Analysis of Hierarchical Multi-Content Text Classification Model on B-SHARP Dataset for Early Detection of Alzheimer's Disease

Renxuan A. Li
Computer Science
Emory University
Atlanta, GA, USA
albert.li@emory.edu

Ihab Hajjar
Neurology
Emory University
Atlanta, GA, USA
ihajjar@emory.edu

Felicia Goldstein
Neurology
Emory University
Atlanta, GA, USA
fgoldst@emory.edu

Jinho D. Choi
Computer Science
Emory University
Atlanta, GA, USA
jchoi31@emory.edu

Abstract

This paper presents a new dataset, B-SHARP, that can be used to develop NLP models for the detection of Mild Cognitive Impairment (MCI) known as an early sign of Alzheimer's disease. Our dataset contains 1-2 min speech segments from 326 human subjects for 3 topics, (1) daily activity, (2) room environment, and (3) picture description, and their transcripts so that a total of 650 speech segments are collected. Given the B-SHARP dataset, several hierarchical text classification models are developed that jointly learn combinatory features across all 3 topics. The best performance of 74.1% is achieved by an ensemble model that adapts 3 types of transformer encoders. To the best of our knowledge, this is the first work that builds deep learning-based text classification models on multiple contents for the detection of MCI.

1 Introduction

Alzheimer's Disease (AD) is a progressive neurodegenerative disorder that is associated with memory loss and declines in major brain functions including semantic and pragmatic levels of language processing (Vestal et al., 2006; Ferris and Farlow, 2013). Traditional cognitive assessments such as positron emission tomography or cerebrospinal fluid analysis are expensive and time-consuming (Fyffe et al., 2011). This may cause delay in treating AD, known to be irreversible and incurable (Korczyn, 2012), and put an increasing pressure on public health, especially for seniors whose life expectancy is rapidly growing yet are more likely to develop AD. Thus, it is crucial to find a more intelligent way of detecting AD in the earliest stage possible (Karr et al., 2018).

Mild Cognitive Impairment (MCI) is considered the first phase that patients start having biomarker evidence of brain changes that can eventually lead to AD (Albert et al., 2011). MCI involves subtle language changes from impairment in reasoning that may not be noticeable to people other than friends and relatives. Because of this, the detection of MCI is a much more challenging task than detecting dementia (Suzman and Beard, 2011). Recent studies in NLP have shown that it is possible to detect early stages of AD by analyzing patients' language patterns; however, most previous works have focused on the detection of dementia instead and researches tackling the detection of MCI have been based on relatively small datasets (Section 2).

This paper presents a new dataset that comprises three types of speech segments from both normal controls and MCI patients (Section 3). Then, a hierarchical text classification model is proposed, which jointly learns features from all three types of speech segments to determine whether or not each subject has MCI (Section 4). Individual and ensemble models using three types of transformer encoders are evaluated on our dataset and show that different transformer encoders reveal strengths in distinct types of speeches (Section 5). We believe that this work takes the initiative of deep learning-based NLP for detecting MCI that will be broadly beneficial to global public health.

2 Related Work

Only few studies have tackled the detection of MCI using NLP.[1] Asgari et al. (2017) conducted interviews with (27_C, 14_M), and developed SVM and random forest models on their transcribed speeches. Beltrami et al. (2018) conducted three speech tasks with (48_C, 32_M, 16_D), and analyzed phonetic and linguistic features of their speeches and transcripts. Fraser et al. (2019) conducted 3 language tasks with (29_C, 26_M), and built a cascade model to learn multimodal features such as audio, text, eye-tracking. Gosztolya et al. (2019) conducted question answer-

[1] $\#_C$: the number of normal controls,
$\#_{M/D/A}$: the number of MCI / Dementia / AD patients.

Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, pages 358–365
December 4 - 7, 2020.

