Detecting Dementia from Written and Spoken Language

by

Vaden Masrani

BSc., The University of British Columbia, 2014

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)

December 2017

Abstract

This thesis makes three main contributions to existing work on the automatic detection of dementia from language. First we introduce a new set of biologically motivated *spatial neglect* features, and show their inclusion achieves a new state of the art in classifying Alzheimer's disease (AD) from recordings of patients undergoing the Boston Diagnostic Aphasia Examination. Second we demonstrate how a simple *domain adaptation* algorithm can be used to leveraging AD data to improve classification of mild cognitive impairment (MCI), a condition characterized by a slight-but-noticeable decline in cognition that does not meet the criteria for dementia, and a condition for which reliable data is scarce. Third, we investigate whether dementia can be detected from *written* rather than spoken language, and show a range of classifiers achieve a performance far above baseline. Additionally, we create a new corpus of blog posts written by authors with and without dementia and make it publicly available for future researchers.

Lay Summary

Difficulty producing language is a well known sign of early onset dementia. This has led to recent attempts to create non-invasive diagnostic tools that detect dementia from samples of a patient's language. Our work makes three main contributions to this effort. First, we suggest a new set of biologically motivated "spatial neglect" features that improve our ability to detect Alzheimer's disease from recordings of patients undergoing standard diagnostic exams. Second, we demonstrate how to use Alzheimer's data to detect Mild Cognitive Impairment, a condition for which reliable data is scarce. Last, we investigate whether dementia can be detected from written language, a more difficult task than using spoken language because writers are able to make revisions to the text. We develop a new blog post data set and show our system is able to correctly classify posts at a rate far above baseline.

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 (Point Grey campus), in collaboration with Dr. Thalia Field at the UBC Faculty of Medicine and Dr. Gabriel Murray at University of the Fraser Valley. I was the lead researcher, responsible for the coding, data preprocessing and analysis, plots, concept formation, and first drafts of the manuscripts. Dr. Giuseppe Carenini, Dr. Thalia Field and Dr. Gabriel Murray were responsible for concept formation, draft edits, interpreting the results, and suggestions for improvement. This work originally began as a class project in collaboration with Halldor Thorhallsson and Jacob Chen, both of whom contributed to the feature extraction code in Chapter 3.

Three publications came from this work. The results from *Improving Diagnostic Accuracy Of Alzheimer's Disease From Speech Analysis Using Markers Of Hemispatial Neglect* [25] appear in Chapter 4. The results from *Domain Adaptation for Detecting Mild Cognitive Impairment* [51] in Chapters 4 and 5. The results from *Detecting Dementia through Retrospective Analysis of Routine Blog Posts by Bloggers with Dementia* in Chapter 6. The central findings from each publication appear here, while the plots have all been expanded with more models and metrics for consistency across chapters. The "we" I use throughout this work refers to myself, Giuseppe Carenini, Thalia Field, and Gabriel Murray, the authors of the above publications.

Table of Contents

List of Tables

List of Figures

Glossary

ACC Accuracy

ADOD Alzheimer's disease and related dementias

AD Alzheimer's disease

ASR automatic speech recognition

AUC Area Under the Curve

BDAE Boston Diagnostic Aphasia Examination

CT Computed Tomography

DSM-5 Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition

EDU Elementary Discourse Units

FMS F-measure

KNN K-Nearest Neighbours

MCI mild cognitive impairment

MFCC Mel-frequency Cepstral Coefficient

ML machine learning

MMSE Mini Mental State Examination

MNCD Mild Neurocognitive Disorder

MRI Magnetic Resonance Imaging

NLP Natural Language Processing

OPTIMA Oxford Project to Investigate Memory and Aging

SVM Support Vector Machine

Acknowledgments

Only convention prevents me from listing multiple names on the title page, as this thesis is not the product of one person. Having Giuseppe Carenini as a supervisor felt like cheating; he was generous with his time and with his experience, and was sincerely interested in seeing his students succeed. It is rare to find such kindness coupled with technical expertise, and I can only hope to emulate those qualities in the future.

I must also express my gratitude to Thalia Field and Gabriel Murray, who both played pivotal roles in this research. Thalia, for your vast medical knowledge and for your careful comments to multiple drafts I sent your way. Gabe, for your technical help and for making time to be a second reader of this work. It's been a pleasure and a privilege publishing with you both.

To Halldor Thorhallsson, Jacob Chen, Kimberly Dextras, Robbie Rolin, Louie Dinh, Meghana Venkatswamy, Giovanni Viviani, Dilan Ustek, Antoine Ponsard, Kuba Karpierz, Neil Newman, Daniel Almeida, Jordon Johnson, and the rest of my fellow graduates, thank you. Thank you for being beer-callers, brainstormers, sounding boards, project members, debate partners, mountain hikers, and friends. May you all have happy careers spent solving interesting problems.

To my dear friends in Vancouver, Calgary, Tokyo, Perth, Seattle, and Portland, you all mean the world to me. These last two years would have been gray and bland without you. We've made enough memories to last several lifetimes, and I hope we make enough to last several more.

Second to last but not second to least, my wonderful family whose emotional[1] support was unfailing over these last few years. I wouldn't be writing these words

[1] And occasionally financial.

had my parents not kindled in me a love of exploration. Curiosity is the fuel of science and they made sure to fill up the tank.

And finally, Halina, my life's longest love. We grew up together and will grow old together. You make everything better.

Dedication

To Halina, without whom I would have neither the skill nor the will to do much of anything.

Chapter 1

Introduction

Every year, Canadians spend $10.4 billion caring for persons with dementia. Alzheimer's disease (AD), which accounts for 60% - 80% of all dementia diagnoses, is projected to become a trillion dollar disease worldwide by 2018 which places it among the most financially costly diseases in developed countries [11, 59]. Although there is not yet a cure for AD, researchers believe early detection will be key to preventing, slowing, and stopping the disease [3].

Of the 47 million people who live with dementia today, only approximately 25% receive a formal diagnosis [35]. A diagnosis of dementia may involve repeated medical follow-up, interviews with patients and caregivers by trained health care professionals, and detailed cognitive assessment [3]. Blood tests and neuroimaging, which can be distressing to elderly patients, are also often used to rule out other causes of dementia-like symptoms. 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 [35]. What is needed is a diagnostic tool which is non-invasive, inexpensive and easy to administer, so patients in developed and developing countries can receive care and plan for their future, as well as to make lifestyle choices that can slow the progression of the disease [3].

One promising avenue, and the focus of this thesis, is to develop an automated tool that makes a diagnosis by detecting changes in language. While AD is characterized by a decline in many cognitive functions, including "impairments in atten-

tion/concentration, orientation, judgment, visuospatial abilities, executive function, and language," [24] dysphasia[1] has been suggested as being more significant than other symptoms due to its correlation with a decline in noncognitive skills such as hygiene, dressing and eating [66]. Those with AD often have a number of linguistic deficits, including:

- Difficulties finding words
- Diminished vocabularies
- A difficulty recalling the names of everyday objects ("anomia")
- A tendency to speak with repetitions ("echolalia")
- A difficulty producing sounds, syllables and words ("verbal apraxia")

Given the importance of dysphasia in detecting early signs of dementia, researchers are applying advances in machine learning (ML) and Natural Language Processing (NLP) to develop a tool that identify dementia based only on a sample of a patient's speech. Such a tool would assist clinicians in making a diagnosis and hopefully would obviate the need for more invasive screening techniques. Further, it could be easily distributed to developing countries via a mobile phone application. Previous work in this area has shown positive preliminary results using language to distinguish between patients with and without dementia, as well as between subtypes of dementia (e.g. AD with and without additional vascular pathology), but most previous work has been limited by small data sets and has focused only on one form of language production, namely spoken language [27, 61]. This work builds upon previous research on speech analysis, and makes three main contributions which are detailed below.

1.1 Contributions

1.1.1 A Novel Feature Set: Spatial Neglect

DementiaBank is a well studied data set of patients undergoing the "Cookie Theft Picture Description Task" component of the Boston Diagnostic Aphasia Exami-

[1] The loss of the ability to produce or understand language.

nation (BDAE). Participants are asked to describe the cartoon image seen in Figure 3.2 and their responses are recorded and manually transcribed. We introduce a new set of features that measure whether a respondent is more perceptive on one side of their visual field than the other, a condition known as spatial neglect. These "spatial neglect" features are effective and simple to extract and we show their inclusion achieves a new state of the art in detecting AD from speech. This study also considered two variations of the spatial neglect features, as well as discourse features, and show they have little or no effect across a range of models trained on the DementiaBank data set.

1.1.2 Domain Adaptation: Using Alzheimer's Data To Diagnose Mild Cognitive Impairment

Training a ML classifier to identify mild cognitive impairment (MCI) is difficult both because those patients with MCI are less symptomatic than those with AD, and because there is less training data available. We show how the available AD data can be leveraged to improve classification accuracy for MCI using domain adaptation techniques. We compare two simple domain adaptation algorithms, AUGMENT and CORAL, and show that only AUGMENT is an effective way to improve the performance of the best classifier in detecting MCI.

1.1.3 Written Language: A New Corpus And Demonstration Of Viability

Most research on automatically detecting dementia from language has been focused on spoken language, but little work has been done on *written* language. Analyzing written language is difficult because the author has the opportunity to delete mistakes and make revisions to the text, as well as receive third party assistance. Depending on the source (e.g. blogs, email, twitter), the author may not be constrained to a single topic as are the participants of the BDAE. Additionally, acoustic and test-specific features cannot be extracted from blog posts as they can from data collected from standardized test.

Despite these difficulties there will be a lot of written data available in the future as a greater number of seniors begin using the internet. We show that a

range of models can determine whether the author of a blog post has dementia at a rate far above baselines. We create a new corpus of blog posts written by either persons diagnosed with dementia or caregivers of persons with dementia, and make it publicly available for further research.

1.2 Reproducibility

The code to reproduce all results and the corresponding plots is available at: https://github.com/vadmas/dementia_classifier and the blog corpus is available at: https://github.com/vadmas/blog_corpus.

1.3 Thesis Overview

To improve readability the three contributions are each given their own chapters (4, 5, 6, respectively) with results concluding each. Chapter 2 goes into the background knowledge required to understand the remainder of the thesis. Specifically we focus of Alzheimer's disease (AD), the most common form of dementia, and mild cognitive impairment (MCI), a condition that often precedes dementia and is the focus of Chapter 5. We also discuss previous work that has been done on automatic detection of dementia from speech. Section 3 discusses the methodology common to each of the experiments. Any deviations from the general methodology are discussed within each chapter. Most of the results, and some of the text and figures, have appeared in three publications which were published over the duration of this thesis: *Domain Adaptation for Detecting Mild Cognitive Impairment* [51], *Detecting Dementia through Retrospective Analysis of Routine Blog Posts by Bloggers with Dementia* [50], and *Improving Diagnostic Accuracy Of Alzheimer's Disease From Speech Analysis Using Markers Of Hemispatial Neglect* [25]. We conclude with Chapter 7 which summarizes the work and outlines areas for future research.

Chapter 2

Background

In this section, we provide an overview of dementia focusing on AD and MCI, the two subtypes of dementia discussed in this thesis. We do not aim to provide a comprehensive review of the current state of medical knowledge of dementia (cf. Association [3]), but rather to provide relevant background information necessary to understand the remainder of this work. We then discuss previous work that has been done using machine learning models to classify dementia from written and spoken language. Discussions on how our research builds upon previous research is contained within Chapter 4, 5, and 6. The reader is presumed to have some familiarity with basic ML models (logistic regression, random forests, Support Vector Machines (SVMS), etc) and therefore the details of each model will not be discussed[1]. The background for domain adaptation and discourse parsing is covered in sections 5.1 and 4.2, respectively.

2.1 Medical Overview

2.1.1 Alzheimer's Disease And Other Dementias

Dementia is an umbrella term for a variety of diseases that cause a decline in cognitive ability beyond that which is expected from normal aging. Symptoms are

[1]For those unfamiliar with common ML models, an overview of the various models used in this thesis is here: https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/

often gradual, irreversible, and significant enough to affect daily functioning. To be classified as a major neurocognitive disorder[2] by DSM-5, a patient must show a "significant cognitive decline from a previous level of performance in one or more cognitive domains":

- Learning and memory
- Language
- Executive function
- Complex attention
- Perceptual-motor
- Social cognition

and:

a) The cognitive deficits interfere with independence in everyday activities

b) The cognitive deficits do not occur exclusively in the context of a delirium

c) The cognitive deficits are not better explained by another mental disorder (eg. depression, schizophrenia) [4]

Language impairment is common to all dementias, although the dysphasia may manifest itself differently depending on the underlying pathology [41, 74]. Table 2.1 (reproduced from [41]) lists the key speech and language impairments that characterize each subtype of dementia.

Symptoms intensify as the disease progresses. In the early stages, a person may have difficulty performing chores around the house and may become noticeably more forgetful, needing prompting to take pills or do other routine daily activities. The individual may also demonstrate personality and mood changes, as well as have difficulty finding words to express themselves. These symptoms worsen through the middle and late stages of the disease until the person is unable to live

[2]Dementia was renamed "major neurocognitive disorder in the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5)"

Types of Dementia	The key speech and language impairments in the early stages of dementia
Alzheimer's disease	- Finding the right word for objects - Naming the objects - Word comprehension - Loud voice
Vascular dementia	- Finding the right word for objects - Naming the objects - Word comprehension - Incomprehensible speech - Decreased complexity
Dementia with Lewy Bodies	- Language disorders include both the symptoms of AD and PDD
Parkinson's disease dementia	- Non-articulated speech - Loss of verbal fluency - Non-grammatical sentences - Slow speech - Soft voice
Frontotemporal dementia, Progressive non-fluent aphasia	- Slow and hesitant speech - Grammatical mistakes - Worsened understanding of complex sentences - Finding the right word for objects - Loss of literacy skills such as reading and writing
Semantic dementia	- Finding the right word for objects - Naming the objects - Word comprehension - A lack of vocabulary - Surface loss of literacy skills
Mixed dementia	- Language disorders include the symptoms of AD, vascular dementia and DLB, or just a combination of two of them

Table 2.1: Speech and language impairments in the individual types of dementia. Table replicated from Klimova and Kuca [41].

without assistance. Late stage dementia is characterized by almost complete aphasia, severe memory loss, and a total reliance on care providers.

Alzheimer's Disease is the most common cause of dementia, accounting for 60% to 80% of cases, while vascular dementia, which is caused by disease or injury to the brain that impedes blood flow, is the second most common, accounting for 10% of known cases. One in nine people aged 65 and older have AD, while about one third of people age 85 and older have AD [3]. Less common forms of include dementia with Lewy bodies, Parkinsons disease, frontotemporal dementia, and Creutzfeldt-Jakob disease. Characteristics of these dementias are summarized in Table 2.2 (reproduced from Kumar et al. [43]).

At present, there is no single test to diagnose dementia. Physicians rely on clin-

Types of Dementia	Characteristics
Alzheimers disease	- Most common type of dementia; accounts for 60 to 80 percent of cases. - Difficulty remembering names and recent events is often an early clinical symptom; apathy and depression are also often early symptoms. - Later symptoms include impaired judgment, disorientation, confusion, behaviour changes, and trouble in speaking, swallowing and walking. - Hallmark abnormalities are deposits of the protein fragment beta-amyloid (plaques) and twisted strands of the protein tau (tangles).
Vascular dementia (also known as multi-infarct dementia or vascular cognitive impairment)	- Considered the second most common type of dementia. - Impairment is caused by decreased blood flow to parts of the brain, often due to a series of small strokes that block arteries. - Symptoms often overlap with those of Alzheimers, although memory may not be as seriously affected.
Mixed type	Characterized by the presence of the hallmark abnormalities of Alzheimers and another type of dementia, most commonly vascular dementia, but also other types, such as dementia with Lewy bodies.
Dementia with Lewy bodies	- Pattern of decline may be similar to Alzheimers, including problems with memory and judgment and behaviour changes. - Alertness and severity of cognitive symptoms may fluctuate daily. - Visual hallucinations, muscle rigidity and tremors are common. - Hallmarks include Lewy bodies (abnormal deposits of the protein alpha-synuclein) that form inside nerve cells in the brain.
Parkinsons disease	- Many people who have Parkinsons disease develop dementia in the later stages of the disease. - The hallmark abnormality is Lewy bodies (abnormal deposits of the protein alpha-synuclein) that form inside nerve cells in the brain.
frontotemporal dementia	- Involves damage to brain cells, especially in the front and side regions of the brain. - Typical symptoms include changes in personality and behaviour and difficulty with language. - No distinguishing microscopic abnormality is linked to all cases. - Picks disease, characterized by Picks bodies, is one type of front temporal dementia.
Creutzfeldt-Jakob disease	- Rapidly fatal disorder that impairs memory and coordination and causes behaviour changes. - Variant Creutzfeldt-Jakob disease is believed to be caused by consumption of products from cattle affected by mad cow disease. - Caused by the misfolding of prion protein throughout the brain.

Table 2.2: Major types of dementia and their characteristics. Table replicated from Kumar et al. [43].

ical interviews with family members and the patient to see if they meet the criteria enumerated in the DSM-5. Magnetic Resonance Imaging (MRI) and Computed Tomography (CT) are often used to rule out treatable or non-dementia pathologies (e.g. brain tumours or cerebrovascular disease) that can cause cognitive decline. Neuroimaging is also used to search for biological markers (e.g. cerebral atrophy or reduced glucose metabolism in the fronto-temporo-parietal and cingulate

cortices) which can add evidence towards a diagnosis [23, 33, 60].

Clinicians also rely on the Mini Mental State Examination (MMSE), a 30-point questionnaire that measures impairment across five cognitive functions: orientation, registration, attention and calculation, recall, and language. [67]. Of a total 30 possible points, a score between 20-24 indicates mild dementia, 13-20 moderate, and $\leq$ 12 severe dementia. Language impairment is assessed by asking the participant to recall the names of a watch and pencil, read and repeat a phrase (separate tasks), and follow a three-stage command [45]. Fine-grained assessment of impairment across different language domains is not included, despite evidence that AD causes semantic, pragmatic, syntactic, and phonological language deficits[3]. Automated analysis of language has been suggested as a promising approach to detecting linguistic impairment across multiple linguistic domains [73].

There are no known cures for dementia, although medication can be used to treat symptoms or slow its progress. Commonly used medication for AD includes donepezil, an acetylcholinesterase inhibitor which increases the concentration of acetylcholine[4] in the brain, and memantine, which targets the glutamatergic system by binding with NMDA receptors to reduce toxicity associated with excessive glutamate[5][69]. Both donepezil and memantine are approved for moderate and late stage AD and provides modest improvement in cognitive function [5, 56, 75]. Other drugs are available [6] for various stages of AD but all are palliative in nature.

2.1.2 Mild Cognitive Impairment

MCI is a defined as a noticeable decline in cognitive function that may - but crucially, may not - lead to an eventual dementia diagnosis. Individuals with MCI (reclassified as Mild Neurocognitive Disorder (MNCD)[7] in the DSM-5 [71]) show

[3]This is likely due to its age. The MMSE was developed in 1975, before much of the research into the effect of AD on language impairment was conducted [73]

[4]Acetylcholine is a neurotransmitter associated with attention and memory.

[5]Glutamate is important neurotransmitter in the brain involved in learning and memory. AD causes an excessive buildup of Glutamate in the brain which kills glutamate receptors (or "NMDA receptors") by overexposure. Memantine and other NMDA receptor antagonists reduce toxicity by binding with NMDA receptors to reduce their exposure to excess glutamate.

[6]See http://www.alzheimer.ca/en/Home/About-dementia/Treatment-options/Drugs-approved-for-Alzheimers-disease

[7]The central difference between MNCD and MCI is that research into MCI mainly involved a cohort

cognitive impairment beyond that which is expected for their age, but which is less severe than dementia and does not significantly interfere with daily activities [57, 71]. Unlike with AD and related dementias, the individual retains the ability to perform functional tasks (e.g. hygiene, eating) but may become slower or less efficient at performing everyday tasks. A patient story from Langa and Levine [47] reads:

> *Mrs J, age 81 years, with hypertension and hyperlipidemia, requested a referral to a neurologist, stating: "I am forgetting things I just heard".*
>
> *Mrs J and her husband began noticing mild memory problems 1.5 years earlier, and report slow progression since. Her husband noticed changes in problem solving and time management. Mrs J was easily distracted and had difficulty remembering recent conversations. She misplaced objects and spent time looking for them; she read and wrote less than before. She repeatedly asked how to do things on her computer and cell phone. Her husband reported that she exhibited no initiative, and that their home seemed more disorganized. She had difficulty planning dinner and her cooking was simpler. Both denied changes in language or speech. She continued to drive locally without accidents but had difficulty remembering directions to familiar places. Mrs J had no hallucinations or delusions. She slept well, her mood was fine, and she exhibited no behavioral problems or personality changes.*
>
> *Functionally, she remained independent in all activities of daily living (ADLs). She had urinary frequency and over the past couple of months she had a few incidents of incontinence, especially when awakening from a nap. In instrumental activities of daily living (IADLs), Mr J had recently taken over paying bills. Finally, even with a compartmentalized pill-box, she occasionally forgot to take her medications (amlodipine 5 mg daily; losartan 50 mg twice daily; and ergocalciferol 1,000 units daily.)*

of elderly patients while MNCD included all age groups [71].

