A Comparative Analysis of Language Models for the Classification of Alzheimer's Disease Based on Connected Speech

Helena Balabin[1,2,*], Laure Spruyt[1], Ella Eycken[1], Ines Kabouche[1], Bastiaan Tamm[1,3], Jolien Schaeverbeke[1], Patrick Dupont[1], Marie-Francine Moens[2], Rik Vandenberghe[1]

[1]Laboratory for Cognitive Neurology, Department of Neurosciences, Leuven Brain Institute, KU Leuven, Belgium
[2]Language Intelligence and Information Retrieval Lab, Department of Computer Science, KU Leuven, Belgium
[3]Processing Speech and Images, Department of Electrical Engineering, KU Leuven, Belgium

Abstract

Early diagnosis of neurological disorders is a fast developing field of applied artificial intelligence (AI). In this context, AI-based language models have been increasingly used to distinguish cognitively healthy individuals from those affected by Alzheimer's disease (AD) based on their connected speech. Yet, it remains unknown how the adaptation of the language models to the language and domain of the connected speech samples impacts the classification results. Here, we construct several classification tasks from Dutch Flemish samples of connected speech from a cohort of 74 healthy controls and 20 subjects affected by AD. First, we compare the classification performance of Dutch and multilingual models as well as models that incorporate long-range context. Additionally, we examine how varying amounts of fine-tuning data from a separate corpus of speech samples affect domain adaptation. We demonstrate that increasing fine-tuning data leads to increased domain adaptation, but it does not necessarily translate into improved classification performance. Furthermore, our findings support the use of language-specific models over multilingual ones, even for multilingual models that were trained to incorporate wider context.

Introduction

Recent developments in artificial intelligence (AI) present promising directions for aiding the diagnosis of neurological disorders such as Alzheimer's disease (AD) (Li et al. 2022). In particular, the classification of AD solely from connected speech using AI-based language models emerges as an alternative approach to standard evaluations that are often based on costly and invasive assessments (de la Fuente Garcia, Ritchie, and Luz 2020). This classification can be based either directly on audio features or on text from transcribed speech. Nonetheless, the characterization of neurological disorders from speech comes with its own challenges, specifically with regard to adapting language models to a different language (other than English) and domain (i.e., connected speech). Therefore, we aim to address the following two questions in this paper:

- Given the choice between a monolingual language model adapted to the input language or a multilingual one specialized to capture long-range context, which of the two strategies is preferred?
- What effects are observed with varying degrees of fine-tuning on domain adaptation to connected speech?

Based on analyzing Flemish Dutch connected speech from $n = 74$ healthy controls and $n = 20$ participants affected by AD, we found that (i) a Dutch word-level model outperforms multilingual sentence- and document-level models, and (ii) unsupervised fine-tuning on a separate corpus through masked language modeling helps with domain adaptation to both the different types of connected speech used in this approach as well as the nuances of the Flemish Dutch language.

Related Work

There are numerous approaches for the classification of AD based on connected speech (Petti, Baker, and Korhonen 2020). Language models have been employed for this goal, most notably in the context of the ADReSS challenge (Luz et al. 2020), and its extensions ADReSSo (Luz et al. 2021) and ADReSS-M (Luz et al. 2023) for audio and multilingual data. However, there are three key limitations. Firstly, most AD patients included in such approaches are in an advanced stage and have clinical diagnoses, as opposed to biomarker-proven cases, limiting the applicability for the pre-clinical or prodromal stages of AD. Secondly, it remains unclear how various language model architectures, training objectives (word-level, sentence-level, document-level) and types of connected speech affect the accuracy of detecting AD, since there is a predominant focus on word-level models for picture description tasks. Lastly, there has not yet been an approach, which leverages domain-specific data for the domain adaptation of language models in the context of AD classification, an alternative to task-specific fine-tuning data that is often not available in clinical practice.

Methods

Based on a cohort of $n = 94$ subjects, we construct binary classification tasks to distinguish cognitively healthy subjects from those affected by AD using different parts from the transcripts of their connected speech. First, we describe the language models used to examine the effect of training language and input format on the classification performance.