		Tokens	Sentences	Nouns	Verbs	Conjuncts	Complex	Discourse
Q_1	Control	186.6 (±60.4)	10.4 (±4.5)	28.1 (±9.6)	30.4 (±11.5)	8.5 (±4.5)	2.3 (±1.7)	8.1 (±5.4)
	MCI	175.6 (±54.5)	9.8 (±4.1)	23.7 (±8.3)	29.3 (±10.4)	8.5 (±4.2)	2.0 (±1.6)	9.2 (±6.0)
Q_2	Control	191.5 (±11.8)	11.7 (±4.7)	41.1 (±13.3)	24.3 (±11.2)	6.6 (±4.5)	3.6 (±2.7)	7.1 (±4.8)
	MCI	178.6 (±11.7)	11.6 (±4.7)	36.7 (±12.1)	23.2 (±10.6)	6.4 (±4.4)	2.9 (±2.3)	8.4 (±5.3)
Q_3	Control	193.4 (±63.4)	12.6 (±5.4)	39.5 (±13.5)	28.4 (±10.1)	8.0 (±4.8)	3.3 (±2.1)	6.1 (±5.5)
	MCI	187.8 (±63.4)	12.7 (±5.1)	36.2 (±13.2)	27.7 (±10.9)	7.2 (±4.2)	2.6 (±2.0)	7.3 (±5.5)
All	Control	**578.1** (±149.8)	34.5 (±10.7)	**110.5** (±27.9)	84.2 (±25.4)	23.5 (±10.1)	**9.3** (±4.5)	21.4 (±13.0)
	MCI	548.7 (±140.6)	34.0 (±10.5)	98.1 (±26.1)	81.2 (±24.1)	22.5 (±9.7)	7.7 (±4.2)	**25.3** (±15.0)
	p	**0.0110**	0.5541	**< 0.0001**	0.1277	0.2046	**< 0.0001**	**0.0006**

Table 1: Average counts and their standard deviations of linguistic features per transcript in the B-SHARP dataset. Complex: occurrences of complex structures (e.g., relative clauses, non-finite clauses), Discourse: occurrences of discourse elements (e.g., interjections, disfluency).

ing sessions with (25_C, 25_M, 25_A), and trained a SVM model using acoustic and linguistic features. All of the previous works were based on fewer than 100 subjects using traditional linguistic features to develop NLP models, compared to our work that is based on 326 subjects and 650 recordings using the latest transformer-based deep neural models.

The task of dementia detection has been more explored by the NLP community. Becker et al. (1994) presented the DementiaBank, that consists of 552 audio recordings describing the picture called "*The Boston Cookie Theft*" from 99 normal controls and 194 dementia patients, that have been used by the following works. Orimaye et al. (2016) presented deep-deep neural network language models using higher-order n-grams and *skip*-grams. Pou-Prom and Rudzicz (2018) leveraged linguistic features and multiview embeddings by applying generalized canonical correlation analysis. Karleka et al. (2018) proposed a model based on convolutional and recurrent neural networks and gave interpretations of this model to explain linguistic characteristics for detecting dementia. Our work is distinguished as:

- We tackle the detection of MCI, not dementia,
- Our documents are multi-contents compared to single-content documents in the DementiaBank.
- Our approach is based on the latest contextualized embeddings compared to the distributional embeddings adapted by the previous works.

3 Dataset

3.1 B-SHARP

Our work is based on data collected as part of the Brain, Stress, Hypertension, and Aging Research Program (B-SHARP).[2] In this dataset, 185 normal controls and 141 MCI patients are selected based on neuropsychological and clinical assessments. Every subject has been examined with multiple cognitive tests including the Montreal Cognitive Assessment (MoCA; Nasreddine et al. 2005) and the Boston Naming Test (BNT; Kaplan et al. 1983), followed by a speech task protocol for recording. 51.5% and 23.9% of the subjects have so far come back for their 2nd and 3rd visits to take new voice recordings, respectively. B-SHARP is an ongoing program; recordings of 20-25 subjects are taken every month; thus, the data is still growing.

