Domain-aware Intermediate Pretraining for Dementia Detection with Limited Data

Youxiang Zhu[1], *Xiaohui Liang*[1], *John A. Batsis*[2], *Robert M. Roth*[3]

[1]University of Massachusetts Boston
[2]University of North Carolina at Chapel Hill
[3]Dartmouth College

{Youxiang.Zhu001, Xiaohui.Liang}@umb.edu
John.Batsis@unc.edu, Robert.M.Roth@hitchcock.org

Abstract

Detecting dementia using human speech is promising but faces a limited data challenge. While recent research has shown general pretrained models (e.g., BERT) can be applied to improve dementia detection, the pretrained model can hardly be fine-tuned with the available small dementia dataset as that would raise the overfitting problem. In this paper, we propose a domain-aware intermediate pretraining to enable a pretraining process using a domain-similar dataset that is selected by incorporating the knowledge from the dementia dataset. Specifically, we use pseudo-perplexity to find an effective pretraining dataset, and then propose dataset-level and sample-level domain-aware intermediate pretraining techniques. We further employ information units (IU) from previous dementia research and define an IU-pseudo-perplexity to reduce calculation complexity. We confirm the effectiveness of perplexity by showing a strong correlation between perplexity and accuracy using 9 datasets and models from the GLUE benchmark. We show that our domain-aware intermediate pretraining improves detection accuracy in almost all cases. Our results suggested that the difference in text-based perplexity values between patients with Alzheimer's Disease (AD) and Healthy Control (HC) is still small, and the perplexity incorporating acoustic features (e.g., pause) may make the pretraining more effective.

Index Terms: Dementia, intermediate pretraining, perplexity

1. Introduction

Detecting dementia via spontaneous speech is faster and less costly compared to conventional cognitive assessment methods that require medical assistance or equipment. The Alzheimer's Dementia Recognition through Spontaneous Speech (ADReSS) challenge [1, 2] prepared spontaneous speech datasets from patients using a picture description task [3], and enable researchers to develop dementia detection models to classify patients with AD and HC or infer their Mini-Mental State Examination (MMSE) scores [4, 5]. Researchers have discovered that transfer learning can significantly enhance the accuracy of detection [6, 4, 5, 7], e.g., using BERT [8]. Transfer learning usually has two implementations: i) use of the output of the pretrained model as a fixed feature extractor, but cannot incorporate any knowledge from the downstream task into the pretrained model; or ii) fine-tune the pretrained model with the downstream dataset, but when the downstream dataset is small, we encounter the overfitting problem. For example, recent ADReSS datasets are small, containing 108 training samples and 48 testing samples. Recent research showed that the benefits from additional intermediate pretraining increase when having additional domain data or unlabeled data from the downstream task [9, 10, 11, 12, 13, 14, 15]. When these data are not available, researchers proposed to find additional domain-similar datasets for pretraining to improve the performance of the downstream task [9, 11]. They proposed to measure the domain similarity between the candidate dataset and the downstream dataset via qualitative or quantitative metrics. With the dataset, depending on the availability of labels, supervised or self-supervised intermediate pretraining can be applied [10].

In dementia research, finding a domain-similar dataset is a challenge. Available dementia datasets are small in size [16, 17], and finding a large speech dataset that is qualitatively relevant to dementia is challenge. In this paper, we aim to address a unique challenge, i.e., how could we implement intermediate pretraining without a large dementia-related dataset?

We first introduce a metric to evaluate the similarity between the candidate dataset and the dementia dataset, and then use the metric to select the candidate dataset for pretraining. Specifically, we calculate pseudo-perplexity [18] by inputting a dataset Y into a model pretrained with dataset X. We consider choosing more similar datasets X and Y will produce a small value of the above pseudo-perplexity. To confirm this, we exploit 9 pretrained models from the GLUE benchmark [19]. We demonstrate a strong correlation between the accuracy of dementia detection and the pseudo-perplexity generated by each pretrained model, showing pseudo-perplexity is an effective metric in selecting pretraining datasets for enhancing dementia detection. In addition, we can select multiple pretrained models using pseudo-perplexity and jointly incorporate them into dementia detection without incurring the overfitting problem.

