An investigation into interactional patterns for Alzheimer's Disease recognition in Natural dialogues

SHAMILA NASREEN

Primary Supervisor: Dr. Matthew Purver

Secondary Supervisor: Professor Patrick G. T. Healey

Department of Electronic Engineering and Computer Science
QUEEN MARY UNIVERSITY OF LONDON

Phd thesis

JULY, 2023

Submitted in partial fulfillment of the requirements of the Degree of Doctor of Philosophy

Declaration

I, Shamila Nasreen, confirm that the research included within this thesis is my own work or that where it has been carried out in collaboration with, or supported by others, that this is duly acknowledged.

I attest that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge break any UK law, infringe any third party's copyright or other Intellectual Property Right, or contain any confidential material.

I accept that the College has the right to use plagiarism detection software to check the electronic version of the thesis.

I confirm that this thesis has not been previously submitted for the award of a degree by this or any other university.

Shamila Nasreen
July, 2023

ACKNOWLEDGEMENTS

Undertaking the journey of PhD has been a truly life-changing experience for me and it would not have been possible to embark on this journey and come out victorious without the support and guidance that I received from people for whom I have momentous respect and gratitude.

I am extremely grateful to my supervisor, Dr. Matthew Purver for his invaluable advice, continuous support, and encouragement during my PhD study. Dr. Mathew helped me in the development of the research topic, method, and materials, while generously lending his vast technical expertise throughout this process.

I also want to express my gratitude to my co-supervisors, Patrick G. T. Healey, and Julian Hough, for their insightful suggestions and reflections. I also want to thank Maria Liakata and Usman Naeem for their guidance as my independent assessor and, for their beneficial and valuable discussions about the project. I gratefully acknowledge the funding received towards my PhD from the Higher Education Commission of Pakistan in the form of their PhD studentship. It is through their kind help and support that I have made my study and life in the UK a wonderful time. Furthermore, I wish to express my gratitude to my examiners, Fasih Haider and Ioannis Patras, for their insightful feedback and generous provision of valuable guidance for future endeavours throughout the examination process.

I'd also like to thank everyone at Queen Mary's Cognitive Science Research group and the Computer Science department who read or reviewed this research and offered suggestions and advice.

A very special note of thanks to Shafaq Murtaza for pushing me to embark on and undertake this journey with continuous support and motivation. I would like to express my gratitude to my parents specially my mother, my brothers, and my sisters. Their encouragement and prayers over the past few years gave me the strength to keep moving forward. Finally, to my husband Mushood Iqbal Ahmed, who has been by my side throughout this PhD, and without whom, I would not have had the courage to embark on this journey in the first place and to my darling son Hamza Mushood for being such a good sport for a single smile of his made me forget all the woes in the world and made it possible for me to complete what seemed to be a laboriously long journey when I took the first step, but surprisingly - or maybe not so surprisingly - an experience of a lifetime.

Abstract

Alzheimer's disease (AD) is a complex neurodegenerative disorder characterized by memory loss, together with cognitive deficits affecting language, emotional affect, and interactional communication. Diagnosis and assessment of AD is formally based on the judgment of clinicians, commonly using semi-structured interviews in a clinical setting. Manual diagnosis is therefore slow, resource-heavy, and hard to access, so many people don't get diagnosed - and therefore using some kind of automatic method would help. Using the most recent advances in deep learning, machine learning, and natural language processing, this thesis empirically explores how content-free, interaction patterns are helpful in developing models capable of identifying AD from natural conversations with a focus on particular phenomena found useful in conversational analysis studies. The models presented in this thesis use lexical, disfluency, interactional, acoustic, and pause information to learn the symptoms of Alzheimer's disease from text and audio modalities.

This thesis comprises two parts. In the first part, by studying a conversational corpus, we find there are certain phenomena that are really strongly indicative of differences between AD and Non-AD. This analysis shows that interaction patterns are different between an AD patient and a Non-AD patient, including types of questions asked from patients, their responses, delay in responses in the form of pauses, clarification questions, signaling non-understanding, and repetition of questions. Although it is a challenging problem due to the fact that these dialogue acts are so rare, we show that it is possible to develop models that can automatically detect these classes.

The second part then shifts to look at AD diagnosis itself by looking into interactional features including pause information, disfluencies within patients speech, communication breakdowns at speaker changes in certain situations, Ngram dialogue act sequences. We found out that there are longer pauses within the AD patients utterances and more attributable silences in response to questions as compared to Non-AD patients. It also showed that using different fusion techniques with speech and text modality has maximise the combination and use of different feature sets showing that these features/techniques can give quite good accurate and effective AD diagnosis.

These interaction patterns may serve as an index of internal cognitive processes that help in differentiating AD patients and Non-AD patients and may be used as an integral part of language assessment in clinical settings.

TABLE OF CONTENTS

List of Tables

Table **Page**

List of Figures

Figure **Page**

CHAPTER 1

INTRODUCTION

1.1 Introduction

Alzheimer's disease (AD) is an irreversible, progressive deterioration of the brain that slowly destroys memory, language, thinking abilities, problem-solving, and eventually the ability to carry out the simplest tasks in Alzheimer's patients daily lives. Difficulty remembering recent conversations and events is found to be an early clinical symptom, later symptoms are impaired communication, poor judgment, disorientation, depression, and behavioural changes.

AD is the most prevalent form of dementia, contributing to 60%-70% among all types of Dementia (Tsoi et al., 2018). It affects approximately 6.7 million Americans with annual costs of care up to $340B in the United States, in 2023 (Association, 2023). According to facts and figures (Association, 2023), age affects how many people have Alzheimer's dementia: 5.0% of people with AD are between the ages of 65 and 74, 13.1% are between the ages of 75 and 84, and 33.3% are over the age of 85. About 200,000 Americans, in total, have dementia that develops before the age of 65. Around 13.8 million people aged 65 and older are anticipated to have the condition by the year 2060 (see figure 1.1). By 2040, there will be 1.6 million dementia sufferers in the UK, with 42,000 of them being under the age of 65 (Wittenberg et al., 2019). Every 65 seconds, someone in the US develops AD while every 3 minutes in the UK someone develops this disease.

Alzheimer's disease (AD) is the sixth most common cause of mortality in the United States and the fifth most common cause of death for people 65 and older. From 2000 to 2014, the number of AD fatalities grew by 89% (Association, 2017). According to current estimates, the number of people over 65 is predicted to triple between 2000 and 2050

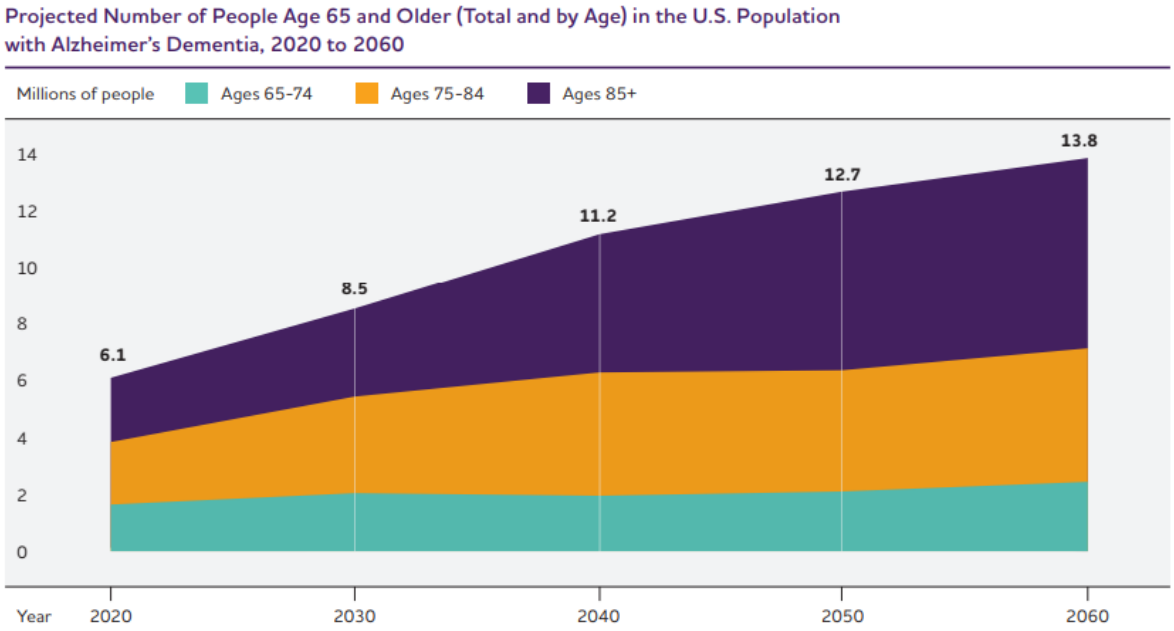


Figure 1.1: The anticipated number of Americans with Alzheimer's disease by 2060.

(World Health Organization, 2015). It is one of the top 10 causes of deaths that cannot be stopped, slowed down, or even treated. Statistics show that deaths due to Alzheimer's disease have increased and the record shows it 145% between 2000-2019 while the death rate caused by the number one death disease (heart-diseased) is decreased by 7.3% (Association, 2023).

There is no single universally accepted medical test for the diagnosis of AD, instead physicians use a variety of ways with the help of a specialist (including a neurologist) to help make a diagnosis. This includes: taking feedback from family members and carers asking about changed patterns in behaviors and thinking, getting family history, conducting cognitive tests with the help of a neurologist, and conducting individuals blood tests and brain imaging (MRI) to check if the individual has high amounts of beta-amyloid, which is an accumulation of protein fragments outside neurons and is one of numerous brain alterations linked to AD.

The criteria for determining whether someone has AD were set by the Alzheimer's Association and the National Institute of Neurological and Communicative Disorders and Stroke (NINCDS) and require that the presence of cognitive impairment needs to be confirmed by neuropsychological testing for a clinical diagnosis of possible or probable AD (Mckhann et al., 1984). The Mini-Mental Status Examination (Folstein et al., 1975), one of the most widely used tests, the Addenbrooke's Cognitive Examination-Revised (ACE-R) (Noone, 2015), the Hopkins Verbal Learning Test (HVLT) (Brandt, 1991), and the DemTect (Kalbe et al., 2004), are all appropriate neuropsychological tests.

Medical diagnoses based on clinical interpretation of patients' history, complemented

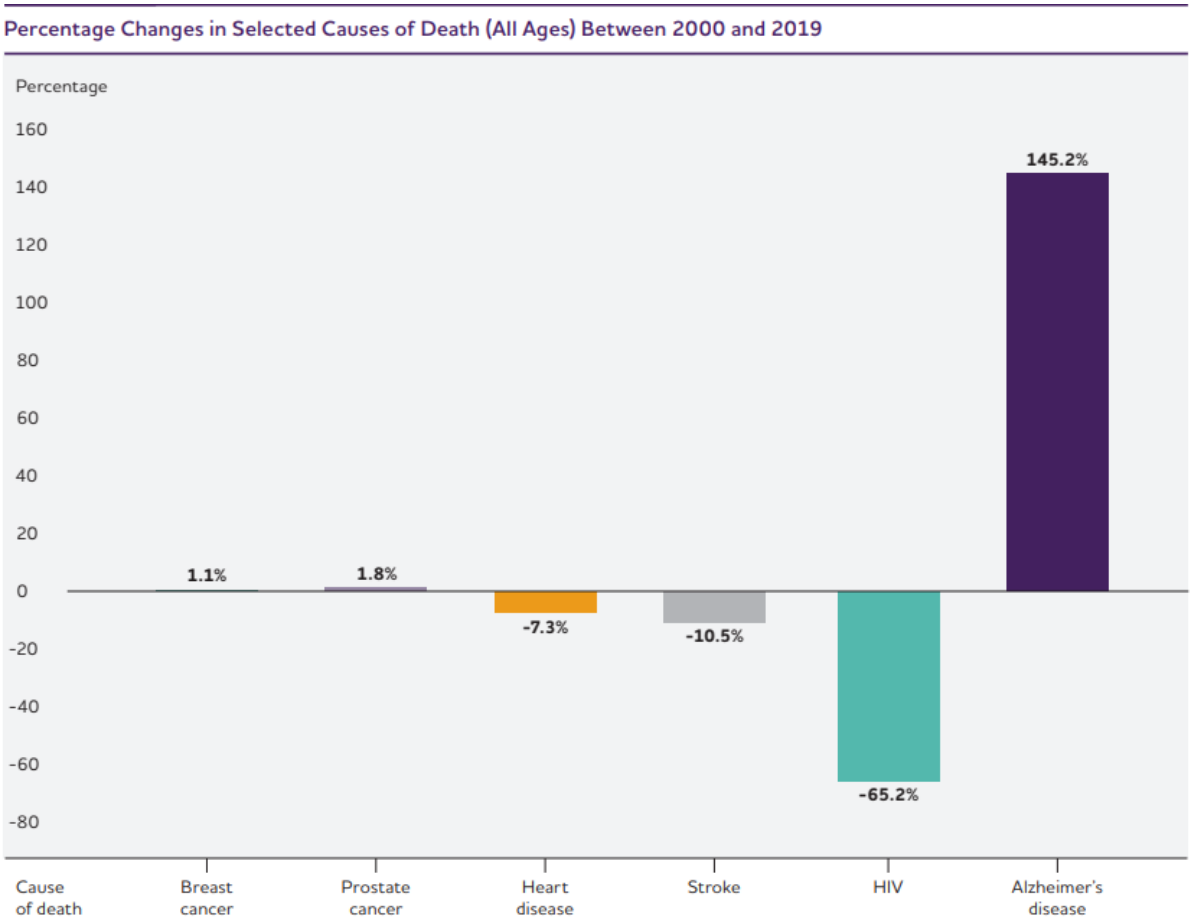


Figure 1.2: Changes in selected death causes (All Ages) by % from 2000 to 2019

by brain scanning (MRI) are time-consuming, stressful, costly, and often cannot be offered to all patients complaining about functional memory. The other alternatives are extensive neurological screening tests that are used for the early diagnosis of Alzheimer's disease and Dementia. These tests as discussed earlier require medical experts to interpret the results and are performed in medical clinics and patients have to visit the clinics for diagnosis.

In order to create tests that are simpler to administer and automate using natural language processing methods, researchers are currently examining how neurological impairment affects patients' speech and language (Fraser et al., 2016a). New methods are required that enhance and accelerate the early diagnosis process, lessen patient distress, and downplay the importance of lengthy, pricy formal testing. Currently, researchers are investigating the impact of neurodegenerative impairment on a patient's speech and language (Asgari et al., 2017) focusing on their interaction during a conversation in natural settings using natural language processing techniques.

The visual description task, which asks patients to describe a picture with simple or

complex settings, is one of the most popular cognitive ability assessments. This project enables repeating speech output because the description must include important elements that are depicted in the picture. As a result, the speech could be evaluated according to how many accurate components are found. The most popular visual description task used in the assessment of neurodegenerative disorders is the Cookie Theft picture description task from the Boston Diagnostic Aphasia Examination (Goodglass et al., 2001). Cookie Theft Picture Description Task and Fluency Test offer valuable information to clinicians in understanding the cognitive and linguistic functioning of individuals, aiding in diagnosis, treatment planning, and tracking changes in cognitive abilities over time. A clinical interview, which comprises structured and semi-structured interviews, is another technique to evaluate decline. (Kopelman et al., 1989).

Previous studies (Orimaye et al., 2017; Rohrer et al., 2010; Ahmed et al., 2013; Fraser et al., 2015) focused purely on language features including lexical, grammatical, and syntactic features extracted from the transcripts of narrative speech and some acoustic features from the audios from DementiaBank dataset and using picture description task, Vocabulary Test and Naming test. Individuals with AD generate more lexical mistakes (Kavé and Levy, 2003), with disfluency, repairs, and word seeking difficulties (Forbes-McKay and Venneri, 2005) and frequent repetitions (Drummond et al., 2015). Pauses present a compensatory strategy in early cognitive decline and Pistono et al. (2019b) found a strong correlation between memory capabilities and the level at which pauses occur. Davis and Maclagan (2010) identified moments of silence during a storytelling task and observed shifts in the function of these pauses, indicating a transition from challenges in word retrieval to struggles in identifying crucial elements within the storyline.

The focus of conversation analysis (CA) on conversation as a collaborative achievement shows that analyzing interaction can offer more insight than analyzing the contributions of the two participants independently. Each contribution to the conversation is based on and in response to the partner's prior contribution. CA studies are qualitative studies that looked into details at what characteristics of conversation might be important to Dementia. These studies particularly focused on phenomena such as interaction among patient-doctor, and response to different types of questions (Elsey et al. (2015); Hamilton (2005); Varela Suárez (2018); Small and Perry (2005)), looking at question answers as adjacency pairs ((Sacks et al., 1978).

Conversational clues are either missing in traditional approaches like picture description tasks or do not capture important conversational phenomena while analyzing the individual's speech. Interactional features, therefore, promise to design models that are non-content-based and non-invasive. Although there are now some computational systems for the interaction-based detection of dementia (Luz et al., 2018; Zhu et al., 2018) , there is still little research on how dementia could affect interactional patterns (Addlesee et al., 2019). In the last couple of years, a body of work has started looking

into dialogue phenomena for the detection of AD/Dementia (Farzana and Parde, 2022; Mirheidari et al., 2019). The research in this thesis contributes to this growing body of research that supports spontaneous speech assessment as a potential approach for the detection of AD through building computational models to capture interaction in natural conversations. This work also seeks to use multimodal techniques to detect AD, by looking into different modalities and concentrating on the best methods and their combinations.

1.2 Aim

Using the most recent developments in representation learning, machine learning, and natural language processing, I aim to develop models that can detect cognitive decline by identifying the differences and similarities that should be taken into account in computational modeling of cognitive decline from spontaneous speech in the current research.

The current work adopts the strategy of analyzing spontaneous speech use as it pertains to a wide range of symptoms of cognitive decline and more fully characterizes the progression of the disorder. In this study, we explore different aspects of the use of spontaneous speech in natural dialogue conversations of AD and Non-AD patients including language that comprises of lexical representation of spoken words and disfluencies; dialogue act sequences capturing the patient-interviewer interaction; and more generic interaction in terms of durational aspects of interaction. This study is carried out in two folds: a corpus analysis followed by building a dialogue act tagger and an AD classification task. We also intended to study some acoustic features in both DA tagging and AD classification that has been demonstrated to be helpful in detecting AD. To keep the scope manageable, we will focus on semi-structured interviews within the community center and aim at developing models capable of identifying AD and interpretable symptoms. There will also be experiments with different fusion techniques with different combinations of features.

Medical diagnosis is resource heavy, expensive, and hard to access for everyone. Other alternatives are extensive screening tests that still need to be performed in clinical setups. Researchers have also created many methods of analysis for recorded and transcribed speech using acoustic, lexical, syntactic, and pragmatic aspects in order to produce assessment tools that are widely accessible. We want to increase the diagnostic utility of conventional neuropsychological language tests so they may be less invasive, affordable and used by a wide range of people. NLP techniques can help in identifying changes in the speech of elderly patients by analysing spontaneous speech. We aimed this work to determine whether investigated interactional patterns of natural conversations can assist in the diagnosis of AD. This study also investigates the impact of these significant

characteristics/identifiable patterns in the dialogues of patients with some combination of language features that helps in distinguishing AD patients from control/Non-AD persons by using NLP methods.

1.3 Research Questions

The research questions this work seeks to examine are:

- **Research Question 1: What are the most significant interactional patterns that are effective in AD diagnosis?**
- **Research Question 2: To what extent, Dialogue Act (DA) tagging and classifying the conversational dialogues give useful interactional patterns about AD patients?**
- **Research Question 3: To what extent can representation learning improve the rare class conversational phenomena that will be distinctive in AD and Non-AD groups?**
- **Research Question 4: Which durational aspects of spontaneous speech along with interaction can serve as distinguishable interactional features between AD and Non-AD groups?**
- **Research Question 5: Can we develop models which combine the most significant interactional aspects of communication with language aspects to maximize the accuracy of AD diagnosis?**
- **Research Question 6. What are the benefits of using various modalities in modelling several forms of features for AD identification?**
- **Research Question 7: What machine learning architecture is more robust in AD identification and to what extent state of the art models identify AD in clinical setups and in more natural settings like community centers?**

1.4 Summary of contributions

Following is a list of contributions that are achieved in this thesis:

1. A rare class DA annotation scheme is developed that is suitable for conversational dialogues to capture the dialogue phenomena and present a comparative analysis of the subject's interaction patterns through a corpus study.

2. A subset of the Carolinas Conversation Collection (CCC) corpus is annotated with the proposed rare class DA annotation scheme. These annotations are available for the research community for further follow-up work and can be used after getting access to the CCC dataset[1].

3. A subset of the CCC corpus now includes an additional collection of conversational pauses with annotations.

4. To capture the dialogue interaction, a rare class dialogue act tagger is developed with deep-learning models that leverage utterance representation with word embeddings, speaker change information, a few previous utterances, and DA's as context to predict the DA for the current utterance. The model utilised both static utterance representation and pre-trained contextualized utterance representation along with acoustic features from speech data. I also fine-tune the pre-trained BERT model on the downstream task on both corpora and achieve the most robust performance (macro F1: 0.58 0n SwDA and 0.48 on CCC). Later it was shown that the challenging setup of rare-class DA labeling for better recognizing rare classes in the CCC data set really helped to detect Alzheimer's disease better.

5. To capture the full conversation context, a conversational level DA tagger is developed using a Conditional Random Field (CRF) to model the sequence of DA tags , following an approach used successfully in general DA tagging work (e.g. (Kumar et al., 2018; Srivastava et al., 2019)). However, this CRF approach did not perform very well with our rare DA classes, suggesting that it is not well suited for our task with its highly uneven class distribution.

6. A set of signal of interaction characterised by specific rare class dialogue acts obtained from rare-class tagger are used in the form of unigram and bigram dialogue act sequences in the AD detection task and it was shown that these rare class dialogue acts feature helped in differentiating between AD and Non-AD group.

7. A set of interactional features is proposed including conversation pauses, for the identification of AD and it showed that they yield high utility in differentiating between AD and Non-AD. The disfluency features are also combined with interactional features and it was shown that the combination of the interaction features with the disfluency features is helpful for the AD classification task in a natural setting. These models were able to achieve performance comparable to the work that is based on content-driven task-specific settings.

[1]Annotations: https://osf.io/8w9z2/?view_only=ee08242870f24ae7ab6754ddf9a0176a

8. Both unimodal and multimodal AD detection methods were tested and the performance is compared and contrasted over several combinations from different feature sets including interactional features, acoustic features, and DA's based features e.g. unigram, bigram, and confusion ratios. These models learn AD markers using speech and text visual modalities. A comprehensive research of fusion strategies for including early fusion, late fusion, and multimodal fusion is also presented. Lastly, it is shown that feature selection along with fusion strategy outperforms when the features are alone used for AD identification.

1.5 Outline of Thesis

Figure 1.3 represents a systematic diagram of all the proposed work presented in this thesis through analysis and AD detection approach using DA tagging and more generic interactional features. Major components are linked to the relevant studies (chapters) and the overview of each chapter is given below.

Chapter 2 In this chapter key symptoms are described that a clinician may consider for identifying AD. The literature review covers several different existing diagnostic approaches primarily focusing on linguistic feature analysis; non-content feature analysis; and conversational analysis. Few dialogue act models with different universally accepted annotation schemes and existing models of DA tagging/classification are also discussed.

Chapter 3 presents our initial corpus study and analysis on a conversational dataset. Our first step towards supporting our hypothesis is that signals of interaction characterized by specific rare dialogue act (DA) classes such as clarification request, signals of non-understanding, repeat questions, distribution of different types of questions, and distribution of pausing at various places are different between AD and Non-AD groups. In particular, different conversation analysis studies are reviewed and interactional cues are analyzed on a more natural and conversational dataset in terms of Dialogue act tags. An annotation scheme is also designed to express the signal of interactions represented as dialogue acts. The distributions of these different DA's in conversational dialogues are explored and provide the base motivation for the experiments presented in the next two chapters.