	Sbj	2nd	3rd	Rec	MoCA	BNT
C	185	100	50	385	26.2 (±2.6)	14.2 (±1.2)
M	141	68	28	265	21.5 (±3.5)	13.4 (±1.5)
Σ	326	168	78	650	24.2 (±3.8)	13.9 (±1.4)

Table 2: Statistics of control (C) and MCI (M) groups. Sbj: # of subjects, 2nd/3rd: # of subjects who made the 2nd/3rd visits, Rec: # of voice recordings, MoCA/BNT: average scores and stdevs from MoCA/BNT. Note that subjects with the 2nd/3rd visits take one/two additional recordings; thus, Rec = Sbj + 1·(2nd) + 2·(3rd).

Table 2 shows the statistics of the control and the MCI groups in B-SHARP. Note that when subjects make multiple visits, there is a year gap in between so that subjects generally do not remember so much from their previous visits. Thus, speeches from the same subject are not necessarily more similar than ones from the other subjects. In fact, most speeches across subjects, regardless of their groups, are very similar when they are transcribed since all subjects follow the same speech protocol in Section 3.2.[3]

3.2 Speech Task Protocol

A speech task protocol has been conducted to collect recordings of both control and MCI subjects

[2] B-SHARP: http://medicine.emory.edu/bsharp

[3] A.3 compares B-SHARP with the DementiaBank in details.

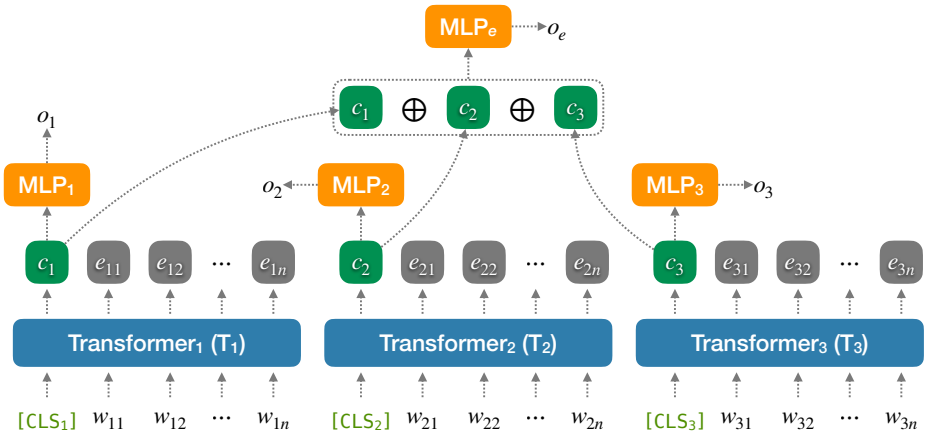

Figure 1: Overview of hierarchical transformer to combine content features from the three types of speech tasks.

who are asked to speak about $\mathtt{Q}_1$: daily activity, $\mathtt{Q}_2$: room environment, and $\mathtt{Q}_3$: picture description for 1-2 minutes each. All subjects are provided with the same instructions in A.2, and visual abilities of the subjects are confirmed before recording. To reduce potential variance, the subjects are guided to follow similar activities before $\mathtt{Q}_1$, located to similar room settings before $\mathtt{Q}_2$, and shown the same picture in Fig 2, "*The Circus Procession*", for $\mathtt{Q}_3$.

The collected voice recordings are automatically transcribed by the online tool called Temi.[4] Table 1 shows linguistic features about our dataset analyzed by the open-source NLP toolkit, ELIT.[5] Transcripts from the control group depict significantly higher numbers of tokens, nouns, and complex structures while transcripts from the MCI group gives significantly more discourse elements, implying that the control subjects are more expressive while the MCI subjects include more disfluency in their speeches.

