Text Classification by Contrastive Learning and Cross-lingual Data Augmentation for Alzheimer's Disease Detection

Zhiqiang Guo[1], **Zhaoci Liu**[1], **Zhen-Hua Ling**[1]*, **Shijin Wang**[2], **Lingjing Jin**[3], **Yunxia Li**[3]
[1]National Engineering Laboratory for Speech and Language Information Processing,
University of Science and Technology of China, Hefei, China
[2]iFLYTEK Research, iFLYTEK Co., Ltd., Hefei, China
[3]Shanghai Tongji Hospital, Tongji University School of Medicine, Shanghai, China
{gzq,zcliu8}@mail.ustc.edu.cn, zhling@ustc.edu.cn,
sjwang3@iflytek.com ,{lingjingjin,doctorliyunxia}@163.com

Abstract

Data scarcity is always a constraint on analyzing speech transcriptions for automatic Alzheimer's disease (AD) detection, especially when the subjects are non-English speakers. To deal with this issue, this paper first proposes a contrastive learning method to obtain effective representations for text classification based on monolingual embeddings of BERT. Furthermore, a cross-lingual data augmentation method is designed by building autoencoders to learn the text representations shared by both languages. Experiments on a Mandarin AD corpus show that the contrastive learning method can achieve better detection accuracy than conventional CNN-based and BERT-based methods. Our cross-lingual data augmentation method also outperforms other compared methods when using another English AD corpus for augmentation. Finally, a best detection accuracy of 81.6% is obtained by our proposed methods on the Mandarin AD corpus.

1 Introduction

Alzheimer's disease (AD) is a worldwide popular neurodegenerative disease and the major cause of dementia. It is estimated that 32.8 million people worldwide living with AD in 2015, resulting in economic costs of 604 billion US dollars, and the number of patients is expected to almost double every 20 years (Prince et al., 2016). With no cure has been found yet, in order to allow effective intervention in early stages to delay onset of the disease, early detection of AD is crucial (Dubois et al., 2016). Nonetheless, nowadays common methods of AD diagnosis, including cognitive tests and biomarker detection, are too time-consuming and expensive to cover all potential patients. Thus, developing an effective and convenient method for early AD detection becomes a valuable research topic.

Language impairment generally appears at the early stages of AD (Morris, 1996). Suffering the disease, patients have significant worse performance than controls on a standard aphasia test, specifically in the areas of verbal expression, auditory comprehension, repetition, reading, and writing (Murdoch et al., 1987). The performances of AD patients on some language tasks, such as picture description and sentence repetition, are distinctive from healthy people, which give us a hint to utilize these tasks to detect AD. Other advantages of these language tasks include their simple procedure and quick process. One usually completes the picture description task within 5 minutes and needs few instructions.

A lot of researches have applied machine learning approaches to automatic AD detection based on language tasks. Fraser et al. (2016) extracted total 370 features considering part-of-speech, syntax, acoustics, as well as other aspects of linguistics, and obtained a best average accuracy of 82% for binary classification using the 35 top-ranked features on the DementiaBank corpus. Wanker et al. (2017) designed a single feature by building two trigram language models from the speech transcriptions of AD patients and controls respectively, and resulted in an EER of 77%. Several later studies (Fraser et al., 2019; Fritsch et al., 2019; Guo et al., 2019) continued to show the effectiveness of language model features.

AD detection based on speech transcriptions is essentially a text classification task. Various models have been proposed to solve general text classification tasks, such as convolutional neural networks

*Corresponding author: Zhen-Hua Ling.

Proceedings of the 28th International Conference on Computational Linguistics, pages 6161–6171
Barcelona, Spain (Online), December 8-13, 2020

(Kim, 2014), recurrent neural networks (Liu et al., 2016), and hierarchical attention networks (Yang et al., 2016). Recently, the advent of large-scale self-supervised pretraining has played a central role in the progress of natural language processing (NLP) research (Devlin et al., 2019; Liu et al., 2019a; Yang et al., 2019). In particular, BERT (Devlin et al., 2019) greatly improved the performance of several general NLP benchmarks such as GLUE (Wang et al., 2018) and SuperGLUE (Wang et al., 2019). Fine-tuning BERT is a straightforward and powerful solution to text classification. However, in the AD detection case, it faces a major challenge, which is data scarcity.

Data scarcity is always an issue of building automatic AD detection models, and the difficulties in recruiting eligible subjects make it hard to overcome. It constrains model capacity as powerful models usually have a large number of trainable parameters and tend to overfit small datasets. The DementiaBank corpus (Becker et al., 1994) contains only 256 cases labeled as "possible" or "probable" AD, which is still larger than most of the similar datasets in other language, such as Japanese (Shibata et al., 2016), Portuguese (Aluísio et al., 2016), and Mandarin Chinese (Lai et al., 2009). The Mandarin corpus used in this study contains 104 samples of AD patients, which is also much smaller than DementiaBank.