Chapter 4 Drawing on the observations of the corpus study in Chapter 3, an experiment is presented in which spoken utterances with different previous contexts are examined. A rare class dialogue act tagger is developed to identify the appropriate DA for the spoken utterances from both patients and interviewers. Current utterance along with immediate preceding utterance and previous DA are found to be

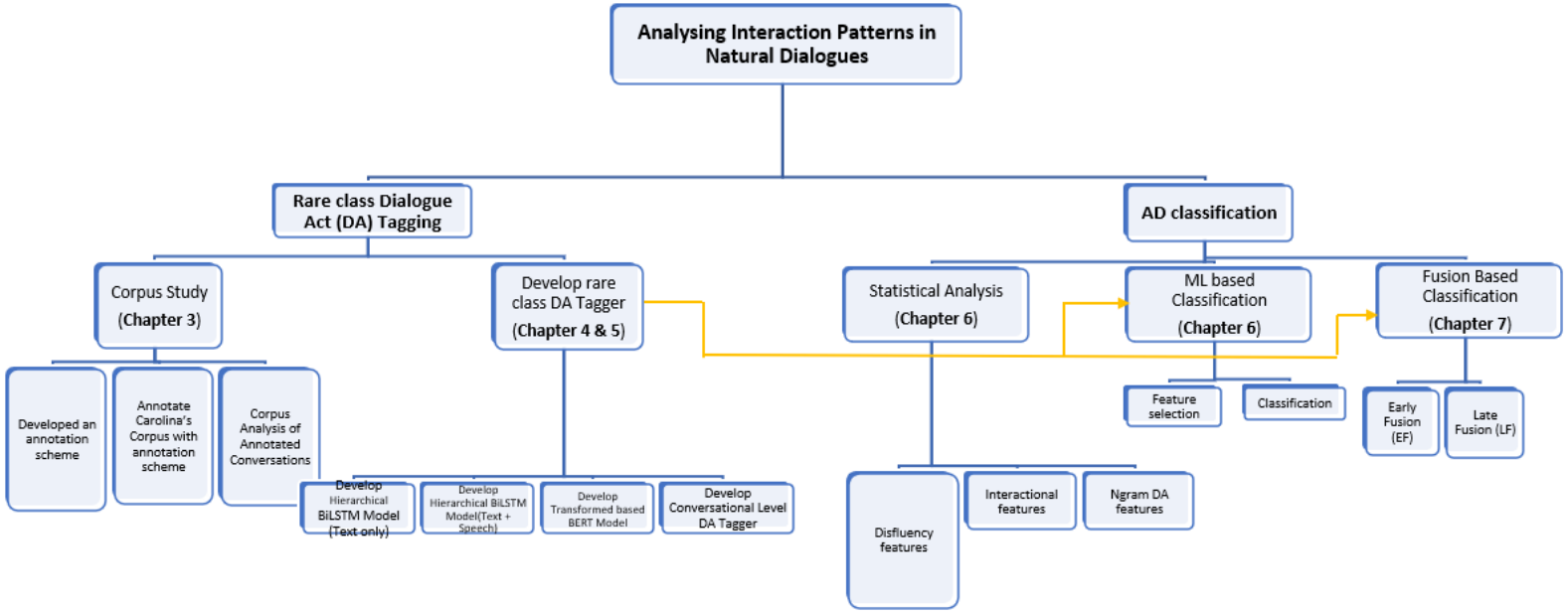


Figure 1.3: Systematic Diagram of proposed work. The yellow lines shows that the results from developed DA taggers are used in diagnosis in chapter 6 and 7.

useful in predicting the rare class DA for the current spoken utterance. Different utterance representations including Glove embeddings and embeddings from language models (ELMo) are examined with different context settings and results are presented.

Chapter 5 extends experiments on building rare class dialogue act tagging from chapter 4. Based on experiments in chapter 4, some rare DA classes such as clarification request, declarative questions got poor performance. To improve the class-wise performance of certain classes, acoustic features from speech are added. Static word embedding and contextualized word embeddings such as Bidirectional Encoder Representations from Transformers (BERT) are also explored and a comparative analysis on the results is presented. These contextualized utterance representation along with acoustic features outperform for certain DA classes. An other experiment is performed to build a conversational level dialogue act model to capture the full dialogue as context. This is further supported by using Conditional Random Fields (CRF) to capture the dependency between the sequence of dialogue acts.These predicted DA classes are used as unigram and bigram features in next chapter for the AD classification task.

Chapter 6 focuses on dialogue acts features from previous chapter along with more interaction features between patient-interviewer conversations and disfluency features for detecting cognitive decline in spontaneous speech in the context of Alzheimer's disease diagnosis.The ability to distinguish between the functionality of various types of pauses is important for analyzing conversations and can also be beneficial for detecting AD. An annotation protocol for various functions of pauses is also presented and natural conversations are annotated based on this annotation scheme. A statistical analysis is performed on interactional features, disfluency feature and DA features. Then, traditional Machine learning approaches are used to automatically detect the AD from the earlier stated feature set.

Chapter 7 In chapter 7, a second experiment is presented in which data from various modalities is used including text and speech, taking into account the different combinations of features including acoustic features from speech, interactional features obtained in chapter 6, language features and DA's sequences obtained from chapter 5 using different fusion strategies. A comprehensive analysis of different fusion strategies with different combinations of features is also presented.

Chapter 8 concludes the thesis with main findings resulting from the work. Referring back to the relevant literature it outlines the developments made within the work and answers to the research questions, alongside the limitations of the study and proposals for our research's future direction.

Appendix A contains the annotation protocol for dialogue act tagging for initial study in chapter 3 and later used in building rare class dialogue act tagger.

Appendix B contains annotation protocol for pauses types along with examples of different pauses types that will be used in experimental work in chapter 6. It also contains annotation for acoustic features from speech data for experimental work in chapter 7

Appendix C have statistics about the responses generated against different types of questions.

Appendix D records the detailed statistical analysis results for DA's unigram and bigram features.

1.5.1 Associated Publications

- The corpus study of analysing interaction on conversational dialogues discussed in chapter 3 was presented at SemDial 2019 in 23rd WORKSHOP ON THE SEMANTICS AND PRAGMATICS OF DIALOGUE in London, in the paper (Nasreen et al., 2019).
- The experimental work on building a rare class dialogue act tagger to capture the important cues found useful through corpus study was presented in the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL 2021) in the paper (Nasreen et al., 2021a).
- The experimental work based on extracting the interaction feature along with disfluency features used in AD detection (presented in chapter6) was submitted and published in Frontier in Computer science section human media interaction in Special edition 'Alzheimer's Dementia Recognition through Spontaneous Speech' in 2021 in paper (Nasreen et al., 2021b).
- The Experimental work presented in chapter 7 based on utilizing multimodal information from spoken language and from speech for the detection of AD was published in INTERSPEECH 2021 in the paper (Nasreen et al., 2021c).

CHAPTER 2

BACKGROUND

This section provides the background necessary for understanding the proposed work. As this Ph.D. research focuses primarily on the study of improving the diagnosis of Alzheimer's disease from natural conversations, most of this section is dedicated to providing an overview of the amount of research in NLP for Dementia diagnosis based on properties of an individual's language in isolation. Section 2.1 covers a brief overview of the studies that have looked into lexical, and acoustic features in AD sufferer's narrative speech, section 2.2 discussed non-content features in the spontaneous speech of patients, and then the conversational analysis that has been done so far in section 2.3. In the next section, recent developments that have started looking into interaction patterns in natural conversations will be discussed. Recently people started looking into dialogue act-based tags as a kind of conversational feature to capture the interaction and in light of this, I will look into state-of-the-art methods for work on dialogue act tagging.

Many clinical studies have proven that language deficit is present in many neurodegenerative diseases and this language impairment plays an important role in the identification and diagnosis of AD and primary progressive Aphasia. Although the most prominent impairment in AD is concerned with episodic memory, language impairment and combination of some other cognitive disorder are also present with the progression of the disease (Boschi et al., 2017). Word comprehension and verbal fluency impairment, as well as anomias and semantic paraphasias, are the most common language deficits reported (Forbes-McKay and Venneri, 2005), alteration in discourse planning at a pragmatic level (Chapman et al., 1998), while Phonological and syntactic processing is relatively spared, at least in early stages (Kavé and Levy, 2003). Language impairment in AD is worse and more pervasive as the disease develops with speech restricted to echolalia and verbal stereotypes(Ferris and Farlow, 2013).

2.1 Linguistic Feature Analysis of Narrative speech

In this section,a discussion on previous research in AD diagnosis that has mainly focused on language impairment using a different kinds of linguistic and acoustic features in an interviewed based setting will be presented. In general, patients are given a narrative narration task to see how well they can develop a story that includes a series of events, actions, and semantic units (subjects and objects). Thus, the number of objects, subjects, events, and activities stated in sequential or temporal sequence can be used to measure how thorough a picture description task is. These traits make it simple to analyze lexical and semantic skills, grammatical complexity (de Lira et al., 2011), discourse and pragmatic information (Drummond et al., 2015).

Orimaye et al. (2017) argued that developing an automated diagnosis model with low-level language cues gleaned from verbal utterances could aid in the detection of Probable AD in a large population. They used lexical features, syntactic features, and N-gram as semantic features obtained from 99 probable AD patients and 99 control participants from the Dementia Bank Corpus. This dataset includes transcripts of interviews of patients with interviewers based on the description of the Cookie-Theft picture component of the Boston Diagnostic Aphasia Examination (BDAE). The task of describing Cookie theft picture has been considered clinically pertinent in determining a linguistic deficit in both Dementia and Aphasia (Rohrer et al., 2010). Lexico-semantic features were applied byThomas et al. (2005) to identify impairments at the word and content levels. Part-of-speech categories, such as open-class and closed-class words, or more particularly nouns, verb, etc., can be used to categorize words. The author showed that the average rate of occurrence for each part-of-speech category can be used to identify difficulties in accessing a specific word class or to analyze the lexical distribution of words produced.

Ahmed et al. (2013) demonstrated that essential characteristics of connected speech that are used to analyze longitudinal profiles of impairment in Alzheimer's disease include syntactic complexity, lexical content, speech production, semantic content, idea density, and idea efficiency. With narrative speech, Fraser et al. (2015) used POS tags, syntactic complexity, information content and psycholinguistic features from transcripts, and acoustic features from speech to discriminate between possible and probable AD.

Fraser et al. also worked on detecting the presence or absence of depression in AD patients using a variety of features including syntactic complexity, acoustic characteristics, fluency, and the information content (Fraser et al., 2016c). They combined the three additional psycholinguistic features of arousal, valance, and dominance with the textual features (part-of-speech tags, psycholinguistic measures, parse constituents, measures of complexity, vocabulary richness, and informativeness) and acoustic features (fluency measures, voice quality features, MFCCs, and measures of periodicity and symmetry) to extract features from the transcripts.

Kavé and Dassa (2018) showed how Dementia severity is associated with language features and information content in a picture description task in the Hebrew language other than English. This study covers only the grammatical and lexical aspects of language using ten features. In comparison to the control group, it was discovered that those with AD produced more frequent words, a lower type-token ratio, a smaller percentage of content words, and more pronouns in relation to nouns and pronouns. Another study showed that AD patients with semantic variant produce less noun (Fraser et al., 2014) and these are often replaced by the pronoun (Jarrold et al., 2014) as compared to non-AD patients. A simplification of syntax denoted by a reduction of the mean length of utterance and a decrease in syntactic complexity was reported in AD patients by Fraser et al. (2015). Lexical/semantic disorder could be a factor in the lack of content, which would lead to fewer information units (Boschi et al., 2017). However, this is usually studied within particular language tasks (e.g. the Cookie Theft picture description task of the DementiaBank Pitt Corpus[1]).

Most of the studies discussed in this section have focused on a fixed task of picture description of Cookie theft picture from Pitt Corpus of Dementia bank at fixed times and all these approaches are based on guided interviews between interviewer and participants in the form of narrative speech. As these experimental studies are not done in natural settings, and are restricted in terms of time and quantity of information and hence can't be administered frequently. As these studies focused on the content of task being used, makes the models/classifiers quite domain-dependent. One way to deal with domain-dependent task problems is to use non-content features using spontaneous speech. In this connection, conversational dialogue is the primary area of human natural language use and can help in studying the effects of AD on dialogue and interaction might, therefore, provide more generally applicable insights.

[1] http://talkbank.org/DementiaBank/

Paper	Dataset	Features	Groups	Methods
(Orimaye et al., 2017)	DementiaBank	linguistic features including syntactic features, lexical features, and N-grams	AD and HC	Independent *t* test, Manny Whitney U test, and SMO a variant of SVM
(Ahmed et al., 2013)	OPTIMA	Linguistic feature including syntactic complexity, lexical content, speech production, fluency, semantic content	MCI, AD and HC	Statistical analysis using: Independent samples t-tests, Mann-Whitney U-tests, Wilcoxon signed-rank, repeated measures ANOVA
(Fraser et al., 2015)	DementiaBank	Linguistic features(POS, syntactic complexity, psycholinguistics,vocabulary richness, information content), and acoustic features	AD and HC	Multilinear logistic regression
(Fraser et al., 2016c)	DementiaBank	part-of-speech tags, parse constituents, psycholinguistic, complexity, vocabulary richness, informativeness, and item acoustic features	AD depressed and AD non-depressed	Logistic regression and SVM
(Kavé and Dassa, 2018)	Data in Hebrew language	grammatical, lexical aspects of language and information content	AD and HC	Independent samples t-tests, and Pearson correlation for correlation among MMSE and language features
(Jarrold et al., 2014)	Data from NIH funded project performed at the UCSF Memory and Aging Center	acoustic features, POS features and LIWC word count	AD, HC, other Dementia	Logistic Regression, Multi-layered Perceptrons(MLP), Decision trees

Paper	Dataset	Features	Groups	Methods
(de Lira et al., 2011)	Data collected on Picture description task of a Dog story	Microlinguistic features including lexical errors and syntactic index	AD and HC	Student's t-test (t) followed by the Mann-Whitney test (U) and logistic regression analysis
(Luz et al., 2018)	CCC	Speech characteristic including avg dialogue duration, avg turn duration, norm turn duration, avg words per minute, etc	AD and Non-AD	Ada boosting Model(additive Logistic regression), Decision Trees, Random Forest
(Tóth et al., 2018)	Data collected in Hungarian language	Acoustic features (hesitation ratio, speech tempo, length and number of silent and filled pauses, length of utterance)	HC and MCI	Naive Bayes, SVM, Random Forest
(Roark et al., 2011)	Data from Layton Aging & Alzheimer's Disease Center	language features: POS, Idea density, content density and speech features: total pauses, pause duration, pause rate, phonation rate, Locution time	HC and MCI	
(Singh et al., 2001)	Data collected at Bristol Memory Disorders Clinic	Temporal characteristics of Speech: verbal rate, Phonation rate, Mean pause duration, stand pause rate	HC and DAT	Mann-Whitney U Test, PCA, Discriminant function analysis

Table 2.1: Summary of literature review on previous studies on Dementia

2.2 Non Content features in Spontaneous Speech

In addition to studies discussed in section 2.1, there exist more studies that focus on spontaneous speech and experimental setups in natural settings and based on generalized tasks using non-content features. In the context of conversational analysis studies, non-content features refer to aspects of communication that are not directly related to the semantic or informational content of the conversation. Instead, these features focus on various elements that contribute to the structure, dynamics, and social aspects of the interaction. Some examples of non content features includes pauses, silences, prosody, turn taking patterns and disfluecy. Disfluency refers to interruptions, hesitations, or disruptions in the smooth flow of speech that are not directly related to the semantic content of the communication. Disfluencies represent instances where speakers exhibit pauses, fillers, repetitions, or corrections during their utterances. With the help of a combination of open-ended, closed-ended, choice questions, disfluencies and distribution of pauses, the interviews are used to elicit spontaneous speech output between speakers. (Boschi et al., 2017).

Luz et al. (2018) worked on the automatic detection of Alzheimer-type Dementia based on characteristics of spontaneous dialogues between AD patients and interviewers recorded in natural settings. They worked on the Carolina conversation Collection (CCC) dataset (Pope and Davis, 2011) and extract features like speech rate, and dialogue turn-taking statistics including; dialogue duration, Dialogue duration text to speech (TTS), Average turn duration, total turn duration, Avg turn duration TTS, average no of words, average words per minute and show that this can build a predictive statistical model for the presence of AD. Pakhomov et al. (2013) assessed the changes in cognitive functions based on different characteristics of language and speech from the spontaneous speech and showed that silent pauses, filled gaps, false starts, and longer pause duration are major disfluencies in cognitively impaired persons.

Another study that permits analysis of spontaneous speech based on three different tasks to observe the earliest detectable indicators of cognitive decline in MCI distinguishing them from control persons. This study consists of three tasks including an immediate recall test after watching one minute video, a spontaneous speech about the previous day's activities, and a delayed task (Tóth et al., 2018). The authors performed the statistical analysis and showed most of the acoustic factors (speech temp, hesitation ratio, articulation rate, silent pause, duration of utterance, and pause per utterance ratio) between the two groups are significantly different.

(Ogata et al., 2009) focused on a reliable detection technique that can handle both filled and silent pauses in spontaneous Japanese speech, which increased the performance of a speech recognizer by identifying and handling both filled (lengthened vowel) and silent (unfilled) pauses. Roark et al. (2011) described a system that uses spoken responses from MCI and controls participants during a neuropsychological exam. From

the audio and transcripts of the spoken narrative recall task, a range of markers are identified including speech features such as a pause in the speech, filled pauses, total pause duration, pauses per retelling, standardized pause rate, Total phonation time, phonation rate, verbal rate and language features such as; words/clause, Frazier/word, tree nodes/word, content density.

Singh et al. (2001) present a mechanism to quantify the speech deficit in the spontaneous speech of probable Dementia type Alzheimer's diseases (DAT). They conducted a semi-structured interview with eight DAT patients and eight healthy persons and ask questions about themselves using mostly open-ended questions. Using Mann-Whitney U analysis, Principal Component Analysis (PCA), and Discriminant Analysis, they demonstrated that there is a significant difference between the groups based on transformed phonation rate (TPR), mean duration of pauses (MDP), and standardized phonation time (SPT).

In this section, I will present a few studies that have used different non-content features from the spontaneous speech of AD sufferers and healthy/Non-AD participants. It would be more advantageous to look into more detailed characteristics of dialogue like what types of questions are being asked, how the responses are made, and the role of the carer in their conversation in combination with non-content features of dialogue to get more insight.

2.3 Conversational Analysis

Work in the conversation analysis (CA) tradition has looked in more detail at what characteristics of dialogue with dementia might be important. Jones (2015) performed a conversational analysis on the audio-recorded telephonic conversation of an Alzheimer's patient with her daughter and son-in-law to capture particular aspects of language that are affected due to cognitive impairment. The timing of speech (such as overlapping speech and pauses inside and between speakers' turns) was the main topic of discussion. The author showed that the diminishing capacity to communicate doesn't lie solely in semantic and cognitive impairment but are interactional difficulties when conversant with other people.

Jones et al. (2016) provide a CA study on the dialogue between patients and clinical experts during initial visits to a specialist clinic. Their conversation analysis is based on the interactional behavior of patients on the questions of neurologists including: a) how long it takes them to respond to a question, b) their ability to respond to compound questions, c) their ability to answer questions about personal information like age, d) the amount of detail they elaborate due to their history of memory loss, and e) their ability to show working memory during interaction. The result of this study showed that the patients with ND show significantly different behavior having delayed and undetailed

responses, unable to track all parts of compound questions: usually answer one part and forget what was the other one.

Elsey et al. (2015) highlighted the role of carer while examining triadic interactions among a doctor, a patient, and a companion. They establish differential conversational profiles which distinguish between non-progressive functional memory disorder (FMD) and progressive neuro-degenerative Disorder (ND), based on the interactional behavior of patients responding to neurologists' questions about their memory problems. Features include difficulties responding to compound questions, giving detailed explanations of examples and answering questions about personal information, time is taken to respond, and frequent "I don't know" responses.

Questions present an interesting testing ground when exploring effectiveness of communication between caregivers and people with AD. Sacks et al. (1978) formalized the question and its answer sequence as a type of adjacency pair, in which the first utterance represents the question and the second one is an answer. Hamilton (2005) has explored the use of questions in conversation with a patient of AD over four years. They found that yes-no questions are responded to much more frequently than open-ended questions i.e. Wh-questions. GOTTLIEB-TANAKA et al. (2003) used a similar approach by using yes–no and open-ended questions in a conversation between family caregivers and their spouse with AD during different activities of daily life. TYes-no questions are used by caregivers far more frequently than open-ended ones (66 % vs. 34 %, respectively), and there are also fewer communication failures when using yes-no questions.

Patients' ability to complete the question-answer adjacency pair is preserved until the last stages, however, the number of answered questions, preferred answers, and relevant answers start to decrease. Varela Suárez (2018) observed the Dementia's patient ability to respond to different types of questions including close-ended questions, open-ended questions, and multiple-choice questions. Shenk (2011) found in a pilot study that questions that require answers from episodic memory, particularly of recent events, perform worse than those that require answers from semantic memory. The distinction between a yes-no question and open-ended questions is less useful than between episodic and episodic semantics. The answers to some closed questions, such as "You really like roses, don't you?," may be longer due to related to semantic-episodic memory. The AD patient can respond with a clear yes/no or they may elaborate further. Similar to this, lengthy responses may be obtained from open-ended questions that tap semantic-episodic memory, such as "What do you enjoy doing in the afternoon?". These findings are also further supported by Small and Perry (2005) who reported that persons with AD had been more successful in responding to a caregiver's question (yes–no or open-ended) when they asked for information from semantic memory then when it is asked to recall information related to an event, specific place, or time (episodic memory). Kopelman et al. (1989) argued that autobiographical and personal semantic memory show a consistent pattern of impairment among amnesic patients and healthy participants. In a parallel

study of a larger population conducted on Alzheimer's patients and Korsakoff patients using a similar approach, Kopelman (1989) showed that both groups exhibit retrieval deficits in their remote memory.

Davis et al. (2014b) have explored the interactions between visitors and residents with dementia and discussed three conversational intervention techniques: go ahead, quilting, and indirect questions to prolong the communication. In this study, they have investigated the effect of using different types of questions and the lengths of the pauses before responses for different types of questions and also showed that pauses length are larger in question and answer than other turns (Davis et al., 2014b).

As it is evident from the literature that language production is affected in AD sufferers and the interaction patterns can be used to detect the changes using natural language processing. As the studies (Shenk, 2011; Varela Suárez, 2018; Jones et al., 2016; GOTTLIEB-TANAKA et al., 2003; Kopelman, 1989) have shown that questions, answers, and utterances/response behavior helps in diagnosis purpose, however, these studies are based on qualitative analysis. There is a need of computational models to be built that can focus on these interaction behaviors such as question-answers to detect AD. Recent developments in AD detection have started to capture interaction in terms of questions and answers using dialogue act modeling (Farzana et al., 2020; Farzana and Parde, 2022; Nasreen et al., 2019, 2021a). In the next section, we will discuss some of the state-of-the-art literature for DA tagging and three annotation schemes used in dialogue act tagging work.

2.4 Dialogue Act Models

The ability to model and recognise discourse structure is an important step toward working with spontaneous dialogue and the initial analysis stage involves the identification of dialogue acts (DA). DA tags are labels that are assigned to each utterance in conversational dialogues. DA's represent the meaning of utterances at the level of illocutionary force (Stolcke et al., 2000). Classifying utterances and assigning DA is very useful in many applications including answering questions in conversational agents, summarizing meeting minutes, improving speech summarization, resolution of ambiguous communication (Sridhar et al., 2009), and assigning proper DAs in dialogue-based games.

2.4.1 DA Tagging/classification

Traditional Machine learning approaches have been investigated and have achieved state-of-the-art performance for DA Classification using a domain-independent DAMSL dialogue act annotation scheme. A Hidden Markov Model that utilised numerous lexical and prosodic features as input was one of the first effective machine learning models

for DA recognition (Stolcke et al., 2000). Stolcke et al. (2000) utilized HMM, which has a 71% accuracy and employs word sequences within sentences and utterances as well as dialogue act sequences over utterances. Other significant work includes Conditional Random Field (Zimmermann, 2009), Support Vector Machines (SVM) and Hidden Markov Models (HMM) (Surendran and Levow, 2006) and Bayesian Networks (Grau et al., 2004). Sridhar et al. (2009) employed maximum entropy for automatic dialogue act tagging in the Switchboard corpus utilizing lexical, syntactic, and prosodic cues (features). Most recent advances in Artificial Neural networks(ANN) and Deep Learning have led to new approaches that increased the performance, particularly Recurrent Neural networks (RNN) (Papalampidi et al., 2017; Ortega and Vu, 2017) and Convolutional Neural Networks (CNN) (Kalchbrenner and Blunsom, 2013; Lee and Dernoncourt, 2016). The context of earlier utterances directly affects a particular utterance and its related DA. For example, keeping the previous utterance as a question helps in predicting that the next utterance 'yeah' is the answer and not a backchannel. This view has led the neural network approaches to consider contextual information like preceding utterances and previous DA along with the semantic content of current utterances. Bothe et al. (2018) used a context-based approach to classify dialogue acts by employing a character-level language model utterance representation, and RNN-based architecture to learn context. They used up to 4 previous utterances as context and got the best accuracy of 77.34% with three previous utterances as compared to previous work of (Stolcke et al., 2000) with accuracy 71% and (Kalchbrenner and Blunsom, 2013) with an accuracy of 73%. Ortega et al. (2019) also presented a unique technique for context modeling in DA classification that pairs convolutional neural networks (CNNs) with conditional random fields (CRFs), and tested it on the MrDA and SwDA datasets, and got an accuracy of 84.7% on MrDA and 74.6% on SwDA with two utterances as context. According to several researchers, both dependencies between successive utterances and consecutive DAs are elements that affect natural discourse (Kumar et al., 2018).

