Early Dementia Diagnosis from Spoken Language using a Transformer Approach

Alexander Aslaksen Jonasson, Alfred Wahlforss, Jonas Beskow and Joakim Gustafsson
Department of Speech, Music and Hearing, KTH Royal Institute of Technology
aljonass, alfredwa, jkgu, beskow@kth.se

Background

- Early detection of dementia is necessary for development of efficient treatments against further progression. An early diagnosis requires advanced, expensive and extensive examinations, which are unfeasible for the majority of people with cognitive complaints.
- There is considerable evidence that dementia affects speech, also in very early stages. Analysing the speech output of patients could therefore be a potential tool for creating cost effective diagnostics tools.
- New transformer models such as BERT brought forth a new paradigm for natural language processing. They vastly improve accuracy on many linguistic tasks.

Data Overview

- The Pitt corpus from DementiaBank was used, which includes transcripts from subjects (Ss) performing the Cookie Theft task. Samples are taken from a study of AD conducted between 1983 and 1988.
- After excluding Ss with MCI, 275 were left with a total of 512 data samples (269 AD, 243 Control), with some Ss performing the test up to three times. The Pitt corpus also includes audio samples from each patient visit.
- Google's Automatic Speech Recognition (ASR) model was used to automatically generate transcripts from these. These differ from the linguist transcripts as they also include the physician's utterances.

Method

- We use 10-fold cross validation to reduce bias.
- Each Ss participated in the cookie theft task between one and three times. Thus, in order to avoid training a model on one sample from one subject and also evaluating the model on another sample from the same subject, the samples were grouped together on a per-patient basis. As such, the samples were divided into train-test splits according to patients rather than samples.
- The 275 Ss were split into 10 sets, averaging 27 Ss per split.
- We evaluate three different models; BERT with a maximum token input length of 256, referred to as BERT256, and one with maximum token input length of 512, referred to as BERT512, and RoBERTa with a maximum token input length of 512, referred to as RoBERTa512.

Result

Results for different metrics and models. **Our model outperforms SOTA.**

Model	B256	B512	R512	Guo et al., 2019[1]
Accuracy	84.96%	85.55%	**86.72%**	85.4%
Precision	85.82%	84.45%	**90.69%**	
Specificity	84.36%	81.89%	**90.53%**	
Recall	85.50%	**88.85%**	83.27%	

B refers to BERT, R to RoBERTa. 256 and 512 refers to the maximum input length used.

Best results on different transcript types. ASR performs well.

Transcript	Linguist	ASR
Accuracy	**86.72%**	83.59%
Precision	**90.69%**	86.56%
Specificity	**90.53%**	86.01%
Recall	**83.27%**	81.41%

Conclusion

- The transformer architecture is suitable for classification of text samples from the Cookie Theft Task. RoBERTa achieved the highest, accuracy, precision and specificity.
- Higher performance was achieved when using the linguist transcriptions compared to ASR transcriptions.
- Only slightly worse performance using ASR, may yield better results with more modern recording equipment, and with samples in which physician's voice is excluded.
- Newer equipment combined with ASR and transformer models may offer a comparatively inexpensive method for large scale screenings to find individuals early on, for whom further, more advanced evaluation is desirable.