In order to deal with the issue of data scarcity, this paper designs a contrastive learning method to learn effective representations for text classification. First, we rely on BERT extracting contextual word embeddings for every sample in the corpus. Then, a weighted sum of word embeddings, i.e. a paragraph vector, is obtained by contrastive learning and finally a similarity-based decision is made for classifying test samples. Since embedding features of samples are fixed as input and there are no model parameters in the similarity-based decision, this method is expected to ease the problem of data scarcity. Additionally, an autoencoder-based cross-lingual data augmentation method is proposed to augment the Mandarin corpus with a larger English one. Since merging the BERT-based paragraph vectors of both languages directly to build a classifier is unsuitable, we train an autoencoder-based model to obtain the latent representations shared by both languages, where a large general-domain parallel corpus is utilized. Then, the encoded paragraph vectors of both languages can be combined to estimate the classifier for Mandarin transcriptions.

Our contributions in this paper are twofold. First, a contrastive learning method is designed which studies effective representations for AD detection based on BERT embeddings. Experimental results show that this method achieves better detection accuracy than conventional CNN-based and BERT-based methods by 3.9% at least on our Mandarin corpus. Second, an autoencoder-based method is further proposed to augment original Mandarin corpus with a larger English one, which outperforms other compared data augmentation methods by 1.4% at least in our experiments.

2 Related Work

2.1 Fine-tuning Pretrained Language Models on Small Datasets

Typically, pretrained language models are first trained on large corpora, and then fine-tuned on downstream tasks by taking the model parameters as a starting point, while adding one task-specific layer learned from scratch. Despite its simplicity and ubiquity in modern NLP, this process has been shown to be brittle (Devlin et al., 2019; Phang et al., 2018; Zhu et al., 2019), where fine-tuning performance can vary substantially across different training episodes, even with fixed hyperparameter values. Dodge et al. (2020) gave a systematic study of this phenomenon and we also observed the brittleness in fine-tuning BERT on our AD corpus. We suspect the brittleness is due to the small size of our AD corpus, especially considering the large amount of parameters of BERT. So instead of directly fine-tuning BERT, the feature-based approach (Devlin et al., 2019) is followed in this paper, i.e. we take BERT as a feature extractor to get fixed embedding features, and then build models on top of it.

2.2 Autoencoders for Learning Bilingual Representations

Autoencoder is an unsupervised neural network that learns how to efficiently compress and encode data by measuring the loss of reconstructing the original data from the encoded representations (Baldi, 2012). Chandar AP et al. (2014) and Lauly et al. (2014) both applied autoencoders to learn bilingual word representations. They showed that by simply learning to reconstruct the bag-of-words representations

	Mandarin		English (DB)	
	HC	AD	HC	AD
Number	104	104	242	256
Gender(F/M)	60/44	54/50	155/87	159/97
Age	67.6(9.3)	74.2(9.6)	64.8(7.7)	72.9(8.1)
Education	11.6(3.2)	10.3(4.0)	14.2(2.6)	12.4(3.2)
MMSE	28.4(1.7)	18.1(5.9)	29.1(1.1)	18.8(5.9)

Table 1: Demographics of participants in our Mandarin corpus and the English DementiaBank (DB) corpus, where MMSE stands for the score of mini-mental state examination.

of aligned sentences within and between languages, high-quality representations can be learnt without word alignments. Our work is inspired by them. The differences are that our inputs of autoencoders are the weighted sums of pretrained word embeddings, i.e., paragraph vectors, and our method aims to learn a function that can map paragraph vectors of two languages into a common space, while maintain their representation capabilities. Bag-of-words training is also not used in our work.

2.3 Utilizing Cross-lingual Data for AD Detection

Some studies have already worked on utilizing cross-lingual data for AD detection. Fraser et al. (2019) used information units to transform cross-lingual transcriptions into the same domain, and then introduced a new set of concept-based language model features. A best AUC of 0.89 was achieved by using a multilingual training set. Li et al. (2019) proposed a method to learn a correspondence between independently engineered lexicosyntactic features in two language, which was based on linear regression and required a large parallel corpus of out-of-domain data. Their method outperformed both monolingual and machine translation-based baselines.

Different from their studies, we utilize BERT to get monolingual representations instead of designing hand-designed features. Besides, we propose an autoencoder-based model to learn bilingual paragraph vectors, while Fraser et al. (2019) chose to transform raw data and Li et al. (2019) adopted linear regression to find the map function between two sets of features.

3 Datasets

The Mandarin corpus used in this paper was collected at the Shanghai Tongji Hospital (Liu et al., 2019b). All participants were with the complaint of memory impairment and underwent a comprehensive neuropsychological battery, including the mini-mental state examination (Tombaugh and McIntyre, 1992). Diagnosis was made according to the core clinical criteria to dementia of NIA-AA established in 2011 (Sperling et al., 2011). The corpus contains the transcriptions of the Mandarin speech from 208 participants on the *Cookie Theft* picture description task, 104 of them were healthy controls and the other 104 were diagnosed as AD patients. This task asked participants to say whatever happened in the *Cookie Theft* picture as much as possible, and allowed encouragement from interviewers when participants had difficulties. Speech recordings were manually transcribed following the CHAT (Codes for the Human Analysis of Transcripts) protocol (MacWhinney, 2000), same as DementiaBank. The average length of dialogue samples in the this corpus is 127.4 characters with a standard deviation of 61.7 characters. Demographics of participants in the corpus are shown in Table 1.