Webb et al. (2005) create a dialogue act classifier based on intra-utterance features and used N-grams as cue-phrases based on the likelihood of an N-gram occuring within a DA retaining only those with 'predictivity' value over a specific threshold using VERBMOBIL Corpus and Switchboard Corpus. Another simple but effective approach is by using the probabilistic method of utterance representation and DA classification using RNN architectures without considering any context (Duran and Battle, 2018). The probabilities associated with each DA are represented by a vector, and the utterance representations are created from the probability distribution over all DAs for each word in the utterance. Their method achieved an accuracy of 75.48% on the SwDA corpus and compared to utterance representation with a word embedding approach and achieved an accuracy of 73.68%.

Bidirectional Encoder Representations from Transformers (BERT) enabled several NLP tasks to outperform competing models. Maltby et al. (2023) achieved an accuracy

of 89% with the BERT model on the MrDA dataset and with an ensemble of acoustic features and lexical features, the accuracy is further increased to 90%. Chakraborty et al. (2020) compared the performance of deep learning models such as CNN and LSTM with the BERT model and showed that BERT outperform both models with an F1 score of 0.84.

It is evident that classifying utterances into relevant DAs is useful in many applications. In a dyadic interaction between a subject with senile dementia of the Alzheimer's type (SDAT) and a healthy subject, speech acts (DAs) are employed to observe the various communication patterns. (Ripich et al., 1991). Ripich et al. (1991) performed a comprehensive analysis, distributed the speech acts into categories; requestive, assertive, performative, responsive, and expressive and he showed the differences in words per turn, shorter turns, more nonverbal responses, more requestive speech acts, and fewer assertive speech acts with SDAT than healthy participants. Classifying utterances with DA's also helps in looking different patterns that can relate to cognitive impairments. Another study showed the distribution of DA's in natural conversations between AD patients and Non-AD patients and showed that AD sufferers produced more signal-non-understanding and clarification requests as compared to Non-AD sufferers (Nasreen et al., 2019). In a similar line of research, Farzana et al. (2020), demonstrated that dialog act (DA) sequences can be used to identify dementia because they may be able to capture relevant interaction patterns.

2.4.2 Dialogue Act Annotation Schemes

2.4.2.1 MapTask Annotation scheme

One of the first DA taxonomies was developed for the task-based corpus Maptask (Anderson et al., 1991), where speakers must communicate vocally to reproduce a route printed on one participant's map on the other participant's map and this scheme distinguishes between *initiate moves* and *response moves*. This scheme consists of five response moves (acknowledge, reply-y, reply-n, reply-w, and clarify) and six initiative moves (instruct, explain, check, align, query-yn, and query-w). This taxonomy is not intended to capture all human behavior during conversations because it is particularly task-specific and does not scale to conversations that are not task-focused.

2.4.2.2 DAMSL Annotation scheme

Discourse Annotation and Markup System of Labeling (DAMSL) (Core and Allen, 1997) was the first annotation taxonomy for DA tagging that was not task-based. The complete annotation manual is available online [2] for detail of sub-categories speech acts.

[2] http://www.cs.rochester.edu/research/cisd/resources/damsl/RevisedManual/RevisedManual.html

This annotation scheme proposes a taxonomy that covers semantics aspects (opinion, statements, preferences, etc), syntactic aspects (yn-question, wh-question, declarative questions, etc) and behavioural aspects(hedge, signal non-understanding, backchannels, and conventional opening, closing etc) of conversations. This is most widely used annotation scheme in dialogue act tagging is the Switchboard corpus(Jurafsky et al., 1997).

The discourse annotation and markup system of labeling (DAMSL) tag-set is being supplemented by the discourse tagset in its current form. The corpus is annotated using the DAMSL tagset with approximately 60 fundamental tags or classes that together result in 220 distinct labels. These 220 labels are then combined by Jurafsky et al. (1997) into 43 key classes, such as *Statements, Backchannels, Questions, Agreements, Apology,* etc.

2.4.2.3 ISO 24617-2 standard Annotation scheme

The ISO 24617-2 (Bunt et al., 2010), the international ISO standard for DA annotation, represents the first domain-independent, task-independent annotation scheme that can be used for cross-corpora DA mapping. It is a multidimensional annotation scheme of communicative functions, consisting of the nine dimensions (Task, feedback, Time Mang., Turn Mang., contact mang., discourse structuring, own communication mang, partner comm. mang., and social obligation mang.) and the recognition of dialogue act qualifiers for certainty, conditionality, and sentiment. DialogBank corpora (Bunt et al., 2016) is a publicly available corpus constructed according to ISO 24617-2 annotation scheme.

Based on the above-mentioned annotation schemes, it is decided to choose the SwDA-DAMSL annotation scheme for the following reasons:

1. Maptask annotation scheme is task-specific and does not capture all human behavior in natural conversations like backchannels, while the DAMSL scheme provides not only semantic and syntactic aspects of conversation but also it captures behavioral aspects of the dialogue.

2. ISO standard annotation schemes are domain-independent, task-independent, and multi-dimensional but their tagset related to questions and answers is very generic there are three tags (Prop Q, SetQ, and choiceQ), and very few for answer category. DAMSL tagset scheme provides a comprehensive set of questions and answers types that are suitable for the proposed research.

2.5 Computational models of interaction for Dementia

Recent research has shown that spontaneous speech data obtained naturally are reliable indicators for AD identification in discourse with a clinical diagnosis. Researchers started to investigate if Alzheimer's disease (AD) behavioral symptoms could be identified through dialogue interaction from everyday conversations. Luz et al. (2018) build a predictive model for automatic AD detection from dialogue interaction using non-content features extracted from natural conversations. An automated analysis of conversations inspired by CA-based features is performed on the conversational dataset of patients and neurologists to differentiate between patients with neurodegenerative memory disorder (ND) and functional memory disorders (FMD) (Mirheidari et al., 2019). They built a model With the help of lexical, acoustic, and CA-inspired features, and achieved classification rates of 90.0 % for neurologist-patient data and 90% for IVA-patient conversations. Their CA-based feature set includes information about the number of turns, the average length of turns, patient recall of memory loss, the number of unique words in a turn, the patient's response of "dunno," and the typical number of filler, empty, unique phrases(Mirheidari et al., 2017).

de la Fuente Garcia et al. (2019), examined through a study that builds on the PREVENT-Dementia project, which extracted dialogue features like repair, turn-taking patterns, backchannel behavior, pauses, and prosodic and content features from participant speech during natural conversations. Li et al. (2022) proposed a unique diagnosis architecture made up of an ensemble AD detection module and a proactive listener module. The ensemble AD detection module integrates four classifiers based on audio, language, disfluency, and interaction and uses utterance and dialogue data to diagnose AD from spontaneous speech.

Farzana and Parde (2022), also investigated whether interaction patterns are helpful for AD diagnosis. They coupled the AD identification task with interaction captured through features extracted from the DA tagging-based model. These approaches will be further discussed in the background section of chapter 6 and chapter 7.

2.6 Existing Datasets For Dementia/Alzheimer's

2.6.1 DementiaBank

The University of Pittsburgh's Alzheimer research department uses DementiaBank as a component of the TalkBank project. It contains corpora covering multiple cognitive skills and in many languages. Pitt Corpus is an English language corpora in DementiaBank that contains narrative speech with the common Cookie Theft picture description

task.The interviewer asks the patient to describe its contents and is allowed to occasionally urge or prompt the subject. Elderly controls, individuals with probable and possible AD, and individuals with different dementia diagnoses were all participants. The dataset contains 208 patients from the Dementia group and 104 control participants. In addition to the Cookie theft picture description task, this corpus also contains the Fluency task, Recall task, and sentence construction task for the Dementia group only.

From our research's objective point of view, the limitation of this dataset are:

- Task-specific, a fixed task like the description of a picture or story-telling
- Transcripts are a narrative description of a picture description(Cookie theft picture) in which the examiner asks the patient "tell me everything you see in this picture?" and encourages only when they are not talking.
- Lack of spontaneous dialogues in natural settings.

Part of the DementiaBank particularly has been separated out and used in ADReSS challenge from the transcripts and speech recordings of picture descriptions elicited from participants through the Cookie Theft picture description task (Luz et al., 2020)

2.6.2 Alzheimer's Dementia Recognition through Spontaneous Speech (ADReSS challenge)

The ADReSS Challenge is an initiative that focuses on advancing research in the field of Alzheimer's disease and related dementia detection through the analysis of spontaneous speech (Luz et al., 2020). In order to define a common task that allows various methods of AD recognition in spontaneous speech to be compared, the ADReSS challenge aims to provide a benchmark dataset of spontaneous speech that has been acoustically pre-processed and balanced in terms of age and gender. The data includes transcripts of spoken picture descriptions and speech recordings obtained from participants using the Boston Diagnostic Aphasia Exam's "Cookie Theft" picture (Goodglass et al., 2001). The CHAT coding scheme was used to annotate transcripts . A basic voice activity detection technique based on signal energy threshold has been used to segment the recorded speech for voice activity. There are 2122 voice segments from 78 AD participants and 1,955 speech segments from 78 non-AD subjects in the segmented dataset.

The ADReSS challenge defines two distinct prediction tasks: (a) the MMSE prediction task, which requires researchers to construct regression models of the participants' speech in order to predict their scores on the Mini-Mental State Examination (MMSE); and (b) the AD recognition task, which requires researchers to model participants' speech data in order to perform a binary classification of speech samples into AD and non-AD classes.

2.6.3 Alzheimer's Dementia Recognition through Spontaneous Speech-audio only (ADReSSo challenge)

The ADReSSo-2021 challenge aims to define a common shared task that allows various methods of AD identification in spontaneous speech to be compared, giving a benchmark dataset of spontaneous speech that has been acoustically pre-processed and balanced in terms of age and gender. The ADReSSo Challenge is focused on three challenging automatic prediction tasks that are relevant to society and healthcare. These tasks include the identification of Alzheimer's disease, the inference of cognitive testing score(MMSE) and the prediction of cognitive decline (Luz et al., 2021).

The ADReSSo Challenge employed two separate datasets: (a) collection of speech recordings of persons without AD (controls) and patients with an AD diagnosis describing the Cookie Theft picture from the Boston Diagnostic Aphasia Exam (Goodglass et al., 2001). (b) a collection of speech recording showing individuals with Alzheimer's disease completing a category (semantic) fluency test at their baseline visit for prediction of cognitive decline over two years. There were 237 audio recordings in dataset for AD classification and severity detection, and the state of the subjects is assessed based on the MMSE score. A score of 25-30 on the MMSE is regarded as normal, a score of 21-24 as mild, a score of 10-20 as moderate, and a score of <0 as severely impaired. The second dataset for the disease prognostics challenge, which involves predicting cognitive decline, was derived from a longitudinal cohort study that included individuals with AD. This task involves classifying patients into 'decline' or 'no-decline' groups based on speech samples that were gathered for a verbal fluency test at baseline.

2.6.4 Carolina Conversations Collections (CCC)

Carolina Conversation Collection was a project that started in 2008 for collecting conversations of elderly people and it was initially supported by people that live in North and South Carolina[3].CCC is an interactive, password-protected, user-friendly online collection of audio and video recordings about the health of adults over 65 in natural conversations that have been transcribed (Pope and Davis, 2011).The patients spoke twice a year with linguistics students or academics who conducted the interviews. The CCC is a systematic collection of two cohorts: Cohort one contains 200 conversations with 125 elderly persons with different medical conditions recorded twice a year; once with the research and once with the community person at home or in community settings. Cohort two includes 400 conversations with 125 patients of AD who spoke twice at least with the linguistic research students. This collection includes some questions that apply to everyone and others that are collection-specific for people with particular health conditions, ailments, and dementia sufferers who have cognitive impairments.

[3] https://carolinaconversations.musc.edu/

The interviewer asked general questions about their breakfast, lunch, the weather, their family and kids, and questions about specific events like Thanksgiving day, how they celebrate their birthday, Christmas, etc. However, all the conversations are not made available.

The advantages of CCC over the DementiaBank dataset discussed in 2.6.1 are:

- collection of semi-structured spontaneous dialogues for AD and Non-AD patients.
- provides a framework for the analysis of natural conversations that are not based on fixed tasks (like picture description in DementiaBank) and contain relatively free conversational interaction.

Dataset Names/Demographic Details	Pit Corpus from Dementia Bank	Carolinas Conversations Collection(CCC)
Different types of Dementia Patients	208	125
Control participants	104	125
Age Range	46-80	Greater or Equal to 65
Interview type	Structured	Semi-structured

Table 2.2: Demographic Detail of Existing datasets

2.7 Summary

It has been found that most of the work on the diagnosis of AD seems to focus on language features including semantic, lexical, syntactic, and information content features. Most of these methods rely on specific tasks like picture descriptions and extracted from are narrative descriptions of the task. Fewer studies also exist that include conversational analysis of spontaneous speech with a focus on interactional features. Furthermore, these studies are based on the interaction between the interviewer and patients with the aim to involve the patients in communication. Studies also have performed quantitative analysis to observe the AD patient's behavior on their ability to respond to open-ended and close-ended questions but none of these studies have computationally coded these schemes.

These CA studies also showed that the presence of AD affects the production of questions, their use, and their responses, but all focus on specific types of questions including yes-no, wh-questions, multiple choice questions, particular language tasks, and specific semantic and lexical features of language. As far as we are aware, none of these studies have extended this approach to look into specific aspects of non-understanding or

inability to respond: e.g. non-understanding signals, clarification requests from patients, and repetition of questions asked from patients by using dialogue acts. Recently, Farzana and Parde (2022) started looking into patterns of interaction between patients and examiners in terms of DA's with more generic interaction features. However, while some studies have looked at the general use of interactional differences in AD diagnosis (see e.g. Luz et al., 2018), these use models which are not interpretable in DA terms, making it hard to provide useful output to clinicians or carers. There are also studies focusing on modeling the CA-based features (e.g. (Mirheidari et al., 2019)), but some of the features are based on a predefined set of the questionnaire (e.g. related to memory) in a more task-specific setting.

Here, it is suggested that building computational models that take into account DA tagging based DA's sequences that capture patient-interviewer interaction, conversational pauses along with disfluency and general interaction features could be helpful in distinguishing AD patients from other groups, while giving outputs that are directly interpretable in terms of the clinical findings above. In this connection, I will first perform a corpus study in the next chapter to find meaningful interaction between patients and interviewers based on spontaneous dialogues from a conversational dataset.

CHAPTER 3

LOOKING AT INTERACTION PATTERNS: A CORPUS STUDY

As chapter 2 has already shown that there is a lot of evidence exists that AD affects the distribution of different types of questions, different responses and clarification requests, etc (Jones et al., 2016; Varela Suárez, 2018; Small and Perry, 2005; Davis et al., 2014b), however, there are very fewer quantitative studies exits that is why I intended to conduct a corpus study to check whether the distributions in terms of different types of questions, responses, signal non-understanding, clarification requests and some forms of pauses are different in AD and Non-AD group? This study is both quantitative and its dialogue acts make the conversations interpretable to clinicians and carers in several ways: Dialogue act based features help in identifying specific communication patterns, such as repetitive speech, use of clarification request, signal non-understanding, difficulties in topic maintenance, or challenges in turn-taking. Recognizing these patterns allows clinicians and caregivers to understand how Alzheimer's disease may impact the structure and flow of conversations. Understanding the DA's used by Alzheimer's patients allows clinicians and caregivers to tailor their communication approaches. They can adopt strategies that align with the individual's communication strengths and challenges, fostering more effective and supportive interactions. DA's based features can serve as an objective assessment tool for clinicians. The quantifiable nature of these features allows for a more standardized and systematic evaluation of communication abilities, supporting diagnostic and treatment decision-making.

In this chapter, I will present the analysis performed on the Carolina Conversation Collection corpus, basic terminologies used in annotation and the annotation scheme that is followed, inter-annotation agreement, statistical tests that are employed, and the temporal variables that are considered for the analysis. In the last section, results are

discussed and findings of first corpus study are presented [1].

3.1 Corpus study research questions

An initial corpus study analysis will be performed with the premise that the presence of AD affects the production of speech including questions, their use, and their responses. I hypothesize that the use of different question types such as binary yes-no questions (in the interrogative or declarative form), tag questions, and alternative ('or') questions will differ between groups; and the signals of non-understanding, back-channels in question form and clarification requests should be more common with AD patients.

As said in section 1.3, the general research question for this chapter is: which interactional pattern in terms of questions and responses is more distinctive between AD and Non-AD patients? In more detail, we are going to examine this question through the following five sub-questions:

Q 1: Is the distribution of question types asked from patients different between AD sufferers and non-AD sufferers?

Q 2: How often do signals of non-understanding, clarification requests, and back-channel questions occur in dialogues with an AD sufferer compared to those without one?

Q 3: Is the distribution of simple-repeat and reformulation questions different for conversations with an AD sufferer compared to those without one?

Q 4: Are the responses on each question type significant for AD and Non-AD groups?

Q 5: Is the pauses behavior including pauses between patient utterances, between interviewer utterances, between patient to the interviewer, and from interviewer to the patient significant between the two groups?

3.2 Methodology

3.2.1 Corpus

I intend to investigate the behavior of AD patients based on the interaction patterns including questions and responses, pauses within utterances, and between turns observed in a corpus of dialogue. For this purpose, we used the Carolina Conversation Collection (CCC), collected by the Medical University of South Carolina (MUSC) Pope and Davis (2011). The reason for choosing this dataset for our initial study is that this dataset contains more interactive semi-structured interviews with spontaneous speech, the

[1] This chapter is based on work published in SemDial,2019 (Nasreen et al., 2019).

communication is dyadic, and there is no fixed task like picture description. The clinical and demographic variables include age range, disease diagnosed, occupation prior to retirement, gender, and level of education (in years) available online. The dataset consists of audio, video (only a few were available), and transcripts that are time-aligned. The identity of patients and interviewer is anonymized keeping in mind security and privacy concerns.

As this dataset includes only older patients with diagnosed dementia, it can only allow us to observe patterns associated with AD at a relatively advanced stage, and not directly tell us whether these extend to early-stage diagnosis. However, it has the advantage of containing relatively free conversational interaction, rather than the more formulaic tasks in e.g. DementiaBank. Much of the quantitative and computational work discussed in the literature in chapter 2 is based on the Cookie Theft Picture description task from DementiaBank.

A portion of this corpus is used for our corpus analysis including 10 random samples of dialogues with patients with AD (7 females, 3 males) and 10 patients (8 females, 2 males) with other diseases including diabetes, heart problems, arthritis, high cholesterol, cancer, and leukemia but not AD. These groups are selected to match the age range, to compare the different patterns of interaction and to avoid statistical bias. The demographic data of the participant is given in table 3.1.

	AD patients (N=10)	**Non-AD patients (N=10)**
Age range	60-90	70-90
Year of Education(range)	9-16	8-16

Table 3.1: Age-range, education (years) for AD and Non-AD patients

This portion comprises 2554 utterances for the AD group and 1439 utterances for the Non-AD group, with a total of 3993 utterances from 20 patients with 23 dialogue conversations.

The CCC transcripts are already segmented at the utterance (turn) level and the word level, and annotated for speaker identity (patient vs. interviewer); however, no DA information is available. We used only the utterance level layers; transcripts were available in ELAN format and we converted them to CSV format. We then manually annotate the transcripts at the utterance level with dialogue act information, pauses, repeat questions and the previous utterance responded to.

3.2.1.1 Ethical Considerations

Online access to the dataset was obtained after gaining ethical approval from the Ethics Research Committee (ERC) of the Queen Mary University of London (QMUL) vide QMERC2019/04 (See Appendix A.6) and MUSC institution hosting this research and complying with MUSC's requirements for data handling and storage.

3.2.1.2 A dialogue session sample from the Corpus Study

The following dialogue has been extracted from the corpus, Interviewer (***Int***) asks different questions from Patient (***Pat***) about her and her husband and the conversation continues with her husband job and the places she visited with him.

Speaker	Utt	Sub-utt	Dialogue
Int	11	1	– how are you?
Pat	12	1	- I'm Mrs. Mason.
Int	13	1	- I'm glad to see you.
Pat	14	1	– uh
	14	2	– did you ever know Preacher Mitchum?
Int	15	1	- I've heard Preacher Mitchum's name.
Pat	16	1	that was my husband.
			silence three seconds
	16	2	he, I don't know where he is now,
	16	3	he's off in some foreign country or something preaching.
Int	17	1	- for heaven's sake.
	17	2	– what countries did he go to?
Pat	18	1	I don't know.
	18	2	– I don't know where he is.
Int	19	1	– well, that's busy.
Pat	20	1	can't keep up with it.
Int	21	1	– hey.
			silence three seconds
	21	2	so, he goes off preaching
	21	3	– and you stay here.
Pat	22	1	– m'am?
Int	23	1	- he goes off preaching
	23	2	- and you stay here?
Pat	24	1	yeah.

Table 3.2: A sample dialogue from CCC

As shown in the dialogue excerpt, there are two questions asked by *Int* from *Pat* about 'What countries his husband visited?' (utterance **17**) and he goes for preaching

and you stay here? (utterance **21**) and two pauses of three seconds. The first question was responded with I don't know where he is and the second question is responded with a signal non-understanding (utterance 22) indicating she is unable to perceive the question, so *Int* repeats the question (utterance 23). There are two long pauses of three seconds as well in the sample.

This sample illustrates how a natural dialogue conversation comprises questions, answers, and pauses between the same and different participant's utterances. We argue that these structures of questions, answers, different signals, and pausing phenomena may serve as useful interaction patterns that could be later used for AD detection.

3.2.2 Basic Terminologies

Throughout this report, specific terms are used for particular question types and response types and use these in our annotation procedure. Following SwDA-DAMSL tagset (Jurafsky et al., 1997), I used **qy** for **Yes-No** questions and **qy^d** for **Declarative Yes-No** questions. Declarative questions ((^d)) are statements that have the pragmatic function of questions but lack the syntax of a question. I used **qw** for **Wh-questions** which includes words like *what, how, when, etc.* and **qw^d** for **Declarative Wh-**questions. Yes-No or Wh-questions are questions that do not have only pragmatic force but have a syntactic and prosodic marking of questions or interrogative in nature. We used **^g** for **Tag questions**, which are simply confirming questions that have auxiliary inversion at the end of statement e.g. (*But they're pretty, aren't they?*). For **Or questions** which are simply choice question and aids in answering the question by giving choices to the patients are represented by **qr** e.g (- *did he um, keep him or did he throw him back?*).

I used the term the **Clarification question** for questions that are asked in response to a partial understanding of a question/statement and are specific in nature. These clarification questions are represented by **qc**. **Signal non-understanding** is generated by a person in response to a question that they have not understood and are represented by **br**. **Back-channel Question (bh)** is a continuer that has a question-like form and takes the form of a question. Back-channels are more generic than clarification questions and often occur in the form (*e.g really? Yeah? do you? is that right? etc.*).

When the response to a Yes-No question is just a yes including variations (e.g. *yeah, yes, huh, yes, Yes I do, etc.*), it will be represented by **ny** and when there is a yes plus some explanation, it will be represented by **ny^e**. **na** is an affirmative answer that explains without the yes or its variation. **nn** is used for **No-answers** and **nn^e** is used for an explanation with No answer. *P* is used for patient and *I* is used for the interviewer. The complete annotation protocol along with guidelines is given in Appendix A.

3.2.3 Annotation scheme

The original SWDA-DAMSL tagset for the Switchboard corpus contains 43 dialogue act tags (Jurafsky et al., 1997). Our initial manual includes dialogue act tags from the DAMSL tagset and our own specific new dialogue act tags. Our focus of this corpus study is particularly on dialogue acts for different types of questions and their possible responses, so 17dialogue act are taken from the SWDA-DAMSL tagset, collapsing all other tags into a single '***other***' tag, and 2 new tags are introduced. These new tags are for clarification questions **(*qc*)** and answers to Wh-Questions **(*sa*)** and were required to distinguish key response types. The ability to tag specific clarification questions is

Type	**Tag**	**Example**
Yes-No Question	qy	Did you go anywhere today?
Wh-Question	qw	When do you have any time to do your homework?
Declarative Yes-No Question	qy^d	You have two kids?
Declarative Wh-Question	qw^d	Doing what?
Or Question	qr	Did he um, keep him or did he throw him back?
Tag question	^g	But they're pretty aren't they?
Open-ended question	qo	And uh -how do you think -that work helps you?
Clarification request	qc	Next Tuesday?
Signal Non-understanding	br	Pardon?
Backchannel in question form	bh	Really?

