Domain adaptation and question-answer pooling for Aphasia modelling

Uwe Reichel[1], **Monica Gonzalez Machorro**[1,2], **Lisa M. Ehlen**[3], **Pascal Hecker**[1,4], **Dorothea Peitz**[3], **Cornelius Werner**[3,5], **Felix Burkhardt**[1,6], **Christian Kohlschein**[7], **Florian Eyben**[1,8], **Björn Schuller**[1,2,9,10]

[1]audEERING GmbH, [2]Chair of Health Informatics, TUM University Hospital, [3]Dep. of Neurology, Medical Faculty, RWTH Aachen University, [4]Hasso-Plattner Institute, [5]Dep. of Neurology and Geriatrics, Johanniter Hospital Stendal, [6]TU Berlin, [7]Accenture GmbH, [8]Agile Robots, [9]Imperial College, [10]Munich Center for Machine Learning

ureichel@audeering.com

Abstract

In this study, we examine the impact of domain adaptation and question-answer pooling on text-based aphasia prediction with standard and medically specialised BERT models for a German corpus. Modelling tasks comprise aphasia type classification as well as multitask regression of communicative, semantic, and syntactic skills. We found that domain adaptation before finetuning as well as question-answer pooling increased performance for the standard but not for the specialised models on all classification and regression tasks.

1 Introduction

Aphasia is a language impairment due to brain damage, after a stroke, traumatic head injury, brain tumours, or progressive neurological conditions. Depending on the brain regions affected, aphasia is featured differently. The most common types of aphasia are: global, amnesic (anomic), Wernicke's and Broca's aphasia (Caplan, 2003; Ardila, 2010). In Broca's aphasia, patients typically exhibit phonemic substitutions and have a non-fluent speech pattern. Wernicke's aphasia is characterised by an effortless but nonsensical speech. Global aphasia combines aspects of both Broca's and Wernicke's aphasia. Amnesic aphasia is primarily characterized by word retrieval and naming problems. Aphasia subtype classification is not straightforward and it is common that various aphasia types co-exist (Fridriksson et al., 2018).

Effective evidence-based therapy consists of high-intensity Speech-Language Therapy (SLT) which has been shown to improve linguistic capabilities (Peitz et al., 2024). However, this needs to be based on detailed diagnostics using appropriate tests. In German-speaking countries, the most common test used for aphasia diagnosis and monitoring is the standardised Aachen Aphasia Test (AAT) (Huber et al., 2013; Huber, 1983). This comprehensive test is designed to assess various aspects of language function, including comprehension, expression, repetition, and naming skills. It also provides information of probabilistic aphasia subtype and severity (Kohlschein et al., 2018). It consists of an examination of spontaneous language and five subtests. A 10-minute semi-structured interview, recorded during therapy, is rated in six domains: communicative behaviour, articulation/prosody, automatised language, semantics, phonology and syntax (Kohlschein et al., 2017). However, AAT is time-consuming and its result depends highly on the rater (Kohlschein et al., 2018), which usually is a highly trained speech and language therapist. An automatic aphasia diagnosis based on the AAT could help reduce waiting periods for patients and clinicians' burden as well as provide personalised remote rehabilitation strategies.

Prior work employing Machine Learning (ML) methods has explored aphasia and its subtype classification using connected speech, derived either from manual transcripts or Automatic Speech Recognition (ASR) systems (Fromm et al., 2022). These studies have focused on feature-based supervised methods, including traditional discourse features (e.g., syntactic complexity, proportion of nouns, verbs, adjectives) or embeddings by end-to-end approaches using large pre-trained models. Zusag *et al.* reported an F1 score of 0.84 for detecting amnesic aphasia, 0.77 for identifying Broca aphasia; and 0.78 for Wernicke aphasia using a Support Vector Classifier (SVC) and linguistic features (Zusag et al., 2023). Dunfield *et al.* employed sentence representation similarity features to capture symptoms of fluent aphasia and found a correlation of 0.61 with the Western Aphasia Battery-Revised Aphasia Quotient (Dunfield and Neumann, 2020). These features include question-answer similarity, closest question-answer pair identification, and binary sentence pair classification. The latter was obtained using BERT to predict the likelihood of a given sentence pair being related (Dunfield and