*Corresponding Author. Email: helena.balabin@kuleuven.be
Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Model	Task					
	Autobiography	Description of a day	News event	Object descriptions	Picture description	Combined
RobBERT	0.861	0.841	0.869	0.888	0.788	0.868
XLM-RoBERTa	0.832	0.814	0.869	0.875	0.848	0.833
S-XLM-RoBERTa	0.851	0.828	0.843	0.866	0.807	0.856
Longformer-XLM	0.786	0.810	0.870	0.875	0.818	0.843

Table 1: **Model comparison classification accuracies.** This table shows the average classification accuracy across five test splits for each interview part and each of the proposed language models. For each task, the best performing model is underlined.

Fine-tuning proportion	Task					
	Autobiography	Description of a day	News event	Object descriptions	Picture description	Combined
20%	0.875	0.807	0.875	0.882	0.798	0.866
40%	0.877	0.841	0.862	0.879	0.818	0.875
60%	0.868	0.821	0.869	0.882	0.828	0.869
80%	0.856	0.814	0.862	0.878	0.788	0.859
100%	0.874	0.834	0.862	0.888	0.808	0.865

Table 2: **Domain adaptation classification accuracies.** This table shows the average classification accuracy for each task and each fine-tuned variant of RobBERT based on using different proportions of the available fine-tuning data. Again, the best performing model per task is underlined. A fine-tuning proportion of 0% corresponds to the RobBERT results from Table 1.

Then, we investigate how the inclusion of domain-specific fine-tuning data affects the performance.

Data

A total of $n = 94$ participants was recruited from from several cohorts from the Laboratory for Cognitive Neurology (LCN) at KU Leuven, mainly from the Flemish Prevent Alzheimer's Disease Cohort KU Leuven (F-PACK) (Schaeverbeke et al. 2021) and Alzheimer Research Centre KU Leuven (ARCK) cohorts. Each subject undergoes extensive assessment, including structural MRI, neuropsychological testing, and interviews to elicit connected speech. Based on AD biomarker testing, we group the combined cohort into $n = 74$ healthy controls and $n = 20$ cases with AD. For our healthy group, we combine amyloid positive and negative cases of cognitively healthy subjects, and our AD cases cover both the prodromal and early dementia stage. For this analysis, we focus on the subjects' text data derived from the interviews. The interviews are conducted in Flemish Dutch, and consist of the following five parts:

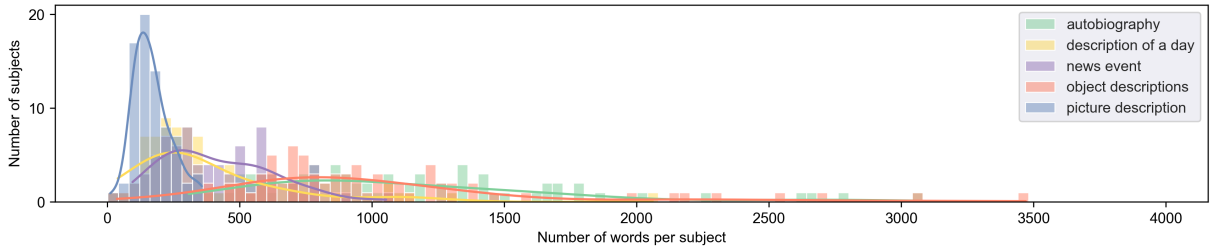
1. **Autobiography:** In the first part of the interview, participants are asked to describe five events from different decades of their lives.
2. **Description of a day:** Here, subjects are asked to recite a day of their last week in detail.
3. **News event:** Participants are asked to recall a recent news event of their choice.
4. **Object descriptions:** Based on a list of 20 concrete and abstract objects, subjects are asked to describe them in as much detail as possible.
5. **Picture description:** For comparison of our analyses with previous studies (e.g., Mueller et al. 2016; Nevler et al. 2017; Mirheidari et al. 2019), we include the Cookie Theft picture description task (Roth 2011).

Each interview part is recorded in audio form, and transcribed to text using an automatic speech recognition (ASR) system based on Poncelet and Van Hamme (2023) and trained on the Spoken Dutch Corpus (Oostdijk 2000). For each interview part, we characterize its total word length for all healthy participants as well as those affected by AD. Both groups have similar input length distributions for all interview parts (see Figure 1a and 1b). Additionally, another LCN corpus comprising interviews following the aforementioned structure from an independent group of $n = 56$ healthy controls is exclusively employed for the purpose of fine-tuning language models.