The English DementiaBank (DB) corpus was used for data augmentation in our experiments. The DB corpus contains 498 samples from 292 participants taking the *Cookie Theft* picture description task, 242 of these samples were labeled as healthy controls and the other 256 were labeled as "possible" or "probable" AD, which we refer as an AD group. Each sample has an audio recording together with its corresponding transcription. Only the transcriptions are used in this study. The average length of dialogue samples in the corpus is 122.7 words with a standard deviation is 66.9 words. Demographics of participants in this corpus are also shown in Table 1.

In our proposed cross-lingual data augmentation method, a parallel Mandarin-English corpus is nec-

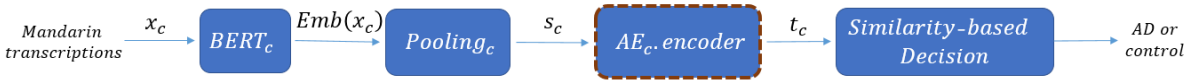

Figure 1: The classification procedures of our proposed methods. The $AE_c.encoder$ module is conducted only when our autoencoder-based cross-lingual data augmentation method is applied.

essary to train the autoencoders for deriving the paragraph representations shared by both languages. The OpenSubtitle corpus (Lison and Tiedemann, 2016) was adopted as the parallel corpus in our implementation. We utilized the Mandarin-English subset of this corpus, which contains 9.9 million lines of dialogue.

4 Methods

4.1 Contrastive Learning for Text Classification

The procedures of our classification model for AD detection is shown in Figure 1. Let x_c denote a sample of the Mandarin corpus. We first use the Chinese BERT model to extract embedding features of x_c and obtain $\mathbf{Emb}(x_c) \in R^{n \times d_{emb}}$, where n is the sequence length of x_c and d_{emb} is the dimension of embeddings. Then, a pooling layer is applied which converts $\mathbf{Emb}(x_c)$ into a single vector $\boldsymbol{s}_c \in R^{1 \times d_{emb}}$, i.e., the paragraph vector for representing this sample. Specifically, this pooling layer employs a trainable memory vector $\boldsymbol{m}_c \in R^{d_{emb} \times 1}$ and calculate $\boldsymbol{s}_c$ as

$$\boldsymbol{w}_c = softmax(\mathbf{Emb}(x_c)\boldsymbol{m}_c), \tag{1}$$

$$\boldsymbol{s}_c = \boldsymbol{w}_c^\top \mathbf{Emb}(x_c). \tag{2}$$

Afterwards, a similarity-based decision is made based on the pooling output to predict the AD or control label of the sample. We calculate the mean cosine similarity between $\boldsymbol{s}_c$ and the paragraph vectors of all AD instances in the training set. The mean cosine similarity between $\boldsymbol{s}_c$ and the paragraph vectors of all control instances in the training set is also calculated. Finally, the sample x_c is classified into the class with higher mean similarity. Since there are no model parameters in the similarity-based decision and BERT is not fine-tuned in our classification model, only the memory vector $\boldsymbol{m}_c$ needs to be estimated at the training stage.

In order to train the memory vector $\boldsymbol{m}_c$ for obtaining powerful paragraph vector representations, we introduce **contrastive regularization**. First, N_1 AD instances $\{x_i^{ad}\}_{i=1}^{N_1}$ are sampled from the training set X_c and a pooled paragraph vector $\boldsymbol{s}_i^{ad}$ is obtained for each of them. Then, for each x_i^{ad}, N_2 AD instances $\{x_{i,j}^{ad}\}_{j=1}^{N_2}$ and N_2 control instances $\{x_{i,j}^{ct}\}_{j=1}^{N_2}$ are further randomly selected from X_c. Their paragraph vectors $\boldsymbol{s}_{i,j}^{ad}$ and $\boldsymbol{s}_{i,j}^{ct}$ are also calculated. Under contrastive regularization, $\boldsymbol{s}_i^{ad}$ is expected to be similar with all $\boldsymbol{s}_{i,j}^{ad}$ and dissimilar to $\boldsymbol{s}_{i,j}^{ct}$. Therefore, the memory vector $\boldsymbol{m}_c$ is estimated by minimizing

$$CR_c = \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} \max(Sim(\boldsymbol{s}_i^{ad}, \boldsymbol{s}_{i,j}^{ct}) - Sim(\boldsymbol{s}_i^{ad}, \boldsymbol{s}_{i,j}^{ad}), 0), \tag{3}$$

where $Sim(\cdot, \cdot)$ calculates the cosine similarity between two vectors.