We further propose a novel sample-level pretraining technique to monitor each pretraining step. Specifically, in each pretraining step, we decide to accept (or reject) the update by the samples depending on whether the conditions set on pseudo-perplexity are met (or not met). Our goal is to include the samples for pretraining that make the model better perform the downstream task. While calculating pseudo-perplexity in each step is time-consuming, we further incorporate the information units (IU) to define a new IU-based pseudo-perplexity, and the time of calculation is largely reduced. Information units have been long studied in dementia research as an effective linguistic feature [20]. Our contributions can be summarized as follows:

First, we introduce a new way of using pseudo-perplexity to evaluate the domain-similarity between a candidate dataset and the dementia dataset. We found a strong correlation between the pseudo-perplexity and the accuracy of dementia detection, thus enabling us to implement intermediate pretraining by the guidance of pseudo-perplexity in absence of dementia-related datasets.

Second, we define and evaluate a new IU-pseudo-perplexity metric to reduce the calculation complexity. This new metric is also shown to be more accurate than pseudo-perplexity because information units are known as an effective linguistic feature.

Third, we propose a new sample-level pretraining technique where the model decides to accept or reject the update in each pretraining step based on the change of the perplexity. This technique achieves the best performance in almost all cases.

2. Background

BERT [8] is pretrained with large-scale datasets Wikipedia and Bookcorpus using self-supervised learning. It consists of two parts: transformer encoders and classification layers. The transformer encoders take tokens as input and output the hidden representations for each token. The classification layers take hidden representations as inputs and produce the task-related outputs using self-supervised learning. When using fine-tuning, the classification layers are replaced with new ones and trained using supervised learning. When using the fixed feature extractor, hidden representations are used for training a new classifier.

GLUE [19] is well-known benchmark for natural language understanding. It consists of 9 datasets in various domains, including miscellaneous, movie reviews, news, social QA questions, Wikipedia and fiction books. In this paper, we will evaluate these pretraining datasets and models in dementia detection.

ADReSS [1] datasets were collected from AD and HC patients. The patients were required to verbally describe the Cookie Theft picture, and their speech were recorded and transcribed by a human. The datasets are labeled with AD and HC labels. We will use the transcripts from ADReSS 2020, and strictly follow the use of ADReSS training and testing datasets for calculating accuracy.

3. Perplexity-based domain similarity

Selecting datasets similar to a given downstream dataset is crucial for intermediate pretraining. Measuring the similarity of two datasets can use a qualitative method. For example, the emotion domain is relevant to sarcasm, and thus emotion detection model can be implemented as an intermediate pretraining for sarcasm detection model [11]. Available dementia datasets are small in size, and we hardly find a large speech dataset that is qualitatively relevant to dementia. Measuring the similarity of two datasets can also use a simple quantitative method, e.g., calculating the overlapped vocabulary [9]. The vocabulary-based method focuses on the vocabulary-level data content from the speech task but ignores the relation among words and other linguistic features, which may be specifically important to dementia detection. We aim to propose a new quantitative method to find large datasets that are similar to the given dementia dataset.

3.1. Pseudo-perplexity

Perplexity measures how well a language model models a dataset. Given a model and a dataset, a perplexity value can be calculated. Conventional perplexity fits autoregressive models running in a single-direction manner but does not fit BERT running in a bidirectional manner [21]. We thus exploit pseudo-perplexity specifically proposed for bidirectional BERT. In the calculation of the pseudo-perplexity, we use the [MASK] token of BERT to mask one token at each time and calculate the corresponding cross-entropy loss with the transformer encoders and self-supervised classification layers of BERT. Denote all tokens of a sample as $\boldsymbol{W}$, and denote $\boldsymbol{W}$ with t-th token masked as $\boldsymbol{W}_{\setminus t} = \left(\boldsymbol{w}_1, \cdots, \boldsymbol{w}_{t-1}, [\text{MASK}], \boldsymbol{w}_{t+1}, \cdots, \boldsymbol{w}_{|\boldsymbol{W}|}\right)$. The cross-entropy (CE) loss of each token in a sample are added to obtain a pseudo-loglikelihood (PLL) score, as shown below