Population-based studies estimate MCI to be prevalent in between 12-18% in people over the age of 60 [58]. Annually, 8 to 15% of those with MCI will progress to dementia while the rest will either revert to normal cognition or will stay mildly impaired [58]. MCI can be due to neurodegenerative diseases (most commonly AD) or reversible causes, including psychiatric illness or metabolic disturbances including thyroid disease or vitamin B12 deficiency [32].

MCI has no single cause and no single treatment. Early diagnosis is important because it allows patients to test for potentially treatable causes (e.g. major depressions or vascular risk factors) and make modifications to their lifestyle to slow the onset of dementia [58]. As well, early diagnosis can lead to confirmatory diagnostic testing for conditions such as AD, and can better allow for planning for social supports and closer medical follow-up. Unsurprisingly, MCI is more difficult to detect than dementia since the symptoms are less severe; the MMSE has a sensitivity of 88.3% (95% CI, 81.3 to 92.9) and a specificity of 86.2% (95% CI, 81.8 to 89.7) for detecting dementia, but only a sensitivity of 45 to 60% and specificity of 65 to 90% for detecting MCI [47, 49]. In addition, MCI has not been studied as extensively as AD and therefore there is less clinical data available with which to train a machine learning model.

2.2 Automatic Detection Of Dementia

With advances in ML models and NLP an interest has emerged in training machine learning models to automatically detect Alzheimer's disease and related dementias (ADOD) from language. We discuss previous work done on both spoken and written text below, and, in chapters 4, 5, and 6, comment on how this research differs from previous work.

2.2.1 Spoken Language

There has been success in using lexical and acoustic features derived from speech to diagnose ADOD. Ahmed et al. [2] determined features that could be used to identify dementia from speech, using data collected in the Oxford Project to Investigate Memory and Aging (OPTIMA) study. These researchers used a British cohort of 30 participants, 15 with AD at either MCI or mild stages, and 15 whose

age and education matched healthy controls. They found that language progressively deteriorates as AD progresses and suggested using semantic, lexical content, and syntactic complexity features to identify cases.

Rentoumi et al. [61] then used a Gaussian naive Bayes Classifier with lexical and syntactic features to distinguish between two subtypes of dementia; AD with and without additional vascular pathology. They achieved a classification accuracy of 75% on 36 transcripts from the OPTIMA data set.

Orimaye et al. [54] expanded on this work by using a similar feature set to distinguish between ADOD and healthy patients. They performed a comparison of five machine machine learning classifiers - SVMS with a RBF kernel, Naive Bayes, Decision trees, Neural Networks, and Bayesian Networks - on the larger DementiaBank data set (sample size = 484) and found SVMS had the best performance with a F-measure score of 74% [54].

In 2014 Fraser et al. [26] compared different feature sets that could be used in discriminating between three different types of primary progressive aphasia (a subtype of frontotemporal dementia). They concluded that a smaller relevant subset of features achieves better classification accuracy than using all features and highlighted the importance of a feature selection step. They also showed how psycholinguistic features, such as frequency and familiarity, were useful in detecting primary progressive aphasia. In later work Fraser et al. [27] used logistic regression to achieve state-of-the-art of 81.92% in distinguishing individuals with AD from those without. Their experiments were run on the DementiaBank data set described in Section 3.1.1 and they found optimal performance when 35-50 features are used, consistent with their previous work [26].

Researchers have also looked at automatically detecting MCI, a harder task than detecting AD - both because limited data is available and because MCI is less symptomatic than AD. Roark et al. [63] demonstrated the viability of this task using transcripts and audio recordings of patients undergoing the Wechsler Logical Memory I/II test. This test involves a patient twice retelling a short story, once immediately after the story was told and again after a 30-minute delay. Roark et al. [63] extracted two broad sets of features; "linguistic complexity" features that measure the complexity of a narrative, and "speech duration" features that include number of pauses, pause length and pause-to-speech ratio. Using SVM's,

they achieved a maximum area under the ROC curve (AUC) of 0.74 and concluded that NLP techniques could be used to automatically derive measures to discriminate between healthy and MCI subjects. Tóth et al. [76] made a step towards fully automated detection of MCI by determining the effect of automatic speech recognition (ASR) compared to manual transcriptions. They showed classification results worsen slightly when using ASR, although they still achieved an F-measure of 85.3 on Hungarian patients. König et al. [42] found similar results using their end-to-end system on a cohort of French speaking seniors, and positive results have also been found with Greek speakers [65].

2.2.2 Written Language

Early signs of dementia can be detected through analysis of writing samples as well [40, 48, 62]. In the "Nun Study" researchers analyzed autobiographies written in the US by members of the School Sisters of Notre Dame between 1931-1996 [68]. Those nuns who met the criteria for dementia had lower grammatical complexity scores and lower "idea density" in their autobiographies. Surprisingly, the measure of idea density in autobiographies written by nuns in their 20's was predictive of dementia in late life [40].

Le et al. [48] performed a longitudinal analysis of the writing styles of three novelists: Iris Murdoch who died with AD, Agatha Christie (suspected of having AD), and P.D. James (normal brain aging). Measurements of syntactic and lexical complexity were made from 51 novels that collectively spanned the authors' careers. Murdoch and Christie exhibited evidence of linguistic decline in later works, such as vocabulary loss, increased repetition, and a deficit of noun tokens [48].

Hirst and Wei Feng [31] studied the question of whether a model trained to recognize authorship would recognize text written by an author late in their life if that author had dementia. They used "authorship attribution" and "authorship verification"[8] methods on the works of Iris Murdoch and Agatha Christie (AD) and P.D. James (control). They hypothesized than in the case of the AD authors, an SVM classifier would not be able to attribute (or verify) text written in the late

[8] Authorship attribution is a multi-label classification problem: Given a set of authors and an unknown text, determine the author. Authorship verification is a binary classification: Given an author and a text, determine if the text was written by the author.

stage of the authors career to the author because of changes in writing style due to dementia. Their results were inconclusive as they found changes in the control author (P.D. James) as well, indicating that authors' styles change naturally (or perhaps intentionally) as a result of age.

Despite evidence that linguistic markers found in writing samples can predict dementia, it appears that no attempts have been made to train models to classify dementia based on writing alone. To date, it is not clear whether systems that can detect dementia from spoken language will work for written language, given that audio and test-specific feature groups are not available from unstructured written text as they are from data collected from patient undergoing standard diagnostic exams.

Chapter 3

Methodology

In this section we detail the experimental methodology common to all subsequent experiments. Any deviations from the general procedure described below (such as the blog data set used in Chapter 6) are discussed within each chapter. The full pipeline is seen in Figure 3.1.

3.1 Data Set

We use two data sets for this work: one consists of samples of spoken language and the other consists of samples of written language. We detail the DementiaBank data set, which we use in Chapter 4 and 5, below. The data set of written samples is discussed in Section 6.1.

3.1.1 Dementiabank

For spoken samples we used the DementiaBank data set, a publicly available data set that consists of transcripts and recordings of English-speaking participants describing the "Cookie Theft Picture," a component of the Boston Diagnostic Aphasia Examination [29]. A patient is asked to describe the cartoon image in Figure 3.2 and their answer is recorded and manually transcribed - including false starts, pauses, and paraphasia[1] - and then segmented into utterances. An utterance

[1] "Paraphasia" is a type of language error associated with aphasia. It is characterized by unintended syllables, words, or phrases that result from an effort to speak. Examples include "lelephone" for "telephone" or "ragon" for "wagon".

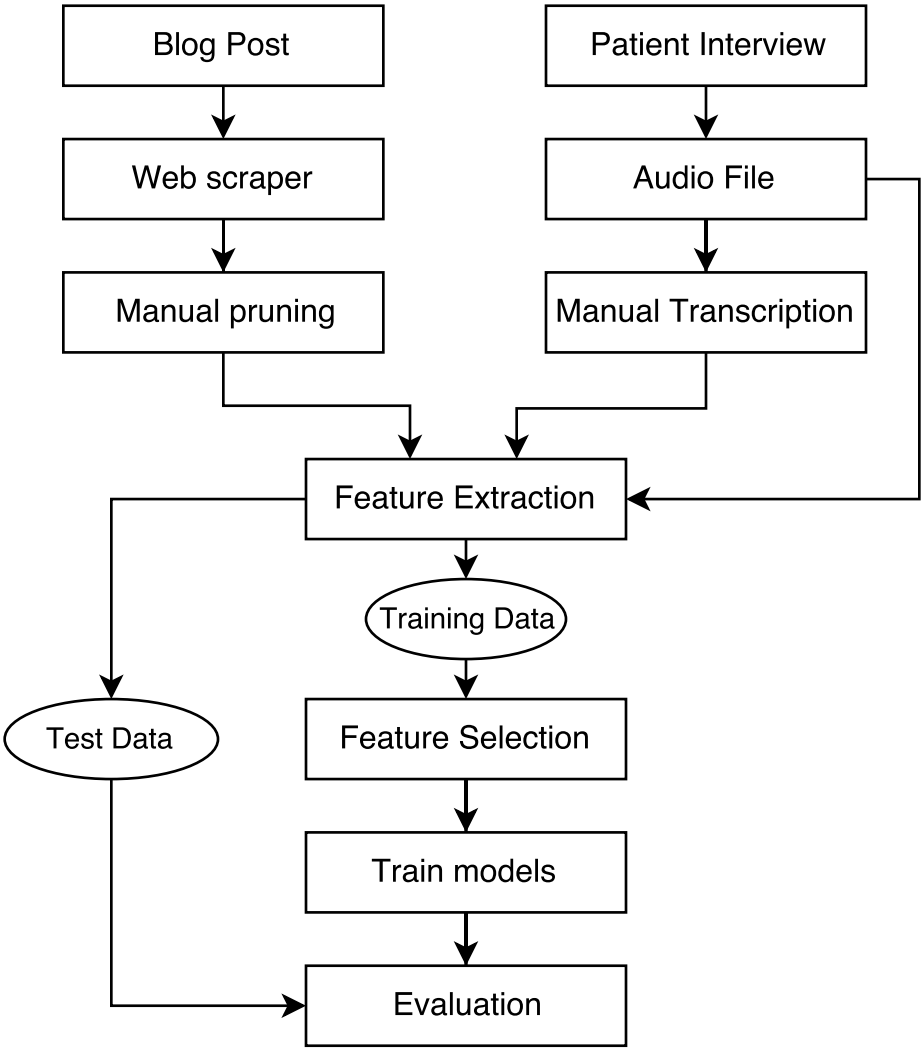

Figure 3.1: Processing pipeline from clinical interview to evaluation. We perform a 10-fold cross validation for the evaluation stage. Experiments use either the the blog data set (left) or the DementiaBank data set (right) but not both.

Diagnosis	Patients	Samples	Mean Words	Mean Age	Gender (F/M)
AD	169	257	104.98 (s=59.8)	71.72 (s=8.47)	87/170
MCI	19	43	111.09 (s=55.8)	69.39 (s=8.09)	27/16
Control	99	242	113.56 (s=58.5)	63.95 (s=9.16)	88/154

Table 3.1: Demographics of DementiaBank Dataset

is defined as a unit of speech bounded by silence. An example response is seen in Figure 3.3.

DementiaBank consists of 309 samples from 208 persons with dementia and 242 samples from 102 healthy elderly controls. A patient can give multiple interviews. Ages ranged from 45 to 90 with interviews conducted between 1983 and 1988. Medical re-review was done up to five years after the end of the study. Of the 208 persons included in the study, 181 patients were diagnosed with probable/definite Alzheimer's disease (AD) and seven with vascular dementia. Some patients were discarded due to misdiagnoses or on other clinical grounds. Of the 309 interviews with dementia patients, 43 samples were classified as mild cognitive impairment (MCI) and 256 samples as possible/probable AD. The remaining interviews were not used in this study. Demographic information about the DementiaBank samples used in this study is listed in Table 3.1.

3.2 Features

In addition to the age of the patient, which is a known predictor of dementia not leveraged in previous work [28], a total of 314 lexical and acoustic features were extracted and divided into eight groups. All features listed below have appeared in previous work, most notably from Fraser et al. [27]. A full list of all features appears in Table A.1 in the appendix.

3.2.1 Parts-Of-Speech (15)

We use the Stanford Tagger[2] to capture the frequency of various parts of speech tags (nouns, verbs, adjectives, adverbs, pronouns, determiners, and so forth). Frequency counts are normalized by the number of words in the utterance and we

[2]Available at: http://nlp.stanford.edu/software/tagger.shtml

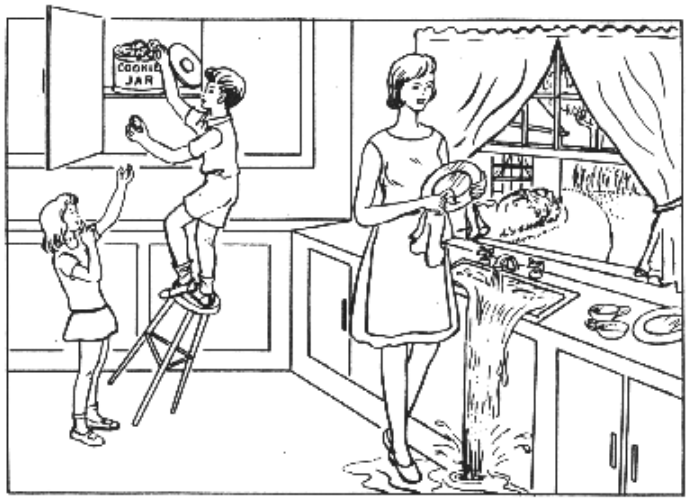

Figure 3.2: Cookie Theft picture from the Boston Diagnostic Aphasia Examination.

um. mhm. alright. there's um a young boy that's getting a cookie jar. and it he's uh in bad shape because uh the thing is falling over. and in the picture the mother is washing dishes and doesn't see it. and so is the the water is overflowing in the sink. and the dishes might get falled over if you don't fell fall over there there if you don't get it. and it there it's a picture of a kitchen window. and the curtains are very uh distinct. but the water is flow still flowing.

Figure 3.3: Manually transcribed sample response from a patient undergoing the Cookie Theft Picture Test.

record the mean across utterances. Disfluencies ("um", "er", "ah"), not-in-dictionary words of three or more letters, and word-type ratios (noun to verb, pronoun to noun, etc) were also counted.

3.2.2 Context-Free-Grammar Rules (44)

These features count how often a phrase structure rule occurs in an utterance, including NP→VP PP, NP→DT NP, etc. We use Penn Treebank tags and parse trees come from the Stanford parser.

3.2.3 Syntactic Complexity (27)

These features measure the complexity of an utterance through metrics such as the depth of the parse tree, mean length of words, mean length of sentences, mean length of T-Units, mean length of clauses, and clauses per sentence. We use the L2 Syntactic Complexity Analyzer[3].

3.2.4 Vocabulary Richness (4)

We calculated four metrics that capture the range of vocabulary in a text. The standard type-to-token ratio (the ratio of vocabulary size to text length $|V|/N$), and the moving-average type-token ratio (MATTR), which is a length-independent metric of lexical diversity [15]. We also record Brunet's index, an alternative length-independent metric of lexical diversity that has appeared in previous work, and Honore's statistic, a metric of lexical diversity based on the counting the number of singleton words appearing in a person's speech [9]

3.2.5 Psycholinguistic (5)

Psycholinguistic features are linguistic properties of words that effect word processing and learnability [64]. We used five psycholinguisic features (numbers in parenthesis indicate the number of words with scores in the database):

- **Familiarity** (3626): A measure associated with how familiar a word is ("monad" has a low score while "breakfast" has a high score).
- **Concreteness** (1372): A measures of how concrete or abstract a word is ("however" has a low score while "December" has a high score).
- **Imagability** (4829): A measure of how easily one can conjure a mental image of a word ("equanimity" has a low score and "beach" has a high score).
- **Age of acquisition** (31104): A measures of how old people are on average when the first learn the word. We use the expanded set from Kuperman et al. [44].

[3] Available at: http://www.personal.psu.edu/xxl13/downloads/l2sca.html

- **SUBTL** (74K): A measure of the frequency with which a word is used in daily life.

Scores for concreteness, familiarity, imagability, and age of acquisition were derived from surveys and crowdsourcing, and the results from multiple studies were aggregated and made publicly available[4]. Participants were asked to rate words on a numerical scale (generally between 1-5 or 1-7, depending on the study) and scores were averaged across ratings. The SUBTL word scores were not derived from crowdsourcing, but instead from television and film subtitles. Word frequencies were calculated from 51 million words across 8388 film and television episodes [8].

3.2.6 Repetitiveness (5)

We vectorized the utterances using TF-IDF and measure the cosine similarity between utterances. We then recorded the proportion of distances below three thresholds (0, 0.3, 0.5), as well as the minimum and average cosine distance.

3.2.7 Information Units (info-units) (40)

Croisile et al. [16] compiled a list of 23 items that can be discerned in the Cookie Theft Picture. These "information units" can be either actions or nouns and examples include *jar*, *cookie*, *boy*, *kitchen*, *boy taking*, and *woman drying*. For each information unit we extracted two features: a binary feature indicating whether the subject has mentioned the item (or one of its synonyms in WordNet), and a frequency count of how many times an item has been mentioned. Three info-units (e.g., "woman indifferent to the children") from Croisile et al. [16] were not included in this work due to the lack of specificity of the info-unit.

For each information unit, we used WordNet to create a set of synonyms, hypernyms and hyponyms that could be used to identify the item. We manually removed inappropriate unigrams (e.g., "irrigate" for "water"). For an information unit to be considered recognized, one of the unigrams in the set must appear in the transcript. In the case of the four action information units (*boy taking, water overflowing, mother washing, stool falling*), both unigrams must be present in a single utterance

[4] Available at: http://websites.psychology.uwa.edu.au/school/MRCDatabase/mrc2.html

and the words must be tagged with the appropriate POS tag. The list of info units and their synonyms is shown in Table 3.2.

3.2.8 Acoustic (172)

Mel-frequency Cepstral Coefficients (MFCCS) are frequently used is speech processing and represent spectral information from the speech signal, using a scale known as the "mel-frequency scale," which is chosen to mimic the way humans perceive audio. MFCCS are calculated by segmenting the signal into short frames and taking the (discrete) fourier transform of each segment. The MFCCS are then calculated from the "mel log powers" of the first 14 coefficients calculated by the fourier transform of each segment[5]. Each segment from the original signal then produces 14 MFCCS, resulting in 14 MFCC distributions. We then calculate the mean, variance, skewness, and kurtosis of the first 14 MFCCS, representing spectral information from the speech signal. We did the same for the velocity and acceleration, where velocity is calculated as the delta between consecutive time steps and acceleration as the double-deltas.

3.3 Feature selection

We follow the recommendation of Fraser et al. [26] and performed a feature selection preprocessing step. Within each training fold, we selected for inclusion the first k features that have the highest absolute correlation with the training labels. The reported performance is the maximum average across all $1 \leq k \leq D$, where D is the number of features. Figure 4.4 shows the importance of feature selection, with the maximum performance achieved for most models between 35-50 feature.

3.4 Models

We used five models from the SKLearn python package, as seen in Table 3.3. We also considering including a multilayer perceptron classifier, but it was later excluded due to badly overfitting the data (despite efforts spent hyperparameter tuning).