As shown in Figure 1, once our autoencoder-based cross-lingual data augmentation method is applied, a separate encoder module is converting the paragraph vector $\boldsymbol{s}_c$ extracted by monolingual BERT into a representation space shared by both languages. A similarity-based decision is then made by the input of bilingual representation $\boldsymbol{t}_c$, with the aid of English AD data. Details of the encoder will be introduced in the next subsection.

4.2 Autoencoder-based Cross-lingual Data Augmentation

The aim of our cross-lingual data augmentation method is to improve the performance of the AD detection method introduced in Section 4.1 on the Mandarin AD corpus by utilizing another English AD

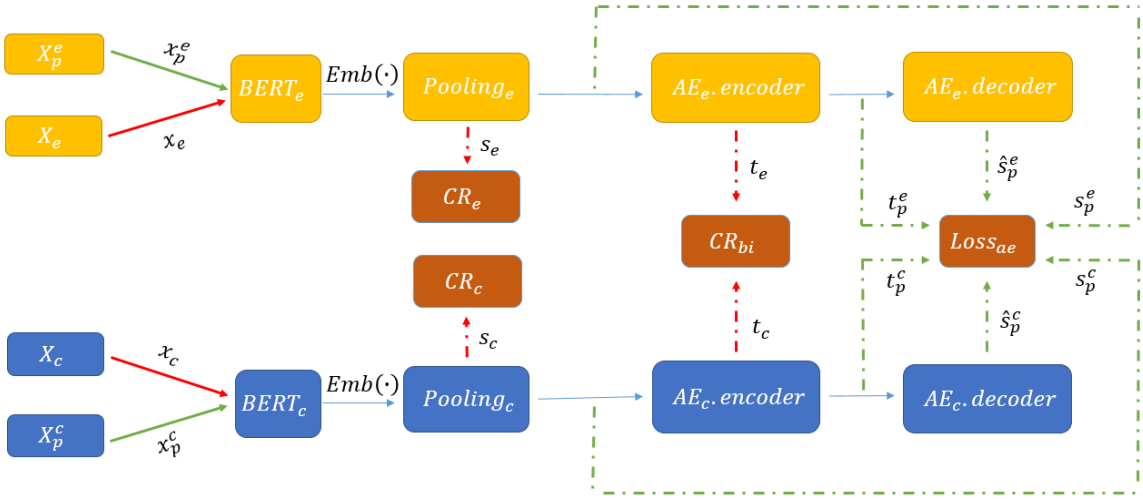

Figure 2: The training procedure of our autoencoder-based cross-lingual data augmentation method. X_e and X_c denote AD datasets in English and Mandarin, respectively. X_p^e and X_p^c stand for the English part and the Chinese part of the parallel corpus X_p. CR is short for contrastive regularizer.

corpus. It is achieved by training a unified model containing two trainable pooling layers $Pooling_e$ and $Pooling_c$ and two autoencoders AE_e and AE_c for English and Mandarin respectively, as showed in Figure 2. The inputs of each autoencoder are the paragraph vectors calculated by monolingual BERT and the trainable pooling layer, which has the same structure as the one introduced in Section 4.1. The two autoencoders are expected to convert the language-dependent paragraph vectors into a latent representation space shared by both languages. Therefore, another corpus containing parallel Mandarin-English sentences is also necessary when training the model.

Let X_p denote the parallel Mandarin-English corpus. For a pair of parallel samples $(x_p^e, x_p^c) \in X_p$, they are first transformed into paragraph vectors $\boldsymbol{s}_p^e$ and $\boldsymbol{s}_p^c$ after BERT and pooling operations. Then, an autoencoder loss for training the unified model is defined as

$$loss_{ae} = ||\boldsymbol{s}_p^e - \hat{\boldsymbol{s}}_p^e|| + ||\boldsymbol{s}_p^c - \hat{\boldsymbol{s}}_p^c|| - \lambda_1 Sim(\boldsymbol{t}_p^e, \boldsymbol{t}_p^c), \tag{4}$$

where $\boldsymbol{t}_p^e$ and $\boldsymbol{t}_p^c$ are encoded representations of $\boldsymbol{s}_p^e$ and $\boldsymbol{s}_p^c$ respectively, and $\hat{\boldsymbol{s}}_p^e$ and $\hat{\boldsymbol{s}}_p^c$ are reconstructed representations of $\boldsymbol{s}_p^e$ and $\boldsymbol{s}_p^c$ respectively. The first two terms in Eq. (4) are conventional autoencoder losses, while the third term with weight λ_1 is a similarity measure that helps model to learn the representations shared by both languages. Here, the cosine similarity is also used for the $Sim(\cdot, \cdot)$ function.

In addition to $loss_{ae}$, three contrastive regularizers are also designed based on the Mandarin and English AD corpora to train the unified model. As shown in Figure 2, CR_c is the same as the one introduced in Section 4.1 which utilizes the Mandarin AD corpus. CR_e is designed accordingly using the English AD corpus. CR_{bi} is a contrastive regularizer that is shared by both languages and uses the encoded paragraph representations of both languages as training data. Therefore, the final loss for training the unified model is defined as

$$loss = \sum_{X_p} loss_{ae} + \lambda_2(CR_e + CR_c) + \lambda_3 CR_{bi}, \tag{5}$$

where λ_2 and λ_3 are weighting coefficients.