Table 3.3: Question types from our proposed tagset with examples from the CCC.

important for our study, as questions asked by the *I* can be followed by a clarification which indicates partial understanding while requesting specific clarifying information (SWDA-DAMSL only provides the ***br*** tag for complete non-understanding). The distinction between answers to Wh-Questions and other, unrelated statements is also important (in order to capture whether the response is relevant: a relevant answer should be different from a simple general statement), but DAMSL tagset provides only a single ***sd*** tag for statements. Different types of questions and their tags are given with examples in table 3.3; a list of response types is given in table 3.4. The complete annotation protocol along with guidelines is given in Appendix A.

Another new addition is the tagging of *repetition* of questions, with or without reformulation. We marked repeat questions as ***simple repeat*** or ***reformulations***, and tagged with the index of the dialogue act (utterance number) they were repeating or reformulating. Questions are repeated either in simple repeat form or reformulation form after a clarification request or signal non-understanding from the *P*'s end. Reformulated repeat questions are slightly changed syntactically but the context remains the same – (see Table 3.5 [2] with utterance 144 and more examples in the Appendix: A).

[2]A and B are participants either *I* or *P* in table 3.5

Type	Tag	Example
Yes answer	ny	Yeah.
Yes- plus expansion	ny^e	Yeah, but they're
Affirmative non-yes answer	na	Oh I think so. [laughs]?
No answer	nn	No
Negative non-no answers	nn^e	No, I belonged to the Methodist church.
Other answers	no	I, I don't know.
Declarative statement wh-answer	sa	Popcorn shrimp and it was leftover from yesterday.

Table 3.4: Answer Types for CCC

Tag	Speaker: Utterance	Text	Repeat Question?
qw	A:15	-Where's she been?	
br	B:16	-Pardon?	
qw	A:17	-Where is she been?	15
qy	A:142	-Well, are you, are you restricted from certain foods?	
br	B:143	-What?	
qy	A:144	-Like, do they, do they make you eat certain foods because of your medication?	142-reformulation

Table 3.5: Examples of repeated questions

The complete list of our proposed tagset along with more explanation and examples is available in appendix A.

3.2.4 Inter-rater agreement

To check inter-annotator agreement, three annotators annotated one conversation between an AD patient and a Non-AD interviewer of 192 utterances. All annotators had a good knowledge of linguistics and were familiar with both the DAMSL dialogue act tag set and the additions as specified above and in the manual. First, all three annotators annotated the dialogue independently by assigning dialogue act tags to all utterances with the 20 tags of interest for this study as shown in Table 3.6 ('other' means the annotator judged another DAMSL act tag could be appropriate apart from the 19 tags in focus). We use a multi-rater version of Cohen's κ (Cohen, 1960) as described by Siegel and Castellan (1988) to establish the agreement of annotators for all tags and also 1-vs-the-rest as shown in Table 3.6 below.[3]

As can be seen, an overall agreement was good (κ=0.842) for all tags and the majority of tags that were tagged by any annotator in the dialogue have $\kappa > 0.67$, with only 'no' getting beneath $\kappa < 0.5$. We judged this test to be indicative of a reliable annotation scheme for our purposes.

[3]The annotation results and scripts are available from https://github.com/julianhough/inter_annotator_agreement.

Tag	**# times annotated**	κ
qy	26	0.758
qw	30	0.895
qy^d	12	0.660
qw^d	3	1.000
^g	2	0.498
qr	0	0
qo	0	0
br	22	0.953
qc	15	0.795
bh	0	0
ny	12	1.000
ny^e	11	0.907
na	8	0.873
nn	1	0
nn^e	6	0.663
no	4	0.497
sa	26	0.637
b^m	4	0.596
other	392	0.896
all tags	576	0.842

Table 3.6: Multi-rater Cohen's κ statistics for one-vs-rest and overall agreement score for one dialogue.

3.3 Temporal Measures

Different variables were selected to quantify both distributions of questions, responses against each question type, and pauses in speech. We use the symbol P for the patient and I for the Interviewer. We will use percentages for the distribution of each question type among the total questions asked from AD patients and then from Non-AD patients. Then we will use the normalized values of each response category against each question type. Finally, we look for pauses by first looking into the total number of pauses for each dialogue conversation for both groups separately. We normalized the number of pauses by the number of utterances.

$$Avg\,No.\,of\,pauses = \frac{Total\,pauses}{Total\,utterances} \tag{3.1}$$

Then we look for the average (Avg) pause rate and average duration of pauses within P utterances.

$$Avg\,No.\,of\,P\,pauses = \frac{Total\,P\,pauses}{Total\,P\,utterances} \tag{3.2}$$

$$Avg\,P\,pause\,Duration = \frac{Total\,P\,pause\,Time}{No.\,of\,P\,pauses} \tag{3.3}$$

We will also look at the pattern of pauses within the interviewer's utterances.

$$Avg\,No.\,of\,I\,pauses = \frac{Total\,I\,pauses}{Total\,I\,utterances} \tag{3.4}$$

$$Avg\,I\,pause\,Duration = \frac{Total\,I\,pause\,Time}{No.\,of\,I\,pauses} \tag{3.5}$$

It will be worth interesting to look for the pauses pattern between turn-taking from *I-P* and from *P-I*. These given below measures are used for pauses occurring while taking a turn from I to P.

$$Avg\,No.\,of\,I\text{-}P\,pauses = \frac{Total\,I\text{-}P\,pauses}{Total\,turns} \tag{3.6}$$

$$Avg\,I\text{-}P\,pause\,Duration = \frac{Total\,I\text{-}P\,pause\,Time}{No.\,of\,I\text{-}P\,pauses} \tag{3.7}$$

These are pauses occurring while taking a turn from P to I.

$$Avg\,No.\,of\,P\text{-}I\,pauses = \frac{Total\,P\text{-}I\,pauses}{Total\,turns} \tag{3.8}$$

$$Avg\,P\text{-}I\,pauseDuration = \frac{Total\,P\text{-}I\,pause\,Time}{No.\,of\,of\,P\text{-}I\,pauses} \tag{3.9}$$

We are also interested in looking at words per dialogue act and words per turn for both AD and Non-AD groups.

$$words\,per\,DA = \frac{Total\,words}{Total\,DA} \tag{3.10}$$

$$words\,per\,turn = \frac{Total\,words}{Total\,turns} \tag{3.11}$$

3.4 Experimental Setup

In SPSS software version 26, different statistical analyses were carried out for the subject discourse (AD and Non-AD) and the interviewer's discourse.

3.4.1 Statistical Analysis

The two subject groups and the interviewer's interactions with the two groups were compared to see how often these discourse elements were present in each group. Percentages, averages, standard deviations, independent t-tests for uneven variance, and Fisher's Exact test were among the statistical techniques used.

3.4.1.1 Fisher's Exact test

We performed the Fisher's Exact test to check whether the two groups are independent in terms of response categories or not. Fisher's Exact test is used to examine the association between the classification groups (Kim, 2017). It is quite useful when the sample size is small and more than 20% of the cells have anticipated frequencies less than 5. We preferred to use this test over the Chi-Square test due to the limitation of sample size and that the frequency value should not be less than five. Our null hypothesis is that 'the responses for each question type are independent of the groups (AD vs Non-AD). The likelihood of rejecting the null hypothesis when it is true is the significance level, denoted as alpha or $alpha$. A significance level of 0.05, for example, suggests a 5% probability of assuming the existence of a difference when none actually exists. We set the $\alpha\ value = 0.05$ as it is commonly used. We calculate the *p-value* to conclude whether our null hypothesis is rejected or not. If the *p-value* $<\alpha$ showing null hypothesis is rejected and the difference between two groups for these response categories is significant and vice versa.

3.4.1.2 Independent sample T-test

To determine whether there are any significant differences between the two groups, we use an independent sample t-test for pauses within the patient utterances, within interviewer utterances, from interviewer to patient utterances, and from patient to interviewer utterances.

Our null hypothesis is that the two groups are independent of these four types of pauses patterns. In order to determine the magnitude of mean differences, we additionally compute the effect size. In a statistical test, this is often calculated after the null hypothesis has been rejected. Effect magnitude is not particularly significant if the null hypothesis is not rejected. We have used Cohen's d effect size measure to calculate the effect size among the two groups given by:

$$d = \frac{\bar{x_1} - \bar{x_2}}{s} \tag{3.12}$$

Here
$\bar{x_1}$: mean of one population (e.g AD group)
$\bar{x_2}$: mean of other population (e.g Non-AD)

s: pooled standard deviation (for two independent samples)
Jacob Cohen defined s, the pooled standard deviation Cohen (2013) as:

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \tag{3.13}$$

s_1 and s_2 are variance for two groups and can be calculated as:

$$s_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n}(x_1, i - \bar{x_1})^2 \tag{3.14}$$

3.5 Results

In this section, we present results that describe the distribution of different question types among two groups (AD vs Non-AD), describes how significant the distribution of response type for each question type among the groups using Fisher's Exact test and pauses pattern among patients utterances, interviewer utterances and between the turn-taking from patient to the interviewer and from interviewer to patient.

Corpus study research question-1:

Is the distribution of question types asked from patients different between AD sufferers and non-AD sufferers?

To investigate the distribution of DA's, we calculated the relative frequency of each question type separately for AD and Non-AD group , and for the P and I within those groups. The percentage frequency of each question type is obtained by dividing the number of each question type asked by interviewer or patient to the total number of utterances spoken by patient or interviewer. (See equation 3.15).

$$Percentage\,of\,each\,questions\,type = \frac{No.\,of\,each\,question\,type}{total\,utterances} * 100 \tag{3.15}$$

A comprehensive analysis of particular types and their distribution between AD and Non-AD P with their I is shown in Figure 3.1 (A). More yes-no questions(qy) are asked by the I from AD Patients than Non-AD patients (6% vs 3.7%) and fewer wh-questions(qw) are asked in the AD group compared to the non-AD group (4% vs 5.4%). Choice questions (qr) are also asked more from AD patients compared to non-AD patients (2% vs 0.3%). These results suggest there is a systematic difference in question distributions; one plausible explanation for this is that AD patients find it easier to answer a simple Yes-No question or a choice question compared to a wh-question. While in Figure 3.1(B) depicts the distribution of these tags for patients, and it can be clearly seen that distribution for clarification requests and signal non-understanding are higher among other tags.

We also compared the distribution of these tags with the Switchboard SWDA corpus, as shown in Figure 3.2. As the CCC is a set of clinical interviews, the percentage of tags

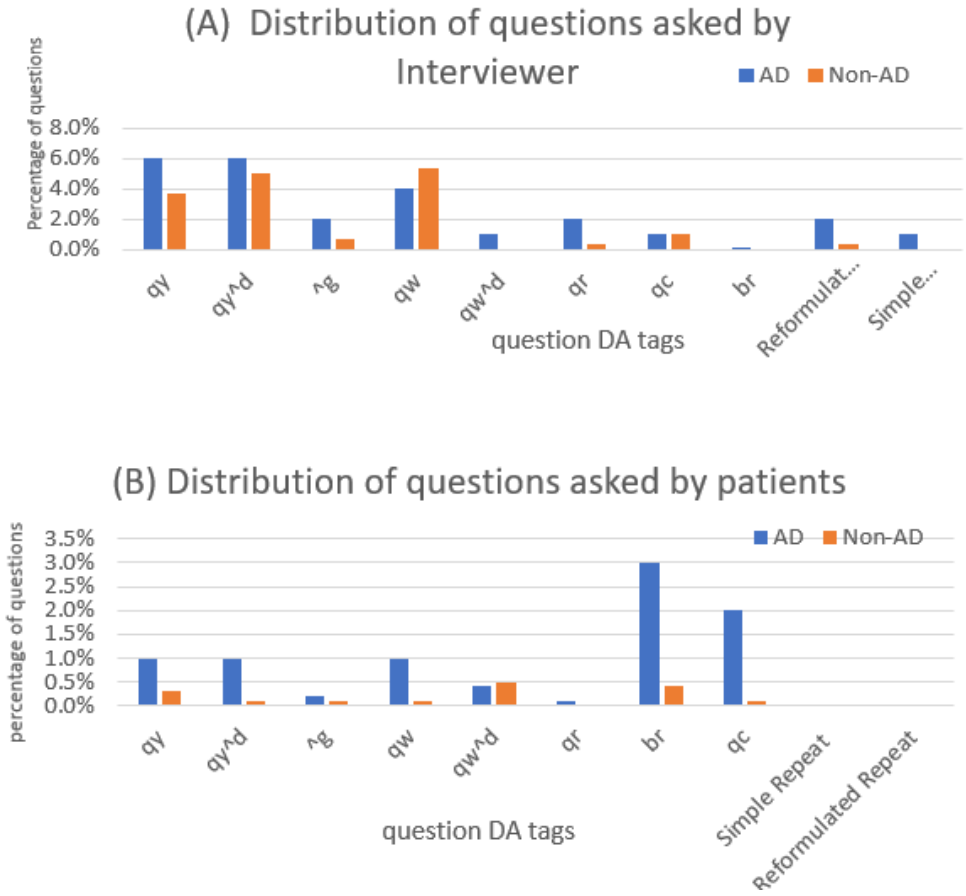


Figure 3.1: (A): Distribution of question tags asked by *I* among AD and Non-AD. (B) questions asked by *P*.

which are questions is higher in this corpus compared to Switchboard. Although simple yes-no questions have almost identical frequencies in both corpora, declarative yes-no, wh-questions, declarative wh-questions, tag questions, and signals of non-understanding are higher in the CCC than in Switchboard. Our new clarification question (qc) tag accounts for 1% for both AD group and Non-AD group tags but is not annotated in SWDA.

Corpus study research question-2

How often do signals of non-understanding, clarification requests, and back-channel questions occur in dialogues with an AD sufferer compared to those without one?

An examination of signals of non-understanding, clarification requests and back-channel requests reveals that the ability to follow and understand questions decreases for AD patients so they produce more signals of non-understanding (e.g *sorry Maam?,*

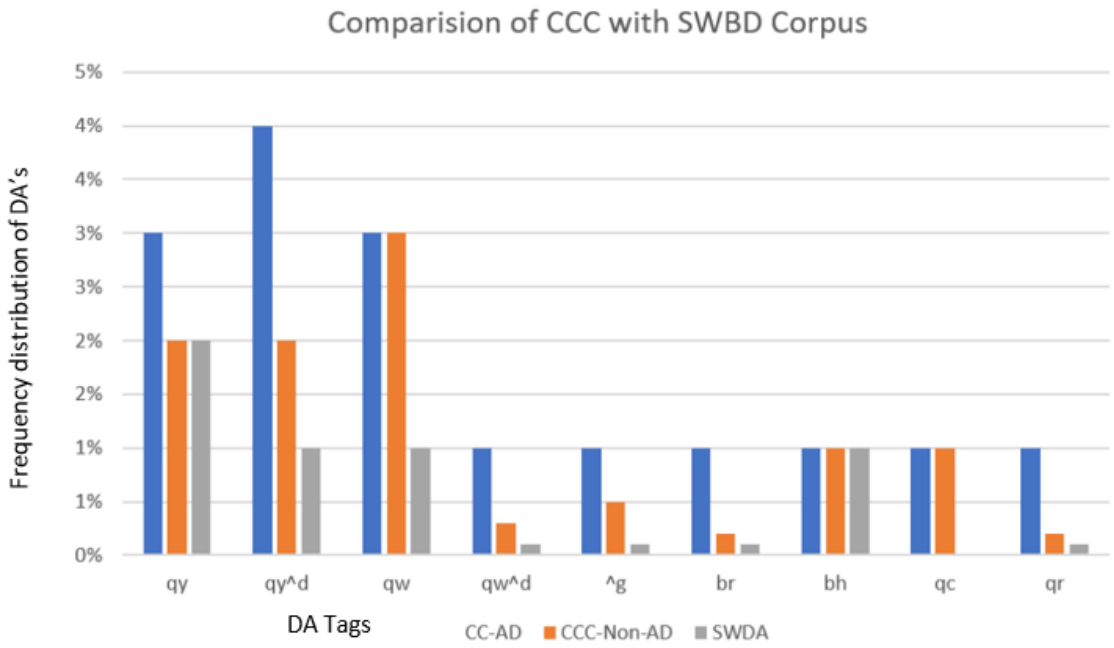


Figure 3.2: Comparison of the relative frequency of DA tags in AD, Non-AD group in CCC and SWDA corpus.

Pardon?, huh?, eh?) when questions are posed to them. On the other hand, signals of non-understanding from Non-AD patients are much less frequent as shown in figure 3.3. The frequency of clarification questions (qc) between the two conversation groups was not systematically different as shown in table 3.7 when utterances from both patient and interviewer are combined, but dealing with them separately, AD patients produce more clarification requests than non-AD patients (2% vs 0.1%)– see figure 3.3.

	AD Group	**Non-AD Group**
Question followed by Signal of Non-understanding	24 (35)	2 (3)
Statements followed by Signal of Non-understanding	11 (35)	1 (3)
Question followed by Clarification Question	8 (34)	1 (11)
Statement followed by Clarification Question	26 (34)	10 (11)

Table 3.7: Occurrences of signal non-understanding and clarification question followed by question/statements.

We further examine how many times signals of non-understanding are issued in response to a question rather than simple statements, and how many times clarification requests are issued in response to questions or a statement/answer. The examination

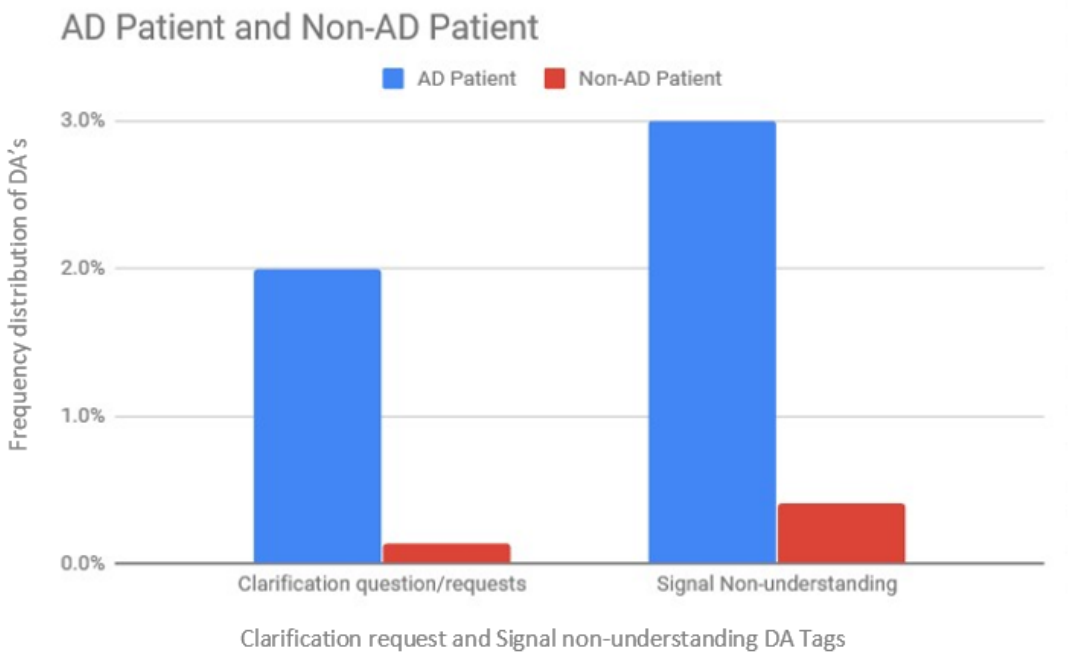


Figure 3.3: Clarification questions and Signal Non-understanding

of data shows that it is not necessary that clarification request is always issued after a question, most of the time they are generated in response to a statement and fewer times after questions are raised.

Out of a total of 35 signal non-understanding, 24 are generated in response to a question of AD Group as shown in table 3.7. 8 clarification questions are asked in response to the question and 26 clarification questions are asked in response to the declarative statement. We suggest that these sequences such as questions followed by signal-non-understanding, questions followed by clarification requests, and statements followed by clarification requests may be attributed towards AD.

Corpus study research question-3

Is the distribution of simple-repeat and reformulation questions different for conversations with an AD sufferer compared to those without one?

Many of the questions are either followed by clarification questions or signal non-understanding, so there will be more repetition of a similar type of question in the case of AD patients. Repeated questions are asked in two variations; either repeated simply or reformulated so that the patient can understand the question properly. In the AD group, 4.7% questions are simple-repeat questions, and 6.7% are reformulated as shown in table 3.8 while for the non-AD group only 2.4% are reformulated questions and there were no repeated questions.

Repeat Type	AD Group	Non-AD Group
Total Question	313	127
Simple-Repeat Question	15(4.7%)	0
Reformulated Question	21 (6.7%)	3 (2.4%)

Table 3.8: Repetition and reformulation of questions for the Non-AD group and AD group.

Corpus study research question-4

Is the responses on each question type significant for AD and Non-AD group?

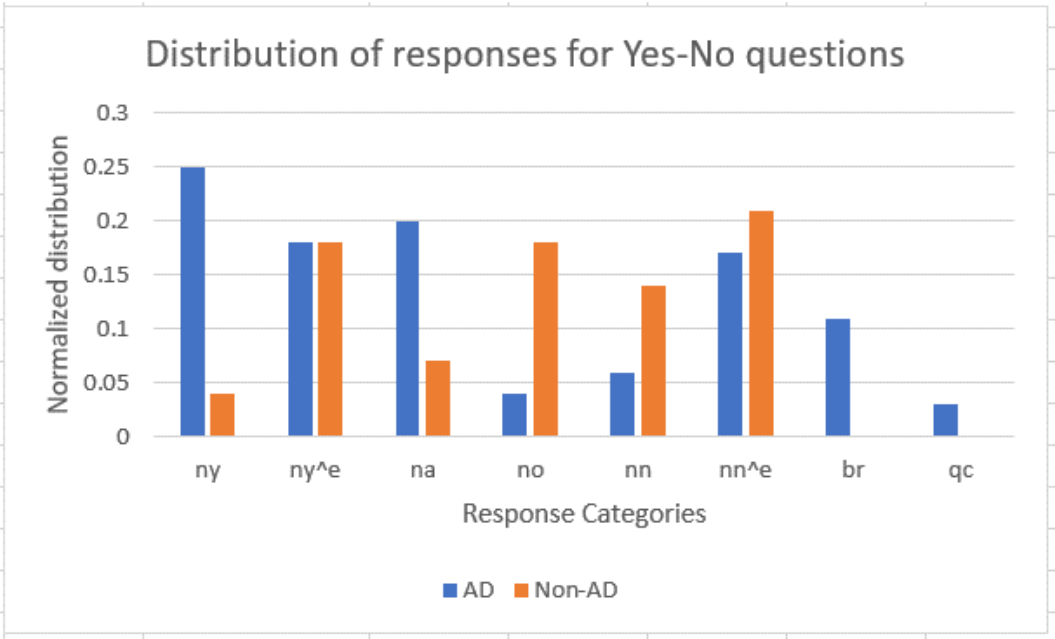


Figure 3.4: Distribution of responses for *qy* questions.

Figure 3.4 presents the statistics for the responses against *qy* questions and it is clear from the figure that there are much more *ny* for AD while *nn* answers and *nn^e* are higher for Non-AD group. For the rest of the question categories see Appendix C Table C.1 which shows the frequency distribution of responses against each question type among the two groups.

Table 3.9 records the responses as signal non-understanding and clarification requests for each question category[4]. Our findings revealed that there are more *br* for AD group than the Non-AD group (0.11 vs 0.0). This means that AD patients are producing these signals of non-understanding and clarification requests even in response to simple yes-no questions. On the other hand for wh-questions and declarative wh-questions, *br*

[4]These responses are normalized by total no of respective category of questions

Question-type	**Response-type**	**AD**	**Non-AD**
qy	br	**0.11**	0.00
	qc	**0.03**	0.00
qw and qw^d	br	**0.12**	**0.05**
	qc	**0.04**	**0.02**
qy^d	br	**0.05**	0.00
	qc	**0.01**	0.00
^g	br	**0.14**	0.00
	qc	0.00	0.00
qr	br	0.00	0.00
	qc	0.00	0.00

Table 3.9: Responses in terms of *br* and *qc* for each question category.

Response against each question type	Fisher's value	Exact significance(2-tail)
qy	21.965	**.002****
qy^d	6.26	.609
^g	4.65	0.54
qr	4.312	1.00
qc	9.209	**0.05***
qw & qw^d	13.536	**.078-**

$^{**}\mathbf{p} < \mathbf{.01}$, $^{*}\mathbf{p} < \mathbf{.05}$, and **-** shows a trend toward significance at $\mathbf{p} < \mathbf{0.1}$.

Table 3.10: Fisher's ExacT-test with $\alpha = 0.05$

are much higher for AD patients (0.12 vs 0.05) than Non-AD patients and *qc* also slightly higher than Non-AD patients. We did not observe any *br* and *qc* in response to choice questions. This reveals that both groups elicit more responses to choice questions and no such signals are generated.