[5] see http://www.practicalcryptography.com/miscellaneous/machine-learning guide-mel-frequency-cepstral-coefficients-mfccs/ for more detail

Info Unit	Synonyms
Boy	boy, son, brother, male child
Girl	girl, daughter, sister, female child
Woman	woman, mom, mother, lady, parent, female, adult, grownup
Kitchen	kitchen, room
Exterior	exterior, outside, garden, yard, outdoors, backyard, driveway, path, tree, bush
Cookie	cookie, biscuit, cake, treat
Jar	jar, container, crock, pot
Stool	stool, seat, chair, ladder
Sink	sink, basin, washbasin, washbowl, washstand, tap
Plate	plate
Dishcloth	dishcloth, dishrag, rag, cloth, napkin, towel
Water	water, dishwater, liquid
Window	window, frame, glass
Cupboard	cupboard, closet, shelf
Dishes	dish, dishes, cup, cups, counter
Curtains	curtain, curtains, drape, drapes, drapery, drapery, blind, blinds, screen, screens
Steal	take, steal, taking, stealing
Fall	fall, falling, slip, slipping
Wash	wash, dry, clean, washing, drying, cleaning
Overflow	overflow, spill, overflowing, spilling

Table 3.2: A list of info units and their synonyms.

3.5 Evaluation

A 10-fold cross validation procedure was used to evaluate each model, where multiple interviews from a given patient were contained either to the training fold or the test fold, but not both. Feature selection took place within each fold: the highest mean performance was returned across all $1 \leq k \leq D$ features. Bar plots show the 90% CI across all folds. We report F-measure, accuracy, and Area Under the Curve (AUC) for most experiments except where the test set is too small to report AUC.

Model	**Hyperparameters**
logistic regression	L2 regularization, alpha = 1.0
K Nearest Neighbors	K = 5
Random Forest	Trees = 100, max depth = 3
Gaussian Naive Bayes	n/a
Support Vector Machine	kernel = 'rbf'
Dummy Classifier	Most frequent label in training fold

Table 3.3: Models and their hyperparameters.

Chapter 4

Evaluating Novel Feature Sets

In this chapter we propose and evaluate two novel feature sets: *Spatial neglect* and *discourse features*, described in sections 4.1 and 4.2 respectively. We consider three variations on the Spatial neglect features - *halves*, *strips*, and *quadrants* - and show that *halves*, the most biologically plausible variant, improves the performance of best classifier. We also show that discourse features have no effect on improving classification accuracy across a range of models.

This chapter builds the earlier research discussed in Section 2.2.1 by evaluating two new feature sets not present in prior work. Our goal in this chapter is to demonstrate the efficacy (or lack thereof) of either or both of the novel feature sets and recommend (or not) their use to future researchers. We discuss both feature sets in detail below.

4.1 Spatial Neglect

Spatial neglect (also: "hemispatial neglect," "unilateral neglect," "hemineglect," "unilateral spacial neglect") is the phenomenon of reduced awareness on one side of the visual field which often occurs as a result of brain damage. Spatial neglect differs from "hemianopia," or blindness over half the field of vision, in that a patient with neglect still has sensation and is able to, for example, detect a bright light on their neglected side. Depending on the extent of the condition, a patient with neglect may fail to notice people or large objects on one side of space, may only

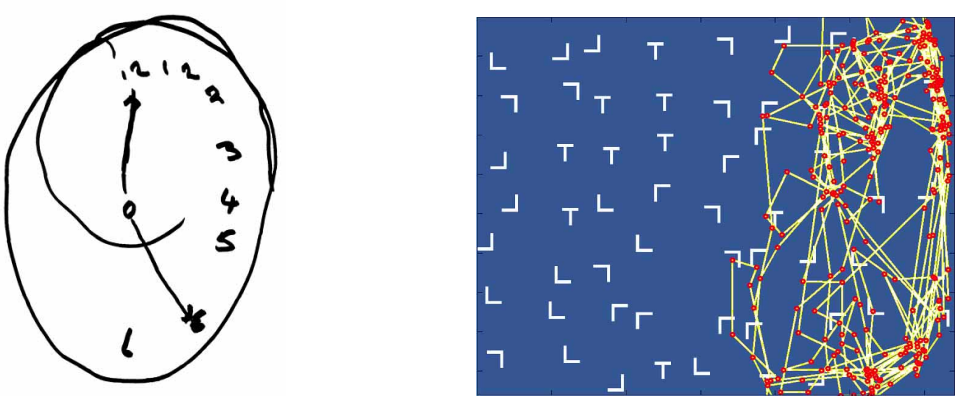

Figure 4.1: Left: A clock drawn by a patient with left-side spatial neglect. Right: Eye movements of a patient with left-side spatial neglect. Patient was asked to search for letter T among Ls. Red dots are fixations and yellow lines are saccadic movements between fixations. Images from Husain [34]

shave or apply makeup to the non-neglected side, or may only draw or examine one half of an image. The left imagine in Figure 4.1 shows a drawing of a clock by patient with spatial neglect [34, 55]. The right image in Figure 4.1 shows the eye movements of a patient with spatial neglect who was asked to identify the letter T among L's.

Previous studies have shown patients with Alzheimer's disease and related dementias (ADOD) exhibit signs of spatial neglect[1] but surprisingly, none of the previous work in automatic detection of dementia have included features which measure neglect [14, 36, 37, 52, 53, 77]. We propose four new features which measure *attention, concentration, repetition* and *perception* in different visual fields. We discuss how each feature is calculated in Section 4.1.2 and discuss the three different spatial partitions we consider in Section 4.1.1.

[1] One study by Kasai et al. [39] disputes these findings. They show that results for one measure of neglect, the "line bisection (LB) task" did not significantly correlate with the results from another measure of neglect, the "left category copying of the Rey-Osterrieth Complex Figure Task (RCFT)" and conclude that Alzheimers patients do not show left unilateral spatial neglect but instead exhibit peripheral inattention (e.g. neglect on both sides). However, their data also shows that patients with AD make "no copying errors" on the left side of the image more often than the right (Figure 3), and also show there are no significant differences between AD and control patients in the LB task anyway, so a lack of correlation may be unsurprising.

4.1.1 Spatial Partitions

We divided the Cookie Theft Image into halves, strips, and quadrants as seen in Figure 4.2. For each division we create a set of info-units contained within, shown in Table 4.1, and used these sets to calculate measures of spatial neglect. An info-unit is included in all divisions it spans (e.g., "girl" is included in SW and NW quadrants), meaning an info-unit can appear in multiple divisions. We consider each partitioning scheme to be its own feature set. Models are trained using the features described in Section 3.2 and one of the *halves*, *strips*, or *quadrant* sets. For each set we also tried adding quadratic features (e.g. $attention^2, concentration^2, attention * concentration$ etc) but in some cases the performance decreased. We report the results of the optimal performance of each feature set, with or without quadratic features.

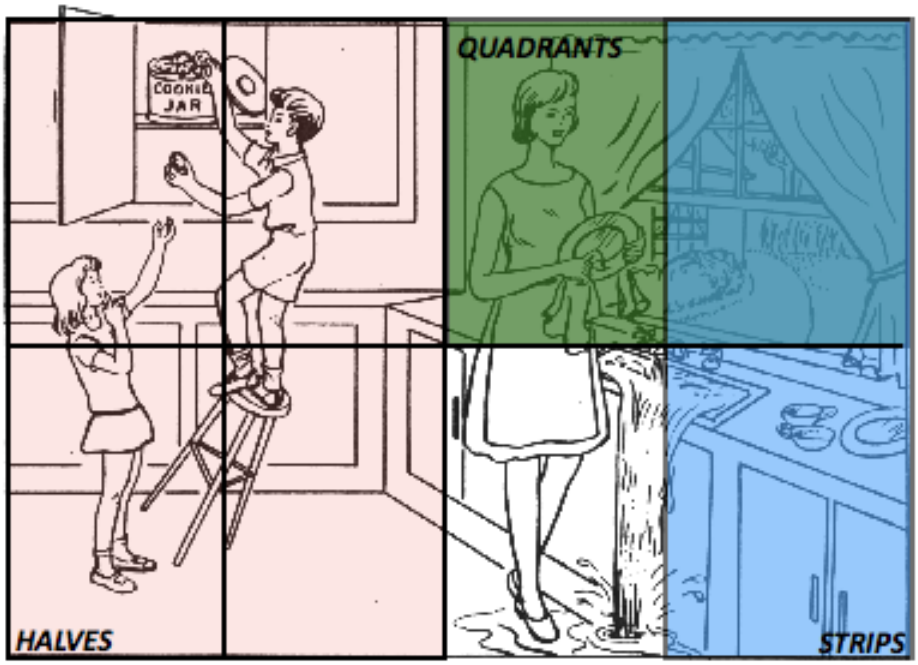

Figure 4.2: We divide the Cookie Theft image into halves (red), strips (blue), and quadrants (green), and create sets of info-unit within each division. For example, the "girl" info-unit is in the left half, far-left strip, and SW and NW quadrants.

4.1.2 Spatial Neglect Features

To measure spatial neglect, for each division we computed four simple metrics using the counts of each info-unit as described in Section 3.2.7. Let S_i be the set of info-units in division i, n_i be the count of mentions of any info-unit in S_i, u_i be the number of unique info-units mentioned in S_i, and n_{all} be the total number of words in the patient's response. Then for division i,

$$attention_i = n_i \tag{4.1}$$

$$concentration_i = \frac{n_i}{n_{all}} \tag{4.2}$$

$$repetition_i = \frac{u_i}{n_i} \tag{4.3}$$

$$perception_i = \frac{u_i}{|S_i|} \tag{4.4}$$

4.2 Discourse Features

One measure of coherence that has been absent in the previous work comes from *discourse analysis.* In a coherent passage, a reader can clearly discern how one sentence relates to the next. A given sentence may *explain* or *elaborate* upon a previous sentence (as this one is doing), or act as *background* for a future sentence. Such relations can be formed on an intra-sentential level as well, with Elementary Discourse Units (EDU) being clause-like units of text which can be related to one another by *discourse relations*. Discourse parsing is the task of segmenting a piece of text into its EDUs and then forming a *discourse tree* with edges corresponding to discourse relations, as seen in Figure 4.3. Features related to the discourse structure of a passage can then be extracted from the discourse tree, as discussed in Section 4.2.2.

Previous work has shown a disparity in the overall discourse ability of patients with ADOD compared to healthy controls [7, 12, 21]. Those with ADOD show a greater impairment in global coherence, have more disruptive topic shift and

Halves	
Left	boy, girl, cookie, jar, stool, cupboard, steal, fall, kitchen
Right	woman, exterior, sink, plate, dishcloth, water, window, dishes, curtains, wash, overflow, cupboard, kitchen
Strips	
Far-left	girl, cookie, jar, stool, cupboard, steal, kitchen, cupboard
Center-left	boy, cookie, stool, steal, fall, kitchen, cupboard
Center-right	woman, exterior, sink, plate, dishcloth, water, window, dishes, curtains, wash, overflow, kitchen, cupboard
Far-right	exterior, window, dishes, curtains, kitchen, cupboard
Quarters	
NE	woman, exterior, plate, dishcloth, wash, window, curtains, kitchen
SE	woman, sink, water, dishes, overflow, cupboard, kitchen
NW	girl, cookie, jar, cupboard, steal, boy, cookie, kitchen
SW	girl, stool, fall, cupboard, kitchen

Table 4.1: List of info-units within each division.

greater use of empty phrases, and produce fewer cohesive ties than controls [18–20, 46]. Discourse parsing has been useful in determining overall coherence in other domains such as essay scoring; thus, we hypothesized that it would also be useful for AD detection [22]. Most recently, Abdalla et al. [1] looked at differences in discourse structure between AD and controls in two data sets, DementiaBank and the Carolinas Conversations Collection [2]. They found significant differences between *elaboration* and *attribution* between the two groups, in both data sets. However, it remains unclear if the inclusion of discourse features improves the accuracy of classifiers.

4.2.1 Discourse Parser: CODRA

CODRA, or "a **CO**mplete probabilistic **D**iscriminative framework for performing **R**hetorical[3] **A**nalysis" is a discourse parser which combines discourse *segmenting*

[2] http://carolinaconversations.musc.edu/about/

[3] "Rhetorical parsing" and "Discourse parsing" are used interchangeably in the literature.

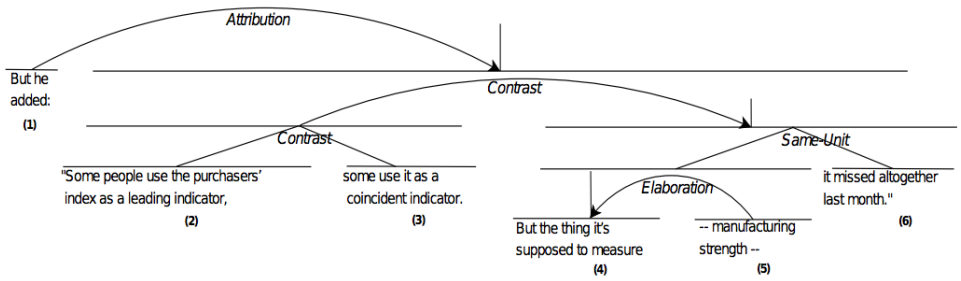

Figure 4.3: Discourse tree for the two sentences "But he added: 'Some people use the purchasers' index as a leading indicator, and some use it as a coincident indicator. But the thing it's supposed to measure - manufacturing strength - is missed altogether last month."' Each sentence contains three EDUs. EDUs correspond to leaves of the tree and discourse relations correspond to edges. (Figure adapted from [38])

- partitioning raw text into EDUs - with discourse *parsing* - the problem of forming a discourse tree from a sequence of EDUs. Most existing parsers perform structure prediction and relation labeling separately, and are therefore unable to make use of sequential dependencies between text segments. CODRA addresses this limitation by implementing a joint model for the two tasks. In addition, CODRA improves on previous discourse parsers by performing inter- and intra- sentential parsing separately with two different probabilistic models. This allows for optimal parsing using the probabilities generated by the forward-backwards algorithm on CRFs (Conditional Random Fields). CODRA significantly out -performed state-of-the-art performance on two data sets in 2015 [38].

4.2.2 Discourse Features

We used CODRA to segment the speech EDUs and identify the relations between them [38]. We counted the number of occurrences of each of the 18 discourse relations (*"attribution", "background", "cause", "comparison", "condition", "contrast", "elaboration", "enablement", "evaluation", "explanation", "joint", "manner-means", "same-unit", "summary", "temporal", "textual organization", "topic change", "topic comment"*), the depth of the discourse tree, the average number of EDUS per utterance, the ratio of each discourse relation to the total number of

discourse relations, and the discourse relation type-to-token ratio.

4.3 Experimental Design

We follow the experimental design discussed in Chapter 3 (See Figure 3.1). Unlike Fraser et al. [26] and Fraser et al. [27] we do not include features related to the number or duration of pauses in the speech. We had tried using forced alignment to determine the time intervals for each word, but found the sound quality was too poor to get reliable results. We also include a single non-linguistic demographic feature - age - alongside the linguistic features with the justification than a non-invasive diagnostic tool would be able to elicit this information easily from a patient.

4.4 Results

To evaluate the relative strength of the four new feature sets (spatial neglect from either halves, strips, or quarters, plus discourse features) we first evaluated the performance of the system without the new feature sets to establish a baseline performance.

4.4.1 Baseline Classification Performance

Figure 4.4 shows the F-measure across a range of models as we vary the number of included features. Features are ordered by absolute correlation with the labels in the training fold and are added in decreasing order (e.g., feature with the highest correlation added first, then second highest, etc.) The coloured shaded regions show 90% confidence intervals. Consistent with Fraser et al. [27], most models peak around 35-50 features and logistic regression outperforms the other models (F-measure: 0.824, and 90% CI=0.798-0.850). This plot highlights the importance of the feature selection step, as the performance of logistic regression would drop beneath 75% had all features been included. Figure 4.5 shows the F-measure, accuracy, and AUC for each model at their optimal number of features. All models substantially outperform the baseline across all three metrics.

We also ran an ablation study where each of the feature groups listed in Sec-

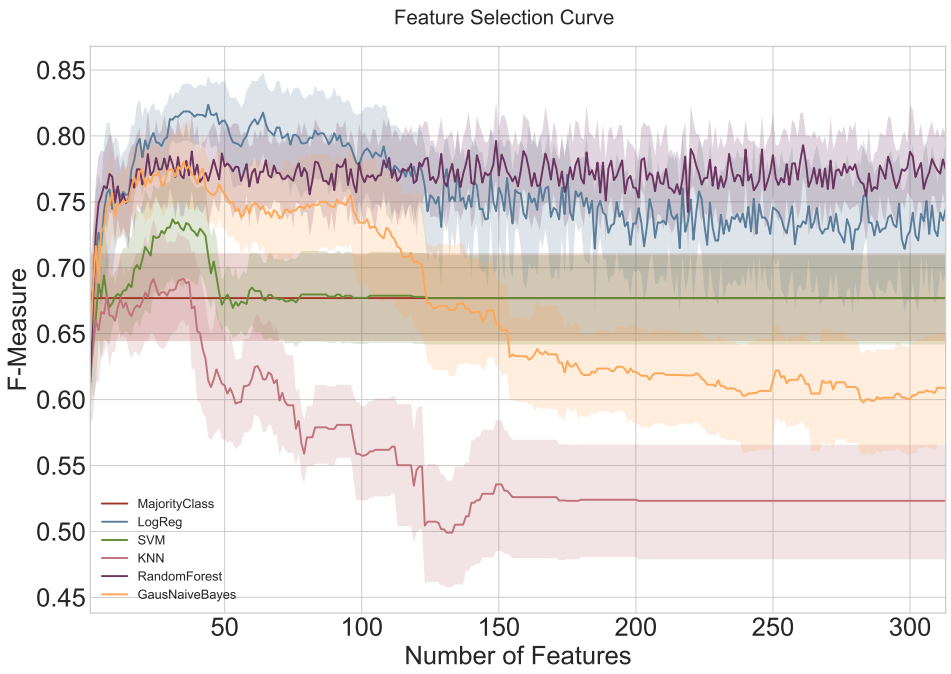

Figure 4.4: F-measure for different models as we vary the number of features included. Dark line shows the mean F-measure across each of the 10-folds and 90% CI are shown in the shaded regions. Features are added in decreasing order of their absolute correlation with the labels in the training fold. Most models reach their maximum performance between 35-50 features and then decline in performance as more features are included. This shows the need to include a feature selection step before training each model.

tion 3.2 were removed ("ablated"), feature selection was redone, and the model was retrained. We then got a measure of the importance of a particular feature group based on how much the performance has changed as a result of the ablation. In Figure 4.6 we see the change in F-measure across all models when a feature group is ablated. A more significant decrease in performance indicates a more important feature group, and the error bars show the 90% CI across the 10 folds. The number at the end of each label indicates how many features were contained within the group.