After training, we obtain encoded bilingual prargraph vectors of both AD corpora as input to make similarity-based decisions instead of Mandarin paragraph vectors only. Specifically, we combine all data from DB corpus and the training data from Mandarin AD corpus to construct the overall training dataset, and then evaluate test samples from the Mandarin AD corpus based on all the training data at the similarity-based decision step. The way of using multi-domain data is in fact the *ALL* method in (Daumé III, 2007), and is adopted as default in the following experiments.

5 Experiments

5.1 Implementation Details

In this paper, we use BERT as our default pretrained language model, while any suitable model can be employed as well. Outputs at the penultimate layer of BERT-base[1] models were taken as embeddings and $d_{emb} = 768$. The input transcriptions contained dialogues of participants and interviewers, for the interviewer's encouragement may provide clues to a participant's cognitive status. Mean accuracy of 10 times 10-fold cross validation was used as the metric of evaluation.

In the case of classification without data augmentation, we set $N_1 = 20$ and $N_2 = 20$ for CR_c, which are sufficiently large for stablely training the model. Learning rate was 1e-5, and we stopped training when the training loss decreased less than 0.01 after 5 epochs. Here an epoch means we iterate over all samples labeled as AD in the training data.

For cross-lingual data augmentation, the encoder and the decoder of both autoencoders contained two feed-forward layers with 768 hidden units at each layer. ReLU is adopted as activation function. The Opensubtitle corpus we used contained 9.9 million lines of movie subtitles. Since each sample in the AD datasets was much longer than one line of movie dialogue, we combined 50 consecutive lines of dialogue into one sample. We set $N_1 = 20$ and $N_2 = 10$ for all three contrastive regularizers, and $\lambda_1 = \lambda_2 = \lambda_3 = 1$ for the final loss function. Batch size was 512 and learning rate was 1e-5. We used 10% of Opensubtitle corpus as a validation set, and stopped training when the validation loss decreased less than 0.01 after 5 epochs.

5.2 Baseline Methods

We compared our contrastive learning method with three baselines, 1) a CNN model proposed by Kim. (2014) for text classification, 2) fine-tuning BERT directly on the Mandarin AD corpus, and 3) fine-tuning BERT but with parameters at first to penultimate layers being frozen.

The CNN model adopted the same structure as in Kim. (2014). We used 300d Glove (Pennington et al., 2014) word vectors as our word embeddings, and padded the input word sequence to length of 512. Filter windows of 3, 4, 5 with 100 feature maps each were used as the convolutional layer, followed by rectified linear units and a max-over-time pooling. Finally, a fully connected layer took the pooled vector as input and had softmax output. The loss was cross entropy. Dropout rate of 0.5, L2 gradient constraint of 3, and mini-batch size of 50 were used.

We fine-tuned BERT on the Mandarin corpus by adding one fully-connected layer on top of the first token [CLS] in the output sequence of BERT. The added layer has a softmax output for classification.

We also compared our cross-lingual data augmentation method with two baseline methods for cross-lingual data augmentation, i.e., multilingual pretrained language model and machine translation .

Multilingual pretrained language model: Multiligual BERT (mBERT)[2] has the same structure and training strategy as BERT, except that the training data of mBERT was from multiple languages. mBERT has got quite good results on some cross-lingual tasks, such as cross-lingual natural language inference. We used mBERT in the same way as we used BERT, i.e., fine-tuning or frozen. The difference was that the data of fine-tuning mBERT also contained the English DB corpus.

Machine translation: Two different translation strategies were implemented in our experiments, Translation-Train and Translation-Test. Translation-Train means translating extra train data into the language of test data, which is translating English into Chinese in our case. On the other hand, Translation-test means translating the training and test data of test language into the language of extra train data. After translation, we used data of both languages to train a unified model through the proposed contrastive learning method, same as we did in the case of no data augmentation.

When fine-tuning BERT and mBERT in frozen or not frozen ways, we followed the same hyperparameters suggested by Devlin et al. (2019). Batch size was 32, learning rate was 5e-5, number of epochs was 3 and dropout probability was always kept at 0.1.

[1] https://github.com/google-research/bert

[2] https://github.com/google-research/bert/blob/master/multilingual.md

		Accuracy (%)
w/o DA	CNN (Kim, 2014)	70.3(4.61)
	BERT(fine-tuning)	68.5(5.26)
	BERT(frozen)	74.3(1.85)
	Our method	**78.2(2.16)**
with DA	mBERT(fine-tuning)	70.9(4.86)
	mBERT(frozen)	75.2(2.36)
	Translation-Train	78.7(1.79)
	Translation-Test	80.2(1.53)
	Our method	**81.6(1.78)**

Table 2: Accuracies of AD detection on the Mandarin AD corpus using different methods. DA stands for cross-lingual data augmentation. Numbers in brackets are the standard deviations of 10 times cross validation.