We then perform the Fisher's exact test to check the significance among two groups (AD vs Non-AD) based on response categories. This demonstrated substantial differences in the occurrence of *qy* and almost significant differences in *qc* between AD patients and Non-AD patients. Fisher's test results of response categories are displayed in table 3.10. Fisher's analysis of the *qy* ^*d*, *qr* and ^*g* question types for response categories showed no difference with both groups consistently responding.

Corpus study research question-5

Is the pauses behavior including pauses between patient utterances, between interviewer utterances, between patient to the interviewer, and from interviewer to the patient is significant between the two groups?

Descriptive statistics for the temporal measures for the AD and Non-AD groups are provided in Table 3.11. Effect size[5] (Cohen's d effect size) and *p-value* for independent T-test are also reported.

	AD		**Non AD**		**T-test**	
	Mean	**SD**	**Mean**	**SD**	**Effect size**	***p value***
Avg No. of pauses	0.078	0.062	0.025	0.031	1.08	**0.016***
Avg No. of P pauses	0.047	0.40	0.030	0.038	0.06	0.17
Avg No. of I pauses	0.049	0.048	0.008	0.013	1.17	**0.014***
Avg No. of I-P pauses	0.02	0.02	0.003	0.005	1.17	**0.03***
Avg No. of P-I pauses	0.02	0.02	0.002	0.006	1.21	**0.005****

$^{**}p < .01$ and $^{*}p < .05$

Table 3.11: Descriptive statistics for participants in the AD group and the non-AD group for temporal measures

There were significant differences in *Avg No. of pauses, Avg No. of I pauses, Avg No. of I-P pauses* (p .05) and *Avg No. of P-I pauses* (p .01) between AD and Non-AD participants. It is obvious from these findings that transition relevant places (TRP) in turn-taking (from I-P or P-I) are more significant among the groups than the pause within the same speaker utterances. There were no significant differences between AD and Non-AD participants in *Avg No. of P pauses* within patient utterances. Despite this, it was decided to keep them for further analysis. Small sample sizes were thought to be the cause of this lack of significance.

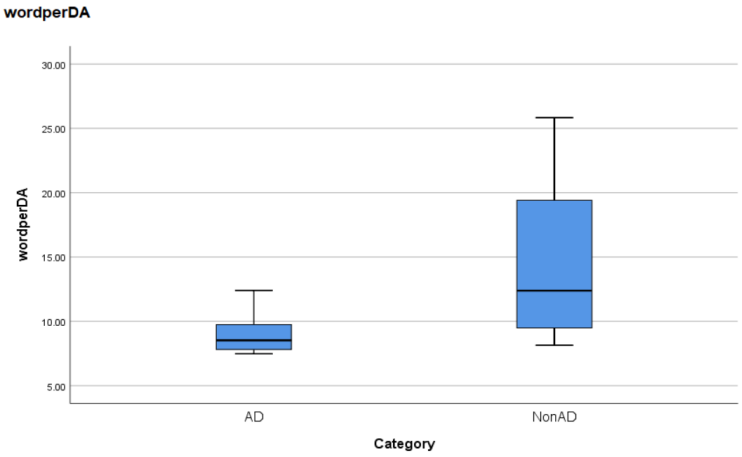


Figure 3.5: Spread of words/DA among two groups

[5]Cohen's d effect size was used. Effect size is modest if d=0.2, medium with (0.5), and larger effect with d= 0.8.

The T-test findings for words per DA act categories revealed that AD patients produce less number of words per DA while speaking less within their utterances (figure 3.5). It is also worth mentioning that AD patients produce shorter utterances containing fewer words per utterance than Non-AD patients as shown in figure 3.6.

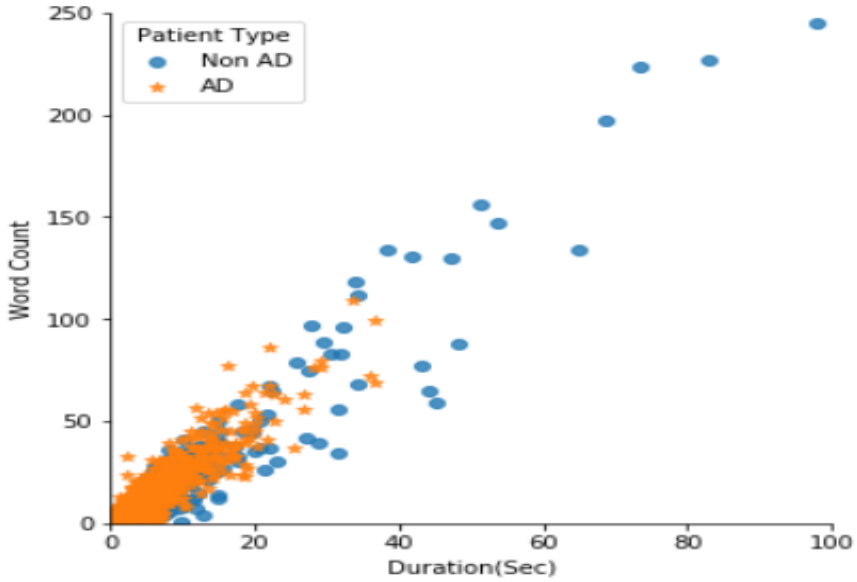


Figure 3.6: Scatter plot of CCC corpus utterance duration by word count

Table 3.12 reports results for the duration of pauses of the patient, interviewer, from interviewer to patient, and patient to interviewer with average duration per participant. There is a significant difference in the *Avg I pause Duration* within their own sub-

	AD		**Non AD**		**T-test**	
	Mean	**SD**	**Mean**	**SD**	**Effect size**	***p value***
Avg P pause Duration	2.7	2.32	1.71	2.41	0.42	0.34
Avg I pause Duration	4.13	3.3	0.95	1.6	1.22	**0.014***
Avg I-P pause Duration	2.8	2.7	1.01	2.47	0.14	0.129
Avg P-I pause Duration	2.92	2.82	0.45	1.42	1.1	**0.023***

$^{**}p < .01$ and $^{*}p < .05$

Table 3.12: Descriptive data for participants in the AD group and the Non-AD group for temporal measures of duration.

utterances and the most important significant difference is *Avg P-I pause Duration* while taking a turn from patient to the interviewer($p<.05$) in the conversation.

3.6 Discussion

This study's main objective was to determine how the interaction between AD and Non-AD groups is different by analysing natural conversations. Our study provides the first analysis of different types of questions asked in conversations with AD patients in the Carolina Conversation Collection (CCC) Corpus. We found that yes-no questions were asked more frequently in the AD sufferer conversations than the Non-AD conversations (6% vs 3.7% of all dialogue acts) and fewer Wh-questions were asked in AD sufferer conversations compared to Non-AD ones (4% vs 5.4%). While our newly introduced tags were not frequent, they are significant in AD sufferer conversations, with 2% of all dialogue acts by AD sufferers being clarification questions and 3% being signals of non-understanding. Similarly, the percentage of choice questions was higher in the AD group which suggests these questions reduce the burden of processing demands by facilitating the patient with all of the information necessary to respond. This is thought to be a particular benefit for people with AD who struggle to access and retrieve words and formulate responses.

The further analysis enables us to characterize that more signal non-understanding and clarification requests are produced in response to the simple yes-no questions and wh-questions from the AD group. We may assume that these sequences of questions followed by these signals of non-understanding and clarification requests may be attributed towards AD. Previous studies discussed in chapter 2 section 2.1 on DementiaBank did not look into this type of interaction, which gives us the motivation to design computational models to capture this phenomenon of natural conversations.

In terms of temporal measures, the results indicate that verbal sequences are shorter in AD discourse with fewer words spoken as compared to the Non-AD group. We also discovered that pauses happen more frequently outside of significant syntactic boundaries. We also accept the following limitations of our study:

- The dialogue sample chosen for the Non-AD group are not very interactive. The chosen samples were narrative descriptions of daily routine with very less interaction with the interviewer.

- Simply considering pauses within patient and interviewer utterances does not convey useful information. For example, long delays with the interviewer's utterances suggest looking at a more functional perspective of pauses behavior. (Feedback from SemDial Presentation). Further studies are now required to explore the possible relationship between temporal measures like the number and duration of short pauses and long pauses among these groups.

- In subsequent analyses, we will work with a larger sample size by considering more patients in both groups as it was felt that patients' number of pauses and

duration of pauses were not significant within their own utterances and these may be due to small sample size.

There is a need to investigate how to build an automatic DA tagger to utilize the benefits of the findings of this corpus study. The existing dialogue act taggers either do not provide all these patterns in terms of clarification requests, signal non-understanding, question types, etc or they provide poor performance on these very rare and less common phenomena. In the next chapter, we are going to try an automatic dialogue act tagger that can tag the utterances of the conversation with the rare dialogue acts to give more insight and interpretation of what is being said within the same Carolina corpus. In the future, we will also explore more compound questions and questions related to semantic and episodic memory. For this purpose, It is also need to do categorization of existing questions within the existing dataset.

CHAPTER 4

AUTOMATIC RARE CLASS DIALOGUE ACT TAGGER

In many applications that require natural language understanding, the identification of DA is a crucial component of understanding the meaning of an utterance. It is assumed that the identification of certain DA's for the patient's utterances in conversations will help in providing cues for the identification of AD. Carolina's CCC corpus is not annotated with dialogue acts and manually annotate patient's conversations to analyze speech act patterns is a time-consuming process. In chapter 2 section 2.4, I discussed different approaches used so far on the task of DA classification. In this chapter, I will discuss proposed approach for building an automatic DA tagger to classify DA's at the utterance level, with pre-possessing steps, features used, experimental setup, and evaluation metrics and in the last section, results of DA tagger performance are presented on both the general SwDA corpus and an AD specific conversational dataset, the CCC Corpus[1].

4.1 Background

Traditional machine learning approaches have been investigated and have achieved state-of-the-art performance for DAs classification using domain-independent dialogue act scheme i.e Discourse Annotation and Markup System of Labeling (DAMSL) tag-set (Stolcke et al., 2000) with a set of 42 DA tags. Stolcke et al. (2000) used a Hidden Markov Model with the intuition that key information lies in both the sequences of words within sentences/utterances and the sequence of dialogue acts over utterances. Improvements have been gained by using Conditional Random Fields (Zimmermann, 2009), cue phrase-based models (Webb et al., 2005), joint classification and segmentation (Ang et al., 2005). Recently deep learning approaches have shown promising results

[1]This chapter is based on work published in SIGDIAL,2021 (Nasreen et al., 2021a)

on the task such as Recurrent Neural Networks (RNN) (Kalchbrenner and Blunsom, 2013), (Ortega and Vu, 2017) and Convolutional Neural Networks (CNN) (Lee and Dernoncourt, 2016). Most recent work sticks with Stolcke et al. (2000)'s original intuition to include contextual information (preceding utterances and their DA roles help predict the current utterance), often via hierarchical models where the higher layers capture DA/utterance sequence information; see e.g. (Raheja and Tetreault, 2019)'s use of a CRF above dialogue-level and utterance level BiLSTMs, achieving state-of-the-art accuracy of 82.9% on the standard SwDA corpus. Unlike traditional models that only consider the past or preceding context, a BiLSTM model, being bidirectional, processes both the past and future utterances to capture a more comprehensive context for assigning dialogue acts to the current utterance. The term "future utterance" refers to the utterances that come after the current utterance in a conversation.

Most advanced studies in deep learning have shown that considering the contextual information i.e. proceeding utterance plays an effective role in predicting the current utterance. For example, keeping the previous utterance as a question helps in predicting that the next utterance 'yeah' is the answer and not a backchannel. This fact is supported by different work including (Ortega and Vu, 2018). However, variants exist; Bothe et al. (2018), for example, consider only a limited number of preceding utterances as a context within an RNN, rather than the full sequence, accuracy is reduced to 77.34% on SwDA but their model, in using only limited preceding context (rather than assuming knowledge of future utterances) is suitable for an incremental online setting.

All these approaches, however, train and evaluate their model assuming that the goal is average performance over a general DA tagset, usually a 42-tag SwDA DAMSL scheme. Few approaches for DAs classifications also utilized fewer classes rather than the full 42 DA tag set. Fuscone et al. (2020) used three dominating DA classes: *Statement*, *Opinion*, and *Backchannel*; Ramacandran (2013) use an 18-tag DAMSL subset. Sridhar et al. (2009) have grouped the 42 classes into 7 disjoint classes based on their frequency and grouped the remaining classes into the 'other' category. They achieved an accuracy of 72.6% on 42 tagset and an accuracy of 82.6% with 7 tagset by considering up to 3 preceding utterances as context. They reported an improved performance of 76% with 42 tagset and 83.1% with 7 tagset considering 3 previous DA's as context on the SwDA corpus. Their model also has the ability to simultaneously observe future utterances inside the conversation. The use of future utterance which restricts its application in dialogue systems where we are unable to foresee the next utterances and only know the context of the previous utterance.

In contrast to the above-mentioned approaches that mainly focus on the 42-tag DA tagset or most common classes (Fuscone et al., 2020; Ramacandran, 2013; Sridhar et al., 2009), I am interested in looking at the rare classes useful for dementia analysis following CA work described in chapter 2 section 2.3. The term "rare class" refers to a dialogue act category that occurs infrequently as compared to frequent classes such as statements

and back-channels but are found very useful in Dementia analysis. The complete list of these DA classes can be found in section 4.3.1 (see Table 4.1). Few studies give details of accuracy on these rarer classes; but Raheja and Tetreault (2019), despite achieving 82.9% accuracy overall, show accuracy of only c.25% for *br* (*signal-non-understanding*, which makes up only 0.1% of SwDA utterances), c.30% for *b^m* (*repeat-phrase*, 0.3% of utterances), c.20% for *qy* (*yes-no-question*, 2%), and <5% for both *qw* (*wh-question*, 1%) and *b* (*backchannel*, a relatively common but important tag).

There are three main reasons that motivate to build our own DA tagger for the rare class phenomenon:

1. Most of the existing research on DA did not make these taggers publicly available. Only a few DA taggers are publicly available such as an ISO standard DA tagger[2] that follows a multi-dimensional ISO standard annotation scheme (Mezza et al., 2018). Another DA tagger Emotional dialogue Act Corpus[3] is available online that used a model trained on dialogue act corpora and used it on an emotion-based corpus to annotate the emotion corpora with dialogue act labels, and an ensemble annotator extracts the final dialogue act labels (Bothe et al., 2018). However, to use this setup, it is needed to send the utterances to their server due to the complex structure representation of the utterances. Because due to patient data and due to privacy concerns, data cannot be shared and sent to the server.

2. The existing research focused on improving the overall accuracy of the tagger and does not provide good performance on rare classes.

3. The existing DA tagger is neither allowed to customize the DA tagset nor to add new tags and hence these taggers cannot be utilized for our research purpose. There were also no existing DA taggers that dealt with only rare classes of the SwDA corpus.

Here, then, our purpose is to improve DA tagging accuracy for the specific DA classes of interest in AD diagnosis, including specific types of questions, answers, and misunderstanding signals, most of which are relatively rare. For this purpose, a context-based hierarchical BiLSTM Model with attention is used, to capture relations at the word level, utterance, and DA level and leverage utterance/DA context information. The advantage of using only a few preceding (left) utterances in context rather than the whole conversation is that it is suitable for dialogue systems in real-time, due to the incremental nature of dialogue. The existing models which take into account the whole conversation can achieve overall higher accuracy on the general DA tagging task so might be expected to improve our rare-class task as well, but require information about future utterances (Li et al., 2018; Raheja and Tetreault, 2019).

[2] https://github.com/ColingPaper2018/DialogueAct-Tagger
[3] https://github.com/bothe/EDAs

4.2 Proposed approach

DA classification is a multi-class problem and to predict DAs for each utterance, several sets of experiments are performed using features from the transcripts of data, to compare a range of models:

- A baseline model using word embedding as text features without any context information
- A Hierarchical BiLSTM model using word embedding and previous utterance representation from context.
- A Hierarchical BiLSTM model using word embedding, previous utterance representation and previous predicted DA tags from context.

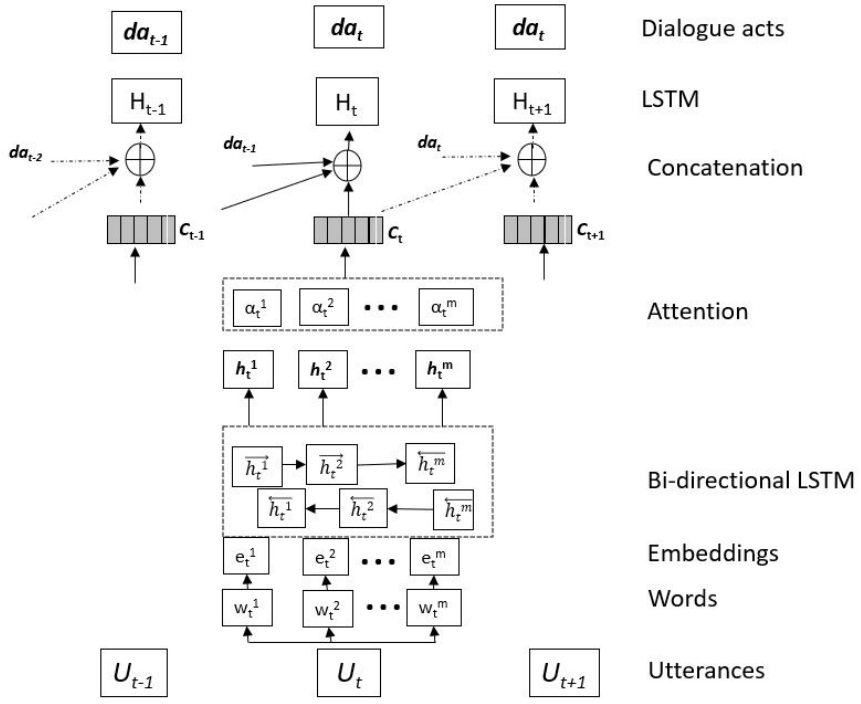


Figure 4.1: Model architecture for DA classification with one utterance and one DA as context.

4.2.1 Model Representation

The task of DA tagging takes the dialogue conversation D as input from a collection of dialogue conversations where D=$\{D_1, D_2, D_3 ... D_n\}$. Each Conversation D is a

sequence of utterances $U = \{U_1, U_2, U_3, ..., U_n\}$ paired with a sequence of DA labels $Y = \{da_1, da_2, da_3, ..., da_n\}$; each utterance $U_t \in U$ is a sequence of words $U_t = \{w_t^1, w_t^2, ..., w_t^m\}$.

Figure 4.1 shows the overall architecture of our model in which U_t represents the current utterance and U_{t-1} represents the previous utterance. Word embeddings are used to extract the lexical feature representations from the transcripts, converting the utterances from word sequences into sequences of word vectors. Usage of different embeddings is compared including randomly initialized embeddings, Glove pre-trained embedding (Pennington et al., 2014), a Glove embedding trained on SwDA and CCC corpus, and contextual Embedding from Language Model (ELMo) (Peters et al., 2018). The embedding matrix maps each word of the utterance into a dense vector representation of the fixed size. Let $\{e_t^1, e_t^2, ..., e_t^m\}$ represents the word embedding sequence for an utterance U_t with m words where $e_t \in R^d$ is the d-dimensional word embedding vector for m words.

$$e_t^i = emb_w(w_t^i) \in R^d \tag{4.1}$$

A Bidirectional Long Short Term Memory (BiLSTM) layer follows the word embedding representation layer, and with this architectur, each LSTM retains the record of three multiplicative units of input, output, and forget gates and self-connected memory cells, which update the data when it is deemed necessary. At each time step t with an *input gate* i_t, a *forget gate* f_t, *an output gate* o_t, a *memory cell* c_t and a *hidden state* h_t, the LSTM units equations are the following:

$$i_t = \sigma_i(W_{ei} e_t^i + W_{hi} h_{t-1} + b_i) \tag{4.2}$$

$$f_t = \sigma_f(W_{ef} e_t^i + W_{hf} h_{t-1} + b_f) \tag{4.3}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tanh(W_{hc} e_t^i + W_{hc} h_{t-1} + b_c) \tag{4.4}$$

$$o_t = \sigma_o(W_{eo} e_t^i + W_{ho} h_{t-1} + b_0) \tag{4.5}$$

$$h_t = o_t \odot \tanh(c_t) \tag{4.6}$$

It comprises of two hidden layers; it first compute the forward hidden vectors $\overrightarrow{h_t^i}$ and then the backward hidden vectors $\overleftarrow{h_t^i}$ and combines $\overrightarrow{h_t^i}$ and $\overleftarrow{h_t^i}$ to generate h_t^t. This layer produces a representation of an utterance as a sequence of corresponding hidden vectors $h_t = \{h_t^1, h_t^2, ..., h_t^m\}$.

$$h_t^i = concat(\overrightarrow{h_t^i}, \overleftarrow{h_t^i}) \tag{4.7}$$

Attention mechanism is then used to weigh these and aggregate them into a single utterance representation. Firstly, the attention score is computed as:

$$s_t^i = W_1 tanh(W_2 . h_t^i + b) \tag{4.8}$$

Here, W_2 is a weight matrix of hidden vectors, and W_1 and b are parameter vectors all learned during training. The score vector S_t is mapped into a probability vector by using softmax functionality:

$$\alpha_t = softmax(s_t) \tag{4.9}$$

Then, attention vector C_t is computed which is a single vector representing the whole utterance U_t.

$$C_t = \sum_{i=1}^{m} \alpha_t^i . h_t^i \tag{4.10}$$

The vector α_t =$[\alpha_t^i]_{i=1}^m$ is a sequence of positive numbers sums to 1, this yields a probabilistic interpretation of Attention. I then concatenate the vector for the current utterance c_t with various combinations of information from previous context: the previous utterance vector c_{t-1}, previous DA (da_{t-1}) (gold-standard or predicted, see section 4.6), and their preceding neighbors c_{t-2}, da_{t-2}. These concatenated vectors are then encoded by a second LSTM (here, a unidirectional left-to-right LSTM is used, rather than bidirectional, to stay compatible with utterance-by-utterance online processing); the resulting sequence of hidden vectors H=$\{H_1, H_2, ..., H_n\}$ is then used to predict da_t, the DA label of the current utterance U_t.

$$da_t = softmax(W_o . H_t + b_o) \tag{4.11}$$

Here, W_o is the output weight matrix and b_o is the bias vector and is learned during training. Softmax is used on the output as this is a multi-class classification problem.

4.2.2 Feature set

4.2.2.1 Current utterance information

For DA classification, our aim is to model the sequential information of utterances. It is worth mentioning that only the transcriptions of the utterances in this study are used, coupled with speaker information (i.e., whether the current utterance is from the same speaker as the prior one or not). We used the lexical feature representations from the transcripts using various word-level representations. Following word embeddings are used to convert the utterances into vectors representations:

Glove Embeddings: The glove is a pre-trained set of embedding trained on a Wikipedia corpus. In order to reduce the reconstruction error between the co-occurrence statistics predicted by the model and the global co-occurrence statistics seen in the training corpus, Glove learns vector embedding from the co-occurrence matrix. (Pennington et al., 2014).

Embedding from Language Model (ELMo): ELMo's are deep contextualized word representations. Elmo embedding differs from traditional embedding in the sense that it does not assign fixed word embedding to each word, Elmo assigns a representation to a word that is the function of the entire input sentence. So same words can have different embedding depending on the context within the sentence. It uses a bidirectional recurrent neural network trained on a particular task to create the embedding (Peters et al., 2018).

4.2.2.2 Previous utterance context information

I will use both the current utterance lexical information and a number of preceding utterances as context and is concatenated as additional features. The advantage of using only a few preceding utterances in context rather than the whole conversation is that it is suitable for dialogue systems in real-time, due to the incremental nature of dialogue.

4.2.2.3 DA history information

Since it is anticipate that the DA tags include useful sequential information, in the first approach, a combination of the history of DA information with the current utterance is used to categorize its DA tag. This is represented as additional features concatenated with the LSTM sentence representation, as shown in Figure 4.1. Following different configurations in this framework are evaluated.

- **Use Gold standard (G) DA labels:** A comparison is made using G labels in both training and testing. Note that using GS labels in testing is not a real testing setup. The purpose of this is only to understand the performance deterioration brought on due to prediction errors and to provide an upper bound.
- **Use G and Predicted (P) labels:** I also use G labels during training and P DA's as context history during testing.
- **Use predicted (P) DA label:** I also used P DA's label in both the training and testing phase.
- **History length:** A comparison is made using DA information from the different number of previous DA labels.