Acoustic, demographic, parts of speech and *info-units* are the most important

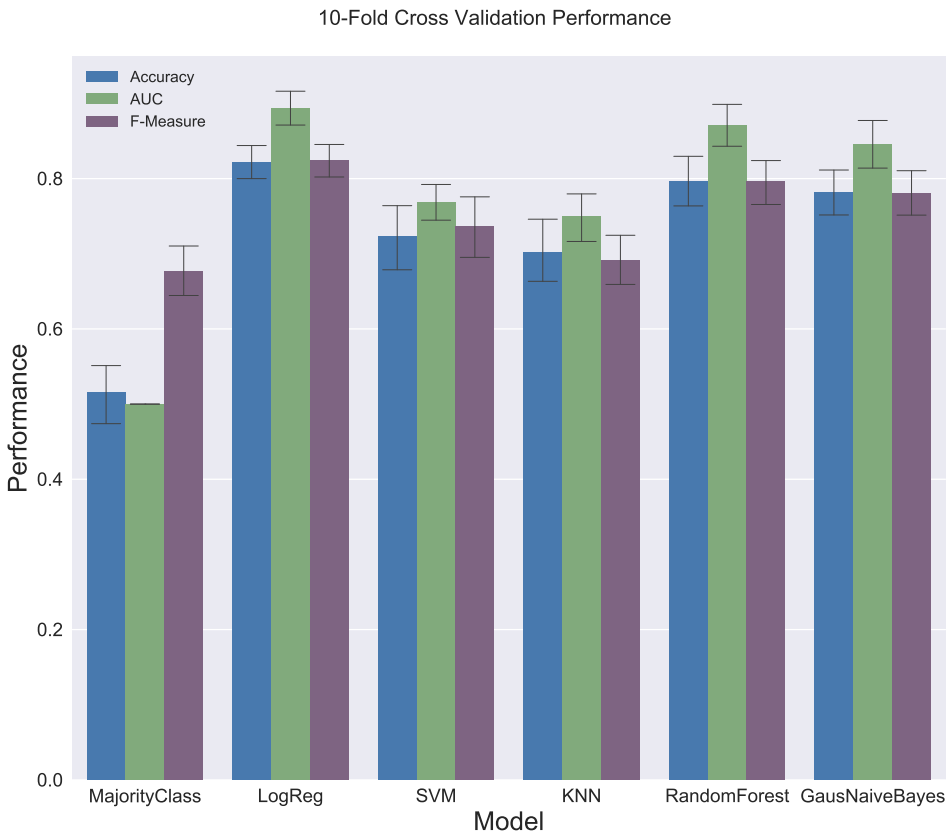

Figure 4.5: We show mean F-measure, accuracy, and Area Under the Curve (AUC) for each model at their optimum number of features (e.g. the peak performance in Figure 4.4). Error bars for each model show 90% CI across all 10 folds. Logistic regression performs best (ACC: 0.822, 90% CI=0.795-0.848, AUC: 0.894, 90% CI=0.867-0.921, FMS: 0.824, 90% CI=0.798-0.850) and has the tightest error bars across all models.

feature groups, causing a decrease in performance across all models. The other five groups either a minor decrease in performance in some models - or in some cases, even improve the model upon being ablated. This counterintuitive result was also seen in Fraser et al. [26]. The top performing model from Figure 4.4 (logistic regression, blue), decreases in performance in every case. The single demographic feature, age, is highly important, as are the diagnostic-test-specific info-unit features. Vocabulary richness, psycholinguistic, and repetitiveness are

the least important feature groups on this data set. Fraser et al. [26] also found vocabulary richness and repetitiveness unimportant, suggesting that "it is the words themselves, and not the number of different words being used, that is important." However, contrary to our results, they found psycholinguistic features to be very important. This could be due to the differences between data sets. Besides being smaller in size (n=40), and involving a different subtype of dementia (primary progressive aphasia), their task was a narrative retelling task where patients were asked to retell the story of Cinderella. Differences in psycholinguistic markers such as *concreteness* or *imagability* might be more apparent in this case compared to the task studied here, where patients are all asked to describe the same image.

Last, we investigated the relative importance of each feature by showing the mean feature score across all 10 folds. Within each fold, the features are sorted based on their absolute correlation with the training labels. The score for feature i in fold j given feature rank r_{ij} is calculated as:

$$score_{ij} = \begin{cases} 1 - \frac{r_{ij}}{50}, & \text{if } r_{ij} \leq 50 \\ 0, & \text{else} \end{cases} \tag{4.5}$$

Figure 4.7 shows the mean feature score and 90% CI across all 10 folds. A score of 1.0 indicates the feature was ranked first in all folds (ranks are indexed from 0) while a score of 0.0 indicates a feature was not selected within the first 50 features in any fold. The threshold of 50 was chosen because most models reached their maximum performance by 50 features. Note that unlike the ablation analysis, the feature score is agnostic to the model as it is only derived based on the correlations between features and labels.

As with Figure 4.6, features from *acoustic, demographic, parts of speech* and *info-unit* scored highly, as did context-free-grammar features. One feature from the psycholinguistics group - "imagability" - scored highly but this is likely due to its correlation with info-units which are all highly imageable. *Noun phrase to personal pronoun*, a context-free-grammar feature which measures personal pronoun usage, also scores highly in agreement with previous literature that demonstrated patients with dementia have an increased rate of pronoun usage [30]. An unanticipated result is that the most highly ranked feature is not age but *mean word length*

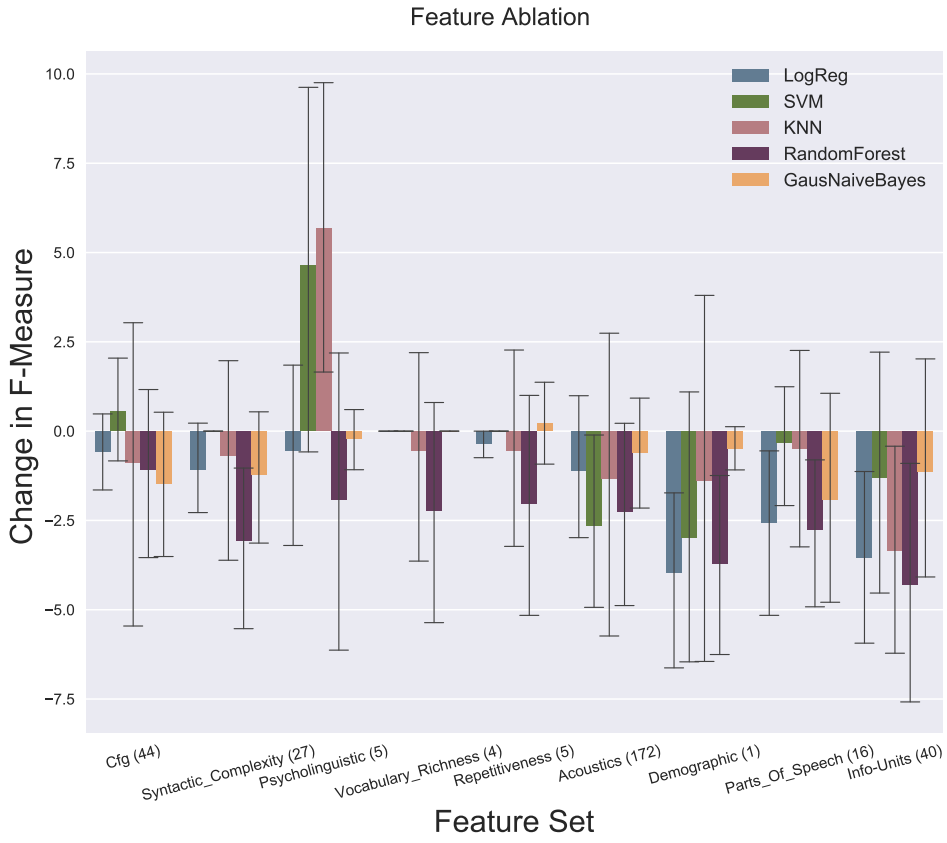

Figure 4.6: This shows the mean change in performance across models when a feature group is removed and the model is retrained. A greater decrease in performance indicates a more significant feature group. The number of features within each group are listed in parenthesis after each group name. *Acoustic, Demographic, Parts of Speech* and *Information Content* groups are important while *Syntactic Complexity, Psycholinguistic* and *Vocabulary Richness* are not. Large error bars indicate that the change in performance varies quite significantly between folds.

which has a perfect score of 1.0. Increased pronoun usage and a bias towards shorter, less imageable words suggests vague and non-specific speech.

4.4.2 Classification Performance With Novel Feature Sets

Now that we have verified the performance of our system using features from prior work, we evaluate the *halves, strips, quarters* and *discourse* feature sets. Each feature set was added in isolation (i.e., separately from the other three sets) to the existing features and the feature selection and model training steps were rerun. Quadratic cross terms were added to each feature set, but resulted in worse performance for *strips, quarters* and *discourse*. Thus these results do not use quadratic terms for those sets. Figure 4.8 shows the performance of each model with the additional feature set. Adding halves features improve the F-measure of the best classifier from 0.824% (90% CI=0.798-0.850%) to 0.846% (90% CI=0.813-0.878%). The *strips* set has the second largest improvement to logistic regression, improving the F-measure to 0.833% (90% CI=0.801-0.866). *Quarters* and *discourse* had a negligible effect on most models.

Figure 4.9 shows the change in performance with the added feature sets. The effect of the *halves* features on the suboptimal models is mixed: *halves* improve K-Nearest Neighbours (KNN) and the confidence interval around SVC is too large to be considered reliable. *Halves* hurts the other two models, Random Forests and Gaussian Naive Bayes. However, when the the quadratic terms are removed, the performance of Random Forests and Gaussian Naive Bayes is no longer decreased by the inclusion of halves features. A plot of the change in performance without quadratic features is shown in supplemental material, as are plots for AUC and accuracy with the new features.

Last we see how the feature score for *halves* compare to other features in Figure 4.10. The highest scoring feature across all features is *perception: rightside*, which measures the fraction of info-units the patient recognized on the right side of the image. *concentration: rightside*, *attention: rightside*, and *perception: leftside* also score highly and have smaller confidence intervals than most other features. In Figure 4.11 we see box plots for the four highest scoring features; *perception: rightside, mean word length, age,* and *noun phrase to personal pronoun*, with control interviews in blue. Respondents with dementia are less perceptive on their right side than healthy controls, they use more pronouns and shorter words on average, and they are older.

4.5 Discussion

In this chapter we proposed and evaluated four feature sets; three that measure spatial neglect across different partitions of the CookieTheft image, and discourse features which measure the overall coherence of a patient's response. We showed that by partitioning the CookieTheft image in two halves and measuring four simple metrics of spatial neglect *attention*, *concentration*, *repetition*, and *perception* (plus their quadratic cross terms), we improve the F-measure of the best classifier, logistic regression, by 2.2% from 82.4% (90% CI=79.8-85.0) to 84.6% (90% CI=81.3-87.8). One spatial neglect feature, *Perception: Rightside*, was more highly correlated with a dementia diagnosis than all other features, including age. Improvements were seen in a number of models, although the addition of quadratic cross terms hurt some suboptimal models. Thus, the inclusion of quadratic cross should be considered model dependent.

Interestingly, the *strips* partition also improved the accuracy of logistic regression (although not as much as *halves*) while *quadrants* did not. This finding agrees with the medical literature which has shown patients with AD tend to exhibit spatial neglect on one side of their visual field [14, 36, 37, 52, 53, 77]. spatial neglect is not known to cause inattention along the horizontal axis (e.g the top or bottom of an image) and therefore the *quadrants* partition did not improve classification performance. Our system was also able to detect other known linguistic deficits of AD patients, namely that they tend to use personal pronouns and shorter words more often than healthy counterparts.

Our main negative finding was that discourse features do not improve classification accuracy across the five models we tested. This is likely due to the structure of the CookieTheft description task. Unlike the narrative retelling task of the Wechsler Logical Memory I/II test which involves a patient retelling a short story, the CookieTheft description task is more of a "checklist" of potential items to be noticed. Therefore there is less opportunity for a response to be coherent (or not) compared to healthy controls. We therefore conclude that while discourse features are not useful in discriminating dementia from controls on the CookieTheft test they may be useful in longer and less structured narratives, such as the blog data set discussed in Chapter 6. In that context, a speaker has an opportunity to use a

larger set of discourse relations to connect one statement to the next.

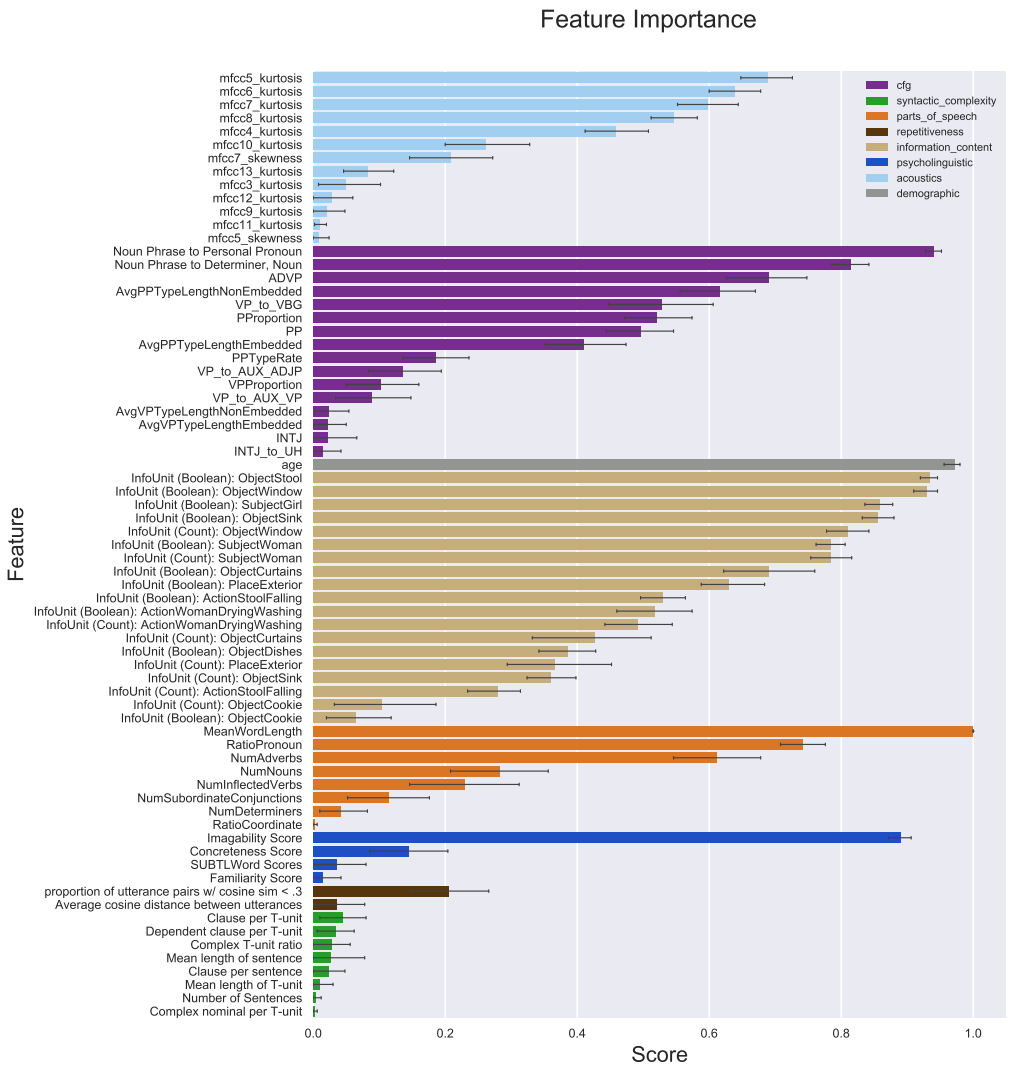

Figure 4.7: Feature importance score is calculated by equation 4.5. A score of 1.0 indicated the feature was selected first in all 10 folds, while a score of 0.0 indicates the feature was not selected within the top 50 features in any folds. Feature ranking does not depend on any particular model and only is based on the correlation between the feature and the binary labels. *Mean word length*, *age*, and *noun phrase to personal pronoun* are the highest scoring features on the DementiaBank data set.

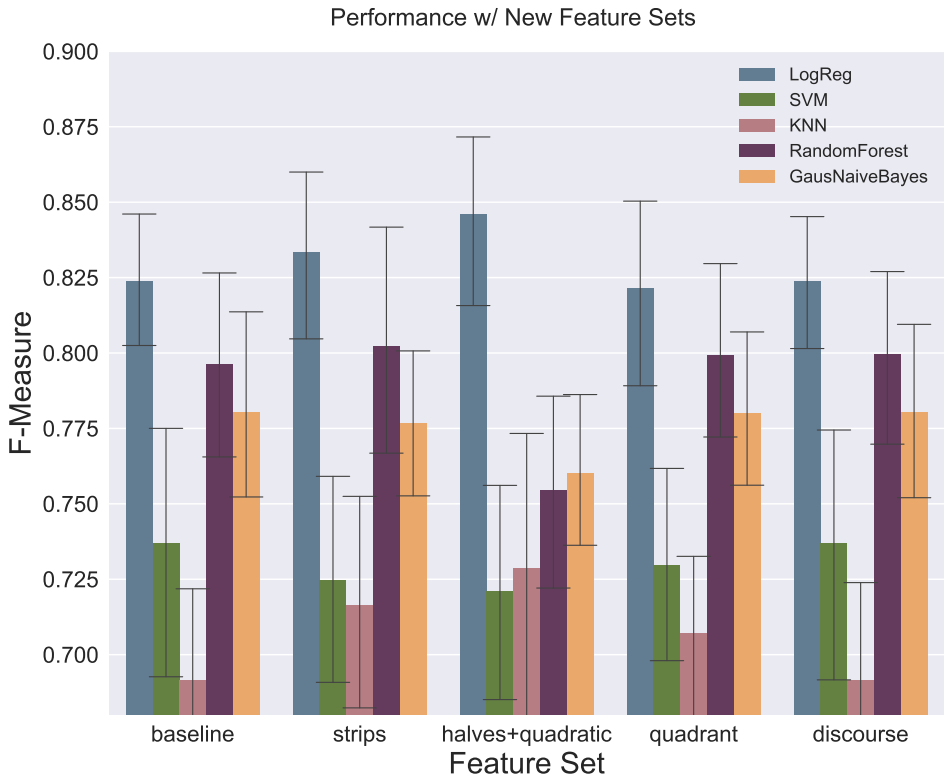

Figure 4.8: For each of the new feature sets we show the mean F-measure across five models. We compare against 'none', which is the performance of the existing system without the new feature set. *halves* improves the best model, logistic regression, from 0.824 (90% CI=0.798-0.850) to 0.846 (90% CI=0.813-0.878). *Strips* improves logistic regression as well, to 0.833 (90% CI=0.801-0.866), although not as much as *halves*. *Quarters* and *Discourse* have negligible effect on the performance of the best classifier.

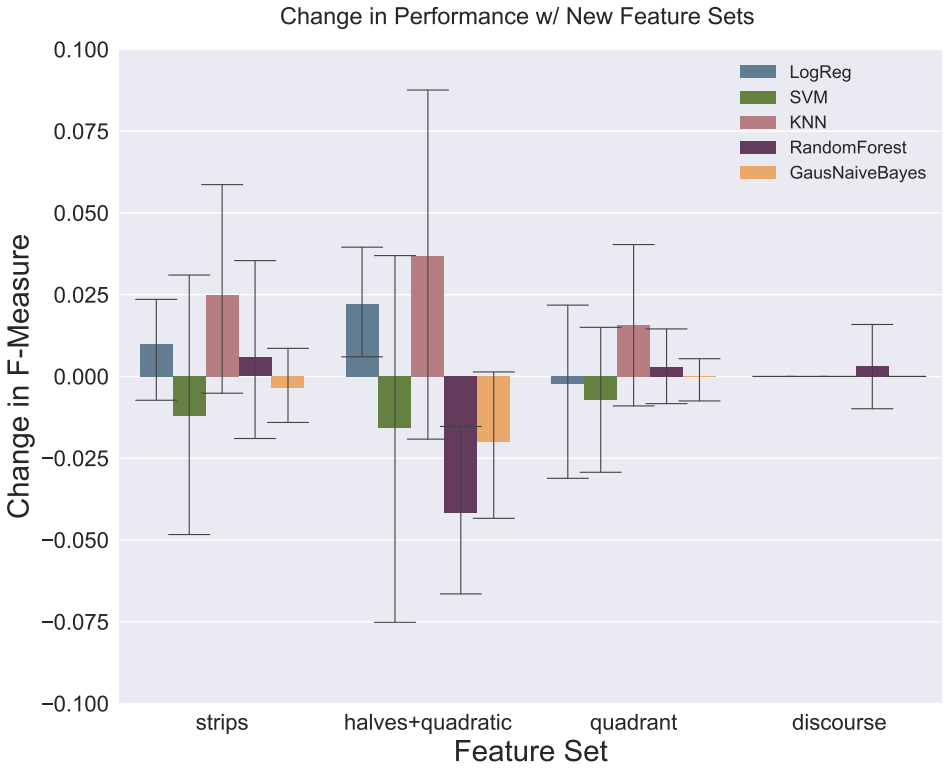

Figure 4.9: For each of the new feature sets we show the change in mean F-measure across five models when the new feature set is added. While *halves* improves the performance of the best classifier (logistic regression) it has mixed results on the suboptimal classifiers. Large error bars indicate the change in performance varies quite drastically between folds. Discourse features have no effect.