4 Hierarchical Transformer

Although transformer encoders have recently established the state-of-the-art results on most document classification tasks, they have a limit on the input size. As in Table 1, the average number of tokens in our input documents well-exceeds 512 when combining transcripts from all three tasks, which is the max-number of tokens that the pretrained models of these transformers allow in general. This makes it difficult to simply join all transcripts together and feed into a transformer encoder. Thus, this section presents a hierarchical transformer to overcome the challenge of long documents while jointly training transcript contents from all three tasks (Figure 1).

Let $W_i = \{w_{i1}, \ldots, w_{in}\}$ be a transcript, where w_{ij} represents the j'th token in the transcript produced by the i'th task $\mathtt{Q}_i$ (in our case, $i = \{1, 2, 3\}$). W_i is prepended by the special token $\mathtt{[CLS}_i\mathtt{]}$ that is used to learn the transcript embedding, and fed into the transformer T_i. The transformer then generates $E_i = \{c_i, e_{i1}, \ldots, e_{in}\}$, where c_i and e_{ij} are the embeddings for $\mathtt{[CLS}_i\mathtt{]}$ and w_{ij}, respectively. $c_i \in \mathbb{R}^d$ is used to make two types of predictions.

First, c_i is fed into a multilayer perceptron layer, $\mathtt{MLP}_i$, that generates the output vector $o_i \in \mathbb{R}^2$ to predict whether or not the subject has MCI based on the transcript from $\mathtt{Q}_i$ alone. Second, the transcript embeddings from all three tasks are concatenated such that $c_e = c_1 \oplus c_2 \oplus c_3 \in \mathbb{R}^{3d}$, which gets fed into another $\mathtt{MLP}_e$ to generate the output vector $o_e \in \mathbb{R}^2$, and makes the binary decision based on the transcripts from all three tasks, $\mathtt{Q}_1$, $\mathtt{Q}_2$ and $\mathtt{Q}_3$.

5 Experiments

There are 650 recordings in our dataset (Table 2), that is rather small to divide into train, development, and test sets. Thus, 5-fold cross-validation (CV) is used to evaluate the performance of our models. Table 5 shows the distributions of the five CV sets for our experiments, where the transcript of each recording is treated as an independent document. Notice that the distributions are calculated based on analysis of the last MLP layer instead of simple majority vote on individual models.

It is worth mentioning that all recordings from the same subject given multiple visits are assigned to the same CV set; thus, there is no overlap in terms of subjects across these CV sets. This allows us to avoid potential inflation in accuracy due to unique language patterns used by individual subjects.

[4]Temi: https://www.temi.com
[5]ELIT: https://github.com/elitcloud/elit

	BERT			RoBERTa			ALBERT		
	Q_1	Q_2	Q_3	Q_1	Q_2	Q_3	Q_1	Q_2	Q_3
ACC	67.6 (±0.4)	**69.0** (±1.2)	67.7 (±0.7)	69.0 (±1.5)	**69.9** (±0.2)	65.2 (±0.3)	67.6 (±1.5)	**69.5** (±0.3)	66.6 (±1.3)
SEN	48.9 (±1.8)	**57.1** (±2.5)	41.5 (±3.6)	44.3 (±4.5)	**55.3** (±1.2)	37.1 (±3.7)	45.9 (±1.9)	**52.2** (±0.6)	37.4 (±3.3)
SPE	80.4 (±1.2)	77.3 (±2.8)	**85.2** (±3.0)	**85.8** (±2.1)	79.7 (±0.7)	84.5 (±3.0)	82.6 (±3.7)	81.4 (±0.3)	**86.8** (±3.3)

Table 3: Model performance on the individual tasks. ACC: accuracy, SEN: sensitivity, SPE: specificity.