4.3 Experiments

4.3.1 DA filtering

To keep our approach as domain- and dataset-general as possible, I start with the standard DAMSL tagset (Stolcke et al., 2000) and adapt it. Based on the clinical studies

Tagset	Label	Example
Yes-No Question	qy	Did you go anywhere today?
Wh-Question	qw	When do you have any time to do your homework?
Declarative Yes-No Question	qy^d	You have two kids?
Declarative Wh-Question	qw^d	Doing what?
Or Question	qr	Did he um, keep him or did he throw him back?
Tag Question	^g	But they're pretty aren't they?
Clarification Question	qc	Next Tuesday?
Signal Non-understanding	br	Pardon?
Backchannel in question form	bh	Really?
Yes answer	ny	Yeah.
Yes- plus expansion	ny^e	Yeah, but they're
Affirmative non-yes answer	na	Oh I think so. [laughs]?
No answer	nn	No
Negative non-no answers	nn^e	No, I belonged to the Methodist church.
Other answers	no	I, I don't know.
Statement answer	sa	Popcorn shrimp and it was leftover from yesterday.
Declarative statement	sd	Me, I'm in the legal department.
Backchannel(continuer)	b	Uh-huh
Repeat phrase	b^m	Ahh, Corn Bread.
Other	Other	I'm sorry

Table 4.1: Dialogue Act Tags with their Labels and Example

described in Section 2.3, 17 specific DA tags of interest from DAMSL are kept; split 2 of them each into 2 sub-categories; and collapse all other tags into a single **other** tag, giving a total of 20 tags. The two new DA tags are **clarification-request** *(qc)* and **statement-answer** *(sa)*: clarification-request *(qc)* is a sub-category of *signal-non-understanding (br)* which requests more specific information (see e.g. Purver et al., 2003; Rodríguez and Schlangen, 2004); while *statement-answer (sa)* is a sub-category of *declarative-statement (sd)* used as an answer to a wh-question *(qw)*, open-question *(qo)* or or-question *(qr)*. The full tagset[4] is shown in Table 4.1.

4.3.2 Datasets

The proposed model will be evaluated on two corpora. First, a standard DA labeled dataset, the **Switchboard Corpus (SwDA)**, a corpus of 1155 five-minute two-speaker telephone dialogue conversations, containing 205K utterances in total. I took all utterances, keeping tag labels included in our selected tagset as shown in Table 4.1, and labeling the rest of the utterances with an **other** tag.

[4]The annotation guidelines are available from https://osf.io/8w9z2/?view_only=ee08242870f24ae7ab6754ddf9a0176a.

Second, the **Carolina Conversation Collection (CCC)** as discussed in chapter 3. We used the same samples of dialogues as used in our initial corpus study. This dataset is transcribed but not annotated with DA tags. We performed manual annotation with DA tags using the SwDA-derived tagset with the set of filtered tags mentioned in table 4.1 above. We annotated 20 conversations with 10 Non-AD patients and 10 conversations with AD patients, giving a total of 30 conversations from CCC[5]. Comparing three annotators on one sample conversation, we achieved an inter-rater agreement of 0.844 (See section 3.2.4).

For the SwDA corpus, we reduced the original 42-tag labels to our reduced tagset. This required manual re-tagging of some *signal non-understanding* utterances with the new subcategory *clarification-request*, and similarly re-tagging some *declarative statement* utterances as *statement answer* (*sa*). To accommodate the new tags in existing data, we performed the following conversions discussed in Section 4.3.2.1 and 4.3.2.2.

4.3.2.1 Conversion of *br* tag

- We manually extract all the signal non-understanding *(br)* and repeat signal non-understanding *(br^m)* sample utterances. We analyzed all those utterances by looking at their previous utterances and the speaker's identity of previous utterances.
  - More generic clarification signals like 'Excuse Me', 'sorry', 'huh', and 'pardon' are labeled with *br* tag.
  - Specific clarification request of part of the previous utterance are labeled with *qc* tag.
- There were total 219 *br* tags and after applying the above-mentioned transformation, there were 97 utterances with *br* tag and 122 with *qc* tag.

4.3.2.2 Rule based classifier for *sa* tag

The second newly introduced tag *sa* is not present in SwDA. We add this tag as a new tag because in SwDA the answer statement to wh-questions, open question, and choice questions are represented with declarative *(sd)*/ opinion statement *(sv)* statements. So it is difficult to distinguish between a declarative statement and an answer statement.

We took 8 conversations from the SwDA corpus containing 27 questions(wh-questions, open-ended questions, and or-questions) and manually re-tagging their answers from

[5]The annotations are available for the research community for further followup work and can be useful after getting access to CCC dataset: https://osf.io/8w9z2/?view_only=ee08242870f24ae7ab6754ddf9a0176a

Speaker	Text	Tag	New Tag
A	What are you learning to be?	qw	-
B	Oh, oh, excuse me?	br	br
A	not fond of Howse	sd	-
A	and I'm not fond of Grieves	sd	-
B	{ F oh} Tom Grieves?	br ^m	qc

Table 4.2: Example for conversion of *br* and *qc* tag in SwDA data.

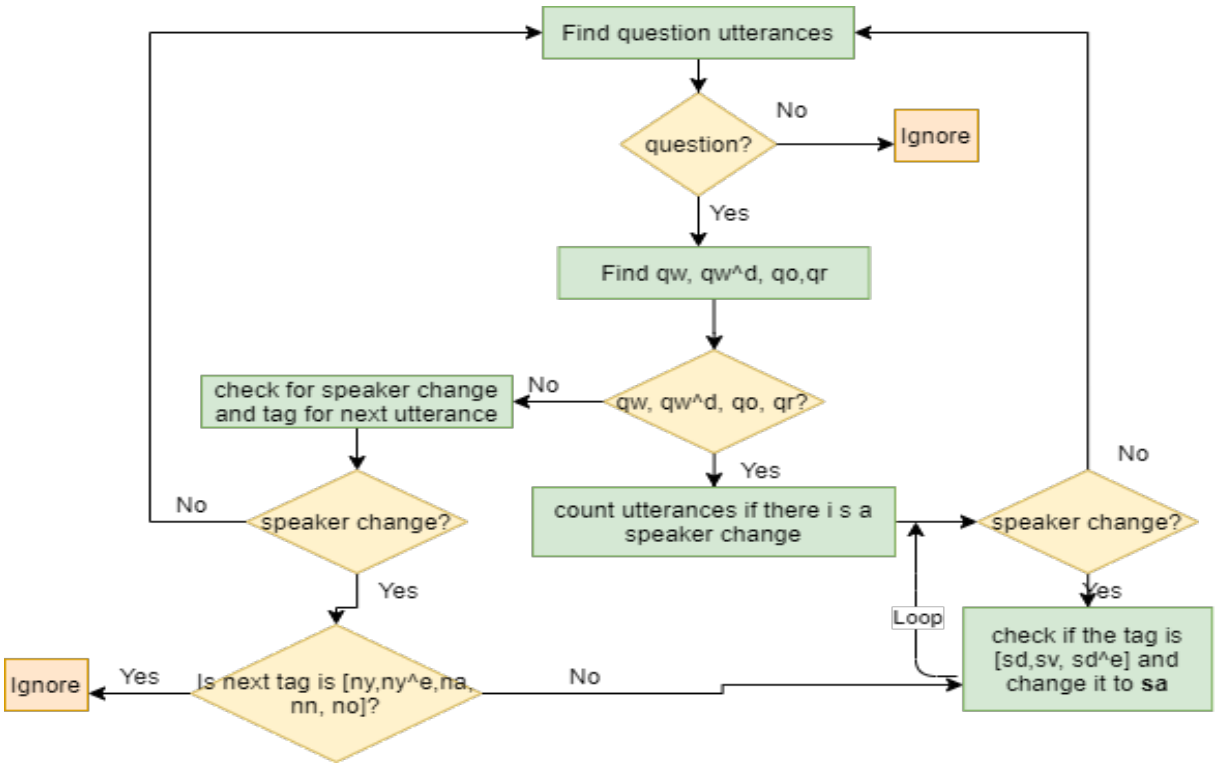


Figure 4.2: Rules for converting declarative statement *(sd)* to statement-answer *(sa)*

sd to *sa* tag. From this, we then built a rule-based classifier to derive simple rules for conversion as shown in figure 4.2 of these statements to *sa* tags. The prediction score for the classification from the rule-based classifier into *sd* and *sa* is reported in Table 4.4. We then used our rule-based classifier to convert statements in response to questions into *sa* statements on the rest of the SwDA corpus. All other tags except the 19 chosen tags, are converted to the '*other*' class. The complete tagset along with its distribution is given in table 4.5. It is obvious from the table 4.5 that most of the classes of interest are very rare in the corpus.

Table 4.6 show the test and train split of both corpora. We trained our model on the SwDA train set and evaluated its performance on test set I (Swda test samples) and

Speaker	Text	Tag	New Tag
A	Do you know anyone that, uh, is, is in a nursing home or has ever been in one? /	qr	-
B	No. /	nn	-
B	But I, my grandparents were looking into it before /	sd ^e	*sa* -
B	so I know what they've said. /	sd	sa
A	Uh-huh. /	b	-

Table 4.3: Example for conversion of declarative statement into statement-answer in SwDA

Class	Precision	Recall	F1
sa	1	0.83	0.90
sd	0.86	1	0.92

Table 4.4: Prediction score for Rule-based classification

Tag	Count	%
b	25395	18%
qy	2996	2.1%
qy^d	820	0.6%
qw	1431	1.0%
qw^d	55	0.04%
qo	459	0.32%
qr	308	0.21%
^g	398	0.3%
bh	776	0.54%
br	97	0.07%
qc	122	0.09 %
na	567	0.4%
no	215	0.15%
ny	2068	1.5%
ng	210	0.14%
nn	934	0.65%
sa	6966	5%
b^m	493	0.35%

Table 4.5: SwDA DA tag count and frequency

test set II (CCC data). Initially, we applied basic reprocessing on the text to remove discourse markers and other annotation symbols. As transcripts from both corpora are transcribed with punctuation marks, we keep { ? , . ´} and the rest of the punctuation are filtered out. Additionally, utterances with *Nonverbal* expressions like *'Chuckles'*,

Dataset	SwDA	CCC
Transcripts	1115	30
Total Utterances	142022	5082
Training Utterances	111356	-
Test Utterances	27840	5082

Table 4.6: Datasets

'Laughter', *'Throat cleaning'*, *'Pauses'*, *'Beeps'* are removed from transcripts as they do not contain any relevant lexical information. Empty speech segments from CCC were also removed. Utterances in the transcript with '+' are concatenated with the previous utterance of the same speaker which reduces the utterance count to 142022.

4.4 Implementation and Evaluation Metrics

In order to predict rare class DAs for each utterance, we set up our model to learn the most pertinent information from text features. We performed a grid search for hyperparameter tuning, changing one hyperparameter at a time while keeping the other ones fixed, on the validation split. With a learning rate of 0.01 and categorical cross-entropy as the loss function for multi-class outcomes, we utilized ADAM (Kingma and Ba, 2014) to train our model. As the classes in our data are highly imbalanced, we use a class-weighted objective function to prevent over-prioritising more common classes; use scikit-learn's StratifiedShuffleSplit (a merge of StratifiedKFold and ShuffleSplit) to preserve the percentage of each class in each fold. Embedding size was set to 100 dimensions for both simple word embeddings and GloVe pre-trained embeddings, with 1024 dimensions for ELMo embeddings. Early stopping was employed to prevent the network from over-fitting, and 20% of the training samples were split for validation.We wait for at least 3 epochs during which the validation set accuracy does not increase. We run the model over 20 epochs and typically, both models, baseline, and context-based hierarchical model took about 8 to 10 iterations. To consider the context for the current utterances, we set the context window up to 3 previous utterances and 3 previous DA's.

We report accuracy, macro-average precision (Prec.), macro-average Recall (Rec.), and macro-average F1 as metrics for multi-class classification. Precision is the ratio of true positive predictions to the total number of positive predictions (true positives + false positives) while recall is the ratio of true positive predictions to the total number of actual positive cases (true positives + false negatives). F1 measure is the harmonic mean of precision and recall. It provides a balanced measure that considers both false positives and false negatives. It is especially useful when there is an imbalance between the classes. These measures of precision, recall and F1 are used to calculate macro-average precision, macro-average recall and macro-average F1. We choose macro-average measures as data

is highly imbalanced and in order to treat all DA classes equally, macro-average will calculate the measure separately for each DA class before averaging the results. With macro-average measures, each class will get equal weight regardless of the sample size of each class. Macro-average precision, recall, and F1 score are calculated according to equation 4.12,4.13, and 4.14.

$$macro-average\ Precison = 1/n \sum_{i=1}^{n} Prec_i \tag{4.12}$$

$$macro-average\ Recall = 1/n \sum_{i=1}^{n} Rec_i \tag{4.13}$$

$$macro-average\ F1\,score = 1/n \sum_{i=1}^{n} F1_i \tag{4.14}$$

Here n is the number of DA classes and i is specific DA from $1..n$.

4.5 Baseline Model

The base model is defined as a single utterance classifications at the sentence level without considering any contextual information either previous utterance or previous DA.

In the next section, results are discussed for the following set of experiments:

1. A baseline model that will not take any contextual information into account across all three different embedding features.

2. A hierarchical BiLSTM model that will consider one preceding utterance as context.

3. Taking into account one previous utterance and its previous DA label as context history

4. Only two previous utterance contexts.

5. Two previous utterances along their DA labels as context history.

6. Three utterances as context information.

7. Three preceding utterances with their DA labels to predict the DA for current utterance

Context	Embedding	SwDA test set				CCC test set			
		Prec.	**Rec.**	**F1**	**Acc.**	**Prec.**	**Rec.**	**F1**	**Acc.**
No context (Baseline)	No Pretrain	0.42	0.47	0.42	0.79	0.33	0.34	0.31	0.50
	Glove	0.44	0.46	0.44	0.83	0.38	0.36	0.32	0.53
	ELMo	0.45	0.55	0.46	0.80	0.37	0.37	0.34	0.52
1 utt only	No Pretrain	0.45	0.57	0.49	0.81	0.44	0.44	0.41	0.55
	Glove	0.48	0.57	0.51	0.83	0.46	0.48	0.43	0.57
	ELMo	0.43	0.54	0.45	0.78	0.40	0.38	0.35	0.52
1 utt & 1 DA	No Pretrain	0.52	0.62	0.56	0.87	0.49	0.45	0.44	0.62
	Glove	**0.55**	**0.62**	**0.57**	**0.88**	**0.48**	**0.47**	**0.45**	**0.62**
	Glove Swda-CCC	**0.57**	**0.61**	**0.58**	**0.88**	**0.51**	**0.48**	**0.45**	**0.66**
	Glove (SP info.)	0.54	0.64	0.57	0.87	0.46	0.46	0.43	0.64
	ELMo	**0.55**	**0.64**	**0.58**	**0.88**	0.47	0.43	0.40	0.62
2 utt only	No Pretrain	0.46	0.53	0.49	0.82	0.37	0.36	0.33	0.53
	Glove	0.48	0.57	0.50	0.82	0.44	0.43	0.40	0.55
	ELMo	0.42	0.45	0.40	0.81	0.40	0.38	0.33	0.51
2 utt & 2 DA	No Pretrain	0.52	0.62	0.56	0.87	0.44	0.45	0.42	0.63
	Glove	0.56	0.59	0.57	0.88	0.48	0.46	0.43	0.69
	ELMo	0.59	0.59	0.56	0.88	0.49	0.43	0.42	0.63
3 utt only	No Pretrain	0.35	0.49	0.40	0.77	0.42	0.33	0.33	0.49
	Glove	0.32	0.43	0.35	0.79	0.35	0.31	0.3	0.51
	ELMo	0.44	0.45	0.39	0.76	0.33	0.38	0.3	0.52
3 utt & 3 DA	No Pretrain	0.51	0.59	0.54	0.87	0.39	0.41	0.37	0.60
	Glove	0.52	0.64	0.56	0.87	0.44	0.45	0.41	0.61
	ELMo	0.51	0.53	0.48	0.88	0.41	0.43	0.36	0.60

Table 4.7: Macro-average precision, recall, F1 score, and accuracy for different contexts with different word embeddings on **SwDA test set** and **CCC test set**.

4.6 Results

Table 4.7 shows the performance of our baseline model (without context) and the proposed models with a range of context settings: with one, two and three previous utterances and previous DA tags as context. Our baseline model yields an average macro F1 score of 0.46 on SwDA test set and 0.34 on CCC test set with ELMo embeddings. Our results improved over the baseline model by adding contextual information of previous utterances and further improved by adding previous DA labels. Our model achieved a macro average F1 score of 0.51 by considering only one utterance as context which is further improved by 6% by considering the previous utterance DA label with an F1 score of 0.57 on SWDA corpus with glove embeddings. Similarly, with ELMo embedding, an F1 score of 0.45 is achieved with an increase of 0.13 in the F1 score by considering the previous DA

label along the preceding utterance with an average of 0.58. The best case scenario, assuming accurate knowledge of words from current and one preceding utterance and one previous DA label results in DA classification with **(Rec.:0.64, F1: 0.58, Acc.: 0.88)** on SwDA test set with Elmo Embedding. Transferring the model learned on SwDA to the AD-specific CCC corpus also gives its best result in this setting: best macro F1 score of 0.45 is obtained on CCC when using one preceding utterance and one DA as context with GloVe embeddings. Using GloVe embeddings trained on the SwDA and CCC data perhaps gives slight improvements over the standard pre-trained GloVe, but they are small (Table 4.7).

I also experimented with different variants of including speaker identity information (e.g. by concatenating speaker ID with DA history); this did not improve results, so results are reported for only for one context setting as an illustration. Overall, these results suggest that the single immediately preceding utterance and DA label have the largest impact on performance: including more context history does not help, and using preceding DAs as well as preceding utterances as context is more effective than using utterances alone. Overall, all context-based techniques significantly outperform the baseline.

Model	**DA**	**Prec.**	**Rec.**	**F1**
1 utt & 1 DA	G	0.55	0.62	0.57
1 utt & 1 DA	P	0.51	0.54	0.49
2 utt & 2 DAs	G	0.56	0.59	0.57
2 utt & 2 DAs	P	0.51	0.52	0.48
3 utt & 3 DAs	G	0.52	0.64	0.56
3 utt & 3 DAs	P	0.58	0.49	0.51

Table 4.8: Comparison of models using gold-standard (GS) DAs label as context vs using predicted (P) DAs as context on SwDA test set. These reported results are macro-averages.

Table 4.7 uses gold-standard contextual DA tag information; this raises the question of whether adding DA information would be less effective when using predictions. Therefore, a comparison is made using predicted (P) DA labels vs. gold-standard (G) DA labels as context when testing, shown in Table 4.8. Better performance achieved when using the gold-standard labels in both training and testing, as expected; on the other hand, when training on gold-standard labels but using previously predicted DAs as context during testing — a more realistic approach in real-time systems —reasonable performance is achieved which improves as the context window increases, suggesting that further improvements may be gained by using more predicted DA labels as context.

4.6.1 Performance on DA classes

Our interest, of course, is not in macro-average figures but in predicting the distribution over the individual DA classes. Therefore, class-wise prediction scores are examined for these DAs labels assigned to utterances. Figure 4.3 shows the recall scores for question tags while scores for answer DAs, signal non-understanding, clarification request, back-channels and other tags are represented in Figure 4.4 and Figure 4.5 indicating higher recall values of DA classes on SWDA than CCC dataset. It is noted that performance exceeds that of Raheja and Tetreault (2019) (see Section 4.1) by a very large margin in all cases. A fairly good recall score is obtained on *qy*, *qw*, *qr*, and ^*g* questions on both corpora. Other classes of yes-answer *(ny)*, affirmative non-yes answer *(na)*, no answer (*nn*), statement answer (*sa*), signal non-understanding *(br)* got fairly good recall scores on both corpora. Our model achieved the best accuracy, macro average recall, and F1 score with context window 1. In this section, class-wise accuracies for different categories of DA are presented.

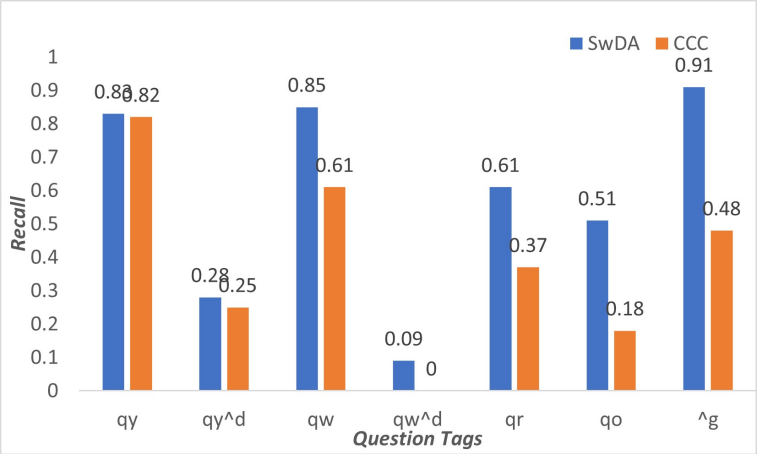


Figure 4.3: Accuracy of questions tags on both SwDA and CCC corpora.

Effect of context: A further experiment is performed to analyse the effect of adding context on the DAs prediction task and presented results for fewer DA classes in Figure 4.6 and Figure 4.7. Yes-answer (*ny*) improved from 0.22 to 0.58 including only one preceding utterance which is further improved by adding the previous DA label with a recall of 0.75. A simple statement 'yes' can be an answer or it could be a back-channel but supplementing with the previous DA label as the yes-no question (*qy*) will help in distinguishing it from back-channel. It has been shown that adding utterance context increases the recall for answer tags e.g *na*, *no*, *ny*, *sa*, and by further adding the DA labels as well have increased the recall across these classes. Statement answers show a significant increase by adding previous utterance and corresponding labels over considering only previous utterance and considering no context at all.

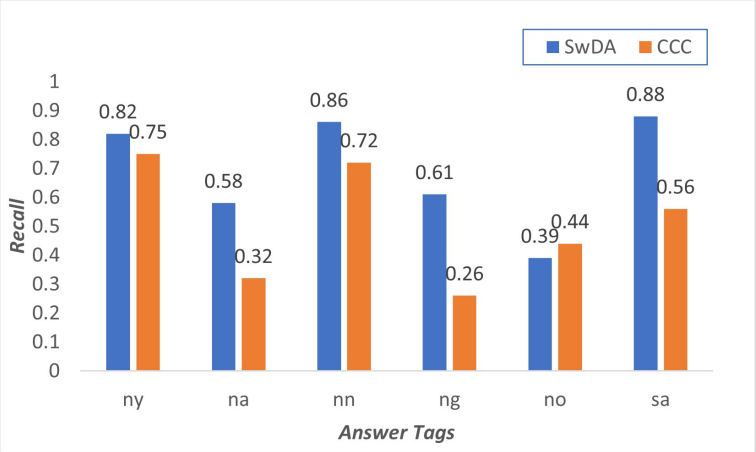


Figure 4.4: Accuracy of Answer tags on both SwDA and CCC corpora.

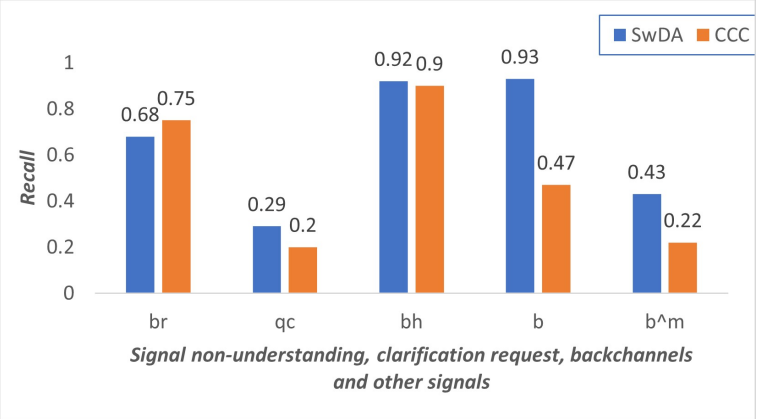


Figure 4.5: Accuracy of other signal tags on both SwDA and CCC datasets.

