Exploring Neural Models for Predicting Dementia from Language

by

Weirui Kong

B.Eng., Zhejiang University, 2017

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

Master of Science

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES
(Computer Science)

The University of British Columbia
(Vancouver)

August 2019

The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, the thesis entitled:

Exploring Neural Models for Predicting Dementia from Language

submitted by **Weirui Kong** in partial fulfillment of the requirements for the degree of **Master of Science** in **Computer Science**.

Examining Committee:

Giuseppe Carenini, Computer Science
Supervisor

Thalia Field, Faculty of Medicine
Supervisor

Richard Lester, Faculty of Medicine
Additional Examiner

Abstract

In this thesis we explore the effectiveness of neural models that require no task-specific feature for automatic dementia prediction. The problem is about classifying Alzheimer's disease (AD) from recordings of patients undergoing the Boston Diagnostic Aphasia Examination (BDAE). First we use a multimodal neural model to fuse linguistic features and acoustic features, and investigate the performance change compared to simply concatenating these features. Then we propose a novel *coherence* feature generated by a neural coherence model, and evaluate the predictiveness of this new feature for dementia prediction. Finally we apply an end-to-end neural method which is free from feature engineering and achieves state-of-the-art classification result on a widely used dementia dataset. We further interpret the predictions made by this neural model from different angles, including model visualization and statistical tests.

Lay Summary

Early prediction of neurodegenerative disorders such as Alzheimer's disease (AD) and related dementias is important in developing early medical supports and social supports, and may identify ideal stages for testing novel therapeutics aimed at preventing disease progression. Changes in speech and language patterns can occur in dementia in its earliest stages and may worsen as the disease progresses. This has led to recent attempts to create automatic methods that predict dementia through language analysis. In addition to features extracted from language samples, previous works have improved the prediction accuracy by introducing some task-specific features. But task-specific features prevent the model from generalizing to other tests. Our work focuses on building classification models without any task-specific features. We explore three approaches and find one such model which achieves state-of-the-art performance. We also perform detail analyses to interpret how the best performer makes a prediction.

Preface

All of the work presented henceforth was conducted in the Laboratory for Computational Intelligence in the Department of Computer Science at the University of British Columbia, in collaboration with Dr. Thalia Field at the UBC Faculty of Medicine. I was the lead researcher, responsible for coding, data preprocessing, result analysis, plots, concept formation and first drafts of the manuscripts. Dr. Giuseppe Carenini and Dr. Hyeju Jang were responsible for concept formation, draft edits, interpreting the results and suggestions for improvement. Dr. Thalia Field was responsible for editing medical related material. The baseline model was implemented by Vaden Masrani, a PhD student at UBC.

A version of Chapter 6 has been accepted as proceedings of the Machine Learning for Healthcare Conference 2019 [Weirui Kong, Hyeju Jang, Giuseppe Carenini, Thalia Field. A Neural Model for Predicting Dementia from Language.]. I was the first author, responsible for all major areas of concept formation, experiment design and analysis, as well as the majority of manuscript composition. Hyeju Jang and Thalia Field contributed to manuscript edits, refining the paper to a large extent. Giuseppe Carenini was the supervisory author on this project and was involved throughout the project in concept formation and manuscript edits.

Table of Contents

Abstract . . . . . **iii**

Lay Summary . . . . . **iv**

Preface . . . . . **v**

Table of Contents . . . . . **vi**

List of Tables . . . . . **viii**

List of Figures . . . . . **ix**

Glossary . . . . . **x**

Acknowledgments . . . . . **xi**

Dedication . . . . . **xii**

1 Introduction **1**

1.1 Contributions 3

1.1.1 Fusing Features from Different Modalities 3

1.1.2 A Novel Feature: Coherence Score 3

1.1.3 An End-to-end Neural Model: Hierarchical Attention Networks 3

1.2 Reproducibility 4

1.3 Thesis Overview 4

2 Related Work 5
2.1 Computational Approaches to Dementia Prediction 5
2.2 Multimodal Learning 6
2.3 Coherence Models 7

3 Datasets 9
3.1 DementiaBank 9

4 Multimodal Embedding for Feature Fusion 11
4.1 Our Joint Embedding Method 11
4.2 Experiment 13
4.2.1 Experiment Settings 13
4.2.2 Experiment Results 14
4.3 Discussion 16

5 A Novel Feature: Coherence Score 17
5.1 A Neural Coherence Model 18
5.2 Experiment 19
5.2.1 Experiment Settings 19
5.2.2 Experiment Results 21
5.3 Discussion 21

6 An End-to-end Neural Model: Hierarchical Attention Networks . . 24
6.1 Hierarchical Attention Networks 24
6.2 Experiment 25
6.2.1 Experiment Settings 25
6.2.2 Experiment Results on DementiaBank 27
6.2.3 Analysis of Effects of Dataset Size 27
6.2.4 Analysis of Attention 28
6.2.5 Evaluation on the Blog Corpus 32
6.3 Discussion 34

7 Conclusions and Future Work 35

Bibliography 38

List of Tables

Table 3.1 Demographics of the DementiaBank dataset. 10

Table 4.1 Results of the multimodal feature embedding evaluation. Numbers in the model name indicate how many features are used after feature selection. Numbers in parenthesis show the change in performance compared to the corresponding baseline. 15

Table 4.2 Results of the shared representation evaluation. 16

Table 5.1 Statistics on the DementiaBank, WSJ and VIST datasets. We compute the number of samples as # Doc., and the average number of sentences per document as Avg. # Sen. 20

Table 5.2 Results on the effectiveness of the coherence feature. The performance metric is accuracy. Numbers in parenthesis show the change in performance. L&A features denote linguistic and acoustic features. 22

Table 6.1 Features used by traditional methods. Info: information unit features. Spatial: spatial neglect features. 26

Table 6.2 Binary classification with 10-fold cross-validation. Note that results of Fraser's model and Masrani's model are from the original papers. 28

Table 6.3 Contingency table (numbers in parenthesis are expectation values). 30

Table 6.4 Blog information as of April 4th, 2017. 33

Table 6.5 Binary classification with 9-fold cross-validation on blog corpus. 33

List of Figures

Figure 3.1 The Cookie Theft picture. 10

Figure 4.1 Neural feature fusion frameworks. 12

Figure 5.1 A document sample and the corresponding entity grid table. The figure is taken from the original paper [32]. S denotes subject, O denotes object, X denotes other and - means the word is absent from the sentence. 19

Figure 5.2 Neural coherence model. The figure is taken from the original paper [32]. 20

Figure 5.3 Distributions of coherence scores for AD patients and healthy controls. x-axis denotes coherence score and y-axis denotes probability density. 22

Figure 6.1 Hierarchical attention network for dementia prediction. 26

Figure 6.2 Test accuracy by varying training data proportions. 29

Figure 6.3 Visualization of attention. 30

Figure 6.4 Attention frequency vs. random frequency. 32

Glossary

AD Alzheimer's disease

BDAE Boston Diagnostic Aphasia Examination

CCA Canonical Correlation Analysis

CNN convolutional neural networks

HAN Hierarchical Attention Networks

MMSE Mini Mental State Examination

ML machine learning

MFCC Mel-frequency Cepstral Coefficient

NLP natural language processing

OPTIMA Oxford Project to Investigate Memory and Aging

RNN recurrent neural network

Acknowledgments

First of all, I would like to thank my supervisors Dr. Giuseppe Carenini and Dr. Thalia Field, who supported me all the way. You are not only first-class scientists, but also kind friends to me. Thank you for your advice and encouragement. I am grateful for working with you, and for the opportunity to learn from you.

I must also express my gratitude to Dr. Hyeju Jang. You taught me a lot about scientific writing and provided tons of detailed improvements for our work. Your enthusiasm for research motivated me.

I want to thank my dear friends, who stood by my side in good and hard times.

To Xiaoxuan Lou, Bowei E and Tonghe Wang, you are the best roommates ever. We lived and studied together, for me I cherish each and every moment. Although we are now in different countries, all the things we have been through build up our brotherhood and nothing can break the bond between us.

To my badminton teammates Mengqi Li, Haocong Shi, Angli Xue, Weigeng Yang, Hang Zhou and Ye Fan, it's a pleasure to fight for trophies with you. I'll always remember how you comforted me when my mistake led to losing the game-deciding point, and the tremendous cheer from you for my nice smashes.

To my childhood buddies Xingyan Chen, Bo Hu, Ziyi Liu, Wangyang Dai, Zhengri Xiong, Hongrui Tu and Yanxin Zhu, you are all special to me. We grew up together in a small town and shared so many pleasant memories. I really appreciate having you such treasures.

Finally, Rong Hu and Qing Kong, my parents who have been trying to give their everything to me. They never went to university, but have the power to create me a world full of love. I am very fortunate and proud to be your son.

Dedication

To my grandma, who means more to me than anything else. Wishing you good health and happiness. Every milestone ahead in my life, I want to celebrate with you.

Chapter 1

Introduction

Dementia is a progressive cognitive impairment caused by neurodegenerative disease, which affects more than 46.8 million people around the world [1]. Among diverse types of dementia, Alzheimer's disease (AD), which accounts for 60% - 80% of all dementia diagnosis, is among the most financially costly diseases in developed countries [8]. Although there is not yet a cure for AD, research suggests that novel therapeutics will be most effective if given early in the disease course [36].