	CNN	BERT_e	RoBERTa_e	ALBERT_e	$\text{B}_e + \text{R}_e$	$\text{A}_e + \text{R}_e$	$\text{B}_e + \text{A}_e + \text{R}_e$
ACC	69.5 (±0.2)	69.9 (±1.1)	71.6 (±1.5)	69.7 (±2.9)	72.2 (±0.7)	71.5 (±1.9)	**74.1** (±0.3)
SEN	49.2 (±0.8)	57.6 (±3.4)	48.5 (±6.1)	46.2 (±8.3)	56.5 (±2.5)	51.7 (±1.3)	**60.9** (±5.2)
SPE	83.5 (±0.9)	77.4 (±4.8)	**87.5** (±1.8)	85.4 (±0.5)	83.1 (±0.9)	86.7 (±3.4)	84.0 (±2.4)

Table 4: Performance of ensemble models. Bert_e/RoBERTa_e/ALBERT_e use transcript embeddings from all 3 tasks trained by the BERT/RoBERTa/ALBERT models in Table 3, respectively. B_e+R_e uses transcript embeddings from both Bert_e and RoBERTa_e (so the total of 6 embeddings), A_e+R_e uses transcript embeddings from both ALBERT_e and RoBERTa_e (6 embeddings), and B_e+A_e+R_e uses transcript embeddings from all three models (9 embeddings).

Three transformer encoders are used, BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2020), and ALBERT (Lan et al., 2019) for our experiments. Every model is trained 3 times and its average performance with the standard deviation are reported.[6]

	CV_0	CV_1	CV_2	CV_3	CV_4	ALL
C_{Rec}	77	77	77	77	77	**385**
M_{Rec}	53	53	53	53	53	**265**
C_{Sbj}	37	37	37	37	37	**185**
M_{Sbj}	27	28	28	29	29	**141**

Table 5: Statistics of the CV sets for our experiments. Rec/Sbj: # of recordings/subjects, C/M: in control/MCI group. CV_i: the i'th set. ALL: $\sum_{i=0}^{4} \text{CV}_i$.

5.1 Performance of Individual Models

Individual models are built by training transcripts from each task separately using MLP_i in Section 4. Table 3 shows the performance of the 3 transformer models on the individual tasks. The performance on Q_2 shows the highest accuracy for all three models, achieving 69.9% with RoBERTa, implying that the room environment task of Q_2, involving many spatial descriptions, are the most effective to distinguish the MCI group. The highest sensitivity of 57.1% is achieved by BERT on Q_2, and the highest specificity of 86.8% is achieved by ALBERT on Q_3. Such a low sensitivity and a high specificity imply that it is easier to recognize the normal controls but not the MCI patients given the short speeches.

5.2 Performance of Ensemble Models

Ensemble models are developed by jointly training multiple transcript embeddings from the individual models using MLP_e in Section 4. Table 4 shows the model performance of the ensemble models. Additionally, results from a model that takes transcripts from the 3 tasks as one input document and trains a convolutional neural network (CNN) are provided for comparison to Karleka et al. (2018).[7] R_e shows 1.7% improvement on accuracy over the RoBERTa model in Table 3 although its sensitivity is worse. Table 6 shows the voting distributions of each task combination; given the samples correctly predicted by RoBERTa_e, we count how often the individual models are correct for those samples by comparing the weights in MLP_e and estimate the percentages. The combination of (Q_1, Q_3) shows the highest percentage of 30%, meaning that 30% of the corrected predicted samples are voted by both Q_1 and Q_3.

Q_1	Q_2	Q_3	$Q_{1,2}$	$Q_{1,3}$	$Q_{2,3}$	$Q_{1,2,3}$
5.8	6.4	2.8	19.5	**30.0**	8.8	26.1

Table 6: Voting distributions of each task combination for RoBERTa_e. Q_i: % of only the Q_i model is correct, $Q_{i,i,j}$: % of all Q_i, Q_i, and Q_j models are correct.

A similar analysis is done for B_e+R_e+A_e although displaying the distributions is quite infeasible since it involves 2^9-1 combinations. Among the samples correctly predicted by B_e+R_e+A_e, 86% are derived from majority votes; in other words, at least 5 out of 9 individual models agree with the predictions. Votes from 6 and 5 models are the largest groups, showing 35% and 28%, respectively. Only 0.21% are agreed by all 9 models. No case of votes from 3 or less models is found, implying that no individual model dominates the final decision of B_e+R_e+A_e.