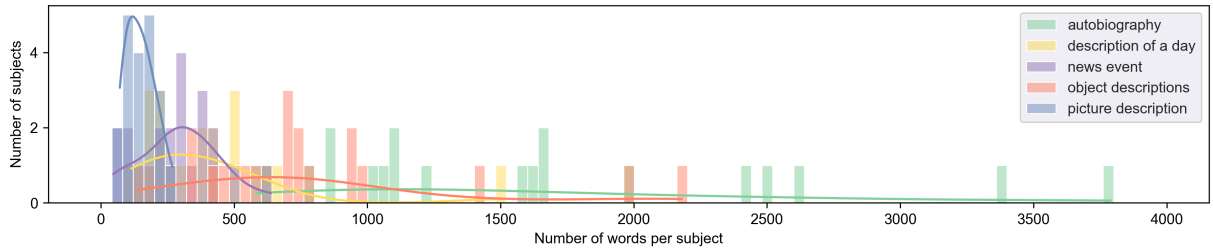
Model Comparison

Here, we vary two criteria for the selection of language models. First, we vary the language used during pre-training to investigate the effect of Dutch versus multilingual language models. Secondly, we aim to include both word-level, sentence-level and document-level models to examine how different strategies for incorporating context in the input affect the results. Taken together, this results in the following models:

- **RobBERT** (Delobelle, Winters, and Berendt 2022), a Dutch word-level language model based on the RoBERTa (Liu et al. 2019) model architecture
- **XLM-RoBERTa** (Conneau et al. 2020), a multilingual transformer-based word-level language model



(a) Word length distribution for $n = 74$ cognitively healthy subjects



(b) Word length distribution for $n = 20$ subjects affected by Alzheimer's disease

Figure 1: **Input length distribution.** The two histograms depict the distribution of the input length measured by the number of words for each interview part across all subjects, grouped by their neurodegenerative status.

- **S-XLM-RoBERTa** (Reimers and Gurevych 2020), an adaptation of XLM-RoBERTa that incorporates inter-sentence context, based on using Sentence-BERT (Reimers and Gurevych 2019) as a teacher model in a multilingual knowledge distillation setting
- **Longformer-XLM**, a multilingual adaptation[1] of the Longformer model (Beltagy, Peters, and Cohan 2020), which can handle inputs of up to 4096 tokens, based on a similar knowledge distillation strategy as Reimers and Gurevych (2020)

For all models, we derive embeddings for each interview part for each subject. Note that in most cases, the input of a given interview part is substantially longer than the maximum input length of the language models (except for the Longformer-based models that can process the entire input at once). This is why we split the input into maximally long chunks, resulting in multiple input examples per interview part and subject for the first three language models.

Domain Adaptation

We employ RobBERT as a base model for various fine-tuning procedures to investigate the effect of domain adaptation to the interview format and Flemish language on the classification performance. The model is fine-tuned in an unsupervised manner using the masked language modeling (MLM) objective for $2,000$ steps using a batch size of 16, and the loss is monitored on a 20% evaluation split of the fine-tuning corpus. To examine how much data is needed to improve model performance, we vary the amount of data used for fine-tuning (i.e., 20%, 40%, 60%, 80% and 100% of the available training data), resulting in a total of five model variants. The amount of data included per subject in each fine-tuning variant is shown in Figure 2. The resulting model variants are then further evaluated on the classification tasks.

[1]https://github.com/MarkusSagen/Master-Thesis-Multilingual-Longformer, accessed on November 16, 2023

Classification

For each interview part, we construct a binary classification task to examine whether we can distinguish healthy subjects from those affected by AD. Additionally, we also group all inputs from all interview parts together for a combined classification task. We add a classification head to each language model, and (further) fine-tune the entire model on the task data for 20 epochs with a batch size of 64 in a stratified five-fold cross-validation setting, using the default optimizer and learning rate scheduler from the huggingface transformers python library (Wolf et al. 2020). To account for class imbalance, we employ class weights in the cross-entropy loss that are inversely proportional to the class distribution. We report the average accuracy across the five test splits.

Results

The classification accuracies for all language models and interview parts are presented in Table 1 and 2. Overall, all models perform best on the object descriptions task.