However, predicting AD especially in its early stages is difficult. A diagnosis of dementia involves clinical opinion based on functional status, cognitive performance on standardized tests and resource-intensive specialized tests, such as lumbar puncture or advanced neuroimaging [30]. In developing countries, access to some or all of these resources may not be available, and this is reflected in the higher than average rates of undiagnosed dementia in those regions. All around the world, only approximately 25% of the 46.8 million dementia population receive a formal diagnosis [1]. Therefore, a non-invasive diagnostic tool that is inexpensive and easy to administer is of great importance to dementia patients, especially those in developing countries.

One promising direction is to design a tool that can assist in prediction of preclinical disease by using automated analysis of language. Language is one of the first facilities afflicted by the disease and subtle changes in language are observed a year or more before dementia is diagnosed, according to longitudinal studies on

people with AD [2]. These changes include, for example, low grammatical complexity, limited vocabulary and frequent word finding problems [21].

Given that linguistic deficits are early signs of dementia, researchers have developed dementia prediction systems based on language by applying machine learning (ML) and natural language processing (NLP). Most prior work built computational models on the dataset DementiaBank [7], a publicly available dataset that contains audio recordings and transcripts of participants (people with dementia and healthy controls) describing the Cookie Theft picture (Figure 3.1). Prior work used not only acoustic features and various linguistic features but also task-specific features such as information units. *Information units* [9] are objects and actions appearing in the picture (e.g., mother, stool, overflowing, and drying), which are usually pre-defined by human experts. Information unit features measure how well a participant captures key concepts in the picture. Based on these task-specific features as well as linguistic and acoustic features, prior models using traditional classification methods such as logistic regression have been shown to give reliable AD prediction [12, 28]. Although the task-specific features are effective for dementia prediction, one major disadvantage is that they are specific to a particular picture. If participants are asked to describe a different picture, information units in the picture need to be re-defined.

The advances in neural networks, especially the recent deep neural models, undermine the necessity of feature engineering. The interactions between neurons, the hierarchical network structure, and an appropriate loss function make the deep models capable of tackling complex tasks even with raw data as input. The power of neural models is likely to make up the absence of some well-designed features. In this thesis, we explore neural models for dementia prediction without using task-specific features. We delve into three different directions and our contributions can be summarized as below.

1.1 Contributions

1.1.1 Fusing Features from Different Modalities

We propose to use a neural network model for combining task-agnostic multimodal features from prior work. Previous work [12, 28] extracted various linguistic and acoustic features for dementia prediction. Then, they combined all these features by simple concatenation. In this thesis, we combine the two groups of features by using the neural multimodal embedding framework [22]. We demonstrate that combining multimodal features in this way allows obtaining performance comparable to prior work using feature selection.

1.1.2 A Novel Feature: Coherence Score

We extract a new type of task-agnostic features by a neural network model. Previous literature [10, 11, 23] has shown that people with dementia tend to have problems with discourse coherence including impairment in global coherence, disruptive topic shift, frequent use of filler phrases, and less use of connective words. Linguistic features in previous studies capture language deficits with respect to certain aspects of discourse coherence at a lexical and syntactic level (e.g., word repetitiveness, syntactic complexity, and vocabulary richness). However, no prior work has attempted to investigate semantic level of discourse coherence for predicting dementia – specifically, the patterns of how people repeat entities to make a coherent speech. In this thesis, we compute discourse coherence scores based on entity transition patterns by using the neural coherence model [32], and apply them for dementia prediction.

1.1.3 An End-to-end Neural Model: Hierarchical Attention Networks

We propose to use a neural network model in an end-to-end manner to avoid any task-specific features and alleviate the problem of manual feature engineering. We apply a neural framework, called Hierarchical Attention Networks (HAN) [38] for the task, and obtain results comparable to traditional models that use task-specific features. By including a demographic feature (age), our model achieves state-of-

the-art performance, improving the classification accuracy of the top-performer traditional method which also uses age, from 84.4% to 86.9%. With the attention mechanism in HAN, we analyze the model predictions, and provide some insights on their interpretation.

We also apply HAN to a dementia blog corpus, and discuss the results in comparison to prior work. In essence, on this corpus of written text, the neural method is not a competitive solution.

1.2 Reproducibility

The code to reproduce experiments in Chapter 4 and Chapter 5 is available at https://github.com/arankong/dementia_classifier and the code to reproduce all results and the corresponding plots in Chapter 6 is at https://github.com/arankong/han.

1.3 Thesis Overview

The rest of the thesis is organized as follows: in Chapter 2 we review prior studies that focus on traditional ML and NLP approaches for dementia prediction. We also provide backgrounds on multimodal learning and discourse coherence models. After that, in Chapter 3 we provide an overview of datasets used for our experiments. Then, the three contributions are each described in their own chapters (4, 5, 6, respectively). Lastly, in Chapter 7 we conclude and suggest some future directions.

Chapter 2

Related Work

Computational approaches for automatic dementia prediction have received growing attentions in recent years. In this chapter, we first discuss previous computational works for dementia prediction (Section 2.1). Then, to provide some background to the neural models used for our studies, we review multimodal learning methods (Section 2.2) and discourse coherence models (Section 2.3).

2.1 Computational Approaches to Dementia Prediction

Prior research has shown that NLP and ML techniques that exploit various features can predict dementia by classifying dementia patients from healthy controls.

Ahmed et al. [3] proposed features that were helpful for identifying dementia from speech, using data collected in the Oxford Project to Investigate Memory and Aging (OPTIMA) study. They found that language was progressively impaired as the disease progressed and suggested using semantic, lexical content and syntactic complexity features for classification.

Orimaye et al. [33] used diverse machine learning methods with lexical and syntactic features to distinguish between dementia patients and healthy adults on the DementiaBank dataset [7]. They compared five different classifiers including support vector machines (SVMs), naive Bayes, decision trees, neural networks and Bayesian networks, and reported that SVMs showed the best performance with a F-score of 74%.

In another study, Al-Hameed et al. [4] extracted acoustic features from the audio files of the DementiaBank dataset, building a regression model to predict Mini Mental State Examination (MMSE) scores used for dementia prediction (ranging from 0 to 30). This work used only acoustic features, and their regression model predicted MMSE scores with a mean absolute error less than 4.

Fraser et al. [12] explored a broad spectrum of both linguistic and acoustic features, demonstrating the necessity of feature selection. They found that optimal performance was obtained when 35-50 features were used, and the performance dropped off dramatically with a feature set size larger than 50. They achieved an accuracy of 81.96% in distinguishing individuals with AD from those without.

As briefly mentioned in the introduction, the DementiaBank is associated with a set of human-defined information units representing key components of the Cookie Theft picture, such as subjects, objects, locations and actions [9]. Upon the information units, Masrani [28] proposed a novel feature group called spatial neglect features. They vertically split the picture into two halves and computed features that measure spatial neglect, e.g., count of mentions of any information unit for each region. Combing their new feature group with linguistic, acoustic, information unit features and the demographic feature (age), followed by a feature selection step, they achieved the accuracy of 84.4%.

Our study differs from previous approaches in that, we aim to build models without any task-specific features, while achieving comparable or even better performance.

2.2 Multimodal Learning

Our attempt to fuse features from different modalities is inspired by the research field of multimodal learning. The multimodal representation can be divided into two types, i.e., joint representation and coordinated representation. The idea of joint representation is to build one common representation for different modalities. The simplest way is to concatenate features from different modalities, which is also known as early fusion. But this naive concatenation does not help us gain any insights into the data. More advanced method is to train a multimodal autoencoder [31], where they adopted an unsupervised training scheme to learn a shared repre-

sentation from different modalities. Besides, Zadeh et al. [39] made a tensor out of the features from 3 modalities, and used the 3D tensor as the joint representation.

The other type of multimodal representation, namely coordinated representation, aims at building different representations for each modality while putting certain constraints on these representations. The constraints include similarity-based methods (e.g., cosine distance), structure constraints (e.g., orthogonality) and correlation maximization like Canonical Correlation Analysis (CCA) [17]. In particular, a learning approach, *joint embeddings*, is very successful in building coordinated representations of two modalities [22].

An ablation study on the DementiaBank dataset shows that the classification accuracy of logistic regression with linguistic features is 0.740, whereas the accuracy drops to 0.713 when both linguistic features and acoustic features are used, indicating that there are more irrelevant features in the acoustic feature group. Therefore, we prefer coordinated multimodal representations to one shared representation for the linguistic modality and acoustic modality. Specifically, we use a similar training process as Kiros et al. [22]'s.

2.3 Coherence Models

Modelling document coherence is an active area of NLP. There are various coherence modelling methods: entity-based models, graph-based models [16], models based on discourse relations [26], models based on distributed sentence representations [25], etc. We use a neural entity-based coherence model [32] for the coherence feature extraction because it is conceptually easy and has obtained high performance in many coherence evaluation tasks. In entity-based coherence models, document is about entities (nouns could serve as entity candidates) and coherence is created by repeated entity mentions [15]. Barzilay and Lapata [6] proposed an entity grid model that computes coherence score based on entity transitions (the grammatical role switches across sentences). Nguyen and Joty [32] made a neural version of the entity grid model. They transformed the grammatical role of each entity in grid into distributed representation, and used convolutional neural networks (CNN) [24] to capture entity transition patterns. Their model achieved high performance in several tasks like sentence ordering and summary coherence

rating. We use this neural coherence model to generate coherence scores for the DementiaBank samples.