Neumann, 2020). Cong *et al.* leveraged Large Language Model (LLM)-surprisals to predict aphasia, its subtypes, and the level of severity. They reported an F1 score of 0.92 for predicting aphasia from healthy controls and 0.79 F1 score for identifying aphasia subtypes (Cong et al., 2024b). In another work, Cong *et al.* further employed surprisal values of LLMs, including GPT-2, Llama2, and BERT, alongside utterance length, to predict aphasia and its subtypes. Their results demonstrated an F1-score of 0.61 for detecting aphasia and 0.86 for classifying its subtypes in Chinese. For English, they reported an F1-score of 0.79 for identifying aphasia and 0.54 for distinguishing its subtypes (Cong et al., 2024a).

The contributions of our work of automatised aphasia assessment are as follows: (1) Aphasia transcripts are atypical on the lexical, syntactic, and semantic level. Such transcripts are usually not contained in the training material of pre-trained models, which might lower their general applicability on such clinical data. We are going to address this potential shortcoming by domain adaptation as described in section 3.2. (2) Relevant information is expected not to be contained only in the patients' answers in isolation but also within the context of the underlying question. We are going to address this contextualization by embedding pooling alternatives as presented in section 3.3.

2 Data

The German dataset was collected within the autoAAT BMBF project. It contains spontaneous speech samples, manual transcripts, and their associated clinical scores from the AAT. Transcripts were anonymised by removing all personal information. This dataset is built on the work presented in (Kohlschein et al., 2018). Many patients provided more than one recording due to repeated treatment cycles.The scores comprise the aphasia type classification and linguistic skills assessment. Aphasia type is categorised into the four classes Amnesic, Broca, Global, and Wernicke; since the project focus is to automatise aphasia diagnosis for tailored SLT, the dataset does not contain a control group. Other types of aphasia, such as primary progressive aphasia or unclassifiable, have been excluded of the analysis due to data sparsity. Linguistic skills are assessed separately in various impairment levels and on an expert-annotated six point scale (with 0 being the most severe and 5 meaning no impairment). This study focuses on three linguistic impairment levels: communicative behaviour (understanding and responding to questions), semantic structure (word finding difficulties and semantic paraphasias), and syntactic structure (sentence completeness and complexity).

The dataset comprises 331 participants, 92 female, 239 male, with a mean age of 53 ± 13 years. The major aphasia types are represented by the following numbers: 105 Global, 70 Broca, 32 Wernicke, and 34 Amnesic. The rest of the participants correspond to the excluded classes. Due to data protection regulations, the dataset cannot be shared. The dataset was split into speaker-disjunct training, development (10%), and test (20%) sets stratified on the aphasia type of each speaker by means of *splitutils* (Reichel, 2024). A random seed of 42 was applied to ensure reproducibility. Texts were cleaned by removing transcriber comments and special annotation symbols. The linguistics skills scales ranging from 0 to 5 were re-scaled to the range $[0, 1]$.

3 Methods

3.1 Modelling variants

For both tasks, aphasia type classification and linguistic skills regression, we started from two different base models: the general-purpose model *dbmdzbert-base-german-uncased* (Devlin et al., 2019) (referred to as *standard encoder* in the following), and *GerMedBERT/medbert-512* (Bressem et al., 2023), which was pre-trained on medical documents for applications in the clinical domain, henceforth referred to as *specialised encoder*.

For each of these encoders, we further created a variant domain-adapted to our specific aphasia dataset as described in section 3.2. Each of these four variants we combined with three different pooling architectures as described in section 3.3. We finetuned each of these 12 model variants on the two clinical tasks with 5 different random seeds, which we describe in section 3.4.

3.2 Domain adaptation

For domain adaptation, we followed the recipe of (Lendvai et al., 2023) applying vocabulary extension and Masked Language Modelling (MLM). We applied a 90/10 speaker disjunct and aphasia-label stratified split of the training partition into MLM training and development partition. Based on the MLM training partition we extended the tokeniz-

ers' vocabularies with the lexical content of the transcripts by adding up to 300 most frequent, yet unknown words with a minimum length of five characters. Subsequently, each base model was finetuned on the MLM task with a standard BertForMaskedLM head. Finetuning was done in 20 epochs with the AdamW optimizer, a learning rate of $2e-5$, a perplexity loss, and a batch size of 16. We kept the best model in terms of the lowest loss for the development set.