[6] Details about the experimental settings are provided in A.1.

[7] We also experiemented with LSTM-RNN and CNN-LSTM models as suggested by Karleka et al. (2018); however, the CNN model gave the highest accuracy on our dataset.

6 Conclusion

This paper presents the B-SHARP dataset, that is the largest dataset for the task of MCI detection feasible to develop robust deep neural models. Our best ensemble model using hierarchical transformer gives the accuracy of 74% to distinguish MCI patients from normal controls that is very promising. We will also explore models to make a longevity analysis per patient with this dataset.[8]

References

Marilyn S Albert, Steven T DeKosky, Dennis Dickson, Bruno Dubois, Howard H Feldman, Nick C Fox, Anthony Gamst, David M Holtzman, William J Jagust, Ronald C Petersen, Peter J Snyder, Maria C Carrillo, Bill Thies, and Creighton H Phelps. 2011. The Diagnosis of Mild Cognitive Impairment Due to Alzheimer's Disease: Recommendations From the National Institute on Aging-Alzheimer's Association Workgroups on Diagnostic Guidelines for Alzheimer's Disease. *Alzheimer's Dementia*, 7(3):270–279.

Meysam Asgari, Jeffrey Kaye, and Hiroko Dodge. 2017. Predicting Mild Cognitive Impairment From Spontaneous Spoken Utterances. *Alzheimer's Dementia*, 3(2):219–228.

James T. Becker, Francois Boller, Oscar L. Lopez, Judith Saxton, and Karen L. McGonigle. 1994. The natural history of alzheimer's disease: description of study cohort and accuracy of diagnosis. *Archives of Neurology*, 51(6):585–594.

Daniela Beltrami, Gloria Gagliardi, Rema Rossini Favretti, Enrico Ghidoni, Fabio Tamburini, and Laura Calzà. 2018. Speech Analysis by Natural Language Processing Techniques: A Possible Tool for Very Early Detection of Cognitive Decline? *Frontiers in Aging Neuroscience*, 10(369):1–13.

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 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.

Steven H Ferris and Martin Farlow. 2013. Language Impairment in Alzheimer's Disease and Benefits of Acetylcholinesterase Inhibitors. *Clinical Interventions in Aging*, 8:1007–1014.

Kathleen C. Fraser, Kristina Lundholm Fors, Marie Eckerström, Fredrik Öhman, and Dimitrios Kokkinakis. 2019. Predicting MCI Status From Multimodal Language Data Using Cascaded Classifiers. *Frontiers in Aging Neuroscience*, 11(205):1–18.

[8] All our resources including source codes and models are available at http://anonymous.

Denise C. Fyffe, Shubhabrata Mukherjee, Lisa L. Barnes, Jennifer J. Manly, David A. Bennett, and Paul K. Crane. 2011. Explaining Differences in Episodic Memory Performance among Older African Americans and Whites: The Roles of Factors Related to Cognitive Reserve and Test Bias. *Journal of the International Neuropsychological Society*, 17(4):625–638.

Gábor Gosztolya, Veronika Vinczea, László Tóth, Magdolna Pákáskid, János Kálmand, and Ildikó Hoffmann. 2019. Identifying Mild Cognitive Impairment and mild Alzheimer's disease based on spontaneous speech using ASR and linguistic features. *Computer Speech & Language*, 53:181–197.

Edith Kaplan, Harold Goodglass, and Sandra Weintraub. 1983. *The Boston Naming Test*. Philadelphia: Lea and Febiger.

Sweta Karleka, Tong Niu, and Mohit Bansal. 2018. Detecting linguistic characteristics of alzheimer's dementia by interpreting neural models. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 701–707, New Orleans, Louisiana. Association for Computational Linguistics.