$$\mathrm{PLL}(\boldsymbol{W}) := \sum_{\boldsymbol{w}_t \in \boldsymbol{W}} \mathrm{CE}\left(\boldsymbol{w}_t \mid \boldsymbol{W}_{\setminus t}\right) \quad (1)$$

To obtain the perplexity on a dataset $\mathbb{W}$, we calculate the pseudo-loglikelihood for samples in a dataset. Denote the total number of tokens as N, we have the pseudo-perplexity as

$$\mathrm{PPPL}(\mathbb{W}) := \exp\left(-\frac{1}{N} \sum_{\boldsymbol{W} \in \mathbb{W}} \mathrm{PLL}(\boldsymbol{W})\right) \quad (2)$$

3.2. IU-pseudo-perplexity

The calculation complexity of pseudo-perplexity is associated with the number of tokens in the dataset. Such calculation can be expensive if the tokens of a dataset are many. To reduce the complexity, we define a new IU-pseudo-perplexity based on information units. Information units have been used as an effective linguistic feature for dementia detection [22, 23]. Specifically, we used 35 information units from previous research [20], denoted by $\boldsymbol{I}$. Then, we mask only the tokens of IUs in the calculation. Denote the number of tokens of IUs as N', we have

$$\text{IU-PLL}(\boldsymbol{W}) := \sum_{\boldsymbol{w}_t \in \boldsymbol{W} \& \boldsymbol{w}_t \in \boldsymbol{I}} CE\left(\boldsymbol{w}_t \mid \boldsymbol{W}_{\setminus t}\right) \quad (3)$$

$$\text{IU-PPPL}(\mathbb{W}) := \exp\left(-\frac{1}{N'} \sum_{\boldsymbol{W} \in \mathbb{W}} \text{IU-PLL}(\boldsymbol{W})\right) \quad (4)$$

Models	PPPL	IU-PPPL	Accuracy
BERT	6.75	30.69	0.83
COLA	7.97	39.74	0.79
MNLI	11.65	81.18	0.75
MRPC	7.42	47.29	0.83
QNLI	8.88	65.99	0.81
QQP	12.21	88.63	0.73
RTE	7.12	34.41	0.83
SST2	8.70	44.78	0.85
STSB	20.19	186.18	0.63
WNLI	6.69	30.94	0.83

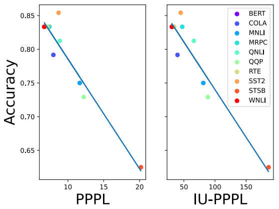

Figure 1: *Correlation between perplexity and accuracy*

3.3. Correlating perplexity with accuracy

We propose to use perplexity to measure the similarity of two datasets: if a model trained on dataset X produces smaller perplexity on dataset Y, then datasets X and Y are determined to be more similar. Our intuition is that using a more similar dataset for pretraining would result in a better performance in the downstream task. To confirm this, we exploited 9 different datasets from the GLUE benchmark and their corresponding models pretrained on BERT. For perplexity calculation, we used fine-tuned transformer encoder and original self-supervised classification layers. For accuracy calculation, we used fine-tuned transformer encoder as a fixed feature extractor to extract features from ADReSS dataset, then trained a Support Vector Machine (SVM) using ADReSS training dataset, and generated accuracy using ADReSS testing dataset. We report PPPL, IU-PPPL and accuracy in Figure 1. We found a strong correlation between PPPL and accuracy and between IU-PPPL and accuracy. We fit regression lines with perplexity and accuracy and obtain small Mean Square Error (MSE) of 0.00367 and 0.00446

