
Education (years)	
M (*SD*)	15.47 (2.76)
Min - Max	8 - 25
Missing (*N*)	10
Aphasia duration	
M (*SD*)	5.22 (4.73)
Min - Max	0.08 - 30.00
Missing (*N*)	3
WAB-R AQ	
M (*SD*)	72.05 (17.88)
Min - Max	10.80 - 99.60
Missing (*N*)	8
BNT-SF	
M (*SD*)	7.26 (4.52)
Min - Max	0 - 15
Missing (*N*)	13
VNT	
M (*SD*)	14.85 (6.26)
Min - Max	0 - 22
Missing (*N*)	11

Note. WAB-R AQ is the Western Aphasia Battery-Revised Aphasia Quotient. BNT is the raw score from the Boston Naming Test-Short Form (Kaplan et al., 2001). VNT is the raw score from the Verb Naming Test (Cho-Reyes et al., 2012).

Table 2

Transcript preparation and prediction examples

Prepared transcript fragment	Ground truth classification	Model prediction probability	Model classification
sɪndərɛrlə **<pʌz>** pretty curl. and her stsɛpsəmʌðɚand stɛtʃɑðɚ no mother was all these ʌðəlɪ wɪnmɪm. and. okay. and she wanted to get all tɑld up for tea prince's sɛləbweʃən. …	0 (non-paraphasia)	0.998	1 (paraphasia)
sɪndərɛrlə pʌz **<pretty>** curl. and her stsɛpsəmʌðɚand stɛtʃɑðɚ no mother was all these ʌðəlɪ wɪnmɪm. and. okay. and she wanted to get all tɑld up for tea prince's sɛləbweʃən. …	0 (non-paraphasia)	0.027	0 (non-paraphasia)
sɪndərɛrlə pʌz pretty **<curl>** . and her stsɛpsəmʌðɚand stɛtʃɑðɚ no mother was all these ʌðəlɪ wɪnmɪm. and. okay. and she wanted to get all tɑld up for tea prince's sɛləbweʃən. …	1 (paraphasia)	0.979	1 (paraphasia)

Note. In the first example, <pʌz> is not a paraphasia since its target ("was") is not a content word.