Chapter 3

Datasets

We use two dementia datasets for this work: one consists of samples of spoken language and the other consists of samples of written language. We detail the spoken one, the DementiaBank dataset below, which is used throughout this thesis (in Chapter 4, 5 and 6). The dataset of written samples is used only in Chapter 6 and will be introduced in Section 6.2.5. Besides, we also use two large corpora for training the neural coherence model and we introduce them along with the coherence model in Section 5.2.1.

3.1 DementiaBank

The DementiaBank corpus was collected for the study of communication in dementia, between 1983 and 1988 at the University of Pittsburgh [7]. It contains interview recordings and manually-transcribed transcripts of English-speaking participants describing the Cookie Theft picture (Figure 3.1). The participants are categorized into dementia patient and healthy control groups. Of the 309 dementia samples, 257 samples are classified as possible/probable AD, and the remaining samples as other types of dementia. Our study uses only the 257 AD samples and 242 healthy elderly control samples. Statistics about the DementiaBank samples used in this study are listed in Table 3.1.

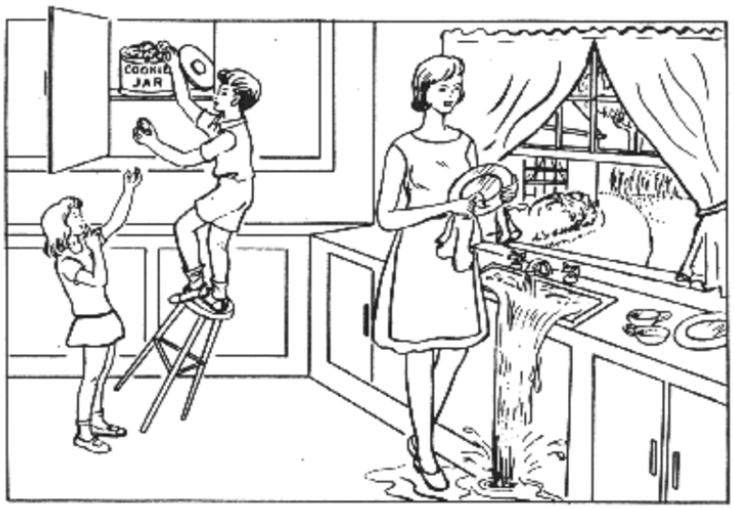


Figure 3.1: The Cookie Theft picture.

Table 3.1: Demographics of the DementiaBank dataset.

Diagnosis	**Samples**	**Mean Words**	**Mean Age**
AD	257	104.98 (s=59.8)	71.72 (s=8.47)
Control	242	113.56 (s=58.5)	63.95 (s=9.16)

Chapter 4

Multimodal Embedding for Feature Fusion

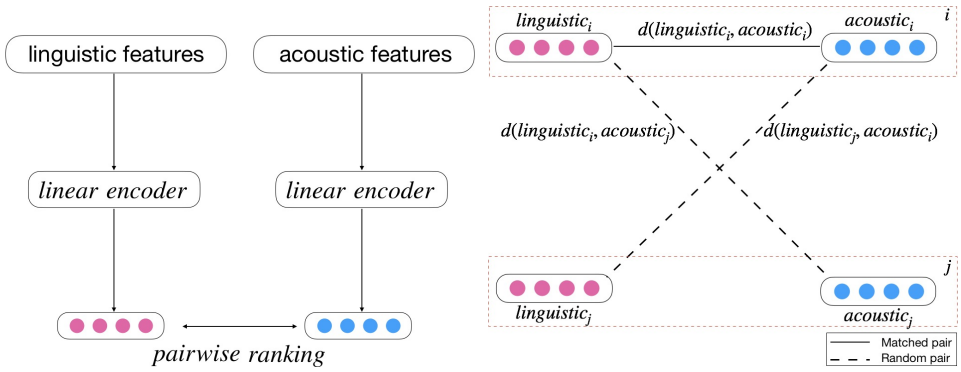
In this chapter, we describe our experiment on using a neural network model for combining existing multimodal features. Prior work [28] showed effectiveness of a variety of linguistic and acoustic features for dementia prediction. To obtain a combined multimodal feature representation, they used a simple concatenation as a fusion mechanism, which is easy to implement. However, this mechanism can wind up being very high dimensional, and could be less effective when features have different frame rates [27]. Here, we propose to use a joint embedding method based on pairwise ranking.

In Section 4.1 we explain our joint embedding method using pairwise ranking in detail. After that, in Section 4.2 we compare the performance of this feature fusion scheme for dementia prediction against the simple concatenation method. We discuss the results in Section 4.3.

4.1 Our Joint Embedding Method

To combine features in different modalities, we use a joint embedding method adapted from [22]. The main idea in this method is pairwise ranking – a matched pair of linguistic and acoustic embeddings, or the original pair of linguistic and

acoustic embeddings from the same data sample[1], should have a shorter distance than random pairs. To implement this idea, our model is composed of three parts: building linguistic representations (embeddings), acoustic representations, and coordinating the two representations. Figure 4.1 shows the model architecture.



(a) Proposed joint embedding method

(b) Pairwise ranking loss to minimize distance between matched pairs of samples

Figure 4.1: Neural feature fusion frameworks.

First, to build linguistic embeddings, we begin with extracting linguistic features from texts as in [28]. These features ($N = 99$ in total) include parts-of-speech ($N = 15$), context-free-grammar rules ($N = 43$), syntactic complexity ($N = 27$), vocabulary richness ($N = 4$), psycholinguistic ($N = 5$), and repetitiveness ($N = 5$). Then, we project these features into the embedding space of dimension m by using an encoder. This encoder is used for linear transformation ($\mathbb{R}^{99} \rightarrow \mathbb{R}^m$), and hence it consists of one linear layer without an activation function. We choose 50 as the embedding size m for our experiment because in [28], the performance of dementia prediction drastically dropped when using more than 50 features at the feature selection step.

In the same fashion, we build acoustic embeddings. We extract acoustic features from audio recordings following [28]. The acoustic features ($N = 172$ in total) were derived from speech samples by using the Mel-frequency Cepstral Coefficient (MFCC) technique [20]. We use another encoder ($\mathbb{R}^{172} \rightarrow \mathbb{R}^d$) for linearly

[1]Note that every sample in the DementiaBank dataset provides one matched linguistic and acoustic embedding pair.

transforming the acoustic features into acoustic feature embeddings. The embedding size d is also set to 50.

After obtaining linguistic and acoustic embeddings, we coordinate these embeddings by using a loss function based on pairwise ranking as in [22]. As shown in Figure 4.1b, given a matched pair of features ($linguistic_i$, $acoustic_i$), the distance in the embedding space between $linguistic_i$ and $acoustic_i$ should be smaller than the distance between $linguistic_i$ and any other acoustic embeddings $acoustic_j$, and the distance between $acoustic_i$ and any other linguistic embeddings $linguistic_j$. For each matched pair, we compute the loss against all other pairs. The loss function for matched pairs ($linguistic_i, acoustic_i$) and ($linguistic_j, acoustic_j$) is defined as

$$\begin{aligned} L(\theta) = &\max\{0, \alpha - d(linguistic_i, acoustic_i) + d(linguistic_i, acoustic_j)\} \\ &+ \max\{0, \alpha - d(linguistic_i, acoustic_i) + d(linguistic_j, acoustic_i)\}, \end{aligned}$$

where θ denotes learnable parameters of our encoders, α is an arbitrary positive constant, $i \neq j$, and $d(x, x')$ is a distance measure. We use cosine similarity as the distance measure in our experiment.

After training this neural model that consists of two encoders using the pairwise ranking loss, we obtain coordinated linguistic and acoustic embeddings from raw linguistic and acoustic features. These embeddings are then used for dementia prediction.

4.2 Experiment

4.2.1 Experiment Settings

The proposed method results in two embeddings: linguistic and acoustic. To evaluate the proposed method, we use the resulting embeddings as features for dementia prediction. Specifically, we build logistic regression classifiers using three different feature groups. The EMBEDDED-L model uses only the linguistic embeddings and the EMBEDDED-A model uses the acoustic embeddings alone. The third model,

EMBEDDED-L&A, uses both linguistic and acoustic embeddings by concatenating them. Note that we do not conduct feature selection for all the models above since our features are already compacted.

We compare these embedding models against three types of corresponding baselines that use task-agnostic features (i.e., linguistic and acoustic features) from prior work. These baselines include BASELINE-L, BASELINE-A, and BASELINE-L&A. BASELINE-L uses only the original linguistic features. BASELINE-A uses the original acoustic features. BASELINE-L&A combines those two types of features by concatenating them. We perform feature selection for these baseline models as in [28]. We use Pearson's correlation coefficients to select top k features.

In addition, we also compare our models to neural network based baselines EMBEDDED-SHARED that build one shared representation for linguistic and acoustic features. To do that, we first combine the two groups of features by concatenating them, and then use an autoencoder to construct a shared embedding. The encoder embeds the concatenated features ($N = 271$, total size of the two feature groups) into a hidden space of dimension h. We use a linear layer with the ReLU activation function for the encoder ($\mathbb{R}^{271} \rightarrow \mathbb{R}^h$). Then the decoder ($\mathbb{R}^h \rightarrow \mathbb{R}^{271}$) reconstructs the combined features from its hidden representation. We compute the L2-norm as the reconstruction loss. The vector in the hidden space is regarded as one shared representation of the linguistic and acoustic features, and we use this vector as features for dementia prediction. We set the dimension of hidden space to 50 and 100, denoting by EMBEDDED-SHARED 50 and EMBEDDED-SHARED 100 respectively.