5.3 Results

Table 2 shows the accuracies of AD detection on the Mandarin AD corpus using different methods. We can see that fine-tuning BERT and mBERT directly achieved the worst performance among all methods without data augmentation and with data augmentation, respectively. Partially fine-tuning BERT and mBERT by freezing most of their parameters obtained much better accuracies than training all parameters. These results indicate that simply fine-tuning large pretrained models on a small AD corpus is not recommended because there are too many parameters to update.

When data augmentation was not applied, the CNN-based text classifier performed better than fine-tuning BERT, but much worse than our method. The reason may be that the CNN model didn't utilize pretrained models and contained much more parameters than our proposed one. Our contrastive learning method outperformed other methods by at least 3.9% on the accuracy of AD detection.

When cross-lingual data augmentation was applied, Translation-Test method achieved an accuracy of 80.2%, which was 2% higher than our method without data augmentation. The promising result perhaps is due to high quality of the English BERT model and intrinsic differences between English and Mandarin, comparing with the Translation-Train scenario. Our autoencoder-based method obtained the best performance among all methods, resulting an accuracy of 81.6% which was better than other methods by 1.4% at least.

5.4 Ablation Studies

5.4.1 Models without Data Augmentation

Several ablated models were built for comparison based on our proposed method.

Monologue: We discards the transcriptions from interviewers and only used the transcriptions from participants as inputs to test the performance of our proposed method.

$[CLS]$ **pooling**: [CLS] is the first and a special token in the input sequence of BERT. Its embedding is popularly used as the pooled representation when fine-tuning BERT on downstream tasks. This model replaced our pooling method introduced in Section 4.1 with [CLS] pooling, i.e., utilizing [CLS] embeddings as paragraph vectors. Its training criterion and decision component were the same as our proposed method.

Logistic regression and **SVM**: These two models replaced the similarity-based decision introduced in Section 4.1 with a logistic regression classifier or a SVM classifier. These classifiers were trained separately after the memory vector for pooling was estimated.

The left part of Table 3 shows the results of these ablated models. By changing input from dialogue to monologue, the accuracy dropped 1.8%, indicating that utterances from interviewers also benefitted our method. Although the structure of our trainable pooling layer was much simpler than [CLS] pooling, our method still performed slightly better than the [CLS] pooling method. When logistic regression and SVM classifiers were built instead of similarity-based decision, the accuracy dropped 1.2% at least.

w/o DA	**Accuracy(%)**	with DA	**Accuracy(%)**
Our method	**78.2(2.16)**	Our method	**81.6(1.78)**
Monologue	76.4(1.96)	- CR_{bi}	80.8(2.59)
$[CLS]$ pooling	77.5(1.89)	- all CRs	73.3(2.28)
Logistic regression	77.0(1.75)	- $loss_{ae}$	79.8(1.75)
SVM	76.5(2.36)	- joint training	80.3(1.70)

Table 3: Results of ablation studies. DA stands for cross-lingual data augmentation.

These results demonstrate the effectiveness of both the pooling layer and the decision component in our proposed contrastive learning method for AD detection.

5.4.2 Models with Cross-lingual Data Augmentation

Several ablated models were built to evaluate the effects of the following components in our proposed cross-lingual data augmentation method.

Contrastive regularizers: First, we removed the cross-lingual contrastive regularizer CR_{bi}, for testing the model's performance with only monolingual contrastive regularizers. Then, all three contrastive regularizers were removed, which means that only the parallel datset was utilized to learn all model parameters.

Autoencoder loss: In contrast to removing contrastive regularizers, the autoencoder loss defined in Eq. (4) was discarded to train an ablated model. Here, the parallel corpus was not necessary anymore.

Training strategy: In our proposed method, the pooling parameters and the autoencoders are trained jointly. An ablated model was built to test the performance of training them separately. First, we used the two AD datasets and the terms of CR_e and CR_c in Eq. (5) to train the pooling parameters of each language respectively. Then, the parallel dataset and the two AD datasets were employed simultaneously to train the autoencoders with the terms of CR_{bi} and $loss_{ae}$ in Eq. (5).

The right part of Table 3 shows the results of these ablated models. We can see that CR_e and CR_c played more important roles than CR_{bi} in our proposed method. The accuracy only dropped 0.8% after removing CR_{bi}, while it dropped 8.3% when removing all three contrastive regularizers. Without autoencoder loss, the accuracy dropped 1.8%. Its performance was still 1.6% better than our proposed method without data augmentation. This implies that our proposed cross-lingual data augmentation method is still effective even when a large parallel corpus is not available. Finally, jointly training pooling parameters and autoencoders achieved better performance than training them separately.