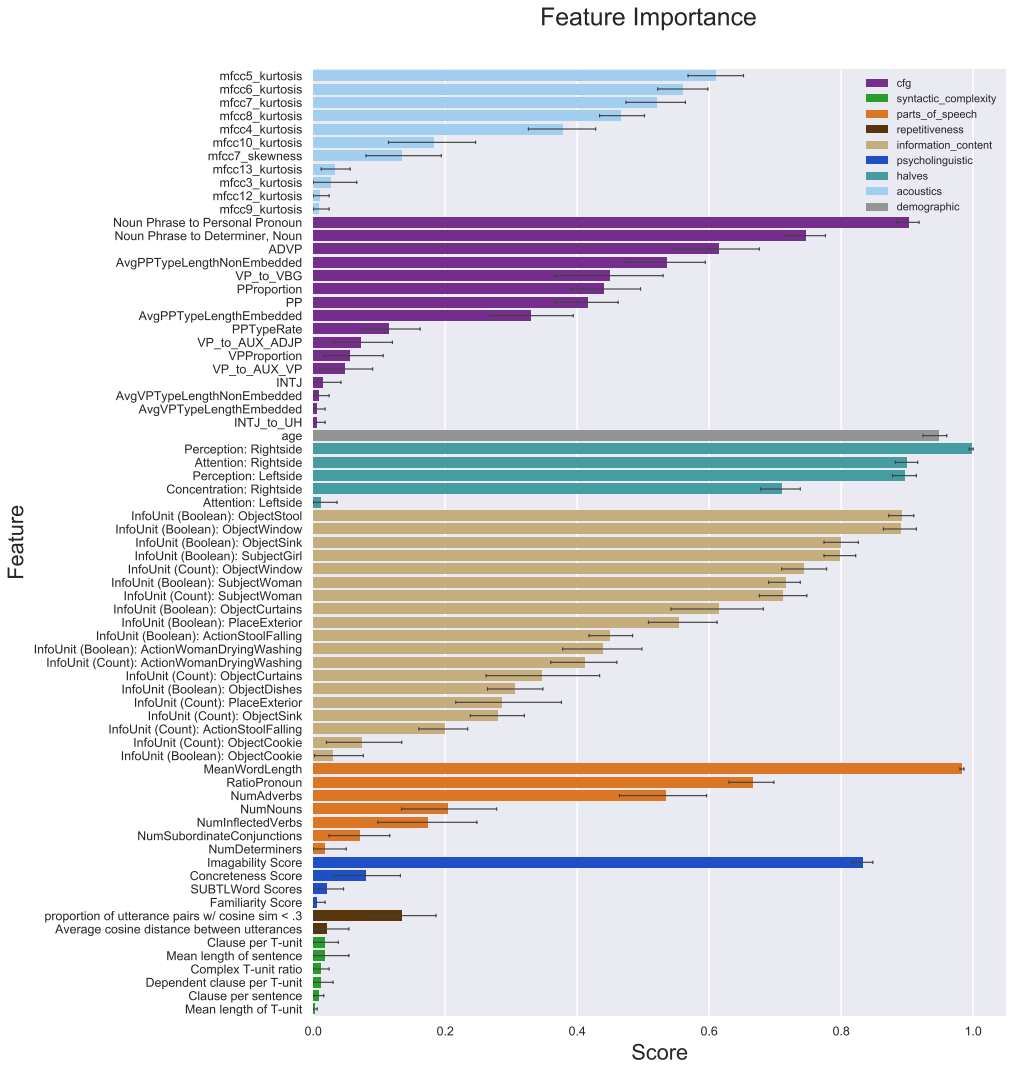

Figure 4.10: Feature importance score is calculated as shown in equation 4.5 with the addition of the *halves* features. A score of 1.0 indicated the feature was selected first in all 10 folds, while a score of 0.0 indicates the feature was not selected within the top 50 features in any folds. *Perception: Rightside* receives an almost perfect score, scoring more highly than *Mean word length*, *age*, and *Noun Phrase To Personal Pronoun* from Figure 4.7. Three other *halves* features, *Concentration: Rightside, Attention: Rightside* and *Perception: Leftside* also score highly.

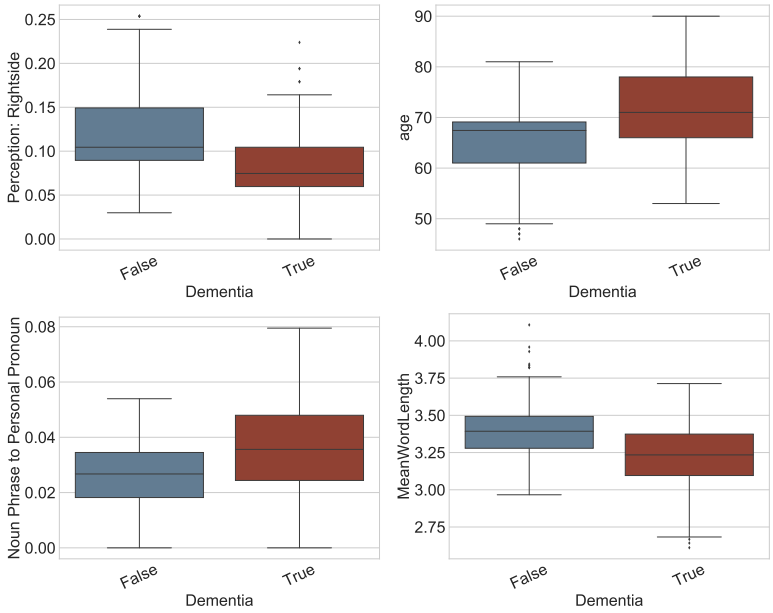

Figure 4.11: Box plots of the top four features from Figure 4.10. Top left shows *right-side perceptivity*, top right shows *age*, bottom left shows *noun phrase to person pronoun* (a measure of how often the patient uses personal pronouns), and bottom right is *mean length of words*. Those with dementia are less perceptive on the right side of their visual field than controls, as well as being older and more likely to use personal pronouns and shorter words.

Chapter 5

Detecting Mild Cognitive Impairment with Domain Adaptation

While much work has focused on Alzheimer's disease (AD), comparatively little attention has been paid to mild cognitive impairment (MCI). Given that it is a more heterogenous condition and is associated with less impairment than AD, people with MCI may not receive medical attention until they develop a more profound cognitive impairment. Thus, there are less available spoken language samples from people with MCI than patients with AD. The relative paucity of MCI data compared to AD therefore makes it difficult to build a diagnostic model for MCI. Given that people with MCI have a greater potential benefit from further assessment and therapy than those who have progressed to dementia, a model that could make optimal use of limited available data could be potentially very useful.

This chapter will demonstrate how *domain adaptation* can be used to exploit available AD data, thereby improving detection of MCI from spoken language samples. We compared two simple domain adaptation algorithms, AUGMENT and CORAL, and show AUGMENT improved upon all baselines. These algorithms are discussed in detail in Section 5.1.1 and 5.1.2. Our work differs from the previous work on MCI described in Section 2.2 in several ways. We use the feature set proposed by Fraser et al. [26] which is larger than the feature sets of Roark et al.

[63], Tóth et al. [76], and Satt et al. [65]. Unlike Roark et al. [63], we used MCI data collected from DementiaBank (as described in Section 3.1.1), where patients undergo a picture description task rather than a narrative retelling task. Most significantly, the goal of our study was different: while previous research in this area has focused on MCI detection (either with manual transcriptions or using automatic speech recognition (ASR)), the goal here was to demonstrate the viability of using domain adaptation algorithm to overcome the lack of MCI data [65, 76]. We begin with a brief discussion of domain adaptation.

5.1 Domain Adaptation

Domain adaptation is a general term for a variety of techniques aimed at exploiting resources in one domain (the *source* domain) in order to improve performance on some task in a second domain (the *target* domain). This is typically done when the target domain has little or no labelled data, while the source domain has a relatively large amount of labelled data, as well as existing models trained on that data. Typically the source data have been annotated for some phenomenon of interest, and the target data relate to another phenomenon that is very similar.

The issue of domain adaptation has received increasing attention in recent years. In work by Chelba and Acero [13], the source model is used to derive priors for the weights of the target model. They employ this technique with a maximum entropy model and apply it to the task of automatic capitalization of uniformly-cased data. They report that adaptation yields a relative improvement of 25-30% in the target domain.

Blitzer et al. [6] introduced Structural Correspondence Learning (SCL), in which relationships between features in the two domains are determined by finding correlations with so-called *pivot* features, which are features exhibiting similar behaviour in both domains. They used SCL to improve the performance of a parser applied to Biomedical data, but trained on Wall Street Journal data.

Daume [17] introduced an approach wherein each feature is copied so that there is a source version, a target version and a general version of the feature. He showed that this straight-forward approach could yield improvement on a variety of NLP sequence labeling problems, such as named entity recognition, shallow

parsing and POS tagging. More recently, Sun et al. [72] proposed CORAL, a method which aligns the second-order statistics of the source and target domain. We have implemented these two approaches, and describe them in more detail in below.

5.1.1 AUGMENT

Daume III's AUGMENT domain adaptation algorithm is simple ("frustratingly" so [17]) and has been shown to be effective on a wide range of data sets. It augments the feature space by making a "source-only", "target-only", and "common" copy of each feature, as seen below.

$$\underset{(n\times d)}{\begin{bmatrix} X_s \\ X_t \end{bmatrix}} \Rightarrow \underset{(n\times 3d)}{\begin{bmatrix} X_s & 0 & X_s \\ X_t & X_t & 0 \end{bmatrix}} \tag{5.1}$$

Here $X_s \in \mathbb{R}^{n_s \times d}$ and $X_t \in \mathbb{R}^{n_t \times d}$ are matrices of source and target data, where each of the n rows is an observation, each of the d column is a feature, $n = n_t + n_s$ and $n_t \ll n_s$. We create three copies of each column: a source-only column with zeros in target rows, a target-only column with zeros in the source rows, and the original column with both target and source entries left untouched. This augmented data set is then fed into a standard learning algorithm.

The motivation for this transformation is intuitive. If a column contains a feature (such as mean word length) which correlates to a diagnosis in both the target and source data (i.e. MCI and AD), a learning algorithm will increase the weight in the common column and reduce the weight on target-only and source-only copies, thereby reducing their importance in the model. However, if a feature correlates to a diagnosis only with MCI data, a learning algorithm can increase the weight of the target-only column (which contains zeros for all the source data) and reduce weight of the original and source-only columns, thereby assuring the feature will be less relevant to the model when applied to Alzheimer's data. By expanding the feature space and padding with zeros, a model can learn whether to apply a given feature on zero, one, or both data sets.

Although not explicitly stated in the original paper, AUGMENT assumes the

model learns a weight vector (e.g. logistic regression, SVM) so as to select the appropriate copy of the feature. Because of this we expect that models that classify without learning weights (e.g. KNN, Naive Bayes, Random Forests) will not improve under AUGMENT's feature transformation.

5.1.2 CORAL

CORAL (**COR**relation **AL**ignment) is another recently proposed "frustratingly easy" [72] domain adaptation algorithm that works by aligning the covariances of the source and target features. The algorithm first normalizes the source data to zero mean and unit variance, and then a whitening transform is performed on the source data to remove the correlation between the source features. A whitening transform is a linear transformation of the feature space such that the covariance of the transformed feature space is the identity matrix. We use PCA whitening on the source data as follows:

$$\Sigma_s = E[X_s X_s^T] - \cancel{E[X_s]E[X_s^T]} = QDQ^T$$
$$W = QD^{-\frac{1}{2}}Q^T$$
$$\hat{X}_s = WX_s$$

That is, we first take the eigenvalue decomposition of the covariance matrix of the (zero meaned) source data. Then we set the whitening matrix to be the eigenvalue decomposition with the negative square root of the eigenvalues. This results in the

whitened source data having an identity covariance:

$$
\begin{aligned}
cov(\hat{X}_s) &= E[\hat{X}_s\hat{X}_s^T] - E[\hat{X}_s]E[\hat{X}_s^T] \\
cov(\hat{X}_s) &= E[WX_sX_s^TW^T] - E[WX_s]E[X_s^TW^T] \\
cov(\hat{X}_s) &= WE[X_sX_s^T]W^T - \cancel{WE[X_s]E[X_s^T]W^T} \\
cov(\hat{X}_s) &= W\Sigma_sW^T \\
cov(\hat{X}_s) &= QD^{-\frac{1}{2}}Q^TQDQ^TQD^{-\frac{1}{2}}Q^T \\
cov(\hat{X}_s) &= QD^{-\frac{1}{2}}D^{\frac{1}{2}}D^{\frac{1}{2}}D^{-\frac{1}{2}}Q^T \\
cov(\hat{X}_s) &= QQ^T \\
cov(\hat{X}_s) &= I
\end{aligned}
$$

Finally, the source matrix is "recoloured" with the correlations from the target data using colour matrix W_t:

$$
\begin{aligned}
\Sigma_t &= E[X_tX_t^T] - \cancel{E[X_t]E[X_t^T]} = QDQ^T \\
W_t &= QD^{\frac{1}{2}}Q^T \\
\tilde{X}_s &= W_t\hat{X}_s \\
cov(\tilde{X}_s) &= E[\tilde{X}_s\tilde{X}_s^T] \\
cov(\tilde{X}_s) &= E[W_t\hat{X}_s\hat{X}_s^TW_t^T] \\
cov(\tilde{X}_s) &= W_tE[\hat{X}_s\hat{X}_s^T]W_t^T \\
cov(\tilde{X}_s) &= W_tIW_t^T \\
cov(\tilde{X}_s) &= QD^{\frac{1}{2}}Q^TQD^{\frac{1}{2}}Q^T \\
cov(\tilde{X}_s) &= QDQ^T \\
cov(\tilde{X}_s) &= \Sigma_t
\end{aligned}
$$

These three steps are shown in Figure 5.1. A model is then trained on the re-coloured source data and used to classify the target data.

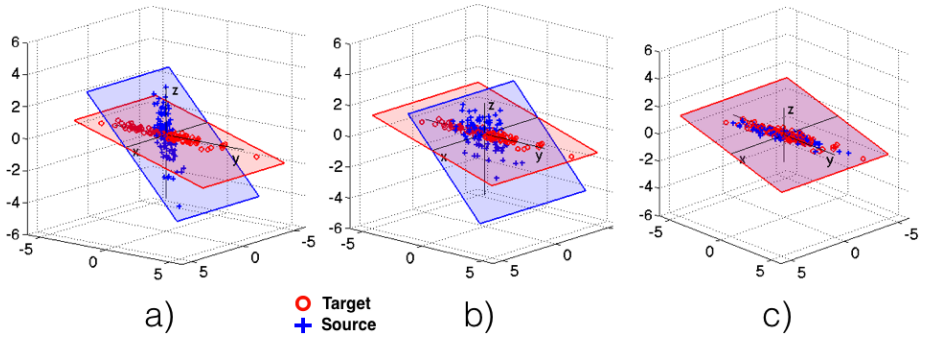

Figure 5.1: The CORAL algorithm is shown in three steps. The target and source data set consist of three features; x, y, z. In a) the source data and target data are normalized to unit variance and zero mean, but have different covariances distributions. b) The source data is whitened to remove the correlations between features. c) The source data is recoloured with the target domain's correlations and the two data sets are aligned. A classifier is then trained on the re-aligned source data. (Figure adapted from [72])

5.2 Data Set

The DementiaBank data set, described in detail in Section 3.1.1, contains 43 MCI samples from 19 patients, 257 possible/probable AD samples, and 242 control samples. We split the data set into "target" data (86 rows, 43 MCI, 41 control) and "source" data (458 rows, 257 possible/probable AD, 201 control). Multiple interviews from a single control patient were contained to either the target or the source data sets, but not both.

5.3 Baseline, Experiments, Results

We followed the experimental designed described in Chapter 3, using an augmented feature space for AUGMENT and CORAL. We compare against three domain adaptation baselines. *Target only* trains the model only using target data, *source only* trains a model only using source data but evaluates on the target data. In the *relabeled source* model, we pool the target data and source data in the training folds and relabel AD to MCI. Along with the domain adaptation baselines we

included one baseline model, *majority class*, which predicts the majority class in the training fold.

The test set contained only MCI data. In the AUGMENT, CORAL, and *relabeled* approaches, each fold of the training set contained a combination of MCI+AD data while the *source only* baseline contained only AD in the training fold. Our goal was to verify whether the accuracy achieved by using these domain adaptation methods outperforms the accuracy achieved by using MCI data alone.

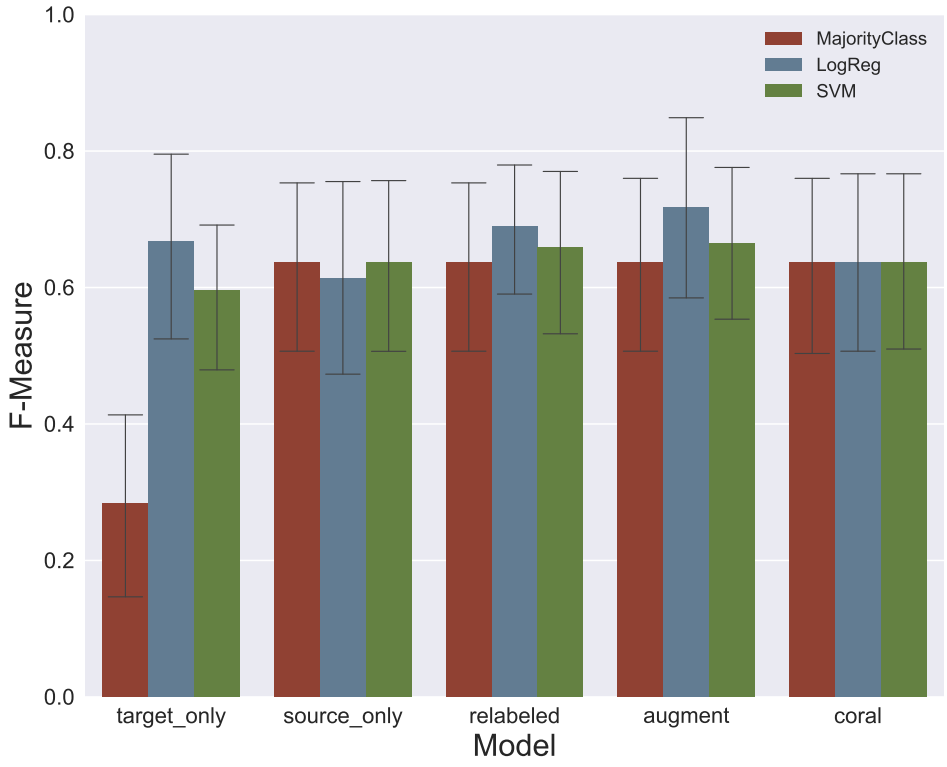

Figure 5.2: Comparison of two domain adaption methods, AUGMENT and CORAL, against three domain adaptation baselines and one model baseline (dummy classifier which predicts the majority class in the training fold). Mean F-measure and 90% CI are shown across 10-folds. Only target data appears in the test fold. AUGMENT with logistic regression outperforms all baselines. CORAL doesn't improve either model above the majority class baseline.

In figure 5.2 we show the F-measure for models with a weight vector (SVM and logistic regression). The AUGMENT domain adaptation algorithm with logistic regression performed best (F-measure: 0.717, 90% CI=0.562-0.871), beating all three domain adaptation baselines and the dummy classifier. AUGMENT also improved the SVM classifier over baselines although the performance (F-measure: 0.664, 90% CI=0.533-0.796) did not match logistic regression. CORAL does not improve either model beyond the simple majority class baseline model, and for logistic regression it results in a worse performance than the *target only* domain adaptation baseline.

Figure 5.3 shows both methods on three models (Naive Bayes, Random Forests, K-Nearest Neighbours (KNN)) which do not classify using a weight vector. As we expected, we see AUGMENT fails to improve any models and actually makes their performance worse than the *target only* baseline. This underscores the importance of only using the AUGMENT method with a model that is able to select, via the weight vector, which of the three copies of the feature to use.

5.4 Discussion

This chapter showed how we can use a simple domain adaptation algorithm, AUGMENT, to use AD data to overcome the scarcity of MCI data. Using AUGMENT we improved the F-measure of logistic regression from 66.7%, (90% CI=50.5-82.9) using MCI data only, to 71.7% (90% CI=56.2-87.1), using both MCI and AD data. AUGMENT requires a simple modification of the target and source feature space, and can be easily extended to incorporate source data from multiple domains.

We also showed that AUGMENT only works with classifiers that learn a weight vector. This is an important caveat that was not explicitly stated in the original paper by [17]. Practitioners should be cautious about applying AUGMENT to models that do not learn a weight vector, such as KNN or Gaussian Naive Bayes, because doing so can actually *decreases* the performance.