We perform 10-fold cross validation following the practice in prior work [12, 28]. Because the weights of our encoders are randomly initialized, we report the average performance on ten different runs as the performance of our model.

4.2.2 Experiment Results

Our experiment results are listed in Table 4.1 and Table 4.2. We first compare models that use embeddings against models that use raw features. As shown in Table 4.1, models using embeddings from our joint method (EMBEDDED-L, EMBEDDED-A, and EMBEDDED-L&A) drastically outperforms models using their

corresponding raw features (BASELINE-L, BASELINE-A, and BASELINE-L&A). This suggests that embeddings from our method contain more predictive information than raw features.

This pattern is also observed after performing feature selection on the baseline models using raw features. Our models EMBEDDED-L and EMBEDDED-A improve over BASELINE-L 50 and BASELINE-A 50 which select 50 important features by using Pearson's correlation coefficient. EMBEDDED-L and EMBEDDED-A also outperform BASELINE-L BEST and BASELINE-A BEST which show best performances among all k values for feature selection. In addition, EMBEDDED-L&A shows improvement over baselines selecting 50 features (BASELINE-L&A 50) and 100 features (BASELINE-L&A 100). EMBEDDED-L&A also shows performance comparable to EMBEDDED-L&A BEST, which selects 47 features. These results indicate that our joint embedding method generates linguistic and acoustic embeddings that perform in a similar degree to the effect of feature selection.

Table 4.1: Results of the multimodal feature embedding evaluation. Numbers in the model name indicate how many features are used after feature selection. Numbers in parenthesis show the change in performance compared to the corresponding baseline.

Models	**Accuracy**	**F-score**
Baseline-L no feature selection	0.728	0.738
Baseline-L 50	0.723	0.732
Baseline-L best ($k = 60$)	0.740	0.747
Embedded-L 50	**0.746** (+0.006)	**0.749** (+0.002)
Baseline-A no feature selection	0.567	0.578
Baseline-A 50	0.499	0.522
Baseline-A best ($k = 152$)	0.601	0.623
Embedded-A 50	**0.615** (+0.014)	**0.625** (+0.002)
Baseline-L&A no feature selection	0.665	0.671
Baseline-L&A 50	0.699	0.702
Baseline-L&A 100	0.635	0.653
Baseline-L&A best ($k = 47$)	**0.709**	0.**719**
Embedded-L&A 100	0.708 (-0.001)	0.708 (-0.011)

Table 4.2 shows the comparison between EMBEDDED-L&A and baselines us-

ing both linguistic and acoustic features. BASELINE-L&A models use raw linguistic and acoustic features with or without using feature selection, and EMBEDDED-SHARED models use a simple autoencoder to transform concatenated features into a shared embedding. As seen from the results, our model EMBEDDED-L&A outperforms all other baseline models using both groups of features. The neural baselines, EMBEDDED-SHARED 50 and EMBEDDED-SHARED 100 are not as competitive as our joint embedding method.

Table 4.2: Results of the shared representation evaluation.

Models	**Accuracy**	**F-score**
Baseline-L&A no feature selection	0.665	0.671
Baseline-L&A 50	0.699	0.702
Baseline-L&A 100	0.635	0.653
Embedded-shared 50	0.677	0.679
Embedded-shared 100	0.666	0.669
Embedded-L&A 100	**0.708**	**0.708**

4.3 Discussion

Our experiment results show that linguistic and acoustic embeddings generated by our joint embedding method are more informative than raw features or these selected important features. The results suggest that the pairwise ranking idea behind our method is capable of adding more predictive information when performing dimension reduction in our neural architecture than using feature selection, especially for the same feature dimension.

However, the improvements over baselines using feature selection with the best k are not huge especially for dementia prediction only using linguistic features. When using both linguistic and acoustic features, the baseline using best k features slightly outperforms our model. This could mean that using our joint embedding method could be considered as an alternative to attempting to find the best k for feature selection.

Chapter 5

A Novel Feature: Coherence Score

Neural network models can be used for devising a new feature type for dementia prediction. In this chapter, we experiment on discourse coherence for dementia prediction. We use an existing neural network based NLP approach for computing discourse coherence.

People with dementia have been reported to show impairment in discourse coherence such as disruptive topic shift, frequent use of filler phrases, and less use of connective words [10, 11, 23]. The linguistic features used in previous computational studies for dementia prediction capture language deficits including some discourse coherence at a lexical and syntactic level (e.g., word repeat, syntactic complexity, and vocabulary richness). However, no prior work has attempted to use the overall coherence level of a speech for dementia prediction. In this chapter, we compute discourse coherence scores using a state-of-the-art neural based coherence model, and apply them for dementia prediction.

In Section 5.1, we briefly explain the coherence model we use for obtaining the discourse coherence feature. In Section 5.2, we evaluate this new type of feature for dementia prediction. In Section 5.3, we discuss the results.

5.1 A Neural Coherence Model

We compute discourse coherence scores using the neural network based coherence model proposed by Nguyen and Joty [32], which operationalizes local coherence of a discourse segment based on the Centering theory [15]. The Centering theory claims that certain entities mentioned in an utterance are more central than others and that this property imposes constraints on a speaker's use of different types of referring expressions. They also argue that the compatibility between centering properties of an utterance and choice of referring expression affects the coherence of discourse. Based on the theory, Nguyen and Joty [32]'s model calculates coherence scores that measure how sentences are bound together to deliver a meaning as a whole by capturing entity transition patterns. We use their model because it shows the state-of-the-art performance among the systems that implement discourse coherence based on the Centering theory.

To capture entity transition patterns, the model first requires an entity grid table of input data. A transition of one entity is defined as the grammatical role switches of the entity across sentences. For example, in a document that consists of three sentences, if an entity is mentioned as the subject of the first sentence, the object of the second sentence, and not mentioned in the third sentence, its transition can be denoted as $\{S, O, -\}$. The transitions of all entities in a document are converted into an entity grid table (see Figure 5.1), and used as the input to the neural network coherence model.

The neural network coherence model is based on a convolutional neural network (CNN) [24] for computing coherence scores of a text in an end-to-end fashion (see Figure 5.2). The intuition of this neural coherence model is that each convolutional filter tries to detect a specific transition pattern (e.g., $\{S\text{-}S\text{-}O\text{-}X\}$ for a coherent text) which is informative for determining the coherence level. The CNN layer performs convolution operation on the transitions of each entity independently, followed by a max-pooling and a linear layer that generates a real-valued score. During training, a pair of documents, i.e., the original document (considered as coherent) and its randomly permuted version (considered as incoherent) are fed to the coherence model at the same time. The model outputs two scores, $\phi(original|\theta)$

	UNIT	PRODUCTS	RESEARCH	COMPANY	PARTS	CONTROLS	INDUSTRY	ELECTRONICS	TERM	CONCERN	AEROSPACE	EMPLOYEES	SERVICES	LOS ANGELES	EATON
s_0	O	–	X	X	–	–	–	–	–	–	–	X	–	–	X
s_1	–	–	–	–	–	–	–	–	S	–	–	–	–	–	–
s_2	–	O	–	–	–	–	X	–	–	–	–	O	O	X	–
s_3	–	–	–	–	X	X	–	X	–	O	X	–	–	–	S

s_0: Eaton Corp. said it sold its Pacific Sierra Research unit to a company formed by employees of that unit.

s_1: Terms were not disclosed.

s_2: Pacific Sierra, based in Los Angeles, has 200 employees and supplies professional services and advanced products to industry.

s_3: Eaton is an automotive parts, controls and aerospace electronics concern.

Figure 5.1: A document sample and the corresponding entity grid table. The figure is taken from the original paper [32]. S denotes subject, O denotes object, X denotes other and - means the word is absent from the sentence.

and $\phi(permuted|\theta)$. A pairwise ranking loss defined as

$$L(\theta) = \max\{0, 1 - \phi(original|\theta) + \phi(permuted|\theta)\}$$

forces the model to produce a higher score for the original document.

After training the model, we compute coherence scores for data samples in the DementiaBank dataset. Then, we use the scores for predicting dementia as features .

5.2 Experiment

5.2.1 Experiment Settings

To train the coherence model, we use three different corpora: the DementiaBank dataset, the Wall Street Journal (WSJ) dataset [34], and the Visual Storytelling

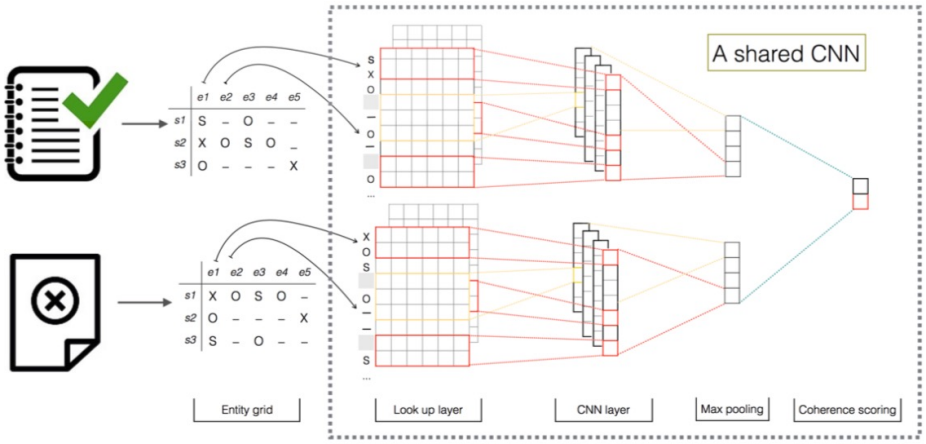


Figure 5.2: Neural coherence model. The figure is taken from the original paper [32].