3.3 Pooling

We applied three types of pooling of the last hidden states of the encoder:

a: answer-only; we extract the embeddings only for the patient's answer and apply mean pooling of these embeddings;

qa-c: answer contextualised by question; we concatenate question and answer with a [SEP] token as for text entailment tasks (Putra et al., 2024), extract the embeddings for this text pair, and apply mean pooling on the answer part of this pair only, which is forwarded to the classification head;

qa-cc: answer contextualised by question plus question-answer coherence; as for *qa-c* we concatenate question and answer. Then, we concatenate the initial CLS token embedding with the mean embedding of the answer. This concatenated pooling we forward to the classification head.

Schematically, the pooling variants can be expressed as follows (the underlined constituents go into the pooling):

a: [CLS] **answer**
qa-c: [CLS] question [SEP] **answer**
qa-cc: [**CLS**] question [SEP] **answer**

We expect *qa-cc* to capture not only answer contextualisation but also question-answer coherence due to the 'semantics' of the CLS token. Since this token had been pre-trained on the next sentence prediction task, it is expected to represent the information the pre-[SEP] text part contains about the post-[SEP] text part, which can be considered as an aspect of text coherence.

In total, we get 12 model variants defined by all combinations of **encoder type** (*standard, specialized*), **domain adaptation** (*yes, no*) and **pooling** (*a, qa-c, qa-cc*). The finetuning of these models on the two downstream tasks is described in the subsequent section 3.4.

3.4 Finetuning

Architecture: To each encoder, we add a two-layer head with a non-linear (tanh) layer and a linear output projection. For classification, this output projection has 4 outputs, one per aphasia type. For multitask regression, it has 3 outputs, one for communicative, semantic, and syntactic skills, respectively.

Hyperparameters: Each model was finetuned in 8 epochs with the AdamW optimizer, a learning rate of $3e-5$ and an effective batch size of 32. For classification, we used the weighted cross entropy loss and unweighted average recall (UAR) as metrics to be maximised on the development set. For regression, we used a Concordance Correlation Coefficient (CCC) loss and CCC metrics for the development set. We kept the models performing best on the development set for further evaluation on the test partition. Finetuning and evaluation was repeated five times with different random seeds (1, 9, 20, 21, 42, generated with *numpy.random.default_rng()*).

4 Results

Figures 1 and 2 show the results in terms of UAR and mean CCC for aphasia type classification and linguistic skills regression, respectively. As an overall tendency for the standard encoder, we observe that domain adaptation as well as question-answer contextualisation slightly improve the performances for classification as well as for regression, but not so for the specialised encoder.

The best aphasia type classification result, a UAR of 0.653 averaged over all random seeds, was obtained with the standard encoder, and the *qa-cc* pooling variant accounting for contextualisation and coherence. For linguistic skills multitask regression, again, the standard encoder this time with the *qa-c* pooling variant for contextualisation only performed best, yielding a mean CCC of 0.755 averaged over all random seeds. Split into the linguistic dimensions it achieved a CCC of 0.738 for communicative, 0.695 for semantics, and 0.831 for syntactic skills prediction.

5 Discussion and Conclusion

We identified two challenges for finetuning pretrained transformer models with aphasia data: First, this text data is rather atypical and usually not part of pre-training datasets. This missing link was addressed by domain adaptation. Second, patients' answers are not only to be seen in isolation but also within context with the corresponding ques-

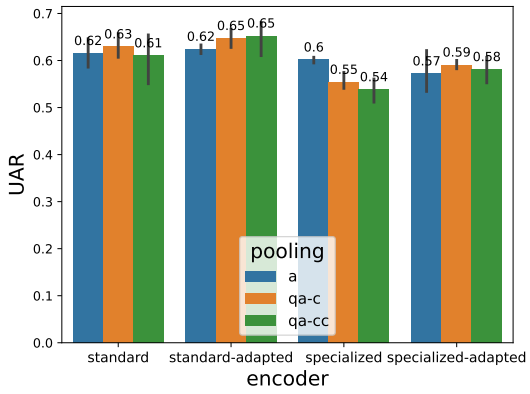

Figure 1: Aphasia type classification results: Unweighted average recall (UAR) values for all encoder and pooling variant combinations (see section 3). Error bars indicate 95% confidence intervals.