The main negative result was the performance of the CORAL domain adaptation method with logistic regression (F-measure: 56.5% 90%CI=40.3-72.8), which is worse than the target-only method. In other words, using CORAL results in a worse performance than not doing domain adaptation at all. It has previously been found

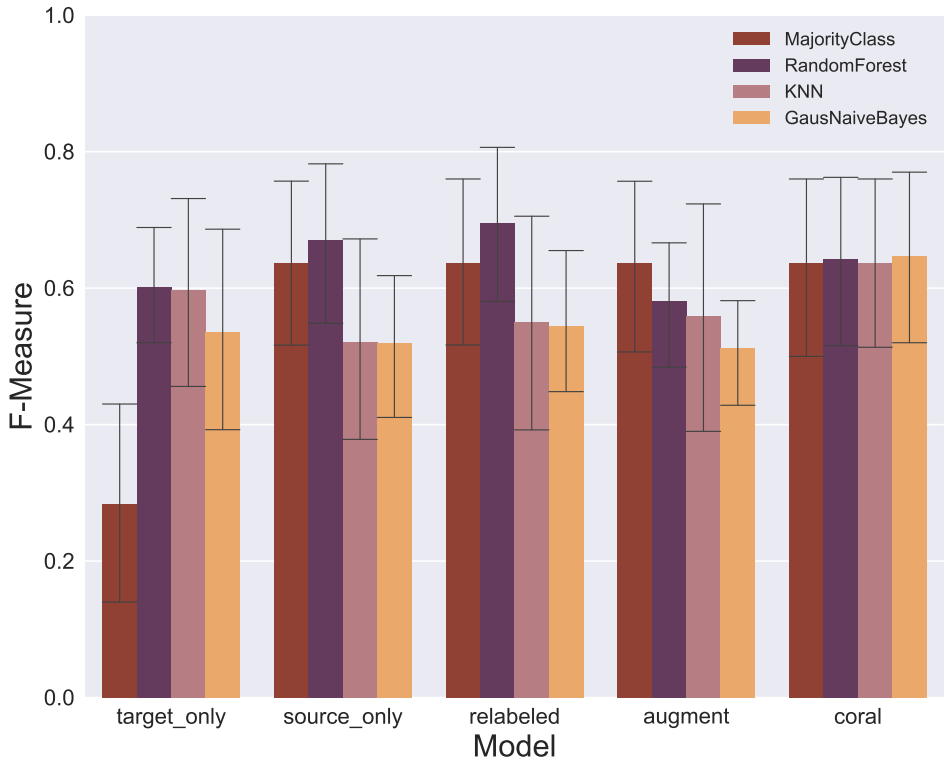

Figure 5.3: Performance of two domain adaption methods, AUGMENT and CORAL, on classifiers that do not learn a weight vector. AUGMENT does poorly in this setting because the models are unable to choose between the "target only", "source only" or "both" version of each feature.

that CORAL does not always work well with boolean features such as bag-of-words features [72]. Info-units, which we see in figures 4.6 and 4.7 to be strong predictors of dementia, are largely boolean.

Chapter 6

Detecting Dementia From Written Text

Chapters 4 and 5 have focused on spoken language collected from patients undergoing a clinical examination. This data is expensive to collect and does not accurately reflect how patients use language in daily life. Perhaps most importantly, as millennials and "iGen's" continue to age and use the internet, the predominant source of language samples from those with dementia will not be spoken, but written.

There were 173 million blogs on the web in 2011, only twenty years since the first website was launched in 1991 [70]. As these bloggers enter their senescence, a percentage of them will be diagnosed with dementia, and a percentage of those will continue to use the internet. There will therefore be a growing data set available in the form of tweets, blog posts, and social media comments with which to train a classifier. Provided these writers have a verified clinical diagnosis of dementia, such a data set would be large, inexpensive to acquire, easy to process, and require no manual transcriptions. Unlike with spoken speech, written text will contain fewer instances of the subject being "flustered" by potential word-finding difficulties and other time-dependent performance issues. This therefore might make it possible to detect subtle lexical, grammatical or pragmatic issues that may be missed from spoken text.

There are downsides to using written language samples as well. Unlike spoken

language, written text can be edited or revised by oneself or others. People with dementia may have "good days" and "bad days," and may write only on days when they are feeling lucid. Thus, written samples may be biased towards more intact language. Furthermore, researchers do not have an audio recording to accompany the text and patients are not constrained to a single topic; people with dementia may have greater facility discussing familiar topics. A non-standardized data set will also prevent the collection of common test-specific linguistic features such as info-units. However, working with a very large data set may be able to mitigate the effects of these limitations. Additionally, since substantial amounts of data can be collected for the same person, more accurate, user-specific longitudinal predictions might be possible

In this chapter we present the first attempt at automatically detecting whether a blog post was written by an individual with dementia. We followed the general methodology described in Chapter 3 with a different data set, described in Section 6.1. The goal was to determine if this task is possible given the constraints listed above, and also to determine if the features most discriminating in the written case are the same as in the spoken case. We make our data set publicly available at https://github.com/vadmas/blog_corpus.

6.1 Data Set

We scraped the text of 2805 posts from 6 public blogs as described in Table 6.1. Three blogs were written by persons with dementia (First blogger: male, Alzheimer's disease (AD), age 72. Second blogger: female, AD, age 61. Third blogger: Male, Dementia with Lewy Bodies, age 65) and three were written by family members of persons with dementia to be used as control (all female, ages unknown). Other demographic information, such as education level, was unavailable. From each of the three dementia blogs, we manually filtered all texts not written by the owner of the blog (such as fan letters) or posts containing more images than text. This left with 1654 samples written by persons with dementia and 1151 from healthy controls. Control blogs were written by children, spouses, or caregivers of seniors with dementia and were selected to control for topic and previous level of writing experience.

URL (http://*.blogspot.ca)	**Posts**	**Mean words**	**Start Date**	**Diagnosis**	**Gender/Age**
living-with-alzhiemers	344	263.03 (s=140.28)	Sept 2006	AD	M, 72 (approx)
creatingmemories	618	242.22 (s=169.42)	Dec 2003	AD	F, 61
parkblog-silverfox	692	393.21 (s=181.54)	May 2009	Lewy Body	M, 65
journeywithdementia	201	803.91 (s=548.34)	Mar 2012	Control	F, unknown
earlyonset	452	615.11 (s=206.72)	Jan 2008	Control	F, unknown
helpparentsagewell	498	227.12 (s=209.17)	Sept 2009	Control	F, unknown

Table 6.1: Blog Information as of April 4th, 2017

6.2 Experimental Design

We followed the general methodology described in Chapter 3 using the blog data set instead of the DementiaBank data set. We use the features described in Section 3.2, with the exception of the acoustic and info-unit feature groups which were not available for blog data. In total we extract 102 features from each blog post with a binary label, indicating whether or not the author has dementia. We performed a 9-fold cross validation across all pairs of blogs with opposite labels. Each test fold contains all posts from one dementia blog and one control blog, and the posts from the remaining four blogs are used in the training fold. As with the previous experiments we run a feature selection step within each training fold, as described in Section 3.3. We report Accuracy (ACC), F-measure (FMS) and Area Under the Curve (AUC) for each model and compare against a dummy classifier that predicts the majority class label in the training fold.

6.3 Results

Figures 6.1 and 6.2 show the feature selection curve and the final peak classification performance, respectively. Unlike with DementiaBank all models reach near-optimal performance near 10 features then the performance either levels off or improves slightly as more features are added. The best model with the blogs data set is K-Nearest Neighbours (KNN) (ACC: 0.728, 90% CI=0.687-0.769, AUC: 0.761, 90% CI=0.714-0.807, FMS: 0.785, 90% CI=0.746-0.823) which slightly beats logistic regression (ACC: 0.724, 90% CI=0.677-0.770, AUC: 0.759, 90% CI=0.689-0.829, FMS: 0.785, 90% CI=0.743-0.827) and had tighter error bars. All

models beat the baseline AUC of 0.50.

We ran the same ablation analysis on the blogs data set as we performed on the DementiaBank (Section 4.4). The results are shown in Figure 6.3. Unlike with the DementiaBank data set psycholinguistic features are the most important feature group, with their ablation causing the performance of all models to drop significantly. Somewhat unexpectedly the removal of the other feature groups causes a slight *improvement* in the best classifier, KNN, although the improvement is within the error bars in all cases, and not seen in logistic regression, the near-optimal classifier.

Figure 6.4 shows the scores for each feature, as calculated by equation 4.5. A score of 1.0 indicated the feature was selected first in all 9 folds, while a score of 0.0 indicates the feature was not selected within the top 50 features in any folds. *SUBTL word score*, which is a measure of how frequently a word is used in daily life, is the most highly correlated with a dementia diagnosis across all 9 folds. The *number of sentences* per post also is highly correlated with a diagnosis. As with the DementiaBank data set, both *mean word length* and *noun phrase to personal pronoun* also score highly. We also observe that the error bands are larger for most features than in the figure 4.7, indicating the correlation between feature and label has a greater dependence on the particular training fold (higher variance in the bias-variance tradeoff) than in the DementiaBank data set.

Finally, Figure 6.4 shows the box plots of the four highest scoring features in: *SUBTL word score*, *number of sentences*, *mean word length*, *noun phrase to personal pronoun*. The top left box plot shows that bloggers with dementia (red) have a higher SUBTL Word Score on average. SUBTL is a measure of how frequently a word is used in daily life, with a higher score indicating a more ordinary word and a lower score indicating a less common one. Scores are derived from television and film subtitles. In the six blogs in our data set, bloggers with dementia use more ordinary (i.e. more frequently occurring) words than their control counterparts. They also tend to write shorter blog posts, and in agreement with the DementiaBank data set, use shorter words and more personal pronouns.

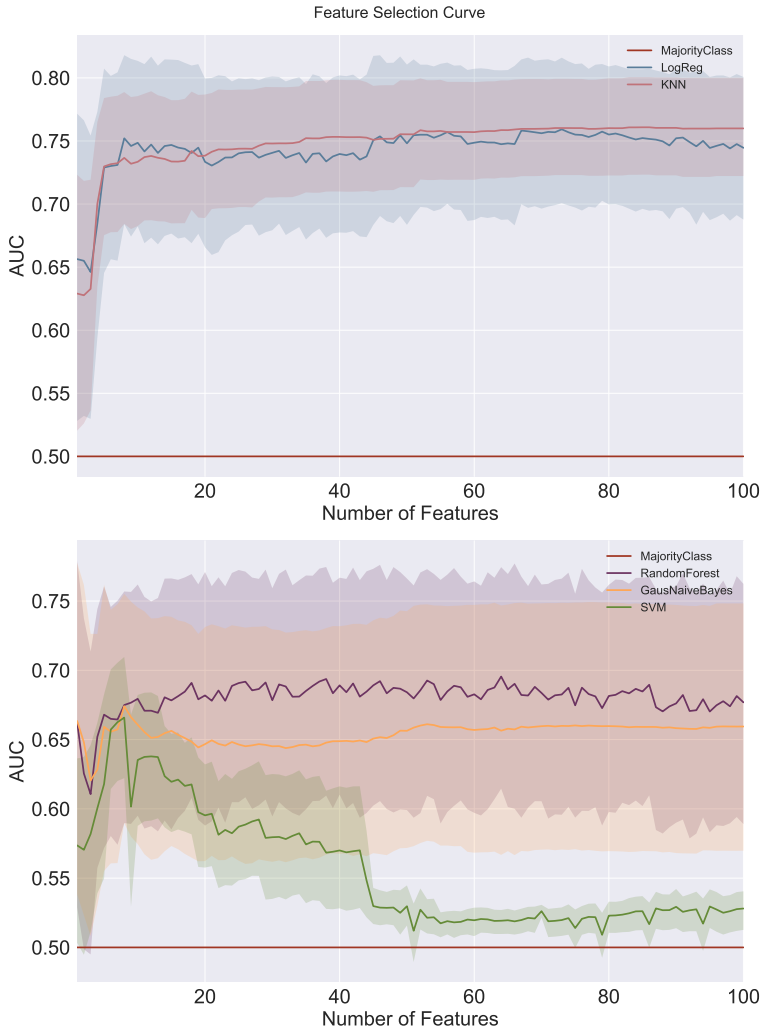

Figure 6.1: We show Area Under the Curve (AUC) for each model as we vary the number of features. Error bars for each model show 90% CI across all 9 folds. We use two plots so error bars are distinguishable. All models beat the dummy classifier (majority class) with the KNN achieving the best performance (ACC: 0.728, 90% CI=0.687-0.769, AUC: 0.761, 90% CI=0.714-0.807, FMS: 0.785, 90% CI=0.746-0.823).

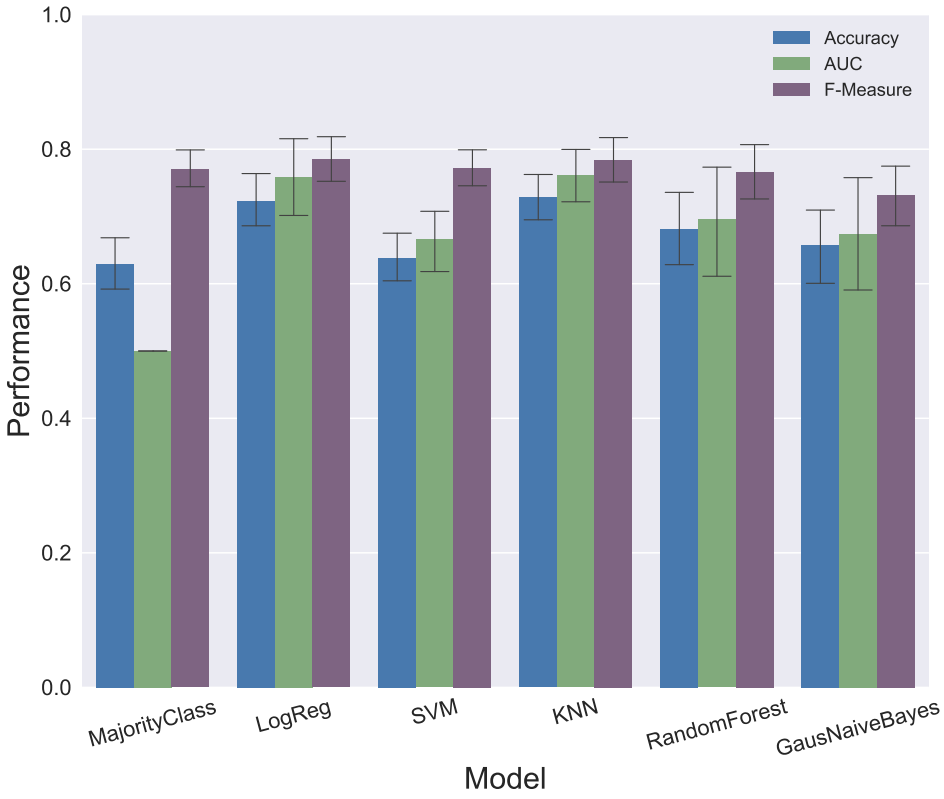

Figure 6.2: We show mean Accuracy (ACC), F-measure (FMS) and Area Under the Curve (AUC) for each model at their optimum number of features (e.g. the peak performance in Figure 6.1). Error bars for each model show 90% CI across all 9 folds. All models beat the dummy classifier (majority class) with the KNN achieving the best performance (ACC: 0.728, 90% CI=0.687-0.769, AUC: 0.761, 90% CI=0.714-0.807, FMS: 0.785, 90% CI=0.746-0.823).

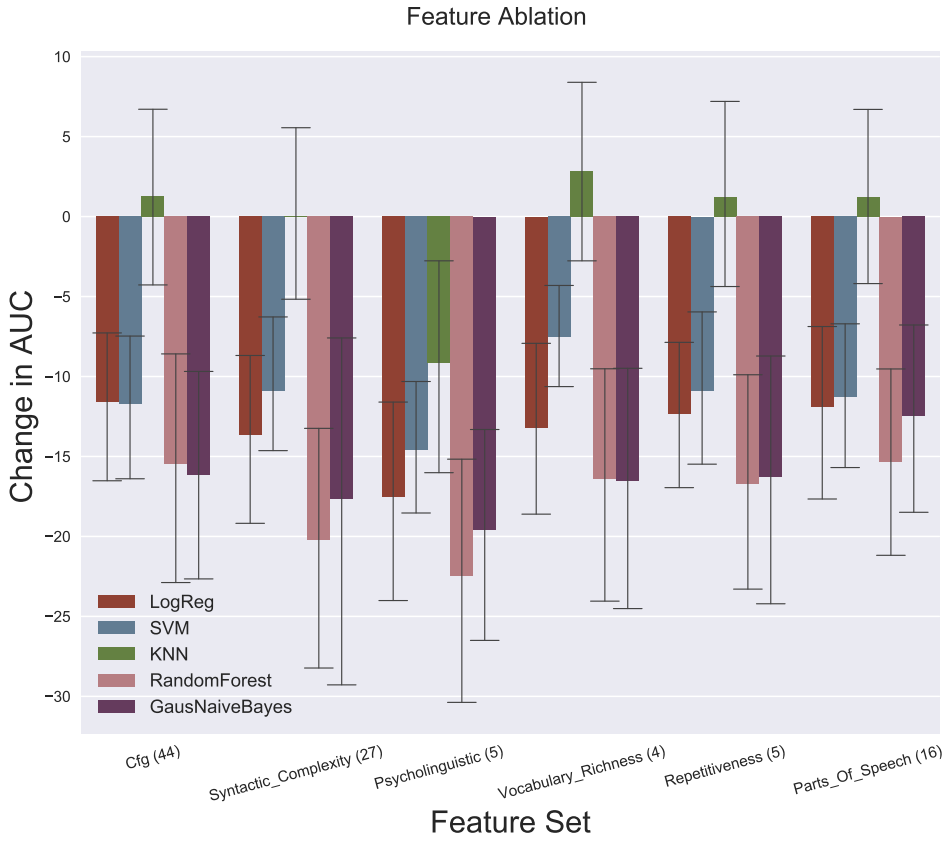

Figure 6.3: As with figure 4.6 we show the mean change in performance across models when a feature group is removed and the model is re-trained. A greater decrease in performance indicates a more significant feature group. The number of features within each group are listed in parenthesis after each name group name. Unlike with the DementiaBank data set all feature groups are important to the prediction accuracy, with the removal of the psycholinguistic group having the greatest deleterious effect across all models.

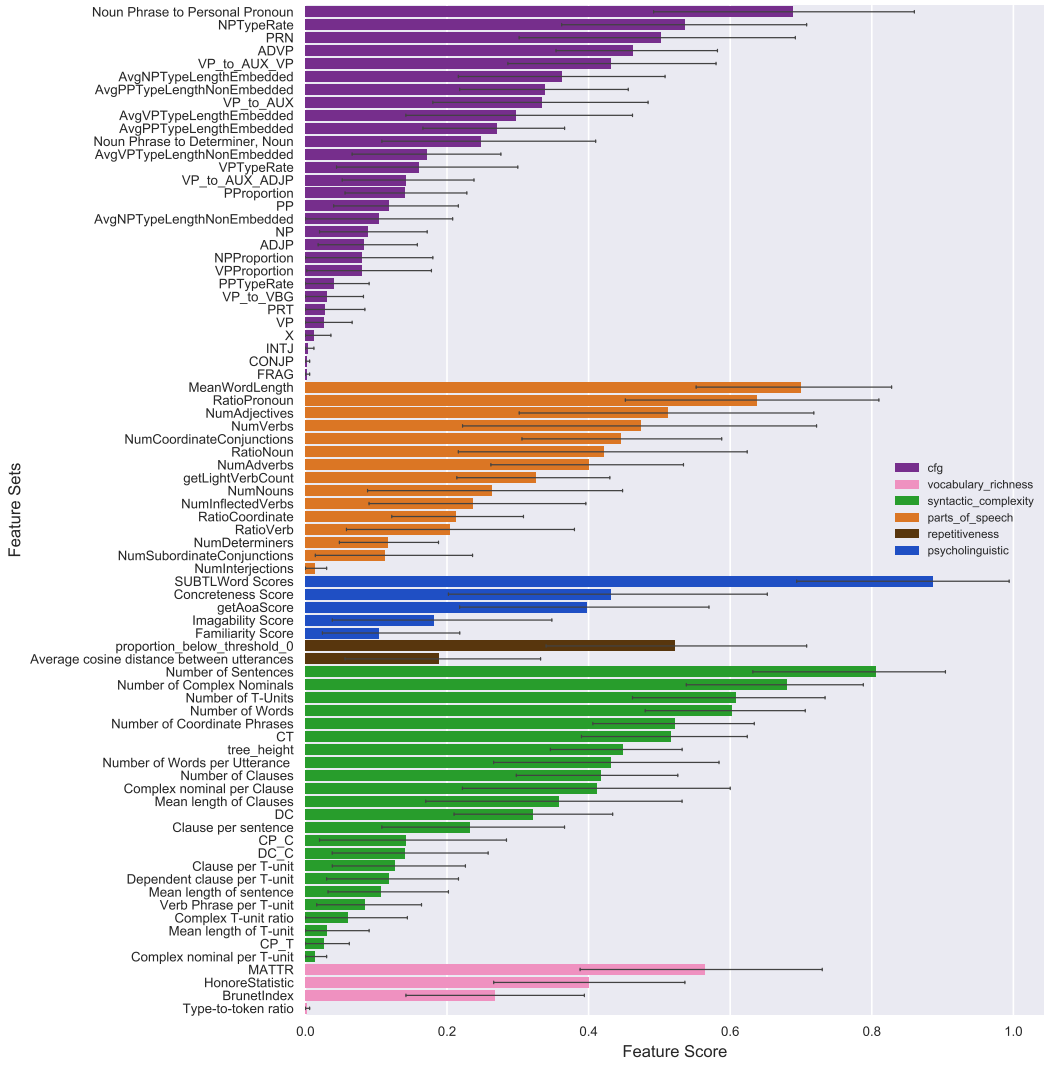

Figure 6.4: Feature importance score for the blog data set, as calculated by equation 4.5. A score of 1.0 indicated the feature was selected first in all 9 folds, while a score of 0.0 indicates the feature was not selected within the top 50 features in any folds. Feature ranking does not depend on any particular model and only is based on the correlation between the feature and the binary labels. *SUBTL Word Score*, *Number of Sentences*, *Mean Word Length*, and *Noun Phrase To Personal Pronoun* are the highest scoring features on the data set.