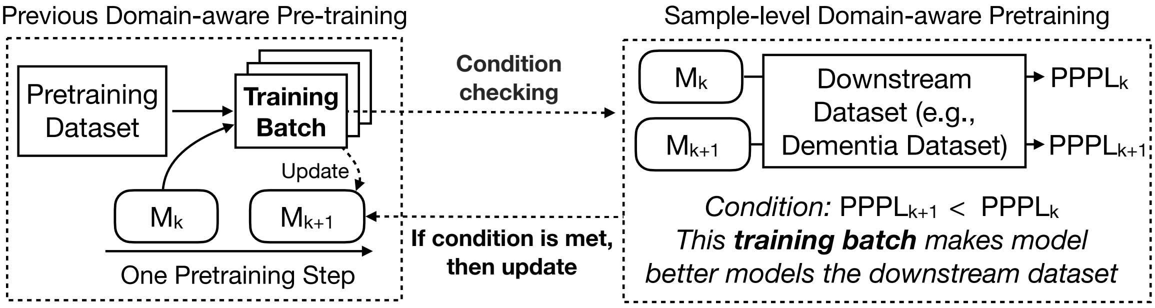
Previous Domain-aware Pre-training
Pretraining Dataset
Training Batch
Update
Mk
Mk+1
One Pretraining Step
Condition checking
If condition is met, then update
Sample-level Domain-aware Pretraining
Mk
Mk+1
Downstream Dataset (e.g., Dementia Dataset)
PPPLk
PPPLk+1
Condition: PPPLk+1 < PPPLk
This training batch makes model better models the downstream dataset

Figure 2: *Proposed sample-level domain-aware intermediate pretraining (condition 1)*

for PPPL and IU-PPPL, respectively. In terms of calculation complexity, one-time calculation of PPPL and IU-PPPL took 94.41s and 11.33s, respectively, using a single V100 GPU.

4. Domain-aware intermediate pretaining

In this section, we propose dataset-level and sample-level domain-aware intermediate pretraining techniques.

4.1. Dataset-level domain-aware intermediate pretraining

We propose to select multiple datasets that are determined similar to a given downstream task to implement an intermediate pretraining. Specifically, we concatenate the hidden representations from different models pretrained with the selected datasets and then feed the joint representation into a SVM built for a downstream task. Note that, there are many datasets and pretrained models, and enumerating all combinations is infeasible. In addition, using testing accuracy to select combinations will overfit the testing set. Our approach aims to select pretraining model based on the ranking of the perplexity. Pretrained models with smaller perplexity are considered to produce better performance. Such perplexity-based selection neither relies on accuracy nor labels of training dataset and would be more beneficial when more unlabeled downstream data is available.

4.2. Sample-level domain-aware intermediate pretraining

In the intermediate pretraining process, not all samples in the pretraining dataset can be used to help the model better model the downstream dataset. We further propose a sample-level domain-aware intermediate pretraining technique to control the pretraining process at a fine-grained level. As shown in Figure 2, in each pretraining step from M_k to M_{k+1}, we calculate the perplexity values PPPL$_k$ and PPPL$_{k+1}$ by inputting the downstream dataset (i.e., dementia dataset D) to the models before and after the update. If the perplexity value decreases after the update, i.e., PPPL$_{k+1}$ $<$ PPPL$_k$, we consider the current batch of samples helps the model to better model the dementia dataset and thus accept the update. Otherwise, we reject the update and move on with the next batch of samples.

ADReSS dataset consists of speech data from patients with AD and HC. We denote the datasets of AD and HC patients as D_{AD} and D_{HC}, where $D_{AD} + D_{HC} = D$. Previous research has exploited perplexity as a feature to design dementia detection [24, 25, 26]. Thus, we consider the perplexity for D_{AD} and D_{HC} may exhibit a difference that can be used for dementia detection. Denote PPPL$_{k,AD}$ and PPPL$_{k,HC}$ as the perplexity values generated by the model at step k with the input of D_{AD} and D_{HC}, respectively. We set five conditions to control the pretraining process.