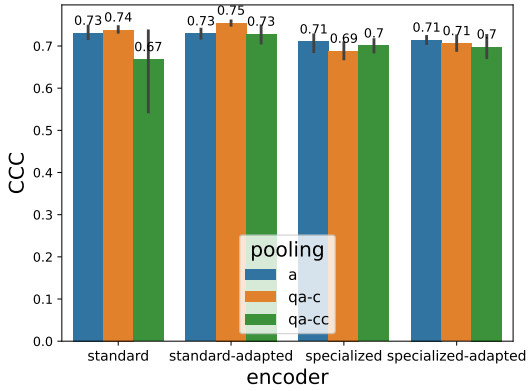

Figure 2: Communication, semantics, and syntactic skills multitask regression results: arithmetic mean Concordance correlation coefficient (CCC) values over the three regression dimensions for all encoder and pooling variant combinations (see section 3). Error bars indicate 95% confidence intervals.

tion. This contextualisation and coherence assessment was addressed by introducing different kinds of question-answer poolings.

For the standard encoder, domain adaptation as well as question-answer pooling turned out to be beneficial for both aphasia type classification as well as linguistic skills regression. Both strategies, by a low margin but consistently, lead to increased performance. As to pooling, for aphasia type classification, joint contextualisation and coherence assessment worked best, for regression contextualisation only lead to the highest performance.

The specialised encoder overall yielded lower performances compared to the standard encoder, which on first sight might appear counter-intuitive. However, the specialised model was not necessarily expected to work better for patient data classification in the first place, since the pre-training material consists exclusively of expert texts from scientific publications and dictionaries, as reported in (Bressem et al., 2023). These documents usually do not include a large amount of patient transcripts, but rather few illustrative examples only. Therefore, this specialised model is well suited for tasks such as clinical expert text classification, but not necessarily for patient transcript classification. One major reason for the overall lower performance of the expert model might be that the specialised pre-training material contains much less variability than the standard encoder's pre-training data, so that it is less capable to extrapolate to that kind of data. Likely due to this shortcoming, the specialised model neither could profit from domain adaptation nor question-answer pooling.

For question-answer pooling, longer error bars were observed for *qa-cc* as opposed to *qa-c* in Figures 1 and 2. This indicates that joint contextualisation and coherence assessment is less stable across random seed variations than contextualisation alone, so that the latter seems to be preferable in terms of model robustness.

To conclude and to give an outlook, our results show that for the given data, aphasia modelling works best with domain-adapted standard BERT models with contextualised mean pooling of the embeddings of patients' utterances. These results were obtained on narrow manual transcripts that preserve linguistic peculiarities relevant for aphasia assessment. For a fully automated aphasia assessment, such transcripts would need to be generated by ASR models, that keep track of clinically relevant utterance characteristics such as disfluencies; see, e. g., (Zusag et al., 2023; Mihajlik et al., 2024; Gohider and Basir, 2024) for such ASR methods. Our next steps thus will include combining automated narrow transcription with our aphasia modelling approach.

Ethics Statement

This research was conducted with strict adherence to ethical standards. The dataset employed in this work was collected under the ethics approval number EK 23-125 by the Ethics Committee of the Medical Faculty of RWTH Aachen University. To further ensure privacy, audio and text data was anonymised removing all personal information.

Data was analysed transparently, avoiding bias and ensuring accuracy.

Acknowledgements

The autoAAT project (*Automatische Auswertung von Spontansprachinterviews des Aachener Aphasie Tests*) is funded by the German Federal Ministry of Education and Research (BMBF), grant number 13GW0489A.

References

Alfredo Ardila. 2010. A proposed reinterpretation and reclassification of aphasic syndromes. *Aphasiology*, 24(3):363–394.

Keno K. Bressem, Jens-Michalis Papaioannou, Paul Grundmann, Florian Borchert, Lisa C. Adams, Leonhard Liu, Felix Busch, Lina Xu, Jan P. Loyen, Stefan M. Niehues, Moritz Augustin, Lennart Grosser, Marcus R. Makowski, Hugo JWL. Aerts, and Alexander Löser. 2023. MEDBERT.de: A comprehensive German BERT model for the medical domain. *arXiv preprint arXiv:2303.08179*. Keno K. Bressem and Jens-Michalis Papaioannou and Paul Grundmann contributed equally.