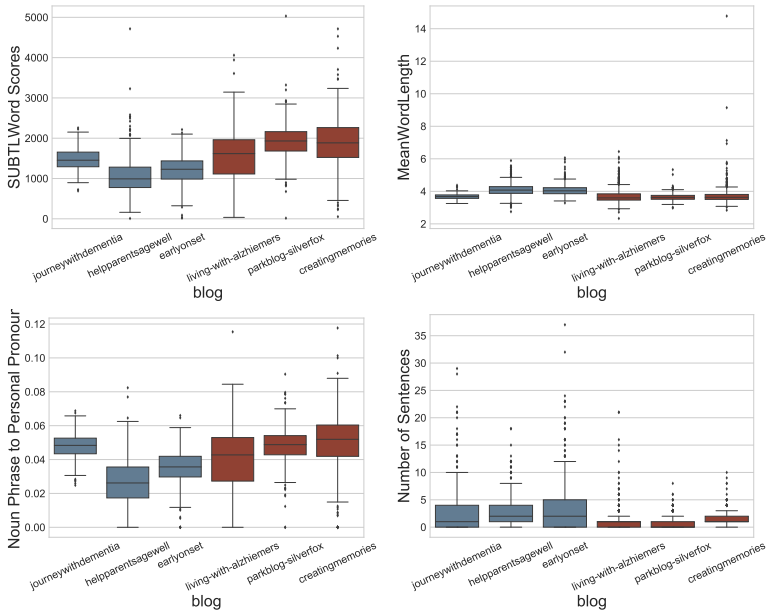

Figure 6.5: Box plots of the four highest scoring features in figure 6.4. *SUBTL Word Score* top left, *Mean Word Length* top right, *Noun Phrase To Personal Pronoun* bottom left, *Number of Sentences* bottom right. Blogs written by persons with dementia are red and controls are blue. As in the spoken case, persons with dementia use the personal pronoun more often and use smaller words on average. Bloggers with dementia also have a higher SUBTL score (indicating an impoverished vocabulary) and write shorter posts.

6.4 Discussion

This chapter demonstrated how dementia can be automatically detected from written text in the form of blog posts. We collected a data set of 2805 blog posts written by either persons with dementia or family members of persons with dementia. We then extracted 102 lexical features from each post and evaluated the performance of five classifiers in detecting whether the author of a post from an unseen blog has dementia. KNN beat the other models, and the baseline classifier, with an AUC of 0.761 (90% CI=0.714-0.807).

We also observed that bloggers with dementia tend to use fewer uncommon

words (as indicated by the higher SUBTL score), and write shorter posts using shorter words, on average. This finding is interesting because it would be difficult for a human reading a single post to detect the simplified language (provided the post was coherent, which from inspection all were), but a higher SUBTL score and shorter word length could be detected automatically given a collection of posts.

Given that the data set consisted only of six blogs, a larger data set is necessary to discern if the change in language we've identified here is in fact due to dementia, or due to the idiosyncrasies of this particular data set. For example, we detected an increased pronoun usage by bloggers with dementia compared to controls. This agrees with previous work and our results from Chapter 4, but it could also be due to the fact that the bloggers in the dementia blogs were writing about their own experience while the authors in the control blogs were writing about someone else's experience, and hence may use less pronouns.

Despite the limitations of a comparatively small data set and the difficulties associated with analyzing written text (including the author's ability to make revisions and make use of third party editors), we have shown it is possible to detect dementia from written text. This opens the door to making use of the upcoming deluge of online text written by seniors suffering from cognitive decline as data on which to train machine learning models.

Chapter 7

Conclusion

Early detection of dementia is important, not only for patients for whom a diagnosis is the first step to receiving adequate support, but for researchers, who say that early detection will be crucial to finding a cure [3]. There have recently been successes in using machine learning and natural language processing techniques to automatically detect dementia from speech. This thesis has made three main contributions towards this effort.

First, we proposed a novel set of features biologically motivated that we call *spatial neglect* features. These features measure whether the respondent is more perceptive on one side of their visual field than the other. We showed their inclusion increases the F-measure of logistic regression from 82.4% to 84.6%. This achieves a new state of the art on the DementiaBank data set, beating the previous state-of-the-art of 81.92%. We considered three different partitions of the CookieTheft image, *halves, strips,* and *quarters* and found that *halves* performs best, in agreement with previous finding in medical literature. Previous work has found that patients with AD show differences in discourse structure and so we also evaluated the effect of discourse features on model performance, but found they had no effect on the DementiaBank data set.

Second, we demonstrated how a simple domain adaptation algorithm can be used to overcome the lack of available mild cognitive impairment (MCI) data. We compared two "frustratingly simple" domain adaptation algorithms that used AD data to improve the accuracy of MCI detection, and found that AUGMENT beats all

baselines and improves the F-measure from 66.7% using only MCI data, to 71.7% using MCI + AD.

Last, we evaluated our framework on written data in the form of blog posts. It is not obvious that a system that can detect dementia from spoken language could do the same for written language, given that one can make revisions to text but cannot do so for recorded extemporaneous speech. We show that a range of models can predict whether the author of a blog post has dementia at a rate far above baselines. KNN achieved a maximum AUC of 0.761 beating the baseline of 0.50 by a wide margin. Additionally we make the blog corpus used in our experiments publicly available for future researchers.

Besides the main contributions listed above, we made some observations that will be useful to practitioners and help guide future work. For practitioners, we recommend the use of spatial neglect features (with a halves partition) as they increases the performance of most models, but recommend including them both with and without quadratic terms in order to determine which performs best. In the case of logistic regression, the addition of the quadratic terms improved the performance significantly but for Random Forests and Gaussian Naive Bayes, the quadratic terms hurt the performance significantly. This is likely due to the fact that many of the quadratic features were uninformative and some models are less capable of dealing with an excess of uninformative features than others[1].

Another observation regards the use of AUGMENT for domain adaptation. As noted in Section 5.3, AUGMENT performs well when a model is able to select between the three copies of a feature (c.f Section 5.1.1) via a weight vector. Models which are unable to do so, such as Random Forests, Gaussian Naive Bayes, and KNN are negatively impacted by the augmentation of the feature space. Practitioners should bear this in mind when they use AUGMENT with their models.

We also reiterate that practitioners should consider whether a substantial portion of their features are sparse or binary before using CORAL. Sun et al. [72] also

[1]For example, Random Forests randomly chooses a subset of the features at each node, meaning that the inclusion of a large number of uninformative features reduces the probability that an informative feature will be selected at each node. Similarly, the Naive Bayes classifier assumes conditional independence between all features, so including uninformative features (e.g. $p(f_d|y_n = 0) \approx p(f_d|y_n = 1)$) will hurt the probability of a label being classified correctly by biasing all probabilities towards 0.5.

found CORAL performed poorly on text data sets and they hypothesized it was due to the lack of correlation between the sparse bag-of-words features. We also suspect the poor performance is due to centering the features as a preprocessing step, as centering destroys sparsity. Better results with CORAL may be also obtained with word embeddings as they are lees sparse.

7.1 Future Work

There are multiple directions we would like to take this work in the future. We discuss the spoken and written data sets separately for clarity.

7.1.1 Spoken

One aspect of the AUGMENT domain adaptation algorithm that was not used in this work is its ability to accommodate data from multiple source domains. This is done via a trivial extension to the standard feature space augmentation (c.f., Section 5.1.1) and would allow us to include source data from patients with vascular dementia, dementia with Lewy bodies, and other non-Alzheimer's dementias along with the AD data used in Chapter 5. Incorporating source data from multiple pathologies could potentially improve our diagnostic capabilities, but this has yet to be shown.

Rather than using AUGMENT to leverage data from multiple pathologies, we could use it to leverage data from multiple diagnostic exams. For example we could potentially improve upon the results in Chapter 4.1 by using AUGMENT with data collected from the Narrative Retelling task from the Wechsler Logical Memory I/II test, or the blog data we discuss in Chapter 6. In these settings discourse features, which Chapter 4 showed were not predictive on the DementiaBank data set, may also be more useful given the narrative structure of the speech samples.

We suspect CORAL performed poorly on the DementiaBank data set because of the boolean info-unit features. A small modification to CORAL, where we align (c.f., Section 5.1.2) only the non-boolean features, could improve CORAL's performance. Another potentially interesting area of future work would be merge AUGMENT and CORAL into a single algorithm by adding a "CORAL aligned" copy of the feature to the AUGMENT feature space.

7.1.2 Written

In Chapter 6 we confirmed it is possible to detect signs of dementia from blog posts. There were a few limitations of our approach that we would like to address in future work. First, the small size of our data set meant we were unable to differentiate between subtypes of dementia (e.g., Dementia with Lewy Bodies and AD). This is not desirable because different pathologies have different symptoms (cf. Table 2.2 and 2.1). We would like to collect a larger data set to allow us to control for types of dementia, as well as demographic information such as age, gender, and education level - information that was not present for all the blogs in our study.

Another limitation of the above work is the unstructured nature of the text. Unlike with the DementiaBank data set, none of the bloggers were constrained to a single topic, beyond the general topic of "living with dementia". We could potentially improve our results by performing a topic clustering preprocessing step on the blog posts. After clustering we could either train a classifier separately for each cluster or include topic membership as a feature.

Topic clustering would also help us to better understand the differences we found in some linguistic markers between bloggers (cf. figure 6.4 and 6.5). We observed bloggers with dementia have a higher SUBTL score (indicating an impoverished vocabulary) and shorter average word length compared to healthy controls. These findings need further investigation to confirm if they are in fact due to dementia-induced aphasia, as medical literature would predict. In Masrani et al. [50] we had looked at the longitudinal change of one of the metrics, the SUBTL word score, to see if the bloggers became more symptomatic as the disease progressed. Results were inconclusive however, with the longitudinal trend of the SUBTL score moving in the direction opposite to what we expected. With topic clustering, we could track the longitudinal changes of certain linguistic markers within each topic, as well as the longitudinal changes in the topics themselves, to better understand the differences in writing style between the writers with and without dementia.

Finally, we hope to explore how aphasia manifests itself on different online platforms. Language is shaped by its environment and linguistic features that are useful in classifying blog posts may not be useful classifying tweets. Today only

34% of seniors use social media [10]. That number will surely rise as the Internet generation reaches adulthood and continues to use instant messenger, to comment on Facebook posts, and to converse on online forums. It therefore behooves us to understand how to detect signs of cognitive decline in these settings.

Bibliography

[1] M. Abdalla, F. Rudzicz, and G. Hirst. Rhetorical structure and alzheimers disease. *Aphasiology*, pages 1–20, 2017. → pages 28

[2] 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. → pages 11

[3] A. Association. 2016 alzheimer's disease facts and figures. https://www.alz.org/documents_custom/2016-facts-and-figures.pdf, 2016. Accessed: 2017-11-13. → pages 1, 5, 7, 62

[4] A. P. Association et al. *Diagnostic and statistical manual of mental disorders (DSM-5®)*. American Psychiatric Pub, 2013. → pages 6

[5] J. Birks and R. J. Harvey. Donepezil for dementia due to alzheimer's disease. *The Cochrane Library*, 2006. → pages 9

[6] J. Blitzer, R. McDonald, and F. Pereira. Domain adaptation with structural correspondence learning. pages 120–128, July 2006. → pages 44

[7] L. X. Blonder, E. D. Kort, and F. A. Schmitt. Conversational discourse in patients with alzheimer's disease. *Journal of Linguistic Anthropology*, 4(1): 50–71, 1994. → pages 27

[8] M. Brysbaert and B. New. Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for american english. *Behavior Research Methods*, 41:977–990, 2009. → pages 20

[9] R. S. Bucks, S. Singh, J. M. Cuerden, and G. K. Wilcock. Analysis of spontaneous, conversational speech in dementia of alzheimer type: Evaluation of an objective technique for analysing lexical performance. *Aphasiology*, 14(1):71–91, 2000. → pages 19

[10] P. R. Center. Technology use among seniors. http://www.pewinternet.org/2017/05/17/technology-use-among-seniors/, 2017. Accessed: 2017-12-06. → pages 66

[11] L. W. Chambers, C. Bancej, and I. McDowell. Prevalence and monetary costs of dementia in canada. *The Alzheimer Society of Canada*, 2016. → pages 1

[12] S. B. Chapman, H. K. Ulatowska, K. King, J. K. Johnson, and D. D. McIntire. Discourse in early alzheimer's disease versus normal advanced aging. *American Journal of Speech-Language Pathology*, 4(4):124–129, 1995. → pages 27

[13] C. Chelba and A. Acero. Adaptation of maximum entropy capitalizer: Little data can help a lot. *Computer Speech & Language*, 20(4):382–399, 2006. → pages 44

[14] M. M. Cherrier, M. F. Mendez, M. Dave, and K. M. Perryman. Performance on the rey-osterrieth complex figure test in alzheimer disease and vascular dementia. *Cognitive and Behavioral Neurology*, 12(2):95–101, 1999. → pages 25, 36

[15] M. A. Covington and J. D. McFall. Cutting the gordian knot: The moving-average type–token ratio (mattr). *Journal of Quantitative Linguistics*, 17(2):94–100, 2010. → pages 19

[16] 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 20

[17] H. Daume. Frustratingly easy domain adaptation. 2007. → pages 44, 45, 50

[18] 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 28

[19] K. Dijkstra, M. S. Bourgeois, R. S. Allen, and L. D. Burgio. Conversational coherence: Discourse analysis of older adults with and without dementia. *Journal of Neurolinguistics*, 17(4):263–283, 2004. → pages

[20] 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, 2016. → pages 28

[21] D. G. Ellis. Coherence patterns in alzheimer's discourse. *Communication Research*, 23(4):472–495, 1996. → pages 27

[22] V. W. Feng. *RST-style discourse parsing and its applications in discourse analysis*. PhD thesis, University of Toronto, 2015. → pages 28

[23] L. K. Ferreira and G. F. Busatto. Neuroimaging in alzheimer's disease: current role in clinical practice and potential future applications. *Clinics*, 66: 19–24, 2011. → pages 9

[24] S. H. Ferris and M. Farlow. Language impairment in alzheimers disease and benefits of acetylcholinesterase inhibitors. *Clinical interventions in aging*, 8: 1007, 2013. → pages 2

[25] T. S. Field, V. Masrani, G. Murray, and G. Carenini. Improving diagnostic accuracy of alzheimer's disease from speech analysis using markers of hemispatial neglect. *Alzheimer's & Dementia: The Journal of the Alzheimer's Association*, 13(7):P157–P158, 2017. → pages iv, 4

[26] K. C. Fraser, G. Hirst, N. L. Graham, J. A. Meltzer, S. E. Black, and E. Rochon. Comparison of different feature sets for identification of variants in progressive aphasia. *ACL*, page 17, 2014. → pages 12, 21, 30, 32, 33, 43

[27] 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, 2015. → pages 2, 12, 17, 30

[28] 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. → pages 17

[29] E. Giles, K. Patterson, and J. R. Hodges. Performance on the boston cookie theft picture description task in patients with early dementia of the alzheimer's type: missing information. *Aphasiology*, 10(4):395–408, 1996. → pages 15

[30] D. B. Hier, K. Hagenlocker, and A. G. Shindler. Language disintegration in dementia: Effects of etiology and severity. *Brain and language*, 25(1): 117–133, 1985. → pages 33

[31] G. Hirst and V. Wei Feng. Changes in style in authors with alzheimer's disease. *English Studies*, 93(3):357–370, 2012. → pages 13

[32] G.-Y. R. Hsiung, A. Donald, J. Grand, S. E. Black, R. W. Bouchard, S. G. Gauthier, I. Loy-English, D. B. Hogan, A. Kertesz, K. Rockwood, et al. Outcomes of cognitively impaired not demented at 2 years in the canadian cohort study of cognitive impairment and related dementias. *Dementia and geriatric cognitive disorders*, 22(5-6):413–420, 2006. → pages 11

[33] A. Hunt, P. Schönknecht, M. Henze, U. Seidl, U. Haberkorn, and J. Schröder. Reduced cerebral glucose metabolism in patients at risk for alzheimer's disease. *Psychiatry Research: Neuroimaging*, 155(2):147–154, 2007. → pages 9

[34] M. Husain. Hemineglect. *Scholarpedia*, 3(2):3681, 2008. → pages ix, 25

[35] A. D. International. Dementia statistics. https://www.alz.co.uk/research/statistics. Accessed: 2017-11-30. → pages 1

[36] S. Ishiai, R. Okiyama, Y. Koyama, and K. Seki. Unilateral spatial neglect in alzheimer's disease a line bisection study. *Acta neurologica scandinavica*, 93(2-3):219–224, 1996. → pages 25, 36

[37] S. Ishiai, Y. Koyama, K. Seki, S. Orimo, N. Sodeyama, E. Ozawa, E. Lee, M. Takahashi, S. Watabiki, R. Okiyama, et al. Unilateral spatial neglect in ad significance of line bisection performance. *Neurology*, 55(3):364–370, 2000. → pages 25, 36

[38] S. Joty, G. Carenini, and R. T. Ng. Codra: A novel discriminative framework for rhetorical analysis. *Computational Linguistics*, 2015. → pages x, 29

[39] M. Kasai, J. Ishizaki, and K. Meguro. Alzheimer's patients do not show left unilateral spatial neglect but exhibit peripheral inattention and simplification. *Dementia & Neuropsychologia*, 1(4):374–380, 2007. → pages 25

[40] S. Kemper, L. H. Greiner, J. G. Marquis, K. Prenovost, and T. L. Mitzner. Language decline across the life span: findings from the nun study. *Psychology and aging*, 16(2):227, 2001. → pages 13

[41] B. Klimova and K. Kuca. Speech and language impairments in dementia. *Journal of Applied Biomedicine*, 14(2):97–103, 2016. → pages viii, 6, 7

[42] A. König, A. Satt, A. Sorin, R. Hoory, O. Toledo-Ronen, A. Derreumaux, V. Manera, F. Verhey, P. Aalten, P. H. Robert, et al. Automatic speech

analysis for the assessment of patients with predementia and alzheimer's disease. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 1(1):112–124, 2015. → pages 13

[43] A. Kumar et al. Dementia: An overview. *Journal of Drug Delivery and Therapeutics*, 3(3):163–167, 2013. → pages viii, 7, 8

[44] V. Kuperman, H. Stadthagen-Gonzalez, and M. Brysbaert. Age-of-acquisition ratings for 30,000 english words. *Behavior Research Methods*, 44(4):978–990, 2012. → pages 19

[45] L. Kurlowicz and M. Wallace. The mini mental state examination (mmse). https://www.mountsinai.on.ca/care/psych/on-call-resources/on-call-resources/mmse.pdf, 1999. Accessed: 2017-11-10. → pages 9

[46] M. Laine, M. Laakso, E. Vuorinen, and J. Rinne. Coherence and informativeness of discourse in two dementia types. *Journal of Neurolinguistics*, 11(1):79–87, 1998. → pages 28

[47] K. M. Langa and D. A. Levine. The diagnosis and management of mild cognitive impairment: a clinical review. *Jama*, 312(23):2551–2561, 2014. → pages 10, 11