Condition 1. PPPL$_{k+1}$ $<$ PPPL$_k$: the model better models the whole dementia dataset.

Condition 2. PPPL$_{k+1,HC}$ $<$ PPPL$_{k,HC}$: the model better models the dataset of HC patients.

Condition 3. PPPL$_{k+1,AD}$ $<$ PPPL$_{k,AD}$: the model better models the dataset of AD patients.

Condition 4. PPPL$_{k+1,AD}$ - PPPL$_{k+1,HC}$ $<$ PPPL$_{k,AD}$ - PPPL$_{k,HC}$: the model have larger perplexity difference on the datasets of AD and HC patients after the update.

Condition 5. PPPL$_{k+1,HC}$ $<$ PPPL$_{k,HC}$ and PPPL$_{k+1,AD}$ $>$ PPPL$_{k,AD}$: the model better models the dataset of HC patients, but is more confused with the dataset of AD patients.

The sample-level domain-aware intermediate pretraining with these conditions can ensure the model produces the desired perplexity results on the downstream task.

5. Experiments

In this section, we introduce the implementation details and results from our proposed pretraining techniques.

5.1. Implementation details

We implemented our models with PyTorch and Hugging Face Transformers. For fine-tuning BERT with the GLUE datasets, we used epoch 3 and batch size 32. We averaged BERT hidden representations through time dimension and input the result to SVM with RBF kernel. The calculation of perplexity and accuracy can be referred to in Section 3.3.

5.2. Results from dataset-level pretraining

Table 2 shows the results from the dataset-level pretraining. To select the domain-similar models, we rank the pretrained models by pseudo-perplexity and IU-pseudo-perplexity from lowest to highest. We combine the representations of two models or three models. For the two-model combination case, we selected any two from the top three. For the three-model combination, we selected any three from the top four. The accuracy result is generated from a SVM model trained and tested using ADReSS datasets. We have the following observations: i) The combined models improve the accuracy; for all of these top perplexity combinations, the accuracy is larger or equal to the worst performance of any single model in the combination. ii) Perplexity is an effective metric to select combinations for strong performance. For example, the baseline accuracy for RTE and MRPC are both 83%, and they are ranked 2 and 3 by pseudo-perplexity. The combination of these two models boosts the accuracy to 85%. Using 1st, 2nd, and 4th models ranked by IU-pseudo-

Table 1: *Results from sample-level pretraining using IU-PPPL*

	Baseline			Condition 1: ALL				Condition 2: HC only			Condition 3: AD only			Condition 4: HC-AD			Condition 5: HC, AD		
Model	HC	AD	Acc	ALL	HC	AD	Acc	HC	AD	Acc	HC	AD	Acc	HC	AD	Acc	HC	AD	Acc
COLA	35.07	45.04	0.79	28.69	27.38	30.06	0.79	26.25	30.08	0.83	27.56	29.81	0.81	28.76	33.67	0.79	29.46	31.96	0.83
MNLI	77.19	85.37	0.75	26.57	26.10	27.05	0.79	22.73	25.91	0.77	24.91	25.66	0.77	29.67	32.22	0.83	27.29	35.57	0.79
MRPC	42.83	52.22	0.83	30.69	29.46	31.96	0.83	28.94	31.41	0.83	29.72	31.52	0.83	29.68	32.19	0.83	29.46	31.96	0.83
QNLI	64.20	67.83	0.81	29.74	28.45	31.10	0.81	28.04	29.60	0.81	29.24	31.71	0.83	46.06	54.58	0.79	28.32	33.82	0.81
QQP	82.17	95.61	0.73	28.39	27.84	28.95	0.81	22.29	23.69	0.81	30.41	31.38	0.81	29.78	34.17	0.79	28.37	33.58	0.79
RTE	33.91	34.92	0.83	29.27	27.80	30.82	0.81	28.71	31.91	0.83	25.92	27.33	0.79	29.48	32.00	0.83	29.09	32.42	0.83
SST2	40.87	49.07	0.85	26.51	25.26	27.82	0.85	27.32	30.33	0.85	29.46	31.96	0.83	37.81	44.07	0.83	29.09	32.93	0.83
STSB	147.70	234.68	0.63	30.68	29.28	32.14	0.83	29.46	31.96	0.83	29.75	31.84	0.83	29.46	31.96	0.83	29.46	31.96	0.83
WNLI	29.98	31.94	0.83	30.63	29.59	31.70	0.83	29.44	31.65	0.83	29.46	31.96	0.83	29.46	31.96	0.83	29.46	31.96	0.83