Justin E Karr, Raquel B Graham, Scott M Hofer, and Graciela Muniz-Terrera. 2018. When Does Cognitive Decline Begin? A Systematic Review of Change Point Studies on Accelerated Decline in Cognitive and Neurological Outcomes Preceding Mild Cognitive Impairment, Dementia, and Death. *Psychology and Aging*, 33(2):195–218.

Amos D Korczyn. 2012. Why Have We Failed to Cure Alzheimer's Disease? *Journal of Alzheimer's Disease*, 29(2):275–282.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. *arXiv*, 11942(1909).

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2020. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv*, 1907(11692).

Ziad S. Nasreddine, Natalie A. Phillips, Valérie Bédirian, Simon Charbonneau, Victor Whitehead, Isabelle Collin, Jeffrey L. Cummings, and Howard Chertkow. 2005. The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. *Journal of the American Geriatrics Society*, 53(4):695–699.

Sylvester Olubolu Orimaye, Jojo Sze-Meng Wong, and Judyanne Sharmini Gilbert Fernandez. 2016. Deep-Deep Neural Network Language Models for Predicting Mild Cognitive Impairment. In *Proceedings of*

the IJCAI Workshop on Advances in Bioinformatics and Artificial Intelligence, pages 14–20.

Chloé Pou-Prom and Frank Rudzicz. 2018. Learning multiview embeddings for assessing dementia. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP'18, pages 2812–2817, Brussels, Belgium. Association for Computational Linguistics.

Richard Suzman and John Beard. 2011. Global health and aging.

Lindsey Vestal, Laura Smith-Olinde, Gretchen Hicks, Terri Hutton, and John Hart Jr. 2006. Efficacy of Language Assessment in Alzheimer's Disease: Comparing In-Person Examination and Telemedicine. *Clinical Interventions in Aging*, 1(4):467–471.

A Appendix

A.1 Experimental Settings

Table 7 shows the configuration of the transformer encoders in Section 5. The base pre-trained models are used for all encoders.

Transformer	L	AH	IC	HC	P
BERT	12	12	768	768	108M
RoBERTa	12	12	768	768	125M
ALBERT	12	12	768	128	12M

Table 7: Configurations of the BERT, RoBERTa, and ALBERT encoders for our experiments. L: # of layers, AH: # of attended heads, IC: # of input cells, HC: # of hidden cells, P: # of parameters.

Individual Models For training the BERT and RoBERTa models, the batch size of 5, the learning rate of $5 \cdot 10^{-6}$, and the gradient clip of norm 0.5 are used with the Adam optimizer. A dropout rate of 0.15 is applied to all layers. For the ALBERT model, the batch size of 8 is used. All three models are trained for 30 epochs.

Ensemble Models For training the two model ensembles, B_e+R_e and A_e+R_e, the batch size of 72 and the learning rate of $5 \cdot 10^{-5}$ are used with the Adam optimizer for 200 epochs. A dropout rate of 0.25 is also applied. For training the B_e+A_e+R_e model, the dropout rate is set to 0.3.

A.2 Speech Task Protocol

Table 8 describes the instructions provided to the subjects for the three speech tasks in Section 3.2.

Task	Instruction
Q_1	I would like you to describe to me everything we did from the moment we met today until now. Please try to recall as many details as possible in the order the events actually happened where we met, what we did, what we saw, where we went, and what you felt or thought during each of these events.
Q_2	I would like you to describe everything that you see in this room.
Q_3	I am going to show you a picture and ask you to describe what you see in as much detail as possible. You can describe the activities, characters, and colors of things you see in this picture.

Table 8: Instructions of the 3 speech tasks, Q_1, Q_2, Q_3, provided to the subjects.

Figure 2 illustrates the image of the picture called "*The Circus Procession*" for the picture description task, Q_3, copyrighted by the *McLoughlin Brothers* as part of the Juvenile Collection.

Figure 2: The picture of "*The Circus Procession*" used in the B-SHARP dataset.