Model Comparison

In general, the Dutch RobBERT model performs best across the classification tasks, achieving an average accuracy of 0.866 on the combined input. In particular, it substantially outperforms the multilingual XLM-RoBERTa model. Moreover, when further comparing the multilingual variants, there is no clear improvement of the sentence- or document-level models over the word-level one. However, the XLM-RoBERTa and Longformer-XLM models outperform Rob-

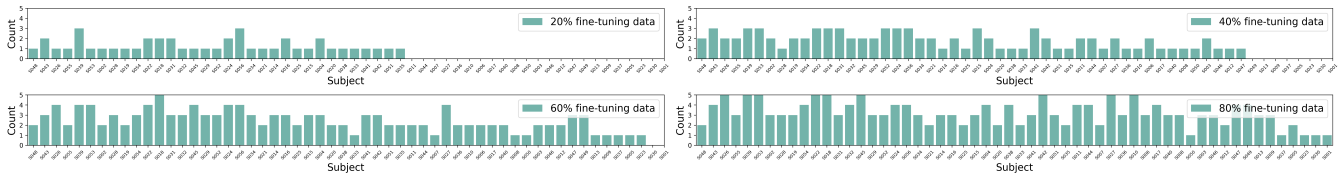


Figure 2: **Subject distribution of fine-tuning corpus.** This plot illustrates the number of interview parts included per subject for each of the fine-tuning partitions.



Figure 3: **Evaluation loss.** For each of the fine-tuned RobBERT variants shown in Table 2, we monitor its loss during its $2,000$ training steps on a separate evaluation dataset.

BERT on the news event and picture description tasks, respectively.

Domain Adaptation

As illustrated in Figure 3, a larger proportion of fine-tuning data results in improved domain adaptation, as measured by the loss of the different fine-tuned RobBERT variants on the evaluation dataset. Notably, a proportion of only 20% indicates signs of overfitting (i.e., an increase in the evaluation loss over time), while proportions of 40% or more seem to mitigate this issue. Nonetheless, improved domain adaptation does not necessarily improve downstream classification performance as shown in Table 2. While the RobBERT model that has been fine-tuned on 40% of the available fine-tuning data results in the best overall classification performance (also compared to the model variants presented in Table 1), further increases of the proportion of fine-tuning data do not lead to improved classification performances.

Discussion

To distinguish healthy subjects from those affected by AD, we applied a wide range of language models to several classification tasks constructed from different kinds of Flemish Dutch connected speech. We found that all investigated models perform best on the object description task. This could be linked to the substantially longer available input for this task (see Figure 1). However, the impact of a longer input length on the model performances is less clear for the autobiographical part, speaking against an explanation of the differences in model performance based on purely the difference in input length between the two groups. Moreover, the Dutch RobBERT model outperforms all multilingual variants on three of the interview parts as well as the combined data. Still, the XLM-RoBERTa model performs better on the picture description task, which can be attributed to its generalisation over multiple languages that might allow for a better representation of descriptions of a visual setting. We also do not see a clear improvement of the S-XLM-RoBERTa and Longformer-XLM models over the word-level XLM-RoBERTa one. Specifically for the sentence-level S-XLM-RoBERTa model, we suspect that the averaging procedure (see Reimers and Gurevych 2019) might result in a noisy representation compared to using the `[CLS]` token in word-level models. Lastly, we show that fine-tuning leads to better domain adaptation. Taken together, we believe that adapting a pre-trained model to variations in language (Flemish versus standard Dutch) and domain (i.e., the interview format) is more beneficial to obtaining nuanced representations of the subjects' connected speech rather than using specialized models for long-range context, especially in applications where the required labelled data for language-specific sentence-level fine-tuning is not available.

Conclusion

In this paper, we investigated how the choice of different language models affect the ability to distinguish between healthy controls and AD patients using transcripts of their connected speech. Our results indicate that for these native Flemish Dutch speakers, a Dutch word-level model outperforms multilingual sentence- and document-level models, and that unsupervised fine-tuning on a separate connected speech corpus helps with domain adaptation.

Acknowledgements

This research was supported by funding from the Research Foundation - Flanders (Fonds Wetenschappelijk Onderzoek, FWO) grant 1154623N. The authors would like to thank Lauren Nicole DeLong for her feedback on the manuscript.

References

Beltagy, I.; Peters, M. E.; and Cohan, A. 2020. Longformer: The Long-Document Transformer. *arXiv:2004.05150 [cs]*.

Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Grave, E.; Ott, M.; Zettlemoyer, L.; and Stoyanov, V. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8440–8451.

de la Fuente Garcia, S.; Ritchie, C. W.; and Luz, S. 2020. Artificial Intelligence, Speech, and Language Processing Approaches to Monitoring Alzheimer's Disease: A Systematic Review. *Journal of Alzheimer's Disease*, 78(4): 1547–1574.

Delobelle, P.; Winters, T.; and Berendt, B. 2022. RobBERT-2022: Updating a Dutch Language Model to Account for Evolving Language Use. *arXiv:2211.08192 [cs]*.

Li, R.; Wang, X.; Lawler, K.; Garg, S.; Bai, Q.; and Alty, J. 2022. Applications of artificial intelligence to aid early detection of dementia: A scoping review on current capabilities and future directions. *Journal of Biomedical Informatics*, 127: 104030.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692*.

Luz, S.; Haider, F.; de la Fuente, S.; Fromm, D.; and MacWhinney, B. 2021. Detecting cognitive decline using speech only: The ADReSSo challenge. *arXiv preprint arXiv:2104.09356*.

Luz, S.; Haider, F.; Fromm, D.; Lazarou, I.; Kompatsiaris, I.; and MacWhinney, B. 2023. Multilingual Alzheimer's Dementia Recognition through Spontaneous Speech: a Signal Processing Grand Challenge. *arXiv:2301.05562 [cs, eess]*.

Luz, S.; Haider, F.; Fuente, S. d. l.; Fromm, D.; and MacWhinney, B. 2020. Alzheimer's Dementia Recognition Through Spontaneous Speech: The ADReSS Challenge. In *Interspeech 2020*, 2172–2176.

Mirheidari, B.; Pan, Y.; Walker, T.; Reuber, M.; Venneri, A.; Blackburn, D.; and Christensen, H. 2019. Detecting Alzheimer's Disease by estimating attention and elicitation path through the alignment of spoken picture descriptions with the picture prompt. *arXiv:1910.00515 [cs]*.

Mueller, K. D.; Koscik, R. L.; Turkstra, L. S.; Riedeman, S. K.; LaRue, A.; Clark, L. R.; Hermann, B.; Sager, M. A.; and Johnson, S. C. 2016. Connected Language in Late Middle-Aged Adults at Risk for Alzheimer's Disease. *Journal of Alzheimer's disease*, 54(4): 1539–1550.

Nevler, N.; Ash, S.; Jester, C.; Irwin, D. J.; Liberman, M.; and Grossman, M. 2017. Automatic measurement of prosody in behavioral variant FTD. *Neurology*, 89(7): 650–656.

Oostdijk, N. 2000. The Spoken Dutch Corpus: Overview and first evaluation. *Proceedings of LREC-2000*, 2.

Petti, U.; Baker, S.; and Korhonen, A. 2020. A systematic literature review of automatic Alzheimer's disease detection from speech and language. *Journal of the American Medical Informatics Association*, 27(11): 1784–1797.

Poncelet, J.; and Van Hamme, H. 2023. Learning to Jointly Transcribe and Subtitle for End-To-End Spontaneous Speech Recognition. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, 182–189.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3982–3992.

Reimers, N.; and Gurevych, I. 2020. Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 4512–4525.

Roth, C. 2011. Boston Diagnostic Aphasia Examination. In Kreutzer, J. S.; DeLuca, J.; and Caplan, B., eds., *Encyclopedia of Clinical Neuropsychology*, 428–430.

Schaeverbeke, J. M.; Gabel, S.; Meersmans, K.; Luckett, E. S.; De Meyer, S.; Adamczuk, K.; Nelissen, N.; Goovaerts, V.; Radwan, A.; Sunaert, S.; Dupont, P.; Van Laere, K.; and Vandenberghe, R. 2021. Baseline cognition is the best predictor of 4-year cognitive change in cognitively intact older adults. *Alzheimer's Research & Therapy*, 13(1): 75.

Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; Davison, J.; Shleifer, S.; von Platen, P.; Ma, C.; Jernite, Y.; Plu, J.; Xu, C.; Scao, T. L.; Gugger, S.; Drame, M.; Lhoest, Q.; and Rush, A. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*.