Table 2: *Results from dataset-level pretraining*

Models	IU-PPPL rankings	PPPL rankings	Accuracy
WNLI (0.83), RTE (0.83)	1 2	1 2	0.83
WNLI (0.83), COLA (0.79)	1 3	1 4	0.79
RTE (0.83), COLA (0.79)	2 3	2 4	0.81
WNLI (0.83), MRPC (0.83)	1 5	1 3	0.83
RTE (0.83), MRPC (0.83)	1 5	2 3	**0.85**
WNLI (0.83), RTE (0.83), COLA (0.79)	1 2 3	1 2 4	0.83
WNLI (0.83), RTE (0.83), SST2 (0.85)	1 2 4	1 2 5	**0.87**
WNLI (0.83), COLA (0.79), SST2 (0.85)	1 3 4	1 4 5	0.83
RTE (0.83), COLA (0.79), SST2 (0.85)	2 3 4	2 4 5	0.85
WNLI (0.83), RTE (0.83), MRPC (0.83)	1 2 5	1 2 3	0.83
WNLI (0.83), MRPC (0.83), COLA (0.79)	1 5 3	1 3 4	0.79
RTE (0.83), MRPC (0.83), COLA (0.79)	2 5 3	2 3 4	0.81

perplexity, the combined model achieved the highest accuracy of 87%. iii) IU-pseudo-perplexity is slightly more effective than pseudo-perplexity. We averaged the ranking numbers for those combinations with accuracy $\geq 83\%$ (BERT baseline). The average ranking number is 2.52 for IU-pseudo-perplexity, slightly better than 2.57 for pseudo-perplexity.

5.3. Results from sample-level pretraining

In Table 1, the baseline results show the perplexity values of these pretrained models on the dataset from HC is smaller than those corresponding to AD. It means that the baseline intermediate pretrained models are capable of recognizing the difference between AD and HC. Now we examine the results from sample-level pretraining. i) Looking at conditions 1 and 2, we found that sample-level pretraining on ALL or HC improves or maintains the accuracy. We count 2 conditions and 9 models for a total of 18 cases. Compared to the baseline, sample-level pretraining improves or maintains the accuracy in 17 of 18 cases. ii) Looking at conditions 2 and 3, we found that lowering the perplexity for either AD dataset or HC dataset also lowers the perplexity for the other group. This shows that the perplexity values of AD dataset and HC dataset are correlated. iii) Looking at conditions 4 and 5, our goal is to diversify the perplexity of AD and HC datasets. However, in condition 4, the model produces a larger perplexity of HC datasets, and thus the detection accuracy is not significantly increased. In condition 5, our intermediate pretraining technique hardly enlarges the difference in perplexity of the two groups, and the perplexity values are similar to conditions 1-3. In this case, most updates were rejected.