(VIST) dataset [19]. First, we use the training data of DementiaBank for building the coherence model. However, the DementiaBank dataset might be too small for learning the deep neural model. Therefore, we also try larger datasets. We use the WSJ dataset for training as in [32]. Additionally, to experiment with a dataset that is more similar to DementiaBank than WSJ in terms of language style, we use the VIST dataset in which participants were asked to write a story based on a sequence of pictures. The statistics of the three datasets used for model training are shown in Table 5.1. For large datasets WSJ and VIST, we make a 50%: 10%: 40% split for training, validation and test according to [32]. For DementiaBank, we perform 10-fold cross validation.

Table 5.1: Statistics on the DementiaBank, WSJ and VIST datasets. We compute the number of samples as # Doc., and the average number of sentences per document as Avg. # Sen.

Dataset	**# Doc.**	**Avg. # Sen.**
DementiaBank	499	12.8
WSJ	2431	21.8
VIST	50197	5

To follow the pairwise ranking training scheme, we generate 20 random permutations for each document. One permuted version consists of all the sentences of an original document in a rearranged order. The original document is treated as coherent, and its permutations are regarded as incoherent. Then, we use a pair of an original document and its permuted version as input for training.

We set model hyperparameters as suggested by the original paper of the model [32]. We set the number of filters to be 150, max-pooling size to be 6 and entity embedding size to be 100. The dropout ratio is 0.5, the mini batch size is 64, and the optimizer is RMSprop [18]. We use early stopping with a patience setting of 5 epochs.

All trained models on the three datasets reported test accuracy of higher than 75%. Based on these trained models, we compute coherence scores for DementiaBank, and use the scores for dementia prediction as features.

To investigate the effectiveness of the proposed feature, we use logistic regression for classification. We evaluate the performance when coherence score is the only feature, and when it is combined with other task-agnostic features (i.e., linguistic and acoustic features). Our baselines include majority class baseline, a model using only linguistic features, a model using only acoustic features, and a model using both linguistic and acoustic features.

5.2.2 Experiment Results

Table 5.2 reports the results. The first row in the table represents our baseline models without the coherence feature. In particular, 0.515 is the accuracy of the majority class classifier. The coherence score, when being the only feature, can improve the accuracy by as much as 4%. When trained on DementiaBank, the coherence feature has a boost of 0.4% when combined with linguistic features. However, in other cases it has no effect, or even hurts the performance, when combined with other task-agnostic features.

5.3 Discussion

Our new coherence feature does not perform well for dementia prediction although using the feature outperforms the majority class baseline. Here, we investigate the

Table 5.2: Results on the effectiveness of the coherence feature. The performance metric is accuracy. Numbers in parenthesis show the change in performance. L&A features denote linguistic and acoustic features.

Models	No other features	Linguistic	Acoustic	L&A features
Baseline	0.515	0.740	0.601	0.713
DementiaBank	0.555 (+0.04)	0.744 (+0.004)	0.599 (-0.002)	0.711 (-0.002)
WSJ	0.543 (+0.028)	0.734 (-0.006)	0.605 (+0.006)	0.709 (-0.004)
VIST	0.527 (+0.012)	0.734 (-0.006)	0.603 (+0.004)	0.713 (-0)

coherence score feature more closely.

Our assumption behind using the coherence feature is that AD patients would give a less coherent picture description than healthy elderly people. To verify this assumption, we examine the relationship between coherence scores and AD. Figure 5.3 represents the distributions of coherence scores for AD patients and healthy controls.

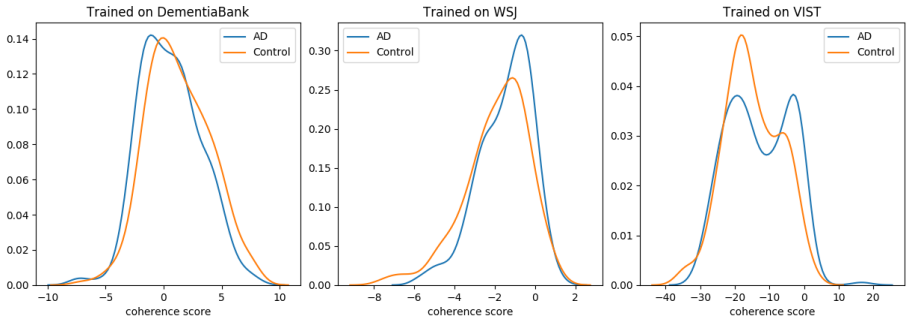


Figure 5.3: Distributions of coherence scores for AD patients and healthy controls. x-axis denotes coherence score and y-axis denotes probability density.

From the graphs, we can see that the distributions of both groups look alike in all three cases, which indicates that the coherence scores do not have much predictive power to distinguish AD patients from healthy controls. Our coherence model is based on the idea that referring to the same entities shows some patterns related

to discourse coherence. However, describing the Cookie Theft picture does not seem to require mentioning the same entity repeatedly too many times. Therefore, it is possible that dementia does not greatly affect this type of coherence which can be captured by the model when describing the Cookie Theft picture.

Chapter 6

An End-to-end Neural Model: Hierarchical Attention Networks

In this chapter we detail our experiments of using Hierarchical Attention Networks (HAN) [38] for dementia prediction. HAN is an end-to-end neural network model, which allows avoiding any feature engineering. It has been very successful in several text categorization tasks like sentiment estimation [40] and topic classification [37].

In Section 6.1 we introduce the original HAN model and our modified version. Then, in Section 6.2 we evaluate HAN models and baselines on the DementiaBank dataset, and test their performance when using only small portions of the dataset. Particularly, in Section 6.2.4 we analyze the information captured by the attention mechanism. We further compare the performance of the HAN model and one traditional model on a written text dataset in Section 6.2.5. Finally in 6.3 we discuss our work on applying HAN to dementia prediction.

6.1 Hierarchical Attention Networks

Figure 6.1 illustrates the overall architecture of the HAN model for dementia prediction. The model input are words from one interview sample (i.e., a description of the Cookie Theft picture). The model output is the probability distribution over two categories, AD and healthy. The model consists of a word sequence encoder, a

word-level attention layer, a sentence encoder and a sentence-level attention layer.

We briefly introduce the functionality of each layer. For more details, refer to [38]. The word encoder uses the bidirectional GRU [5], an efficient implementation of recurrent neural network (RNN). It encodes each word in one sentence into a hidden vector, given the context of other words in the sentence. Then the word-level attention layer puts different weights on each word vector, producing a weighted hidden vector of the sentence. Once we get all the sentence vectors of the input sample, we feed them into another bidirectional GRU, i.e., a sentence encoder. This sentence encoder along with the sentence-level attention layer builds a weighted vector (denoted by v in Figure 6.1) for the whole document, which is the latent representation of an input sample by applying attention mechanism to both word level and sentence level. Finally a linear layer projects v to a 2-dimensional vector, on which a softmax operation is performed. The output is the probability distribution for AD and healthy. Negative log likelihood of the correct label is used as the training loss.

We evaluate the performance of two models, one is the original HAN model and the other incorporates demographic information by concatenating v with the age of the interviewee. Since the scale of age ([50, 90] in our dataset) is much larger than the values of elements of v (typically in [-1, 1]), we standardize the age, making it zero mean and unit variance before concatenating it with v.

6.2 Experiment

6.2.1 Experiment Settings

As in the previous two chapters, we perform 10-fold cross validation. The reported performance is an average across the 10 folds. For evaluation metrics, we compute prediction accuracy, precision, recall and F-score.

The age is an important predictor of dementia according to Gao et al. [14]. Our demographic-based baseline uses only the ages of participants as features to demonstrate the predictiveness of the age feature. In addition to the simple demographic-based baseline, five models are tested for comparison: the model by Fraser et al. [12]; the model by Masrani [28] which obtained the best results among

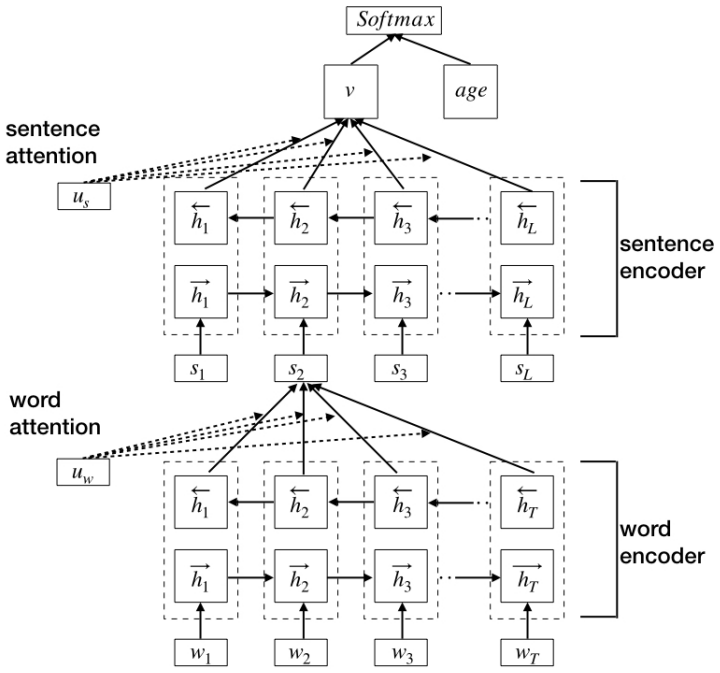


Figure 6.1: Hierarchical attention network for dementia prediction.

previous studies; a bidirectional GRU model; and the two HAN based models mentioned before. In Table 6.1 we list different feature groups leveraged in the traditional methods.