Table 3

Performance metrics across data stratifications

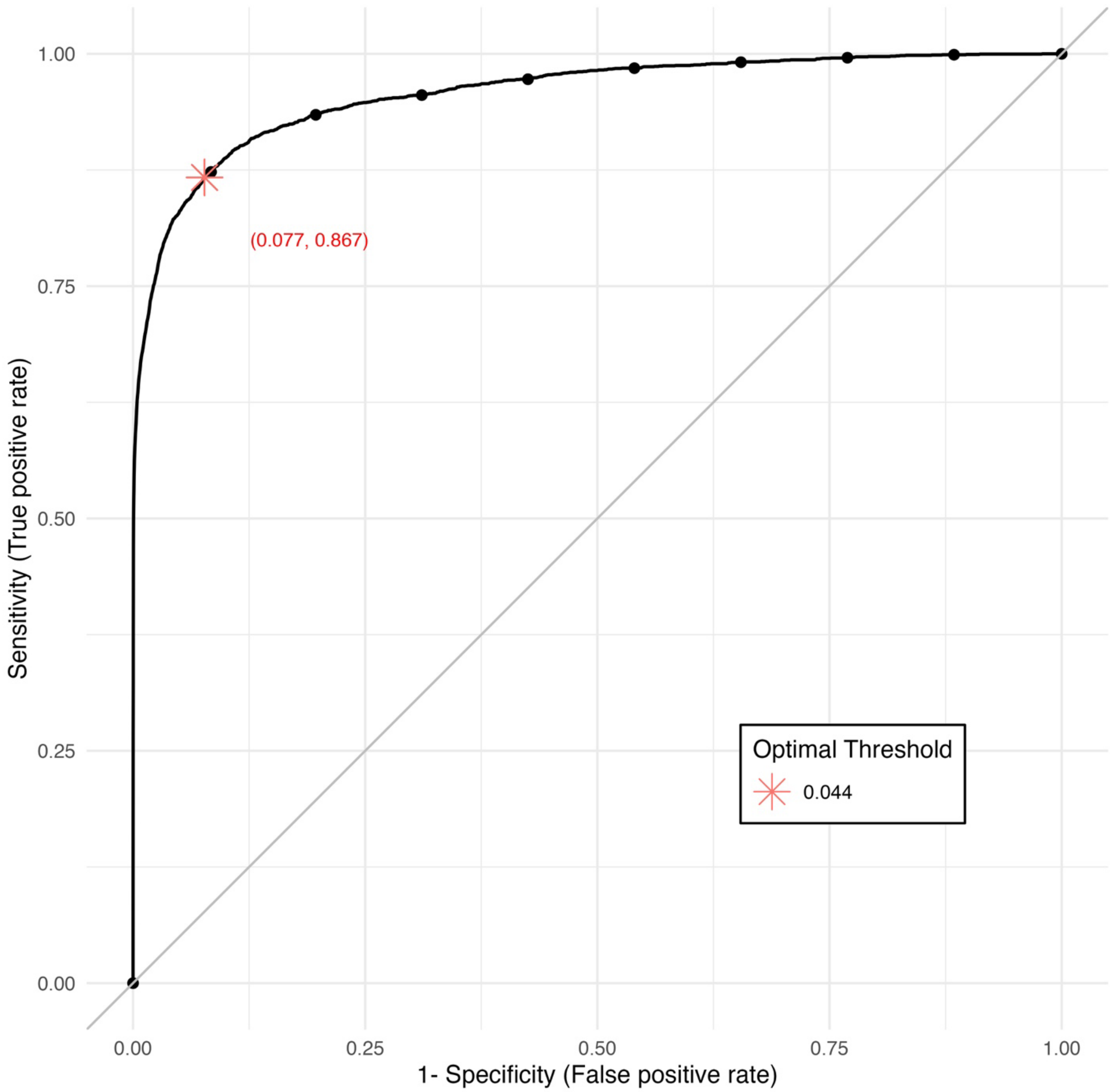
Test set	N sessions	N productions	N paraphasias	Sens	Spec	Pos pred value	Accuracy
All participants	353	93,842	3,107	0.867	0.923	0.278	0.921
WAB-R AQ > median (74.05)	172	54,442	1,189	0.818	0.943	0.242	0.940
WAB-R AQ <= median (74.05)	172	36,911	1,857	0.896	0.892	0.305	0.892
Fluent participants	252	80,036	2,338	0.853	0.925	0.255	0.923
Non-fluent participants	92	11,317	708	0.907	0.903	0.384	0.903
MLU > median (5.41)	177	62,633	1,793	0.852	0.928	0.258	0.926
MLU <= median (5.41)	176	31,209	1,314	0.888	0.913	0.310	0.912

Note. Fluent participants are those with Wernicke's, anomic, conduction, or transcortical sensory aphasia, or those considered "non-aphasic" by the WAB-R. Non-fluent participants are those with Broca's, global, or transcortical motor aphasia. 9 out of 353 total sessions had unavailable WAB-R results and were excluded just from analyses involving WAB-R scores. WAB-R AQ = Western Aphasia Battery–Revised Aphasia Quotient (Kertesz, 2012). MLU = mean length of utterance in words. Sens = sensitivity is TP/TP+FN, spec = specificity is TN/TN+FP, pos pred value = positive predictive value is TP/TP+FP, and accuracy is TP+TN/TP+TN+FP+FN.

Figures

Figure 1

Receiver Operating Characteristic (ROC) curve of the prediction probabilities



Note. Area under the curve (AUC) = 0.957.

Figure 2

Heat map showing prediction probability levels for each production in a sample transcript

the first one sɪləɹɛlə Cinderella . and there . and the the the the kids no don . like her . yeah and and then you know mopping and all of that you know . and and then . what is it called . you know the carriage or something like that . uhhuh and then dancing and all of that you know . and then so . what is it called . carriage you know . it . gone . and and then it . no more . and then the girl the girl good witch . and then you know does it and all of that . and then the girl I mean the the guy you know dancing and all of that you know . and then no more . that . it . and then it the end . I don . know . I mean . oh married . yeah .

Note. Darker highlight represents higher prediction probability. The productions “first”, “one”, “sɪləɹɛlə”, “kids”, “mopping”, “called”, and “witch”, each have prediction probabilities > 0.044 and are classified as potential paraphasias by our model. The two actual paraphasias in this transcript are “sɪləɹɛlə” and “witch”.