5.5 Influences of the Size of Mandarin AD Corpus

An experiment was conducted to explore influences of the size of Mandarin AD corpus on our proposed method with cross-lingual data augmentation. 10 additional models were trained, and each reduced the Mandarin AD corpus by 10% through random sampling. To overcome the randomness of data reduction, each model was trained 5 times and mean accuracy of the 5 trials was reported as its performance measurement. The last model constructed a zero-shot learning scenario, which means that there was no Mandarin AD training data at all and the model had to make decisions by cross-lingual knowledge transfer.

We compared our method with the BERT(frozen) baseline and two machine translation baselines. The results are shown in Figure 3. From this figure, we can see that our method consistently outperformed the other three methods no matter what size the Mandarin AD corpus was. Under the zero-shot scenario, our method still performed much better than random guessing, which demonstrates the effectiveness of our method on cross-lingual knowledge transfer for AD detection.

5.6 Importance of Different Words

Inspired by previous study (Guo et al., 2019) which told different usage of words between AD patients and controls by adopting a proportion test, we conducted an experiment to investigate importance of different words on terms of the performance of our proposed method. Our autoencoder-based model

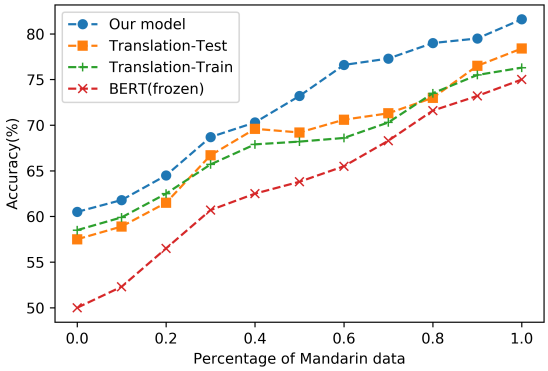

Figure 3: Results of using Mandarin AD corpus with different sizes.

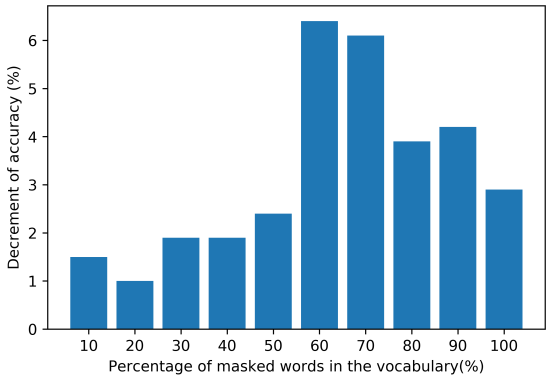

Figure 4: Accuracy decrements of masking every 10% words in the vocabulary. Words with high document frequency get masked first.

with cross-lingual data augmentation was used in this experiment. At the inference stage, 10 groups of decision results were generated which masked the words with high document frequency gradually. The document frequency of a word is defined as the percentage of documents in which the word appears. The higher document frequency a word has, more *common* the word is. For example, words like "and", "the", "is" are typical common words and usually have higher document frequency than words like "faucet" and "puddle". In each group, we found the top 10% words with the highest document frequency in the vocabulary of the Mandarin AD corpus, and replaced their embeddings with the "MASK" embedding of BERT and then removed them from the vocabulary. Afterward, we got the paragraph vectors of test samples and made similarity-based predictions.

Accuracy decrements of masking every 10% words in the vocabulary are shown in Figure 4. We can see that the performance of our method got slightly hurt when masking the words with top 10% or 20% document frequency. The accuracy dropped seriously when the words with top 60% document frequency were masked. This implies that all words didn't contribute equally. Our method relied on the words with middle document frequency more than the ones with high document frequency for AD detection.

6 Conclusions

This paper first presents a contrastive learning method to study effective text representations from small corpus for AD detection. This method contains a BERT feature extractor and a trainable pooling layer. It adopts similarity-based decision, which performed better than building separate classifiers in our experiments on a Mandarin AD corpus. An autoencoder-based cross-lingual data augmentation method is further proposed for utilizing another English AD corpus to improve model performance on the Mandarin one. This method outperformed other cross-lingual data augmentation methods, achieving a best accuracy of 81.6% in our experiments. Exploring more effective pretrained models and better methods for learning bilingual text representations will be the task of our future work.

References

Sandra Aluísio, Andre Cunha, and Carolina Scarton. 2016. Evaluating progression of Alzheimers disease by regression and classification methods in a narrative language test in portuguese. In *International Conference on Computational Processing of the Portuguese Language*, pages 109–114. Springer.

Sarath Chandar AP, Stanislas Lauly, Hugo Larochelle, Mitesh Khapra, Balaraman Ravindran, Vikas C Raykar, and Amrita Saha. 2014. An autoencoder approach to learning bilingual word representations. In *Advances in neural information processing systems*, pages 1853–1861.

Pierre Baldi. 2012. Autoencoders, unsupervised learning, and deep architectures. In *Proceedings of ICML workshop on unsupervised and transfer learning*, pages 37–49.

JT Becker, F Boller, OI Lopez, J Saxton, and KL McGonigle. 1994. The natural history of Alzheimers disease: Description of study cohort and accuracy of diagnosis. *Archives of Neurology*, 51(6):585–594.