Table 6.1: Features used by traditional methods. Info: information unit features. Spatial: spatial neglect features.

Dataset	Methods	Linguistic	Acoustic	Info	Spatial	Age
DementiaBank	Age only	×	×	×	×	✓
	Fraser et al. [12]	✓	✓	✓	×	×
	Masrani [28]	✓	✓	✓	✓	✓
Dementia Blog	Masrani et al. [29]	✓	×	×	×	×

The bidirectional GRU model has the same structure as the word encoder of our HAN model, including the word level attention. Instead of using a sentence

encoder, it builds a document representation via a max-pooling operation across sentence embeddings. The document representation is fed to a linear layer and softmax function to produce the prediction. We consider this bi-GRU model as a baseline to investigate the effect of the hierarchical architecture of the HAN model.

To ensure the best results, all five approaches involve a model selection on the training data, within each step of the 10-fold cross validation procedure. For the first two traditional models, they select k features with the highest Pearson's correlation coefficients between each feature and the binary class in the training set. This subset of features are used for building the classifier. For the bi-GRU baseline and the HAN based models, within the training set we further reserve 10% of the samples for validation. We then train a model on the remaining training samples for many iterations, storing the model parameters after each iteration. The validation data is used for selecting the model that achieves the lowest validation loss.

For the hyper parameters of the HAN models, we set the word embedding dimension to be 300 and the GRU dimension to be 100. The word embeddings are initialized randomly. For training, we use SGD (stochastic gradient descent) with momentum of 0.9 and learning rate of 0.1. The bi-GRU baseline has the same setting as the HAN models. Those hyper parameters are not fine-tuned.

6.2.2 Experiment Results on DementiaBank

Table 6.2 summarizes the results. The HAN model achieves performance of 0.815 in both accuracy and F-score. When combined with the age feature, the HAN-AGE model resultes in a remarkable boost in performance, 2.5% improvement in accuracy and 3% improvement in F-score over Masrani [28]. In addition, the HAN model shows a significant increase in performance compared to the bi-GRU baseline, demonstrating the higher capacity as a result of leveraging hierarchy in HAN.

6.2.3 Analysis of Effects of Dataset Size

In general, training deep neural network models require large data. To investigate if the HAN models are robust to the size of the training data, we evaluated the two

Table 6.2: Binary classification with 10-fold cross-validation. Note that results of Fraser's model and Masrani's model are from the original papers.

Model	Accuracy	Precision	Recall	F-score
Baseline (age only)	0.595	0.591	0.729	0.653
Fraser et al. (no age)	0.820	-	-	-
Masrani (with age)	0.844	-	-	0.846
bi-GRU baseline	0.748	0.750	0.811	0.768
HAN	0.815	0.839	0.818	0.815
HAN-AGE	0.869	0.859	0.904	0.876

HAN-based models and Masrani's model from the last experiment with different proportions of the dataset. We also included a logistic regression classifier with age being the only feature. Figure 6.2 reports test accuracy when we repeated the previous experiment with 5%, 15%, 25%, 50% and 75% of the original DementiaBank dataset. For each proportion setting, we ran 5 independent experiments (randomly selecting the target subset of the data) and computed the mean and standard deviation. We can see that age is very informative, since a majority class classifier would have an accuracy around 0.5. Note that the performance of HAN drops dramatically when limited training data is used, whereas the HAN-AGE model is much less sensitive to the size of training data. The HAN-AGE model maintains a relatively high performance even with only 5% of the data samples.

6.2.4 Analysis of Attention

During the training process, the attention mechanism makes the HAN model learn which words are important in predicting a given label. To explore this information captured by the attention mechanism, we first performed a qualitative analysis by visualizing the hierarchical attention layers on a small subset of our data (see Figure 6.3).

In the visualization, each line represents a sentence. Blue denotes the sentence attention weight and red denotes the word attention weight. Figure 6.3 shows that the model tends to select words like `overflowing`, `stool`, `mother`, and `drying`, and their corresponding sentences. Interestingly, these words belong to

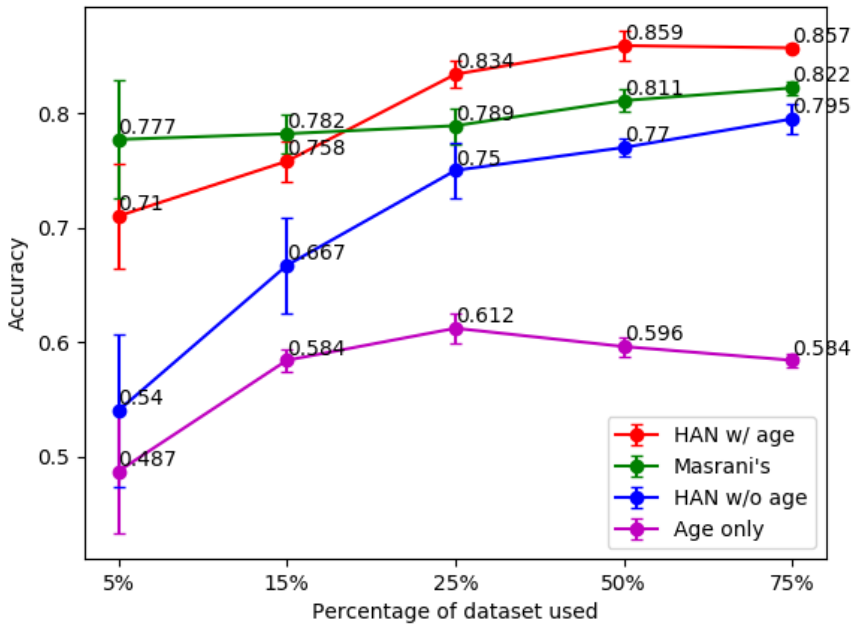
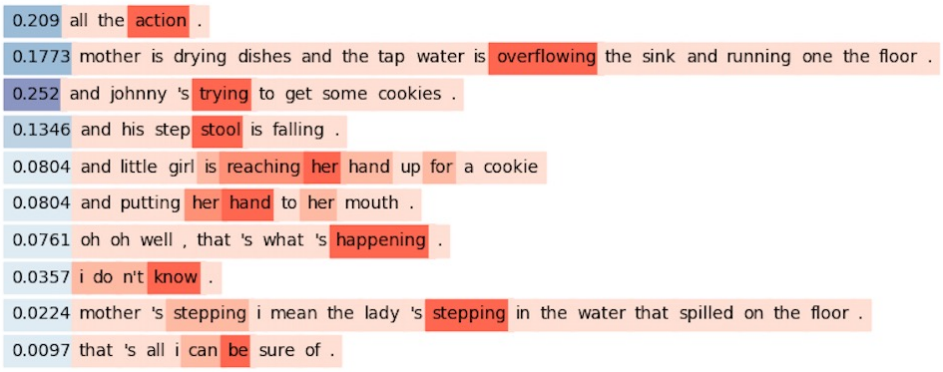


Figure 6.2: Test accuracy by varying training data proportions.

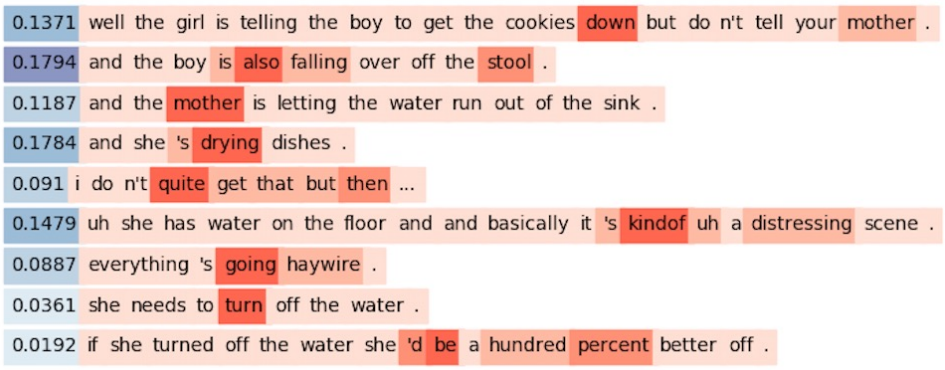
the set of information units defined by human experts for the Cookie Theft picture. To analyze how much information captured by the attention mechanism overlaps with the human defined information units, we performed further quantitative analysis.

In particular, we performed a statistical test to investigate if HAN pays more attention to information unit words, compared to other words. In order to do this, we considered two categories to which every word token[1] in our dataset belongs to: (i) the word is either in the set of information unit words or not (ii) the word is the most attended in its sentence or not. We then went through all the word tokens in the dataset and counted the frequencies of these two categories. Table 6.3 shows the resulting contingency table. Now the χ^2 test can tell us whether the two categories are dependent on each other. More technically, it can tell us whether there is a

[1] A word token is a specific occurrence of a word type in a text, for instance the text "the boy is telling the girl but the girl is not listening" contains 8 word types and 12 word tokens.