A.3 B-SHARP Compared to DementiaBank

DementiaBank is the largest public dataset for dementia detection that comprises recordings for 4 language tasks, picture description, verbal fluency, story recall, and sentence construction, from a large longitudinal study (Becker et al., 1994). Subjects in this study are divided into two groups, normal controls and dementia patients. Among the four tasks, data from only the picture description task can be used for classification since the other tasks give data of dementia patients only.[9] The design of this task is similar to Q_3 in B-SHARP (Section 3.2); each subject is shown "*The Boston Cookie Theft*" picture in Figure 3 to describe for 1-2 minutes.

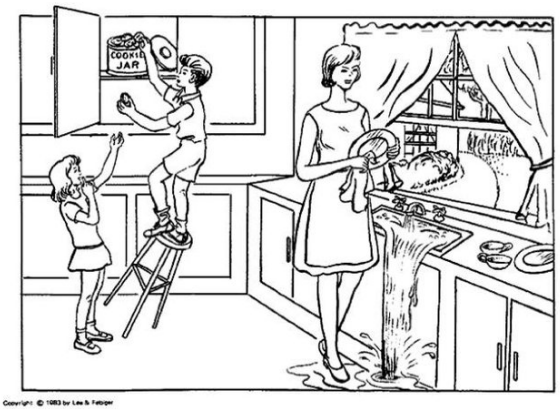

Figure 3: The picture of "*The Boston Cookie Theft*" used in the DementiaBank.

Table 10 shows the statistics of the DementiaBank in comparison to Table 2 in Section 3. Subjects in this study made up to 5 visits compared to 3 in B-SHARP although the number of subjects in each visit is larger in B-SHARP. B-SHARP has ≈100

[9]The verbal fluency task gives 1 audio recording of a normal control, that is still not enough to train classification models.

	Tokens	Sentences	Nouns	Verbs	Conjuncts	Complex	Discourse
Control	124.0 (±59.7)	12.6 (±5.1)	**23.7** (±11.8)	**27.1** (±11.9)	2.8 (±2.8)	1.6 (±1.6)	1.5 (±1.6)
Dementia	114.3 (±61.3)	12.1 (±6.4)	18.7 (±10.4)	23.9 (±12.9)	2.4 (±2.4)	1.4 (±1.4)	**2.8** (±2.9)
p	0.0625	0.3204	$\mathbf{< 0.0001}$	**0.0029**	0.0715	0.1184	$\mathbf{< 0.0001}$

Table 9: Average counts and standard deviations of linguistic features per transcript in the DementiaBank. See the caption in Table 1 for the column descriptions.

more recordings than the DementiaBank, more importantly, B-SHARP is still growing, which makes it the largest dataset for NLP research related to the detection of Alzheimer's Disease. Unlike DementiaBank where 66.2% of the subjects are dementia patients, 43.3% of the subjects belong to the MCI group in B-SHARP; this makes sense because MCI is closer to the preclinical phase that involves a much fewer number of patients reported in general.

Group	Sbj	2nd	3rd	4th	5th	Rec
Control	99	29	28	9	8	243
Dementia	194	53	13	8	3	309
All	293	82	41	17	11	552

Table 10: Statistics of the control and the dementia groups in the DementiaBank. Sbj: # of subjects, i'th: # of subjects who made i'th visits, Rec: # of voice recordings. Note that subject with i'th visits take $(i-1)$ additional recordings; thus, Rec = Sbj + $\sum_{i=2}^{5}(i-1)$'th.

Table 9 shows the statistics of linguistic features in comparison to Table 1 in Section 3. The same tools, Temi and ELIT, are used to measure them. Unlike B-SHARP, the control group in the DementiaBank does not reveal a significantly greater number of tokens than the dementia group. The document size in the DementiaBank is 4.9 times smaller than B-SHARP on average. In both datasets, the noun and discourse counts are significantly different between the control and the other groups.

It is interesting that a significant difference is found in verbs whereas it is not the case for complex structures in the DementiaBank, which is opposite in B-SHARP. This may imply that the verb usage deteriorates as it progresses from MCI to dementia, but more thorough research is needed for further verification.