David Caplan. 2003. Aphasic syndromes. *Clinical neuropsychology*, 4:14–34.

Yan Cong, Jiyeon Lee, and Arianna LaCroix. 2024a. Leveraging pre-trained large language models for aphasia detection in English and Chinese speakers. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 238–245, Mexico City, Mexico. Association for Computational Linguistics.

Yonghao Cong, Amy N. LaCroix, and Jiyeon Lee. 2024b. Clinical efficacy of pre-trained large language models through the lens of aphasia. *Scientific Reports*, 14:15573.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1*, pages 4171–4186.

Katherine Dunfield and Günter Neumann. 2020. Automatic quantitative prediction of severity in fluent aphasia using sentence representation similarity. In *Proceedings of RaPID-2020 at LREC-2020*.

Julius Fridriksson, Dirk-Bart den Ouden, Argye E. Hillis, Gregory Hickok, Chris Rorden, Alexandra Basilakos, Grigori Yourganov, and Leonardo Bonilha. 2018. Anatomy of aphasia revisited. *Brain*, 141(3):848–862.

Davida Fromm, Joel Greenhouse, Molly Pudil, Yiwen Shi, and Brian MacWhinney. 2022. Enhancing the classification of aphasia: A statistical analysis using connected speech. *Aphasiology*, 36(12):1492–1519.

Nada Gohider and Otman A Basir. 2024. Recent advancements in automatic disordered speech recognition: A survey paper. *Natural Language Processing Journal*, 9:100110.

Walter Huber. 1983. *Aachener Aphasie Test (AAT)*. Verlag für Psychologie Dr. C.J. Hogrefe.

Walter Huber, Klaus Poeck, and Luise Springer. 2013. *Klinik und Rehabilitation der Aphasie: Eine Einführung für Therapeuten, Angehörige und Betroffene*. Georg Thieme Verlag.

Christian Kohlschein, Daniel Klischies, Björn Schuller, Tobias Meisen, and Cornelius Johannes Werner. 2018. Automatic processing of clinical aphasia data collected during diagnosis sessions: Challenges and prospects. In *International Conference on Language Resources and Evaluation*.

Christian. Kohlschein, Maximilian Schmitt, Björn Schuller, Sabina Jeschke, and Cornelius J. Werner. 2017. A machine learning based system for the automatic evaluation of aphasia speech. In *2017 IEEE 19th International Conference on e-Health Networking, Applications and Services (Healthcom)*. IEEE.

Piroska Lendvai, Uwe Reichel, Anna Jouravel, Achim Rabus, and Elena Renje. 2023. Domain-adapting BERT for attributing manuscript, century and region in pre-modern Slavic texts. In *Proceedings of the 4th Workshop on Computational Approaches to Historical Language Change*, pages 15–21.

Péter Mihajlik, Yan Meng, Máté S Kádár, Julian Linke, Barbara Schuppler, and Katalin Mády. 2024. On disfluency and non-lexical sound labeling for end-to-end automatic speech recognition. In *Interspeech 2024*, pages 1270–1274, Kos Island, Greece.

Dorothea Peitz, Beate Schumann-Werner, Katja Hussmann, Joao Pinho, Hong Chen, Ferdinand Binkofski, Walter Huber, Klaus Willmes, Stefan Heim, Jörg B. Schulz, Bruno Fimm, and Cornelius J. Werner. 2024. Success rates of intensive aphasia therapy: Real-world data from 448 patients between 2003 and 2020. *Journal of Neurology*.

I Made Suwija Putra, Daniel Siahaan, and Ahmad Saikhu. 2024. Recognizing textual entailment: A review of resources, approaches, applications, and challenges. *ICT Express*, 10(1):132–155.

Uwe Reichel. 2024. splitutils v0.3.0. Zenodo.

Markus Zusag, Lisa Wagner, and Thomas Bloder. 2023. Careful whisper – leveraging advances in automatic speech recognition for robust and interpretable aphasia subtype classification. In *Proceedings of Interspeech 2023*, pages 3013–3017.