(a) Sample id: 059-2 Diagnosis: control Prediction: control



(b) Sample id: 007-3 Diagnosis: AD Prediction: AD

Figure 6.3: Visualization of attention.

statistical significant difference between the expected frequencies (in parenthesis) and the observed frequencies in the two categories.

Table 6.3: Contingency table (numbers in parenthesis are expectation values).

	Most emphasized	**Not most emphasized**	**Total**
Information unit	1481 (823)	7599 (8257)	9080
Non information unit	4889 (5547)	56270 (55612)	61159
Total	6370	63869	70239

The result $\chi^2 = 663, p < 0.00001$ shows that the two categories are dependent on each other, i.e., information unit does affect the attention level, with the number of information unit words being the most emphasized (1481), being much bigger than its expectation value (823). So HAN appears to be able to capture similar information to the one specified by human experts.

Now an interesting question that is still open is whether the attention model is uniformly paying more attention to all the information unit words or it is focusing on a specific subset of the information unit words. To answer this question, we define and compute the *attention frequency* and the *random frequency* for each of the 20 human-defined information units. More specifically, for an information unit word, the *attention frequency* was computed as the number of times it was the word with the highest word attention weight in a sentence. Let S_w denote the set of all sentences containing word w and $weight(c, s)$ be the attention weight of word token c in sentence s, we can formalize the computation of *attention frequency* for word type w as

$$Attention\text{-}Frequency(w) = \sum_{s \in S_w} \mathcal{I}[w = \arg\max_c weight(c, s)],$$

where $\mathcal{I}$ is an indicator function.

In contrast, the *random frequency* was computed as the expected number of times the word would have the highest word attention weight, if weights were assigned randomly within each sentence. Therefore it is defined as follows:

$$Random\text{-}Frequency(w) = \sum_{s \in S_w} \frac{1}{|s|},$$

where S_w denotes the set of all sentences containing word w and $|s|$ is the length of the sentence. The rationale is that if attention weights are assigned at random, a word in a sentence will have the highest attention with probability $1/|s|$.

In Figure 6.4, the x-axis are 20 human defined information units and y-axis shows their respective frequencies. The results indicate that the model does not attend to all information unit words uniformly. It strongly attends to words like `woman, window, stool, sink, water, wash, cookie, exterior`

and `plate`, but pays less attention to words like `dishes, boy, girl`, etc. than what would be expected by their random appearance. Currently, we do not have a satisfactory explanation for why the word attention model is attending more to that specific subset of information unit words.

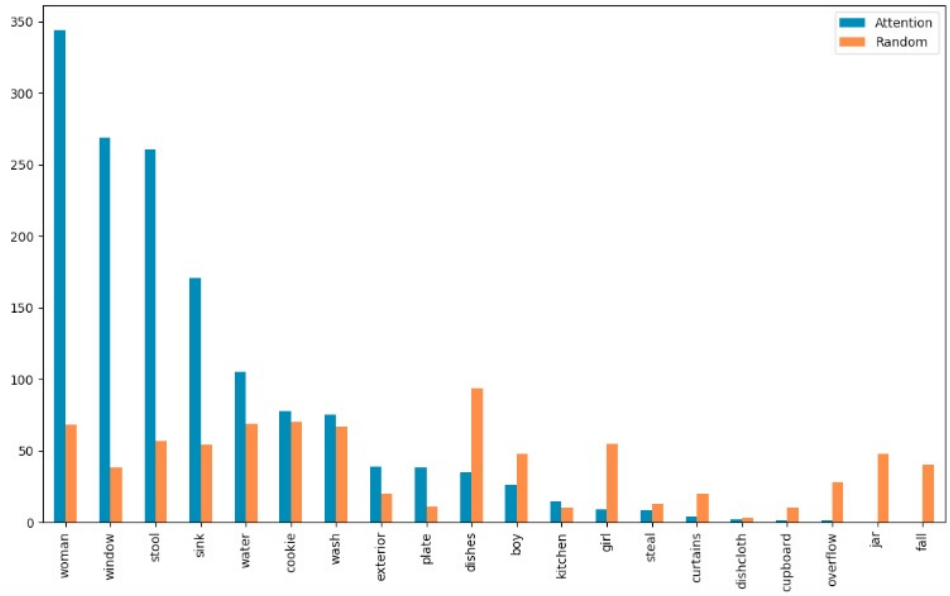


Figure 6.4: Attention frequency vs. random frequency.

6.2.5 Evaluation on the Blog Corpus

We evaluate HAN on the Dementia Blog Corpus to see how it performs on written language. The Dementia Blog Corpus was created by Masrani et al. [29] by collecting blog posts written by authors with and without dementia. In particular, they scraped the text of 2805 posts from 6 public blogs up to April 4th, 2017. Three blogs were written by dementia patients, and three written by family members of dementia patients were used as control. There are a total of 1654 samples written by persons with dementia and 1151 from healthy controls. Table 6.4 summarizes statistics of the Dementia Blog dataset.

We compare our model to Masrani et al. [29]'s, which built and tested traditional models for predicting dementia on the blog dataset using only the linguistic

Table 6.4: Blog information as of April 4th, 2017.

URL (http://*.blogspot.ca)	**Posts**	**Mean Words**	**Diagnosis**	**Gender/Age**
living-with-alzheimers	344	263.03 (s=140.28)	AD	M, 72 (approx)
creatingmemories	618	242.22 (s=169.42)	AD	F, 61
parkblog-silverfox	692	393.21 (s=181.54)	Lewy Body	M, 65
journeywithdementia	201	803.91 (s=548.34)	Control	F, unknown
earlyonset	452	615.11 (s=206.72)	Control	F, unknown
helpparentsagewell	498	227.12 (s=209.17)	Control	F, unknown

features, as shown in Table 6.1. We used 9-fold cross validation as in [29], where each test fold contains all posts from one dementia blog and one control blog, and the posts from the remaining four blogs were used in the training fold. The model selection process was carried out as described in section 6.2.1.

Table 6.5: Binary classification with 9-fold cross-validation on blog corpus.

Model	**Accuracy**	**F-score**
Majority class	0.590	0.742
Masrani et al. [29]	0.724	0.785
HAN	0.579	0.582

The experiment results are summarized in Table 6.5. The traditional model demonstrates that dementia can also be automatically predicted from written text in the form of blog posts. However, the HAN model fails in this task. A key difference that may explain this result, is that the samples in DementiaBank are descriptions of one single picture and so are all about the same topic (i.e., same objects and events, resulting in a corpus vocabulary of 1828 word types). In contrast, samples from the blog data cover a large variety of topics, ranging from regular medical appointments to re-connecting an old friend on Facebook (with a much larger vocabulary size of 27413). The HAN model succeeded in focusing on informative concepts shown in the Cookie Theft picture, with the help of the attention layers. However, for blog data there are no such concepts shared across all blog posts. Thus the data are likely not sufficient to cover such a much larger vocabulary, resulting in the extremely poor performance of HAN. On the contrary, the

traditional machine learning method is quite effective on blog posts, likely because its large human engineered set of features also include features that are not lexically based (i.e., based on words), but instead capture task-independent aspects of language like syntactic constituents and syntactic complexity.

To further explore the large difference in performance between neural and traditional methods on blog data, we conducted an additional experiment. Unlike the original split setting where all posts from the same blog are contained either in the training fold or the test fold, here we shuffle all the posts regardless which blogs they belong to, and divide them into 10 folds for cross validation. In this scenario, posts from the same blog will very likely appear in both the training and testing data, creating a form of data contamination. Not surprisingly, the HAN model is very accurate on this artificial task, with an average accuracy and F-score as high as 0.934 and 0.944, respectively. This could be because HAN captures the writing style and topics of each blogger rather than informative patterns for dementia prediction.

6.3 Discussion

We extend previous work based on traditional machine learning methods and engineered features, by applying a neural model on language samples of elderly people to classify dementia patients from healthy controls. When not including the demographic feature (age), HAN matches the performance of the best model without age. By incorporating age as extra information, the model not only achieves the state-of-the-art performance on the DementiaBank dataset, but can give a decent prediction accuracy even when trained with a small portion of the available data. Visualization and statistical analysis reveal that the attention mechanism of the model manages to capture similar key concepts as the information unit features specified by human experts. Meanwhile, the blog experiment results indicate that HAN is not a universal classifier for predicting dementia from language. In the task where samples are not all about a single topic, a traditional model that exploits linguistic features (e.g., *syntactic complexity, context-free grammar rules*) is a better choice than HAN.

Chapter 7

Conclusions and Future Work

Early prediction of dementia is extremely important, as researchers believe that early diagnosis will be key to slowing and stopping the disease. Currently, a diagnosis is based on clinical expertise and cognitive screening tests, which have limited accuracy in earlier stages of disease, or invasive and resource-intensive testing, such as lumbar puncture or specialized neuroimaging. In this study, we tackled the problem of predicting dementia from language. In particular, we explored neural network models in the direction of avoiding any task-specific features, which could be easily generalized to other language datasets of dementia. This thesis has made three main contributions towards this effort.