Hal Daumé III. 2007. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 256–263.

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 (Long and Short Papers)*, pages 4171–4186.

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. 2020. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *arXiv preprint arXiv:2002.06305*.

Bruno Dubois, Harald Hampel, Howard H Feldman, Philip Scheltens, Paul Aisen, Sandrine Andrieu, Hovagim Bakardjian, Habib Benali, Lars Bertram, Kaj Blennow, et al. 2016. Preclinical Alzheimer's disease: definition, natural history, and diagnostic criteria. *Alzheimer's & Dementia*, 12(3):292–323.

Kathleen C Fraser, Jed A Meltzer, and Frank Rudzicz. 2016. Linguistic features identify Alzheimers disease in narrative speech. *Journal of Alzheimer's Disease*, 49(2):407–422.

Kathleen C Fraser, Nicklas Linz, Bai Li, Kristina Lundholm Fors, Frank Rudzicz, Alexandra König, Jan Alexandersson, Philippe Robert, and Dimitrios Kokkinakis. 2019. Multilingual prediction of Alzheimers disease through domain adaptation and concept-based language modelling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3659–3670.

Julian Fritsch, Sebastian Wankerl, and Elmar Nöth. 2019. Automatic diagnosis of Alzheimers disease using neural network language models. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5841–5845. IEEE.

Zhiqiang Guo, Zhenhua Ling, and Yunxia Li. 2019. Detecting Alzheimers disease from continuous speech using language models. *Journal of Alzheimer's Disease*, 70(4):1163–1174.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751.

Yi-hsiu Lai, Hsiu-hua Pai, et al. 2009. To be semantically-impaired or to be syntactically-impaired: Linguistic patterns in chinese-speaking persons with or without dementia. *Journal of Neurolinguistics*, 22(5):465–475.

Stanislas Lauly, Alex Boulanger, and Hugo Larochelle. 2014. Learning multilingual word representations using a bag-of-words autoencoder. *arXiv preprint arXiv:1401.1803*.

Bai Li, Yi-Te Hsu, and Frank Rudzicz. 2019. Detecting dementia in Mandarin Chinese using transfer learning from a parallel corpus. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1991–1997.

Pierre Lison and Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. *European Language Resources Association*.

Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2016. Recurrent neural network for text classification with multi-task learning. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pages 2873–2879.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019a. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Zhaoci Liu, Zhiqiang Guo, Zhenhua Ling, Shijin Wang, Lingjing Jin, and Yunxia Li. 2019b. Dementia detection by analyzing spontaneous mandarin speech. In *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 289–296. IEEE.

Brian MacWhinney. 2000. *The CHILDES Project: Tools for analyzing talk. transcription format and programs*, volume 1. Psychology Press.

Robin G Morris. 1996. *The cognitive neuropsychology of Alzheimer-type dementia.* Oxford University Press.

Bruce E Murdoch, Helen J Chenery, Vicki Wilks, and Richard S Boyle. 1987. Language disorders in dementia of the Alzheimer type. *Brain and language*, 31(1):122–137.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Jason Phang, Thibault Févry, and Samuel R Bowman. 2018. Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks. *arXiv preprint arXiv:1811.01088*.

Martin Prince, Adelina Comas-Herrera, Martin Knapp, Maëlenn Guerchet, and Maria Karagiannidou. 2016. World alzheimer report 2016: improving healthcare for people living with dementia: coverage, quality and costs now and in the future. *Alzheimers disease International*.

Daisaku Shibata, Shoko Wakamiya, Ayae Kinoshita, and Eiji Aramaki. 2016. Detecting japanese patients with Alzheimers disease based on word category frequencies. In *Proceedings of the Clinical Natural Language Processing Workshop (ClinicalNLP)*, pages 78–85.

Reisa A Sperling, Paul S Aisen, Laurel A Beckett, David A Bennett, Suzanne Craft, Anne M Fagan, Takeshi Iwatsubo, Clifford R Jack Jr, Jeffrey Kaye, Thomas J Montine, et al. 2011. Toward defining the preclinical stages of Alzheimers disease: Recommendations from the national institute on aging-alzheimer's association workgroups on diagnostic guidelines for alzheimer's disease. *Alzheimer's & dementia*, 7(3):280–292.

Tom N Tombaugh and Nancy J McIntyre. 1992. The mini-mental state examination: a comprehensive review. *Journal of the American Geriatrics Society*, 40(9):922–935.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, pages 3261–3275.

Sebastian Wankerl, Elmar Nöth, and Stefan Evert. 2017. An n-gram based approach to the automatic diagnosis of Alzheimer's disease from spoken language. In *INTERSPEECH*, pages 3162–3166.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1480–1489.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5754–5764.

Chen Zhu, Yu Cheng, Zhe Gan, Siqi Sun, Thomas Goldstein, and Jingjing Liu. 2019. Freelb: Enhanced adversarial training for language understanding. *arXiv preprint arXiv:1909.11764*.