6. Discussion

6.1. Diversify perplexity for different labels

The significant difference of perplexity in datasets with different labels can be used to implement accurate classifications, e.g., fact checking [21]. In dementia research, our dataset-level and sample-level pretraining produce different perplexity values for AD and HC. In general, perplexity is lower for HC than for AD, which is intuitive because pretraining datasets are more similar to datasets from HC. However, the difference between AD and HC is small, and thus the improvement by pretraining is limited. One possible reason is that the dataset is generated using a picture description task, so that the AD and HC datasets are domain-similar, and much different from the pretraining datasets. Our current definition of perplexity focuses on the transcripts, which are highly related to the content. To make the proposed pretraining more effective, we envision a new representation to make AD and HC datasets more domain-different. For example, we will study pause representations [4, 27] and generate large-scale text with pause representations from speech [27]. The new perplexity incorporating pause may be diversified between AD and HC datasets, thus improving the corresponding pretraining.

6.2. Other use cases

Domain-aware intermediate pretraining introduces a third way of implementing transfer learning other than using the fixed feature extractor and fine-tuning. We envision our proposed technique can be widely applied to downstream tasks with limited available datasets or limited labeled datasets. For example, in the medical domain, consistent data collection and labeling are expensive, and these datasets are usually much smaller than pretraining datasets. Our proposed techniques can incorporate knowledge from the downstream task into the pretraining process even with a small amount of data and even without any labels. We envision such a new implementation of transfer learning enables wide applicability to tasks with limited data and with no available in-domain or domain-similar datasets.

7. Conclusions

In this paper, we proposed domain-aware intermediate pretraining to address the weakness of the current implementation of transfer learning in dementia detection. We first introduced a pseudo-perplexity to measure the domain similarity and showed its high correlation with accuracy. We further define a new pseudo-perplexity using information units, a proven effective linguistic feature in dementia detection, which has a reduced calculation complexity. Based on the perplexity, we proposed dataset-level and sample-level domain-aware intermediate pretraining: the dataset-level selects the pretraining model using the ranking of perplexity; the sample-level selects samples in each pretraining step to ensure expected perplexity. We demonstrated that domain-aware intermediate pretraining outperformed conventional pretraining and intermediate pretraining. We envision such a new implementation of transfer learning can apply to other downstream tasks with limited data.

8. Acknowledgements

This research is funded by the US National Institutes of Health National Institute on Aging, under grant No. 1R01AG067416.

9. References

[1] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, "Alzheimer's dementia recognition through spontaneous speech: The adress challenge," *arXiv preprint arXiv:2004.06833*, 2020.

[2] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and MacWhinney, "Detecting cognitive decline using speech only: The adresso challenge," *arXiv preprint arXiv:2104.09356*, 2021.

[3] H. Goodglass, E. Kaplan, and S. Weintraub, *BDAE: The Boston Diagnostic Aphasia Examination*. Lippincott Williams & Wilkins Philadelphia, PA, 2001.

[4] J. Yuan, Y. Bian, X. Cai, J. Huang, Z. Ye, and K. Church, "Disfluencies and fine-tuning pre-trained language models for detection of alzheimer's disease." in *INTERSPEECH*, 2020, pp. 2162–2166.

[5] J. Yuan, X. Cai, Y. Bian, Z. Ye, and K. Church, "Pauses for detection of alzheimer's disease," *Frontiers in Computer Science*, vol. 2, p. 57, 2020.

[6] A. Balagopalan, B. Eyre, F. Rudzicz, and J. Novikova, "To bert or not to bert: comparing speech and language-based approaches for alzheimer's disease detection," *arXiv preprint arXiv:2008.01551*, 2020.

[7] Y. Zhu, X. Liang, J. A. Batsis, and R. M. Roth, "Exploring deep transfer learning techniques for alzheimer's dementia detection," *Frontiers in computer science*, vol. 3, 2021.

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[9] S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, and N. A. Smith, "Don't stop pretraining: adapt language models to domains and tasks," *arXiv preprint arXiv:2004.10964*, 2020.

[10] J. Phang, I. Calixto, P. M. Htut, Y. Pruksachatkun, H. Liu, C. Vania, K. Kann, and S. R. Bowman, "English intermediate-task training improves zero-shot cross-lingual transfer too," *arXiv preprint arXiv:2005.13013*, 2020.