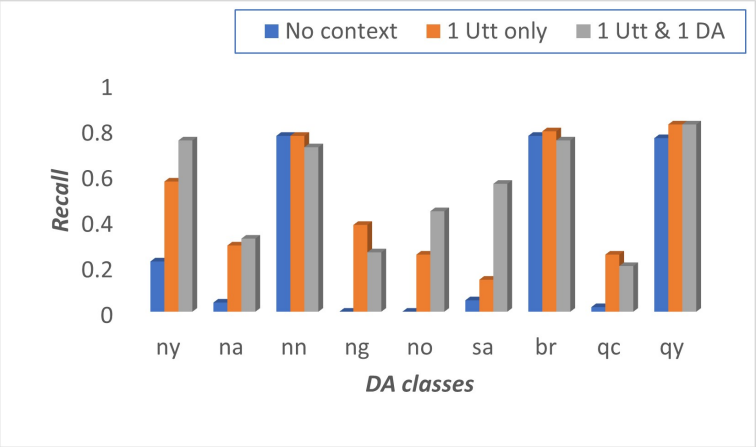


Figure 4.6: Effect of including context on DA prediction on CCC test set

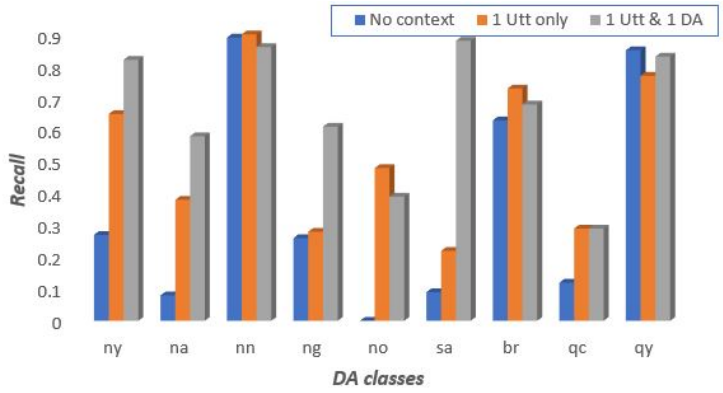


Figure 4.7: Effect of including context on DA prediction on SwDA test set

Error Analysis An error analysis has been conducted to closely look into the poor performance of the model for some DA classes. Poor recall scores are observed for *qw^d* on both corpus and for *qo* questions on CCC. Most of the *qo* and *qw^d* questions are mislabeled with *qw* tag or *other* tag. This is somewhat reasonable, as linguistically the utterances of these classes are quite similar, although the *qw* and *qw ^d* utterances express very specific questions, whereas *qo* utterances tend to be general and brief, they share many syntactic cues which can easily confuse the model. Few *qw ^d* questions also misclassified either *qy^d* or *qy*.

Clarification request *(qc)* recall values are low in both datasets; and upon analysis, it is found that *qc* are somehow confusing with signal non-understanding, wh-questions, and yes-no questions. For example, *qc* utterances with forms such as 'Youre now in what?', 'You must be what?', 'being what?',' what's that?', although requesting clarification in context, are understandably easy to mislabel as *qw*. Encouragingly, including context improved the results better than without using context. Similarly the recall scores for backchannels (*b*) are high for SwDA but lower for CCC. One possible reason could be the different transcription protocols in the two datasets: some transcribers use 'yeah', and 'yup' while others can use the standard form 'yes' to represent a backchannel. Some surface forms of backchannels are also present in the CCC dataset but did not occur in SwDA, and are thus misclassified when testing on CCC. Such examples of some of these cases are shown in Table 4.9 and table 4.10. For instance, the utterance in example 1 seems to be a clarification (*qc*) and is also predicted as *qy*, but its true label is (*qc*). Similarly, an utterance in example 3 underlying text with 'hmm', is a backchannel but predicted as *other*.

SWDA	CCC
key words for back-channel (*b*)	
Uh-huh, okay, yeah, yes, Huh, ah	Huh-uh, okay, yeah, yes, Huh Um-um
Key words for signal non-understanding (*br*)	
Huh?, pardon?, Excuse me?, I am sorry?, Yes? What? Hm?	Huh?, I beg your pardon?, I am sorry?, Ma'm?, Yes?

Table 4.9: Different keywords in transcripts of both corpora for back-channels and signal non-understanding DA labels. Keywords in Blue indicate absence in other datasets.

Example #	Utterance	GS DA	Predicted DA
1	**A:** what's your day like ?	*qw*	*qw*
	B: my day?	*qc*	*qy*
2	**A:** where is that church ?	*qy*	*qy*
	B: fountain hill ?	*qc*	*qy*
3	**B:** on november third , i was ninety four .	*sa*	*sa*
	A: hmm.	*b*	*other*
4	**A:** you're allergic to milk ?	*qy ^d*	*qy*
5	**B:** what do you want to know about her ?	*qw*	*qo*
	A: Everything	*sa*	*sa*
6	so if you was to say to your children they need to do what ?	*qw^d*	*qy*

Table 4.10: Example of utterances of confused pairs (*qc,qy*) and (*qw ^d, qy)* and few more.

4.7 Conclusion

The work in this chapter has presented a DA tagger (a hierarchical BiLSTM model) with a context-based learning approach for the classification of rare class DAs including clarification requests, non-understanding signals, questions, and responses. By using suitable choices of embeddings and the inclusion of contextual history, together with a weighted cost function, our model achieve good performance on these rare classes. The proposed model was assessed on the SWDA and CCC datasets, and the results are presented both with and without context. For SwDA, our model achieved an F1 of 0.58 and a recall of 0.64 when using the immediate preceding utterance and DA label, compared to an F1 of 0.46, a recall of 0.55 without context. It was found that while gold-standard DA information from context gives better performance, the performance using predicted labels can be improved by using longer contextual sequences. The resulting DA tagger utilizes only minimal context of a few preceding utterances and DAs and thus can be easily adopted in real-time systems where one can utilize previous utterances

naturally and not future ones.

In the next chapter, efforts will be made to enhance the DA tagger by incorporating acoustic features from the speech data of both corpora, aiming to further improve class-wise performance. There will be an attempt to modify the model architecture by exploring alternatives, including the use of a CNN layer instead of the LSTM layer to capture longer previous utterance context and incorporating the CRF layer on top to capture the relation between DAs.Exploring improved methods for utilizing attention mechanisms, such as assessing the significance of a token for DA classification based on both its meaning and previous DA history (Tran et al., 2017), will also be attempted. Further, few-shot learning will be explored by adding a few samples from the CCC corpus in training and testing on the rest of the samples.

The purpose of improving these rare-class tags is to use these as interaction features while making classification experiments for AD identification in chapter 6. These predicted DA's unigram and bigram sequences will be used as more specific dialogue features to analyze that either these features show good potential to distinguish between AD and Non-AD patients interaction and whether they can be useful within tools to aid in diagnosis while providing useful, interpretable information about interaction structure, mutual understanding, and question-answering behavior. Phenomena such as clarification requests and signals of non-understanding seem to be quite general across languages and cultures (Dingemanse et al., 2015) and It would be expected that these sorts of conversational features to be more language- and domain-independent than approaches based on vocabulary, syntax, etc for AD diagnosis.

CHAPTER 5

MAKING EXTENSIONS TO RARE CLASS DIALOGUE ACT TAGGER

Dialogue act recognition is important to capture the speaker's intentions in a dialogue system. In Chapter 4, the presented DA tagger is based on transcripts from two corpora. It utilizes lexical information from the current utterance and takes into account fewer previous utterances and DA history in context. This chapter presents further extensions to the existing DA tagger in three ways: first by considering speech data along with transcripts to look at whether adding acoustic features helps in improving the accuracy of certain rare classes that are useful in Dementia Analysis. Secondly, a conversational-level DA tagger will be developed that will consider the entire conversation into account, considering its context, and then comparing its performance with that of our existing DA tagger (from Chapter 4). Finally, a comparison of the results will be conducted, examining the use of contextualized utterance representation with BERT against static pre-trained embeddings. Later in this chapter's discussion section, findings based on experimentation with respect to the research questions investigated are presented.

5.1 Motivation

As discussed in Chapter 4, certain rare classes of interest such as clarification request, declarative wh-questions, and declarative yes-no questions are less common in natural dialogues than other classes such as statements and backchannels. An accuracy of 0.09 was obtained for qw^d from the SwDA corpus and 0 for CCC. Similarly, an accuracy of 0.29 was obtained for clarification request (*qc*) from SwDA and 0.19 for CCC corpus. Poor recall scores were obtained (Figure 4.3 and 4.5) for the qc and qw ^d tags with

both corpora across all models. From CA studies, these classes are useful in identifying different patterns among AD and Non-AD patients. Furthermore, in section 4.6.1, A substantial lexical overlap has been observed among specific DA classes such as (*qc* and *qy,qw^d*). This might be the case since qc utterances are frequently mistaken with qy, ^d, and qw ^d tags since they carry similar linguistic and syntactic characteristics. Although these question types differ in their intent as convey follow-up questions or lack of understanding of specific previous context with part of context repeated, whereas qw ^d and qy ^d are seen out of context settings but they all ultimately seek some sort of information from other partners. It is, therefore, expected that adding acoustic features may help to some extent to improve the performance.

There were also some forms of signal non-understanding and back-channels that were either differently transcribed in both corpora or missing from SwDA, used for training. Here, It is aimed to add a few samples of conversations from Carolina's dataset containing those surface forms that are absent from the SwDA corpus to check how it affects the performance of these rare classes. The idea of extending further experiments is not to improve the overall accuracy of our DA tagger, instead is to enhance accuracy against these rare classes.

The study in Chapter 4 will be extended in-depth to address the following research questions:

Q 1: What kind of acoustic features have been used in literature that helped in the DA tagging task?

Q 2: Which DA classes were performing better with the inclusion of acoustic features at the utterance level?

Q 3: Do contextualize pre-trained embedding more helpful in DA recognition tasks than static pre-trained embeddings?

Q 4: How well fine-tuning the BERT Model for the task of DA recognition improve the results over using pre-trained embeddings ?

Q 5: Does building a conversational DA tagger that takes the full conversation in context better than a DA tagger considering limited context length?

Q 6: Does adding a CRF layer to capture the contextual correlations between DAs help in predicting rare classes in a better way?

5.2 Background

Acoustic features Surendran and Levow (2006) used support vector machines and Hidden Markov models for dialogue act tagging in Map Task corpus by using both text

and acoustic features. They achieved an accuracy of 42.5% with acoustic only, 59.1% with text features, and an improved accuracy of 65.5% with both features. They also got a recall score of 0.01 with acoustic, 0.07 for text, and 0.14 with both feature sets for clarifying utterances. For the reply-w tag, they got a recall score of 0.02 with acoustic, 0.33 with text, and 0.38 with both features. They have shown that combination of acoustic and text features has improved the classification accuracy and recognition of certain classes. Arsikere et al. (2016) used a set of 57 acoustic features for automatic dialogue act tagging on British Call Center (BCC) corpora and SwDA corpus and gained a performance of 79.5%. Pitch, voicing, duration, pausing, intensity, and speaking pace are included in the feature set. They showed that there are significantly more features in the proposed feature set (15/20 features for the BCC corpus and 14/20 features for Switchboard), proving their importance to the performance of the combined feature set. In another study on dialogue act tagging, Ortega and Vu (2017) have utilized both lexical and acoustic features by proposing a lexico-acoustic model (LAM). the acoustic model utilized the 13 *Mel-frequency-spectral coefficients* (MFCC) extracted through the openSMILE toolkit. On thorough analysis, they revealed that acoustic features are useful in three situations: when a dialog act has enough data, when there are no strong lexical cues, and when lexical information is scarce. They obtained an accuracy of 73.6% with lexical, 50.9% with acoustic and 75.1% with the lexico-acoustic model.

There has also been work on dialogue act tagging that uses both lexical and acoustic information from the context as well (Si et al., 2020a). Si et al. (2020a) showed an accuracy of 85.4% with both lexical and acoustic features with a context length of five previous utterances. However, they only discussed those classes that have a proportion greater than 1% in the corpus and suggested that various context lengths are appropriate for different DAs. They only reported a low accuracy on questions with statement format/tune (*dq*) which is possibly due to a low portion reflecting the issue of imbalanced data in deep learning.

The literature discussed so far on dialogue tagging has focused on: improving the overall accuracy over all classes; have utilized all classes; or focused on the most frequent classes. Only a few people have discussed class-wise accuracy of classes including less frequent classes ((Raheja and Tetreault, 2019; Si et al., 2020a; Surendran and Levow, 2006). Duran et al. (2021) reported the F1 score of 23.94 for *qy^d* and an F1 score of 0.0 for *qw^d* with a textCNN model. Building a dialogue act tagger focusing on rare and less frequent classes is a quite challenging task. As certain classes such as *qy^d*, *qw^d*, *qc*, *qw* and *qo* share the same syntactic and linguistic characters and makes it difficult for a model to predict the right tag for the instances with these classes.

Conversational level Modelling for DA tagging Most state-of-the-art DA tagging models used conversational level encoder to capture the context from the utterance representations obtained from the sentence encoders. They used either recurrent neural

network (RNN) (Tran et al., 2017) or BiLSTM ((Si et al., 2020b; Liu et al., 2017). Bidirectional gated recurrent unit used as conversational encoders along with CRF to capture DA sequence dependency by several researchers (Li et al., 2018; Raheja and Tetreault, 2019). Others used the CRF layer on top of Bi-LSTM as a conversation level encoder ((Kumar et al., 2018; Srivastava et al., 2019); CRFs have shown success in a range of sequence-labelling tasks in NLP (see e.g. (Lafferty et al., 2001)).

Pre-trained Language Modelling Bidirectional Encoder Representations from Transformers (BERT) enabled several NLP tasks to outperform competing models. Chakravarty et al. (2019) used a CNN model, an LSTM model with attention, and a BERT model for recognizing dialogues including specific categories of questions and answers on Question-Answers corpora. The BERT model well performed over the LSTM and CNN model with an F1 score of 0.84, while with CNN, it was 0.57 and with LSTM, the F1 score is 0.71. The least frequent classes such as open-ended questions (*qo*:1.01%), choice questions (*or*: 0.73%), statement answers opinions (*so*: 1.01%), declarative wh-questions (*wh-d*: 2.3%) remain undetected with 0.0 recall for CNN model, while with LSTM the recall for *so* is 0.25 and *wh-d* is 0.14 (*qo, or* still remain 0.0). The recall values are improved with the BERT model with 0.56 for *wh-d* and 1.0 for *so* tag. Joukhadar et al. (2021) explored the impact of using BERT for the task of DA identification for the Arabic language. They reported the results in both macro average F1 and weighted micro F1 scores on an Arabian dataset that is very imbalanced in terms of DA's. With BERT Base, the reported macro average F1 score of 0.55 and 0.89 weighted average F1 score. With BERT Large, the macro F1 score is 0.52 while a 0.88 weighted average score is reported.

In a recent study, Noble and Maraev (2021), investigated how well BERT represents the utterances within the dialogue and how well large-scale pretraining and fine-tuning can help in DA recognition tasks. They concluded that BERT fine Tuning (FT) has more accurate results in terms of macro average F1 score with an explanation that at tag level, pretraining has a huge impact on least frequent classes. They also got very poor performance with the pre-trained BERT model on Both Corpora (SWDA and AMI meeting corpus). They achieved a macro F1 score of 47.78 on SwDA and 48.86 with AMI with BERT fine-tuned model. With the pre-trained BERT model, they got a macro F1 score of 7.75 on SwDA and 14.86 on AMI dataset. They stated that Without task-specific fine-tuning, the representations learned through pre-training are simply not performant, indicating a fundamental lack of knowledge relevant to the dialogue context. This contrast with other work (not particularly dialogical in nature), getting much better results with Frozen BERT performing better on fine-tuned BERT model.

Here, then, as stated before, our purpose is to improve DA tagging accuracy for the specific DA classes of interest in AD diagnosis, including specific types of questions, answers and misunderstanding signals, most of which are relatively rare. This chapter will specifically focus on improving the performance for certain rare classes by utilizing

additional features such as acoustic features, trying contextualized BERT embedding to get a better representation of the utterances, and fine-tuning the pre-trained BERT model for the task of DA recognition. Conditional random field (CRF) will be explored to capture the dependency between predicted dialogue acts.

5.3 Proposed Approach

In interactive dialogue systems, dialogue act recognition plays an important role, and research in this area had made great progress over the span of the past few years. Here, following set of experiments are performed to improve the accuracy of certain rare classes that could be helpful later in our downstream task.i.e AD classification.

1. A hierarchical BiLSTM-LSTM model with lexical and acoustic features with some previous contextual information.
2. A feature-based pre-trained language model such as BERT (FB-PRE-BERT) will be used as a sentence encoder.
3. A pre-trained, fine-tuned BERT model (FT-PRE-BERT) for the task of dialogue act tagging.
4. A hierarchical conversational DA tagger to capture longer contextual dependency within the conversation
5. A conversational DA tagger is developed using a CRF to model the sequence of DA tags, following an approach used successfully in general DA tagging work (e.g. (Kumar et al., 2018; Srivastava et al., 2019)).

With FT-PRE-BERT, a simple classification layer is added to the pre-trained BERT model, and all parameters are jointly fine-tuned on the dialogue act tagging task. However, the feature-based approach (FB-PRE-BERT), where fixed features per token basis are extracted from the pre-trained model, is passed to BiLSTM before the classification. These features can then be used as input for the classification of DA's.

5.3.1 Building Bi-modal Hierarchical DA tagger with both lexical and acoustic features

The existing DA tagger built-in chapter 4 is considering lexical aspects of utterance combined with a few preceding utterances as context. However, lexical information is not enough as prosody can also help in identifying speaker intention, and many studies e.g. (see (Ortega and Vu, 2018)) have considered both aspects of conversation in DA tagging. In this section, lexical aspects of utterance are combined with acoustic features

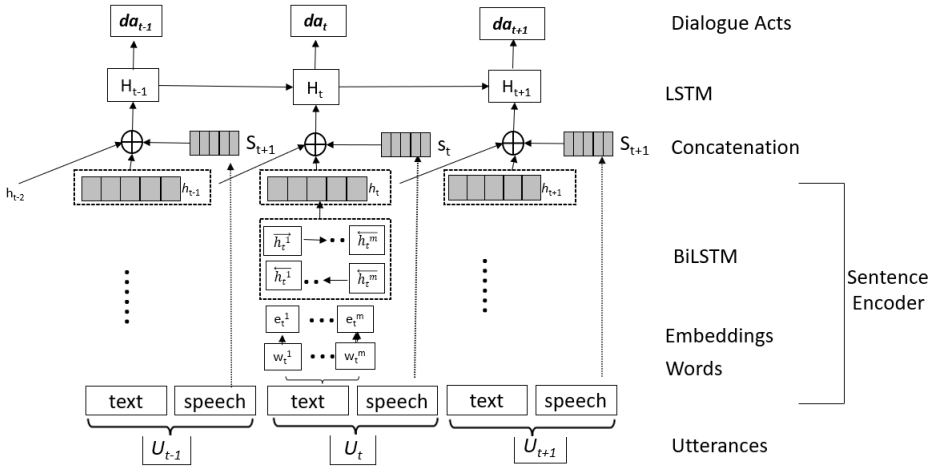


Figure 5.1: Architecture of lexical-acoustic model.

extracted from the previous section 5.3.4. The architecture of the lexical-acoustic DA tagger proposed in this section is depicted in Figure 5.1. The sentence encoder is the key factor of the sentence encoding process and here, two types of sentence encoder are used: the model that is trained in a supervised manner as used in DA classification research and a pre-trained language model (e.g BERT) to generate the sentence encoding at utterance level. For the supervised sentence encoder, to use the lexical information, the text input of each utterance is tokenized and with zero padding represented as a sequence of tokens. Each token is processed in the word embedding layer and transformed into a 100-dimensional word vector representation. A pre-trained Glove (Pennington et al., 2014) model is used here to extract lexical information. These word vectors are then processed by a BiLSTM layer and generate a fixed-size vector that represents the whole utterance and encapsulates the encoding for the entire utterance. To generate sentence encoding from pre-trained language models, Bidirectional Encoder Representations from Transformers is utilized(Devlin et al., 2018) and used the BERT base case version with 768 hidden units and 12 transformer layers and self-attention heads. The encoded utterance conceptually represents the context-agnostic features of the entire utterance. An additional feature i.e change-in-speaker information is concatenated to sequence representations

This sentence encoding vector representation of utterance is then concatenated with preceding context utterances (i.e. lexical information) and the acoustic feature vector of the current utterance. The next layer incorporates the contextual data along with current sentence encoding to be used in the DA classification process. These concatenated vectors of utterance representation are then encoded by a second LSTM (here, LSTM for a unidirectional left-to-right context instead of BiLSTM) is used. A final dense layer is used with softmax that outputs a probability distribution over the rare class DA set

given the current utterance, the final prediction is considered the DA label with the highest probability.

5.3.2 Building a Hierarchical Conversational level DA tagger with longer context

Figure 5.2 shows the overall architecture of our proposed model. The input to the model is different conversations where each conversation C^i consists of a sequence of utterances $U_{t-1}, U_t, U_{t+1}, ...U_{t+m}$. Each word w_k from each utterance U_j is processed by an embedding layer which converts it to a dense vector representation, followed by a bi-directional LSTM which serves as the first encoder of the hierarchical model. The Bi-directional LSTM will produce the utterance representation (h_t) of the utterance U_t by combining the representation of its constituent words. The representation from the last time step was extracted from the sentence encoder since it encompasses the context of all preceding words and previous time steps. The Bi-LSTM layer is employed with a subsequent time-distributed connected layer. This is to process the utterances in each conversation and generate vectors on per utterances basis for each conversation.

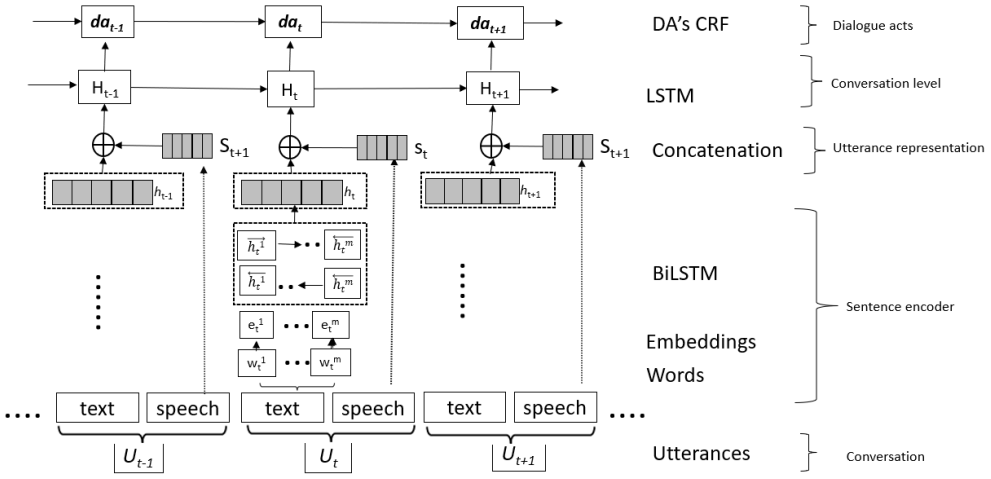


Figure 5.2: An illustration of our hierarchical conversational level DA for rare class tagging.

At this stage, we have the sequence of utterance representation $U_t, U_{t+1}..., U_{t+m}$, each utterance U_t is concatenated with its corresponding speech features S_t. This combined utterance representation will be the input of the next conversation layer which is realized by means of a LSTM layer. The input to the conversation level is the sequence of combined utterance representation $H_{t-1}, H_t, ..., H_{t+m}$ where m is the length of context or number of utterances in a conversation that will be processed at a time. Here, a unidirectional left-to-right LSTM is used, rather than bidirectional, to stay compatible with utterance-by-utterance online processing. The dependency between utterances is well captured

through the conversation-level LSTM encoder while a linear chain CRF can capture the dependency among the DA's. In order to jointly decode the best chain of tags in sequence tagging, it is preferable to look at correlations between DA's in neighborhood tags rather than greedily looking at the tag at each time step. In our experiment, a linear CRF [1] layer is used to obtain the optimal sequence of DA's, while taking into account the contextual correlations between the DA's in a conversation. Previously, CRF has been used in many studies with promising results (Kumar et al., 2018; Si et al., 2020b; Raheja and Tetreault, 2019). A DA-sequence with a specific context length is the model's output.

Unlike the model in section 5.3.1, which uses a 1D array of samples weights for each sample within the training set, here our weight sample is a 2D array because of the temporal nature of data, and to apply a different weight to every time step of every sample within each conversation.

5.3.3 Feature Set

Our feature set comprises three main parts:

1) **Lexical features**: we previously used randomly initialized word embeddings, GloVe embeddings and ELMo embeddings. Here. we are using GloVe pre-trained embedding and BERT pre-trained model representations for each utterance.

2) **Acoustic features** that are extracted are prosody based features ($\mathscr{F}$), energy based features ($\mathscr{E}$) and duration based features ($\mathscr{D}$). These are discussed in detail in section 5.3.4.

3) **Additional features** are speaker identity and speaker change. Speaker change was represented using binary values, depending on whether the speaker of the current utterance remained the same or changed. Additionally, a similarity vector was employed to calculate the similarity between the current utterance and its neighboring utterance through the Cosine Similarity function [2]. The idea is that adding a similarity vector may help the model to be less confused among classes such as *qc* and declarative questions such as *qy^d* and *qw^d*.

5.3.4 Extraction of Acoustic feature

SwDA corpus audios were used to extract the acoustic features. A subset of the SwDA corpus is used for extracting audio features. First of all, the SwDA audios are converted from *.sph* into *.wav* using Sound eXchange utility (*SoX*) and then, left and right channels were separated for both speakers *A* and *B*. The original SwDA corpus provides utterance

[1] Linear chain CRF implementation from tensorflow: https://www.tensorflow.org/addons/api_docs/python/tfa/layers/CRF.