[48] X. Le, I. Lancashire, G. Hirst, and R. Jokel. Longitudinal detection of dementia through lexical and syntactic changes in writing: a case study of three british novelists. *Literary and Linguistic Computing*, 26(4):435–461, 2011. → pages 13

[49] J. S. Lin, E. O'Connor, R. C. Rossom, L. A. Perdue, B. U. Burda, M. Thompson, and E. Eckstrom. Screening for cognitive impairment in older adults: an evidence update for the us preventive services task force. 2013. → pages 11

[50] V. Masrani, G. Murray, T. Field, and G. Carenini. Detecting dementia through retrospective analysis of routine blog posts by bloggers with dementia. *BioNLP*, pages 232–237, 2017. → pages 4, 65

[51] V. Masrani, G. Murray, T. S. Field, and G. Carenini. Domain adaptation for detecting mild cognitive impairment. In *Canadian Conference on Artificial Intelligence*, pages 248–259. Springer, 2017. → pages iv, 4

[52] M. F. Mendez, M. M. Cherrier, and J. S. Cymerman. Hemispatial neglect on visual search tasks in alzheimer's disease. *Cognitive and Behavioral Neurology*, 10(3):203–208, 1997. → pages 25, 36

[53] A. Milner, M. Harvey, R. Roberts, and S. Forster. Line bisection errors in visual neglect: misguided action or size distortion? *Neuropsychologia*, 31 (1):39–49, 1993. → pages 25, 36

[54] S. O. Orimaye, J. S.-M. Wong, and K. J. Golden. Learning predictive linguistic features for alzheimer's disease and related dementias using verbal utterances. In *Proc. 1st Workshop. Computational Linguistics and Clinical Psychology (CLPsych)*, 2014. → pages 12

[55] A. Parton, P. Malhotra, and M. Husain. Hemispatial neglect. *Journal of Neurology, Neurosurgery & Psychiatry*, 75(1):13–21, 2004. → pages 25

[56] E. R. Peskind, S. G. Potkin, N. Pomara, B. R. Ott, S. M. Graham, J. T. Olin, S. McDonald, M. M.-M.-. S. Group, et al. Memantine treatment in mild to moderate alzheimer disease: a 24-week randomized, controlled trial. *The American journal of geriatric psychiatry*, 14(8):704–715, 2006. → pages 9

[57] R. C. Petersen. Mild cognitive impairment as a diagnostic entity. *Journal of internal medicine*, 256(3):183–194, 2004. → pages 10

[58] R. C. Petersen. Mild cognitive impairment. *CONTINUUM: Lifelong Learning in Neurology*, 22(2, Dementia):404–418, 2016. → pages 11

[59] M. J. Prince. *World Alzheimer Report 2015: the global impact of dementia: an analysis of prevalence, incidence, cost and trends*. London, 2015. → pages 1

[60] T. S. Ramachandran, S. Zachariah, and V. Agrawal. Alzheimer disease imaging. *E-medicine. medscape/article*, 336281, 2012. → pages 9

[61] V. Rentoumi, L. Raoufian, S. Ahmed, C. A. de Jager, and P. Garrard. Features and machine learning classification of connected speech samples from patients with autopsy proven alzheimer's disease with and without additional vascular pathology. *Journal of Alzheimer's Disease*, 42(s3), 2014. → pages 2, 12

[62] K. P. Riley, D. A. Snowdon, M. F. Desrosiers, and W. R. Markesbery. Early life linguistic ability, late life cognitive function, and neuropathology: findings from the nun study. *Neurobiology of aging*, 26(3):341–347, 2005. → pages 13

[63] B. Roark, M. Mitchell, J.-P. Hosom, K. Hollingshead, and J. Kaye. Spoken language derived measures for detecting mild cognitive impairment. *IEEE*

Transactions on Audio, Speech, and Language Processing, 19(7): 2081–2090, 2011. → pages 12, 44

[64] T. Salsbury, S. A. Crossley, and D. S. McNamara. Psycholinguistic word information in second language oral discourse. *Second Language Research*, 27(3):343–360, 2011. → pages 19

[65] A. Satt, A. Sorin, O. Toledo-Ronen, O. Barkan, I. Kompatsiaris, A. Kokonozi, and M. Tsolaki. Evaluation of speech-based protocol for detection of early-stage dementia. In *INTERSPEECH*, pages 1692–1696, 2013. → pages 13, 44

[66] E. Schwam and Y. Xu. Cognition and function in alzheimer's disease: identifying the transitions from moderate to severe disease. *Dementia and geriatric cognitive disorders*, 29(4):309, 2010. → pages 2

[67] A. Shahid, K. Wilkinson, S. Marcu, and C. M. Shapiro. Mini-mental state examination (mmse). In *STOP, THAT and One Hundred Other Sleep Scales*, pages 223–224. Springer, 2011. → pages 9

[68] D. A. Snowdon. Aging and alzheimer's disease: lessons from the nun study. *The Gerontologist*, 37(2):150–156, 1997. → pages 13

[69] A. Society. Ebixa (also known as memantine hydrochloride). http://www.alzheimer.ca/sites/default/files/files/national/drugs/drug_ebixa_2008_e.pdf, 2008. Accessed: 2017-11-13. → pages 9

[70] Statistica. Number of blogs worldwide from 2006 to 2011 (in millions). https://www.statista.com/statistics/278527/number-of-blogs-worldwide/, 2012. Accessed: 2017-11-23. → pages 52

[71] G. B. Stokin, J. Krell-Roesch, R. C. Petersen, and Y. E. Geda. Mild neurocognitive disorder: an old wine in a new bottle. *Harvard review of psychiatry*, 23(5):368, 2015. → pages 9, 10

[72] B. Sun, J. Feng, and K. Saenko. Return of frustratingly easy domain adaptation. In *AAAI*, 2016. → pages xii, 45, 46, 48, 51, 63

[73] G. Szatloczki, I. Hoffmann, V. Vincze, J. Kalman, and M. Pakaski. Speaking in alzheimers disease, is that an early sign? importance of changes in language abilities in alzheimers disease. *Frontiers in aging neuroscience*, 7, 2015. → pages 9

[74] D. F. Tang-Wai and N. L. Graham. Assessment of language function in dementia. *Geriatrics*, 11(2):103–110, 2008. → pages 6

[75] P. N. Tariot, M. R. Farlow, G. T. Grossberg, S. M. Graham, S. McDonald, I. Gergel, M. S. Group, et al. Memantine treatment in patients with moderate to severe alzheimer disease already receiving donepezil: a randomized controlled trial. *Jama*, 291(3):317–324, 2004. → pages 9

[76] L. Tóth, G. Gosztolya, V. Vincze, I. Hoffmann, G. Szatlóczki, E. Biro, F. Zsura, M. Pákáski, and J. Kálmán. Automatic detection of mild cognitive impairment from spontaneous speech using asr. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015. → pages 13, 44

[77] A. Venneri, R. Pentore, B. Cotticelli, and S. Della Sala. Unilateral spatial neglect in the late stage of alzheimer's disease. *Cortex*, 34(5):743–752, 1998. → pages 25, 36

Appendix A

Supporting Materials

Parts-of-speech (16)	Number of Nouns, Number of Verbs, Number Of Not-in-Dictionary, Mean Word Length, Number of Adverbs, Number of Adjectives, Number of Determiners, Number of Interjections, Number of Inflected Verbs, Number of Coordinate Conjunctions, Number of Subordinate Conjunctions, Ratio Noun-to-verb, Ratio Noun-to-Noun+Verb, Ratio Pronoun-to-noun, Ratio Coordinate to subordinate conjunctions, LightVerb Count
Context-free-grammar rules (45) (Using Penn Treebank POS Tags)	ADVP_to_RB, INTJ_to_UH, NP_to_DT_NN, NP_to_PRP, ROOT_to_FRAG, VP_to_AUX, VP_to_AUX_ADJP, VP_to_AUX_VP, VP_to_VBD_NP, VP_to_VBG, VP_to_VBG_PP, CONJP, TTR, UCP, VP, Avg NP Type Length Embedded, Avg NP Type Length Non Embedded, Avg PP Type Length Embedded, Avg PP Type Length Non Embedded, Avg VP Type Length Embedded, Avg VP Type Length Non Embedded, WHADJP, WHAVP, WHNP, WHPP, X, FRAG, INTJ, LST, NP Type Rate, VP Type Rate, PP Type Rate, P Proportion, NP Proportion, VP Proportion, NAC, NP, NX, PP, PRN, PRT, QP, RRC, ADJP, ADVP

Syntactic Complexity (27)	Mean Word Length, Mean words per utterance , Mean length of sentence, Mean length of T unit, Mean length of Clauses, Disfluency Frequency, Total Number Of Words, Number of Utterances, Tree height, Complex Nominal per T unit, Complex nominal per Clause, Coordinate Phrase per clause, Coordinate Phrase per T unit, Complex T unit ratio, Clause per sentences, Clause per T unit, Dependent Clause per sentences, Clause per T unit, T unit per sentence, Verb Phrase per T unit, Number of Complex Nominals, Number of Coordinate Phrases, Number of Dependent Clauses, Number of Sentences, Number of T Units, Number of Words, Number of Clauses
Vocabulary Richness (3)	MATTR, Brunet Index, Honore Statistic, Type-to-token ratio
Psycholinguistic (5)	Aoa Score, Concreteness Score, Familiarity Score, Imagability Score, SUBTLWord Score
Repetitiveness (5)	Min Cos Dist, Proportion Below Threshold 0, Proportion Below Threshold 0.3, Proportion Below Threshold 0.5, Avg Cos Dist
Acoustic (172)	mfcc_n_kurtosis (for $1 \leq n \leq 13$), mfcc_n_mean (for $1 \leq n \leq 13$), mfcc_n_skewness (for $1 \leq n \leq 13$), mfcc_n_var (for $1 \leq n \leq 13$), mfcc_n_vel_kurtosis (for $1 \leq n \leq 13$), mfcc_n_vel_mean (for $1 \leq n \leq 13$), mfcc_n_vel_skewness (for $1 \leq n \leq 13$), mfcc_n_vel_var (for $1 \leq n \leq 13$), mfcc_n_accel_kurtosis (for $1 \leq n \leq 13$), mfcc_n_accel_mean (for $1 \leq n \leq 13$), mfcc_n_accel_skewness (for $1 \leq n \leq 13$), mfcc_n_accel_var (for $1 \leq n \leq 13$), energy_accel_kurtosis, energy_accel_mean, energy_accel_skewness, energy_accel_var, energy_kurtosis, energy_mean, energy_skewness, energy_var, energy_vel_kurtosis, energy_vel_mean, energy_vel_skewness, energy_vel_var, fundemental_frequency_mean, fundemental_frequency_var

Information units (info-units) (40)	ObjectCookie (keyword), ObjectCupboard (keyword), ObjectCurtains (keyword), ObjectDishcloth (keyword), ObjectDishes (keyword), ObjectJar (keyword), ObjectPlate (keyword), ObjectSink (keyword), ObjectStool (keyword), ObjectWater (keyword), ObjectWindow (keyword), PlaceExterior (keyword), PlaceKitchen (keyword), SubjectBoy (keyword), SubjectGirl (keyword), SubjectWoman (keyword), ActionBoyTaking (keyword), ActionStoolFalling (keyword), ActionWaterOverflowing (keyword), ActionWomanDryingWashing (keyword), ActionBoyTaking (binary), ActionStoolFalling (binary), ActionWaterOverflowing (binary), ActionWomanDryingWashing (binary), ObjectCookie (binary), ObjectCupboard (binary), ObjectCurtains (binary), ObjectDishcloth (binary), ObjectDishes (binary), ObjectJar (binary), ObjectPlate (binary), ObjectSink (binary), ObjectStool (binary), ObjectWater (binary), ObjectWindow (binary), PlaceExterior (binary), PlaceKitchen (binary), SubjectBoy (binary), SubjectGirl (binary), SubjectWoman (binary)
Demographic (1)	Age
Discourse Features (39)	Comparison, Edu_rate, Topic-Change, Summary, Topic-Comment, Same-Unit, Evaluation, Contrast, Elaboration, Attribution, TextualOrganization, Cause, Explanation, Enablement, Joint, Depth, Background, Temporal, Condition, Manner-Means, Comparison_ratio, Topic-Change_ratio, Summary_ratio, Topic-Comment_ratio, Same-Unit_ratio, Evaluation_ratio, Contrast_ratio, Elaboration_ratio, Attribution_ratio, TextualOrganization_ratio, Cause_ratio, Explanation_ratio, Enablement_ratio, Joint_ratio, Background_ratio, Temporal_ratio, Condition_ratio, Manner-Means_ratio, Discourse_type_token_ratio

Halves Features (9)	Attention: Leftside, Concentration: Leftside, Repetition: Leftside, Perception: Leftside, Attention: Rightside, Concentration: Rightside, Repetition: Rightside, Perception: Rightside, Number of switches from LS to RS

Table A.1: List of all features.

Figure A.1: Plot showing the performance of the *halves* feature set without quadratic terms. The performance of Random Forest and Gaussian Naive Bayes is not hurt in this case as it is in figure 4.9. The performance of logistic regression also decreases without the quadratic terms.

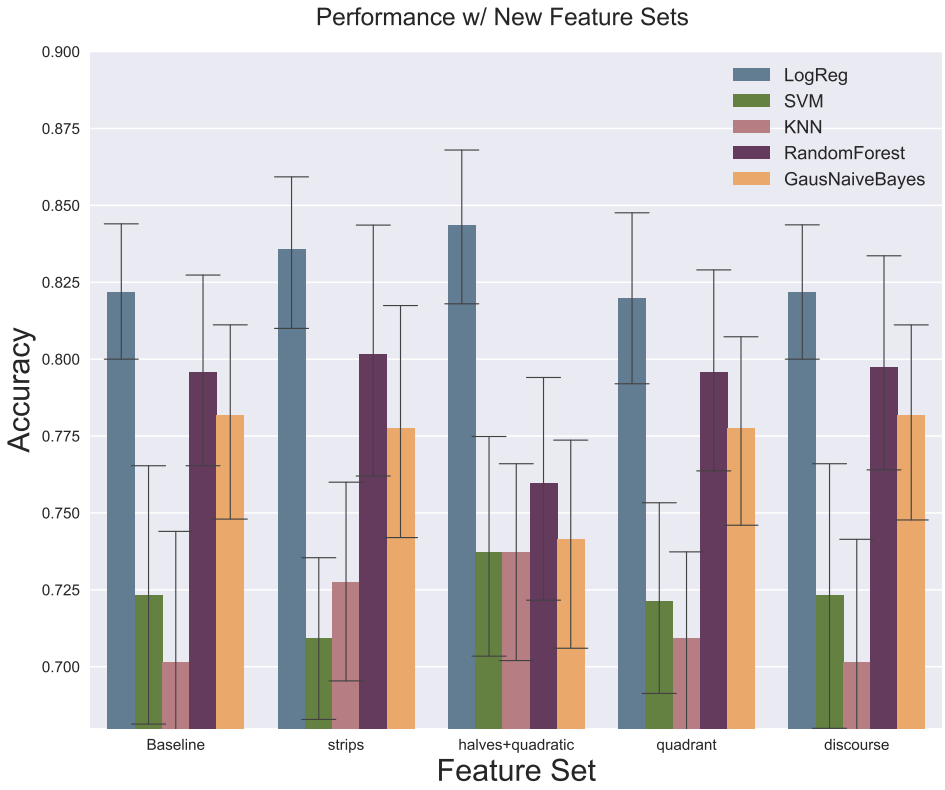

Figure A.2: Accuracy of models with new feature sets.

Figure A.3: Change in accuracy of models with new feature sets.

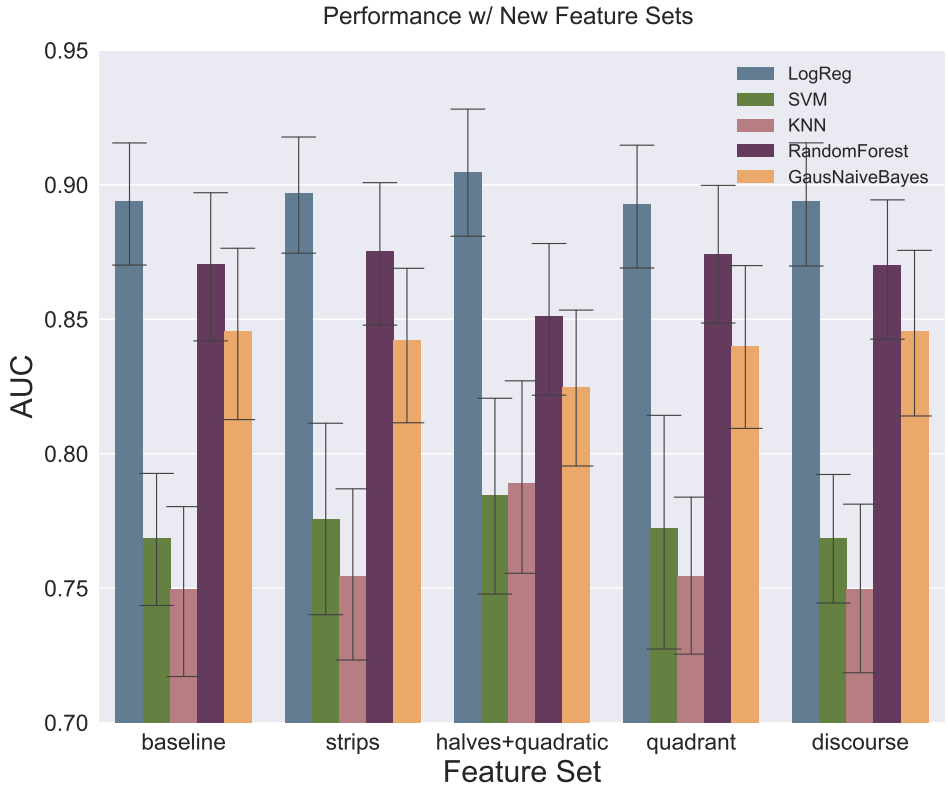

Figure A.4: AUC of models with new feature sets.

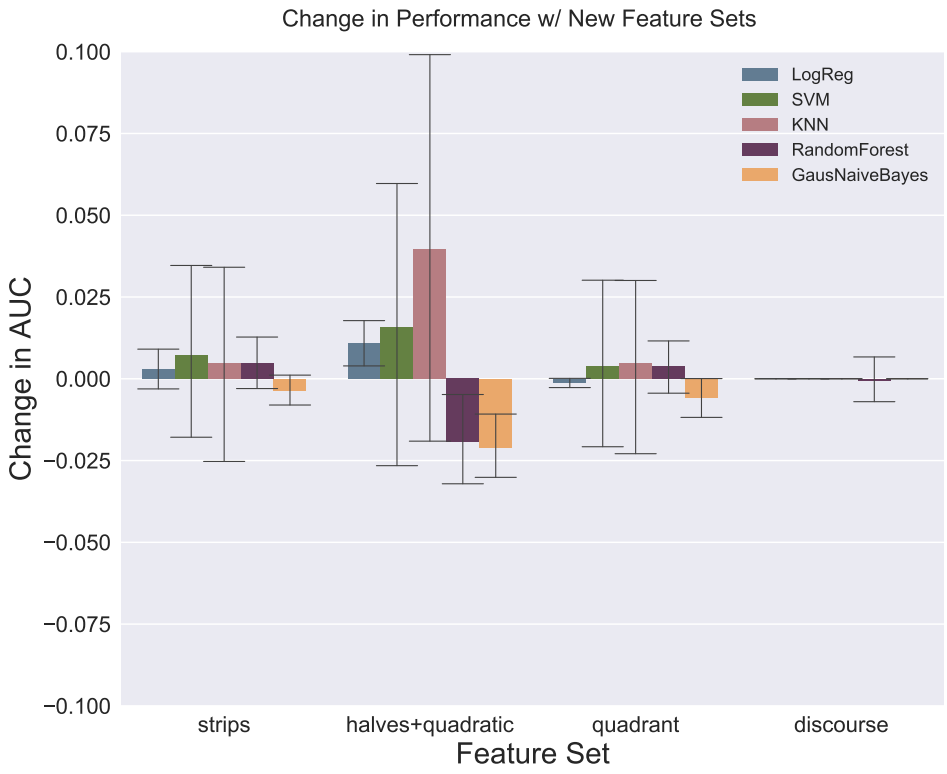

Figure A.5: Change in AUC of models with new feature sets.