First, we proposed to use a joint embedding approach to combine two multimodal task-agnostic feature groups, i.e., linguistic and acoustic features. The experiment results on the DementiaBank dataset showed that our models using the pairwise ranking scheme give performances comparable to baseline models using feature selection.

Secondly, we proposed a novel feature about discourse coherence, which is also task-agnostic for dementia prediction. Unlike previous linguistic features that tried to detect language differences at a lexical and syntactic level, the new coherence feature aimed at capturing the language changes caused by AD in a higher level. We applied a neural coherence model [32] based on the Centering theory to generate coherence scores for the DementiaBank dataset. The logistic regression classifier using the coherence score as the only feature outperformed the majority

class baseline by 4%. However, the coherence feature did not perform well when used together with other task-agnostic features. Our analysis indicated that AD patients and healthy controls do not show much difference with respect to this type of coherence, on the task of describing the Cookie Theft picture, possibly because such task does not seem to require mentioning the same entity repeatedly too many times.

Lastly, we applied Hierarchical Attention Networks (HAN) framework [38] for dementia prediction, which does not require any feature engineering. Our experiments on the DementiaBank dataset showed that HAN obtained comparable results to traditional models that use task-specific features, and the modified HAN-AGE model achieved new state-of-the-art classification performance. In experiments of attention mechanism analysis, we found that the words emphasized by the attention model overlapped but differ from the information units defined by human experts. Further investigation for explaining this difference is left as future work. Moreover, we evaluated the HAN model on a dementia blog dataset. Interestingly, the same neural model did not work well on this corpus of written text, suggesting that dementia prediction from language may require different methods depending on the genre of the source language.

Although our task-agnostic methods were only tested on an English dataset describing the Cookie Theft picture, it could be generalized to other cultures and languages. It would be particularly useful for applying such methods in developing countries, which have an even more pressing need for inexpensive solutions.

Currently, a key limitation of predicting dementia from language is the scarcity of related data sets. The DementiaBank dataset seems to contain sufficient data (257 AD samples and 242 controls) to train neural text categorization models like HAN. However, there are only 5 vascular dementia samples and no sample at all for other types of dementia (e.g., dementia with Lewy body). Automatic prediction of different sub-types of dementia will not be possible until more data is collected.

Moreover, one interesting area of future work would be collecting a dataset with other modalities. Specifically, in a picture description task we could record the facial expressions of participants through a camera and their eye movements through an eye tracker. Similar ideas have been explored by Fraser et al. [13] and Poria et al. [35]. With data sources containing more than just speech, we could for

instance extract new features and apply multimodal learning methods to this new dataset, and might potentially achieve even better performance than what reported in this thesis.

Bibliography

[1] A.D. International. Dementia statistics. https://www.alz.co.uk/research/statistics, 2015. Accessed: 2019-2-13. → page 1

[2] S. Ahmed, C. A. de Jager, A.-M. Haigh, and P. Garrard. Semantic processing in connected speech at a uniformly early stage of autopsy-confirmed alzheimer's disease. *Neuropsychology*, 27(1):79, 2013. → page 2

[3] S. Ahmed, A.-M. F. Haigh, C. A. de Jager, and P. Garrard. Connected speech as a marker of disease progression in autopsy-proven alzheimers disease. *Brain*, 136(12):3727–3737, 2013. → page 5

[4] S. Al-Hameed, M. Benaissa, and H. Christensen. Detecting and predicting alzheimer's disease severity in longitudinal acoustic data. In *Proceedings of the International Conference on Bioinformatics Research and Applications 2017*, pages 57–61. ACM, 2017. → page 6

[5] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014. → page 25

[6] R. Barzilay and M. Lapata. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34, 2008. → page 7

[7] 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*, 51(6):585–594, 1994. → pages 2, 5, 9

[8] L. W. Chambers, C. Bancej, and I. McDowell. *Prevalence and monetary costs of dementia in Canada: population health expert panel*. Alzheimer Society of Canada in collaboration with the Public Health Agency , 2016. → page 1

[9] 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*, 53(1):1–19, 1996. → pages 2, 6

[10] B. H. Davis. So, you had two sisters, right? functions for discourse markers in alzheimers talk. In *Alzheimer Talk, Text and Context*, pages 128–145. Springer, 2005. → pages 3, 17

[11] C. Ellis, A. Henderson, H. H. Wright, and Y. Rogalski. Global coherence during discourse production in adults: A review of the literature. *International journal of language & communication disorders*, 51(4): 359–367, 2016. → pages 3, 17

[12] K. C. Fraser, J. A. Meltzer, and F. Rudzicz. Linguistic features identify alzheimers disease in narrative speech. *Journal of Alzheimer's Disease*, 49 (2):407–422, 2016. → pages 2, 3, 6, 14, 25, 26

[13] K. C. Fraser, K. L. Fors, D. Kokkinakis, and A. Nordlund. An analysis of eye-movements during reading for the detection of mild cognitive impairment. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1016–1026, 2017. → page 36

[14] S. Gao, H. C. Hendrie, K. S. Hall, and S. Hui. The relationships between age, sex, and the incidence of dementia and alzheimer disease: a meta-analysis. *Archives of general psychiatry*, 55(9):809–815, 1998. → page 25

[15] B. J. Grosz, S. Weinstein, and A. K. Joshi. Centering: A framework for modeling the local coherence of discourse. *Computational linguistics*, 21(2): 203–225, 1995. → pages 7, 18

[16] C. Guinaudeau and M. Strube. Graph-based local coherence modeling. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 93–103, 2013. → page 7

[17] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural computation*, 16(12):2639–2664, 2004. → page 7

[18] G. Hinton, N. Srivastava, and K. Swersky. Neural networks for machine learning lecture 6a overview of mini-batch gradient descent. → page 21

[19] T.-H. K. Huang, F. Ferraro, N. Mostafazadeh, I. Misra, J. Devlin, A. Agrawal, R. Girshick, X. He, P. Kohli, D. Batra, et al. Visual storytelling. In *15th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2016)*, 2016. → page 20

[20] James Lyons. Mel frequency cepstral coefficient (mfcc) tutorial. http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/, 2013. Accessed: 2018-11-18. → page 12

[21] D. Kempler. Language changes in dementia of the alzheimer type. *Dementia and communication*, pages 98–114, 1995. → page 2

[22] R. Kiros, R. Salakhutdinov, and R. S. Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014. → pages 3, 7, 11, 13

[23] M. Laine, M. Laakso, E. Vuorinen, and J. Rinne. Coherence and informativeness of discourse in two dementia types. *Journal of Neurolinguistics*, 11(1-2):79–87, 1998. → pages 3, 17

[24] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11): 2278–2324, 1998. → pages 7, 18

[25] J. Li and E. Hovy. A model of coherence based on distributed sentence representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2039–2048, 2014. → page 7

[26] Z. Lin, H. T. Ng, and M.-Y. Kan. Automatically evaluating text coherence using discourse relations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 997–1006. Association for Computational Linguistics, 2011. → page 7

[27] Louis-Philippe Morency, Tadas Baltrusaitis. Tutorial on multimodal machine learning. https://www.cs.cmu.edu/~morency/MMML-Tutorial-ACL2017.pdf, 2017. Accessed: 2019-1-16. → page 11

[28] V. Masrani. Detecting dementia from written and spoken language. Master's thesis, University of British Columbia, 2018. → pages 2, 3, 6, 11, 12, 14, 25, 26, 27

[29] V. Masrani, G. Murray, T. Field, and G. Carenini. Detecting dementia through retrospective analysis of routine blog posts by bloggers with dementia. *BioNLP 2017*, pages 232–237, 2017. → pages 26, 32, 33

[30] F. Nensa, K. Beiderwellen, P. Heusch, and A. Wetter. Clinical applications of pet/mri: current status and future perspectives. *Diagnostic and Interventional Radiology*, 20(5):438, 2014. → page 1

[31] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng. Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 689–696, 2011. → page 6

[32] D. T. Nguyen and S. Joty. A neural local coherence model. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1320–1330, 2017. → pages ix, 3, 7, 18, 19, 20, 21, 35

[33] S. O. Orimaye, J. S.-M. Wong, and K. J. Golden. Learning predictive linguistic features for alzheimers disease and related dementias using verbal utterances. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 78–87, 2014. → page 5

[34] D. B. Paul and J. M. Baker. The design for the wall street journal-based csr corpus. In *Proceedings of the workshop on Speech and Natural Language*, pages 357–362. Association for Computational Linguistics, 1992. → page 19

[35] S. Poria, E. Cambria, N. Howard, G.-B. Huang, and A. Hussain. Fusing audio, visual and textual clues for sentiment analysis from multimodal content. *Neurocomputing*, 174:50–59, 2016. → page 36

[36] H. Posner, R. Curiel, C. Edgar, S. Hendrix, E. Liu, D. A. Loewenstein, G. Morrison, L. Shinobu, K. Wesnes, and P. D. Harvey. Outcomes assessment in clinical trials of alzheimers disease and its precursors: readying for short-term and long-term clinical trial needs. *Innovations in clinical neuroscience*, 14(1-2):22, 2017. → page 1

[37] A. Tsaptsinos. Lyrics-based music genre classification using a hierarchical attention network. *arXiv preprint arXiv:1707.04678*, 2017. → page 24

[38] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy. 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, 2016. → pages 3, 24, 25, 36

[39] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250*, 2017. → page 7

[40] L. Zhang, S. Wang, and B. Liu. Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):e1253, 2018. → page 24