[11] E. Savini and C. Caragea, "Intermediate-task transfer learning with bert for sarcasm detection," *Mathematics*, vol. 10, no. 5, p. 844, 2022.

[12] T. Vu, T. Wang, T. Munkhdalai, A. Sordoni, A. Trischler, A. Mattarella-Micke, S. Maji, and M. Iyyer, "Exploring and predicting transferability across nlp tasks," *arXiv preprint arXiv:2005.00770*, 2020.

[13] Y. Pruksachatkun, J. Phang, H. Liu, P. M. Htut, X. Zhang, R. Y. Pang, C. Vania, K. Kann, and S. R. Bowman, "Intermediate-task transfer learning with pretrained models for natural language understanding: When and why does it work?" *arXiv preprint arXiv:2005.00628*, 2020.

[14] T.-Y. Chang and C.-J. Lu, "Rethinking why intermediate-task fine-tuning works," *arXiv preprint arXiv:2108.11696*, 2021.

[15] C. Poth, J. Pfeiffer, A. Rücklé, and I. Gurevych, "What to pre-train on? efficient intermediate task selection," *arXiv preprint arXiv:2104.08247*, 2021.

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

[17] J. T. Becker, F. Boiler, O. L. Lopez, J. Saxton, and K. L. McGonigle, "The natural history of alzheimer's disease: description of study cohort and accuracy of diagnosis," *Archives of neurology*, vol. 51, no. 6, pp. 585–594, 1994.

[18] J. Salazar, D. Liang, T. Q. Nguyen, and K. Kirchhoff, "Masked language model scoring," *arXiv preprint arXiv:1910.14659*, 2019.

[19] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, "Glue: A multi-task benchmark and analysis platform for natural language understanding," *arXiv preprint arXiv:1804.07461*, 2018.

[20] M. Yancheva and F. Rudzicz, "Vector-space topic models for detecting alzheimer's disease," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2016, pp. 2337–2346.

[21] N. Lee, Y. Bang, A. Madotto, M. Khabsa, and P. Fung, "Towards few-shot fact-checking via perplexity," *arXiv preprint arXiv:2103.09535*, 2021.

[22] B. Croisile, B. Ska, M.-J. Brabant, A. Duchene, Y. Lepage, G. Aimard, and M. Trillet, "Comparative study of oral and written picture description in patients with alzheimer's disease," *Brain and language*, vol. 53, no. 1, pp. 1–19, 1996.

[23] E. Giles, K. Patterson, and J. R. Hodges, "Performance on the boston cookie theft picture description task in patients with early dementia of the alzheimer's type: missing information," *Aphasiology*, vol. 10, no. 4, pp. 395–408, 1996.

[24] Z. Guo, Z. Ling, and Y. Li, "Detecting alzheimer's disease from continuous speech using language models," *Journal of Alzheimer's Disease*, vol. 70, no. 4, pp. 1163–1174, 2019.

[25] N. Linz, J. Tröger, H. Lindsay, A. Konig, P. Robert, J. Peter, and J. Alexandersson, "Language modelling for the clinical semantic verbal fluency task," in *LREC 2018 Workshop RaPID-2: Resources and ProcessIng of Linguistic, Para-Linguistic and Extra-Linguistic Data from People with Various Forms of Cognitive/Psychiatric Impairments*, 2018.

[26] C. Frankenberg, J. Weiner, T. Schultz, M. Knebel, C. Degen, H.-W. Wahl, and J. Schroeder, "Perplexity–a new predictor of cognitive changes in spoken language?–results of the interdisciplinary longitudinal study on adult development and aging (ilse)," *Linguistics Vanguard*, vol. 5, no. s2, 2019.

[27] Y. Zhu, A. Obyat, X. Liang, J. A. Batsis, and R. M. Roth, "Wavbert: Exploiting semantic and non-semantic speech using wav2vec and bert for dementia detection," *Proc. Interspeech 2021*, pp. 3790–3794, 2021.