[2] Cosine similarity function from keras tensorflow is used:https://www.tensorflow.org/api_docs/python/tf/keras/losses/CosineSimilarity

transcripts and DA annotations but it does not contain timestamps at the word level that were useful to extract acoustic features. Timestamps are obtained at the word level from the deep-learning-driven model of incremental detection of disfluency developed by Hough and Schlangen (2017)[3]. Timings information for each utterance is calculated based on the start and end timings of the first and last word of each utterance. OpenSMILE V3.0 (Eyben et al., 2010) was used to extract acoustic features from the audio recordings and the audio features are sampled at 100Hz. Every sample is partitioned into frames of a 10ms window with a shift of 10 ms.

Prosody includes many features that could be computed automatically without word reference information.An attempt has been made to achieve comprehensive coverage of prosodic features, obtained from the study by Shriberg et al. (1998), which plays a crucial role in distinguishing questions from statements, especially declarative and yes-no questions. These are prosodic features based on fundamental frequency (F_0) (e.g raw F_0, final F_0, mean F_0), Root mean square energy (RMS)(e.g mean RMS) and duration (e.g duration of utterance, number of frames in utterance and several frames with f_0). Initially, the F_0 values contain both raw values at the frame level, and then these features are calculated for the whole utterance based on the start and end time of the utterance. Mean f_0 was calculated over all the voiced frames in an utterance to represent the pitch range. Conversation side f_0 mean values are also computed for both speakers and utterance level values are normalized over the conversation side mean values for respective speakers. A minimum f_0 (min_F0) value is also calculated to measure "*good*" F_0 values that are values above this min_F0 threshold. Based on the idea used by Shriberg et al. (1998), min_F0 is also computed and that is a ' good estimate of f0 min is to take the point at 0.75 times the F0 value at the mode of the histogram'.

Additional features are extracted from the last 200 ms region frames called '*end region*' and previous 200 ms region frames from the *end region* called '*penultimate region*'. The contours (rising/falling) in these regions could be indicative of utterance types. A standard zero mean and variance normalization was applied to the acoustic features.

These features are categorized into three categories and a detailed list is given in Table 5.1.

- F0 features ($\mathscr{F}$)
- Duration features ($\mathscr{D}$)
- Energy based features ($\mathscr{E}$)

[3]The timestamp-based SWDA data at the word level is available at: https://github.com/clp-research/deep_disfluency

No	Feature Name	Description
		F0 features ($\mathscr{F}$)
1.	*f0_mean_convside*	Mean *f0* value of the whole conversation side (each speaker)
2.	*f0_max_convside*	Conversation side *f0* maximum value
3.	*f0_max_utt*	Utterance maximum *f0*
4.	*f0_max_n*	Log ratio of values *f0_max_utt* and *f0_max_convside*
5.	*fo_min_utt*	Utterance minimum f0 value that can be below convside *f0_min* (*f0_min* = 0.75* *f0_mode*)
6.	*f0_mean_utt*	Mean of all *f0* values included in utterance
7.	*fo_mean_good_utt*	Mean of *f0* values included in *f0_no_good_utt*
8.	*f0_mean_ratio*	Ratio of utterance mean (*f0_mean_utt*) and convside mean (*F0_mean_convside*)
9.	*f0_mean_n*	Difference between *f0_mean_utt* of utterance and *f0_mean_convside* for *f0* >*f0_min*
10.	*f0_std_utt*	Standard deviation of *F0* values in good utterance
11.	*f0_std_n*	Log ratio of standard deviation of *F0* values in utterance and in convside
12.	*f0_mean_zcv*	means of *f0* values in good utterance normalized by the mean and standard deviation of *f0* in convside
13.	*end_region_f0_mean*	Mean of *f0* values in end region (last 200ms in utterance)
14.	*pen_region_f0_mean*	Mean of *f0* values in penultimate region
15.	*norm_end_f0_mean*	End region *f0* mean normalized by mean and standard deviation from convside f0 value
16.	*norm_pen_f0_mean*	End region *f0* mean normalized by mean and standard deviation from convside f0 value
17.	*abs_f0_diff*	Difference between *F0* mean of end region and penultimate region
18.	*rel_f0_diff*	Ratio of *F0* of end and penultimate end region
19.	*utt_grad*	all points' least squares regression line over utterance
20.	*end_grad*	all points' least squares regression over end region
21.	*pen_grad*	all points' least squares regression over pen region
22.	*reg_start_f0*	First *f0* value of contour, determined by regression line analysis

Table 5.1: F0 feature list

5.3.4.1 F0 features

The list of features based on f_0 is listed in table 5.1. Feature $\mathscr{F}1$ represents the mean f0 value over the whole conversation for each speaker called 'convside' on-wards. This measure is used to normalize the utterance level measures to normalize the differences in the f0 range over speakers. $\mathscr{F}2$ represents the maximum f0 value over convside for speakers. $\mathscr{F}3$ and $\mathscr{F}5$ are maximum and minimum values of an utterance. Utterance mean and mean of good utterance are expressed as $\mathscr{F}6-\mathscr{F}7$. A good utterance count contains several f0 values that are above *f0_min*. Features $\mathscr{F}13-\mathscr{F}16$ represent the mean and normalized mean of *end region* and *penultimate region*. $\mathscr{F}17$ is absolute difference between *end region* and *penultimate region* while $\mathscr{F}18$ is the ratio of f0 between two regions. The least square fit regression line of the f0 contour was computed. However, this may not accurately represent the rising or falling at the end, potentially indicating an overall gradient as falling. To address this, gradients for both the *end region* and *penultimate region* were also calculated.

5.3.4.2 Energy features

Energy features are computed based on standard RMS energy and are listed in table 5.2.

No.	Feature Name	Description
		Energy features ($\mathscr{E}$)
1.	*utt_nrg_mean*	average of *RMS* energy values in utterance
2.	*end_region_nrg_mean*	Mean of *RMS* energy values in end region
3.	*pen_region_nrg_mean*	average of all *RMS* energy values in penultimate region
4.	*norm_end_nrg_mean*	*end_region_nrg_mean* is normalized over *utt_nrg_mean*
5.	*norm_pen_nrg_mean*	*pen_region_nrg_mean* is normalized over *utt_nrg_mean*
6.	*abs_nrg_diff*	difference between the penultimate region's *mean RMS* energy and the end region's *mean RMS*
7.	*rel_nrg_diff*	ratio of mean *RMS* energy of end and penultimate regions

Table 5.2: Energy feature list

5.3.4.3 Duration based features

There are two types of duration features listed in table 5.3. Duration in seconds is expressed by $\mathscr{D}1$ and $\mathscr{D}2-\mathscr{D}4$ represent duration based on correlation with f0 frame counts.

No.	Feature Name	Description
	Duration features ($\mathscr{D}$)	
1.	*utt_dur*	Duration of utterance in seconds
2.	*f0_num_utt*	Count of *f0* frames in utterance
3.	*f0_num_good_utt*	Count of *f0* frames above *f0_min*
4.	*reg_num_frames*	Regression line for the entire utterance was calculated using the number of *f0* frames of contour, ignoring voiceless frames

Table 5.3: Duration based feature list

Dataset	SwDA	CCC
Conversations	583	30
tagset	20	20
Total utterances	99522	7294
(train-test) split conversations	467-116	15-15
Training utterances	80644	3820
Test utterances	18878	3474

Table 5.4: Dataset statistics

5.4 Experiments

5.4.1 Data

The model is evaluated on two previously utilized datasets: 1) **SwDA**, a dialogue corpus featuring conversations between two speakers. For the SwDA dataset, we utilized the version annotated with our 20 rare-class DA tagset. However, our usage of data is constrained by the availability of suitable word-level time stamps, as detailed in section 5.3.4. Currently, experimentation is carried out on approximately half of the conversations from the original standard dataset. 2) **CCC** corpus, a more conversational dataset that contains both transcripts and audio about the health of people over 65 years of age in natural conversations. Here, the dataset is expanded by looking at a larger sample size of 30 patients (15 AD vs 15 Non-AD)[4]. Statistic summary of train and test splits are shown in table 5.4. In both corpora, the classes are highly imbalanced; the majority class is 55.1% while least frequent is 0.04% on SWDA and 38% majority with 0.4% least frequent class on CCC.

[4]In chapter 4, 20 patient's conversations from CCC corpus are used. Due to different sample sizes, results are not directly comparable.

5.4.2 Setup and hyperparameters

The hyperparameters used for our models for both corpora are summarized in table 5.5. For the embedding layer, a 100-dimensional vector size is used to represent each word. For the LSTM layer, the hidden unit size is set to 128. The model is trained using ADAM optimizer (Kingma and Ba, 2014) with a learning rate of 0.01. The training was done with 30 epochs and an early stopping mechanism with the patience of 3 is used. As the classes in our dataset are highly unbalanced, a class-weighted objective function is employed that over-prioritizing more common classes and gives more weights to least common (rare) classes. We particularly used sample weights comprised of two weight dictionaries in our experiments; one weight dictionary for samples of the SwDA dataset and another weight dictionary for the few samples of the CCC dataset. The idea of using two weight dictionaries is to give more weights to the few samples of the CCC dataset that are used in training. The context length n is set up to 3 utterances for the BiLSTM-LSTM model and up to 20 for the conversational DA tagger model. BERT model with an embedding dimension of 768 and 12 hidden layers from the HuggingFace sentence-transformers library[5] is used. For fine-tuning the pre-trained BERT model (FT-PRE-BERT), The trainable parameter are set to true to update the weights for the transformer layers during training.

Due to the imbalanced nature of detests, the scikit-learn's StratifiedShuffleSplit (a merge of StratifiedKFold and ShuffleSplit) was used previously to preserve the percentage of each class in each fold. However, there is a limitation of using StratifiedShuffleSplit and that is: Although the percentage of each class is preserved in each training and testing split making sure that the rare class samples are also part of both training and test split but shuffle split actually shuffles the utterance sequences that distorts the original sequence of utterances in a conversation. This may give poor results when considering a few previous utterances as context because these few utterances may not be the actual utterances in original conversations. Therefore, manual stratification is performed that also preserves the sequence of utterances and will be discussed in the next section 5.4.2.1.

Same metrics are selected as used in chapter 4 to report results for our extended models. The results are reported with accuracy, macro-average precision (*Prec.*), macro-average Recall (*Rec.*), and macro-average F1 as metrics for multi-class classification.

5.4.2.1 Manual stratification on imbalanced dataset

Data imbalance is a common problem with multi-class classification which if randomly split between train and test splits disregards the distribution/proportion of each class resulting in low accuracy on small classes. In particular, it's critical to correctly identify

[5]BERT Model: https://huggingface.co/bert-base-uncased

Model	Hyperparameters	Values
Bi-LSTM	Word embeddings	(GloVe dim. 100)
	LSTM hidden units	128
	Dropout	0.5
	Learning rate	0.01
	Context length (n)	[1-3]
	batch size	256
BERT base uncased	hidden units	768
	Learning rate	2e-5
	batch size	4,16
	context	[1-3]
Conversational Bi-LSTM	Embeddings	Glove (dim. 100)
	LSTM hidden units	128
	batch size	16
	context	[1-20]

Table 5.5: Hyperparameters

classes that are of great relevance yet only occur rarely, like cases of the disease. There have been a few techniques that were used for the imbalance data problem including random over-sampling and random under-sampling (Chawla et al., 2004). The random oversampling method adds the exact duplication of samples of minority/rare classes while under-sampling methods delete random samples of majority classes. However random over-sampling may cause overfitting and under-sampling may reduce some useful clues/information from the dataset. Over-sampling the minority class through Synthetic Minority Another oversampling method is the SMOTE method, which creates additional artificial instances along the line between the minority examples and their chosen nearest neighbors. (Chawla et al., 2002).

As our problem is to build a DA tagger, particularly for rare class DA tags of interest on an imbalanced dataset, which also considers that sequence/order of utterance in conversation is preserved, it is hard to use the above-mentioned techniques of oversampling. Our defined steps distribute the classes in training and testing sets based on conversations (rather than utterances). This approach ensures not only a fair distribution of the rare classes in both sets but also preserves the order of utterances. The steps used for this stratification are listed in Figure 5.3.

The distribution of tags with our manually stratified split method is compared with the sklearn's method *StratifiedShuffleSplit* (which also considers shuffling of utterances) and plotted to compare the difference between the counts of tags in both cases (see Figure 5.4). The y-axis shows the count of each tag class with *log base 10* to fit the larger values.The distribution of tags is also compared with a simple train test split when no stratification is used at all and the comparison is shown in Figure 5.5. The two

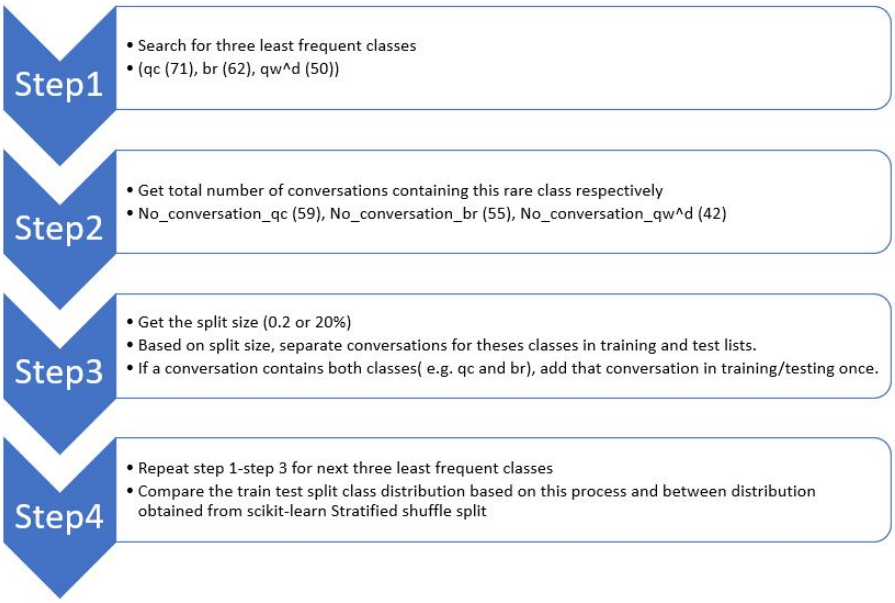


Figure 5.3: Process of stratification for balancing rare classes.

distributions seem quite reasonably similar in Figure 5.4 as compared to distributions in Figure 5.5.

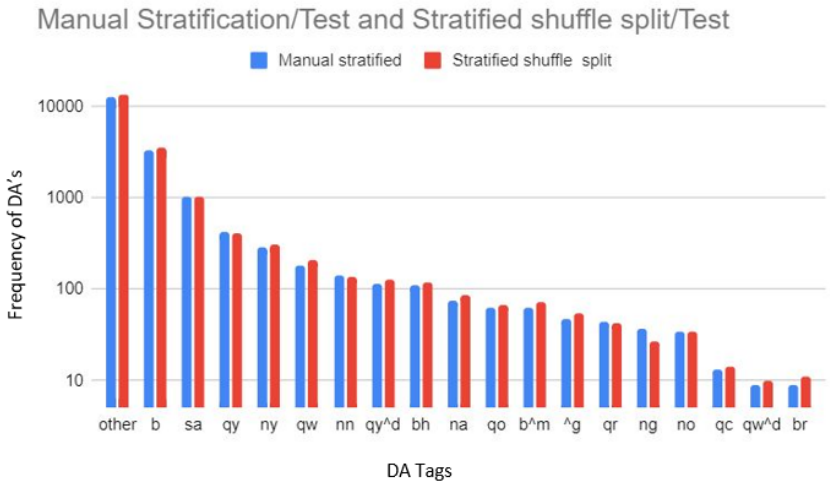


Figure 5.4: Comparison of distribution between manual stratification with stratified shuffle split of tag classes.

Another way to measure the similarity between two distributions is to calculate Jensen-Shannon Divergence (JSD) between the two probability distributions. Suppose P is the probability distribution of tag classes with our manually stratified method, Q is the probability distribution of tag classes with *stratifiedShuffleSplit* method, and $Q1$

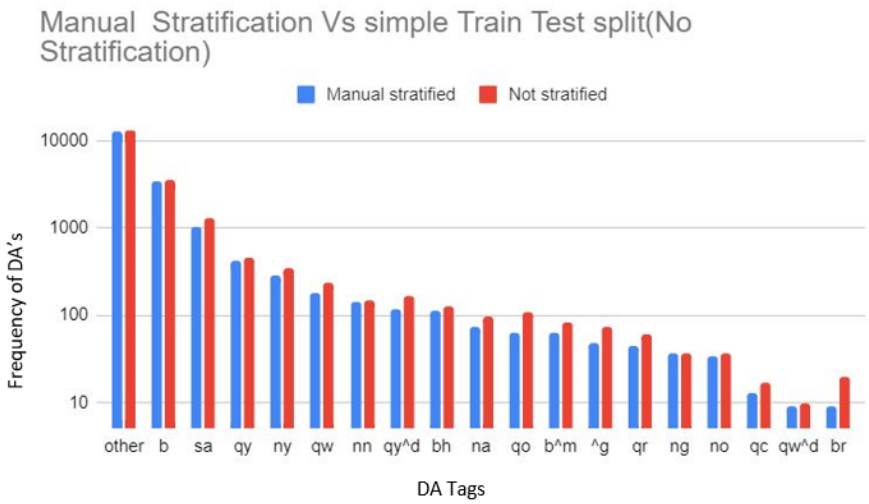


Figure 5.5: Comparison of distribution between manual stratification with a simple split of tag classes without stratification.

Distributions	JSD value
JSD(P,Q)	0.0102
JSD(P,Q1)	0.0302

Table 5.6: JSD values for Comparison

is the probability distribution of tag classes with simple train test method without any stratification.

The JS Divergence can be calculated as follows:

$$JSD(P \parallel Q) = \frac{1}{2}D(P \parallel M) + \frac{1}{2}(Q \parallel M) \tag{5.1}$$

where M can be calculated as:

$$M = \frac{1}{2}(P + Q)$$

which is a mix distribution. The JSD values for both distributions are shown in Table 5.6. The Jensen Shanon divergence between manually stratified and *stratifiedShuffleSplit* is closer to zero (i.e 0.0102) showing that these distributions are quite similar to each other as compared to when compared with simple train-test simple and without any stratification.

5.5 Results

Different experiments were performed to evaluate the different model. In the first subsection, experiments with a hierarchical BiLSTM-LSTM model containing lexical and

acoustic features are shown. The experiments with FB-PRE-BERT and FT-PRE-BERT on the SwDA and CCC corpora are then presented. In the last subsection, results on conversational DA tagger with longer context, with and without CRF are discussed.

5.5.1 Effectiveness of Hierarchical BiLSTM-LSTM with both lexical and acoustic features

Applying a Hierarchical BiLSTM-LSTM model to both lexical and acoustic features with additional features resulted in a macro F1 score of 0.58 on SwDA and 0.40 on CCC. This is much better then the 0.47 and 0.33 (SwDA and CCC) when using only 1 utterance in context (See table 5.7).

Model	Embedding	SwDA test set				CCC test set			
		Prec.	Rec.	F1	Acc.	Prec.	Rec.	F1	Acc.
Current utterance only	Glove	0.42	0.47	0.41	0.82	0.39	0.34	0.27	0.50
1 utt context	Glove	0.50	0.53	0.47	0.84	0.37	0.38	0.33	0.59
1 utt context + 1 DA	Glove	0.53	0.59	0.55	0.89	0.42	0.41	0.34	0.60
1 utt context+ 1 DA + acoustic	Glove	0.53	**0.60**	**0.56**	0.88	0.44	**0.44**	**0.39**	0.61
1 utt context+ 1 DA + acoustic + similarity	Glove	0.55	**0.62**	**0.58**	0.88	0.45	**0.44**	**0.40**	0.62
2 utt context	Glove	0.45	0.50	0.45	0.84	0.31	0.36	0.30	0.54
2 utt context+ acoustic	Glove								

Table 5.7: Accuracy, macro-average precision, recall, and F1 score for different contexts with Glove word embeddings on **SwDA test set** and **CCC test set**. Models without acoustic features are the same models as those used in Chapter 4.

Class-wise results Our interest, of course, is not in overall accuracy figures but in predicting the distribution over the individual rare DA classes. Therefore, class-wise prediction scores are examined for these DAs label assigned to utterances [6] and reported in table 5.8 and table 5.9. The model achieves an accuracy of *br* tag to 0.67 with word embeddings, 1 previous DA, and acoustic features on the SwDA test set (previously br = 0.68 on SwDA and br =0.75 on CCC) and 0.81 on the CCC test set. This is further improved to 0.83 with the inclusion of a similarity vector.

No.	Model	*br*	*qc*	*qw*	*qo*	*qw^d*	*qy^d*	*ny*	*na*	*sa*	*no*	*ng*
1.	1 utt context	0.56	0.23	0.72	0.55	0.10	0.22	0.70	0.43	0.28	0.26	0.05
2.	1 utt context + 1 DA	0.61	0.23	0.65	0.61	0.10	0.11	0.88	0.46	0.90	0.26	0.32
3.	1 utt context+ 1 DA + acoustic	**0.67**	0.15	0.65	**0.65**	**0.22**	**0.22**	0.88	0.35	**0.93**	0.26	**0.41**
4.	1 utt context+ 1 DA + acoustic + similarity	0.56	0.15	0.71	**0.73**	**0.22**	**0.22**	0.89	0.33	**0.91**	0.22	0.35

Table 5.8: Accuracy of some of the rare classes on the SwDA test set.

The accuracy for qc is lower than our previous achieved results in chapter 4 (qc =0.29 on SwDA and qc =0.20 on CCC), the model achieves an accuracy of 0.15 with 1

[6] It is worth mentioning that the train set used in chapter 4 is larger than the train set used in this chapter, so the results of class-wise accuracies are not directly comparable, however, an idea of improvements is obtained.

previous DA and acoustic features on CCC data which is further improved to 0.19 with the inclusion of similarity vector[7]. The model achieves the note-able performance for the *qw^d* tag as it gets 0.22 on SwDA with lexical features, 1 previous DA and acoustic features, and an accuracy of 0.09 on the CCC test split. Previously, the accuracy for *qw^d* tag on the CCC test set was zero. By inclusion of acoustic and similarity vectors helped the model to predict some of their instances with *qw^d*. Previously, Duran et al. (2021) have reported class-wise accuracy of the three most frequent and question types classes and they reported an accuracy of 0.0% for *qw^d* on the SwDA test set.

No.	Model	*br*	*qc*	*qw*	*qo*	*qw^d*	*qy^d*	*ny*	*na*	*sa*	*no*	*ng*
1.	1 utt context	0.83	0.07	0.54	0.07	0.0	0.11	0.67	0.11	0.16	0.07	0.0
2.	1 utt context + 1 DA	0.77	0.09	0.48	0.19	0.0	0.16	0.69	0.22	0.62	0.13	0.11
3.	1 utt context+ 1 DA + acoustic	**0.81**	**0.15**	0.51	**0.26**	**0.09**	0.16	0.81	0.16	0.63	0.15	0.29
4.	1 utt context+ 1 DA + acoustic + similarity	**0.83**	**0.19**	0.54	0.31	**0.09**	**0.17**	0.74	0.16	0.62	0.13	0.25

Table 5.9: Accuracy of some of the rare classes on the CCC test set.

5.5.1.1 Analysis

An analysis is conducted on a few sample examples from the CCC corpus with actual labels and predicted labels for different models and compare the performance in table 5.10. The tags selected are those for which poor performance was observed in Chapter 4. The example 1 with the underlying text 'What do i do in the morning ? ' and its previous utterance text 'so, what's your , what do you usually do in the morning ? ' is providing sufficient information to model [1] predict the *qc* tag accurately. Model [2], [3], [4] also predicts the *qc* tag accurately, however, this example clearly shows that the previous utterance context was sufficient for the model to make prediction. Example 2 with the utterance text 'they're all living?' is misclassified as a declarative yes-no question for model 2 and 3. Adding acoustic features correctly classified this utterance as qc.

Example 5 with the underlying text ' to georgia? ' was predicted with the right *qc* tag when the model is considering previous DA and when the similarity vector is computed from the previous utterance. With the addition of acoustic features model 3 better performs for *qo* class, as in examples 7-9, all three utterances are predicted with *qo* tag while classified with *qy* and *qw* with model [1] and [2]. Declarative questions are also predicted with the right tag with the inclusion of acoustic features (see examples 10-12). The addition of 1 previous DA and acoustic feature also helped in improving the accuracy of negative non-no answers (*ng*) in examples 13-15. Adding a similarity vector helps in better prediction for qc and very few instances of *qw^d* (see examples 6 and 16). It is assumed that different features are appropriate for different DA classes. Adding acoustic features has improved the performance for the question categories such as the declarative yes-no question, open-ended question, and wh-question. Previous DA history

[7]With the smaller train set, the accuracy for qc tag is still comparable (0.20 vs 0.19 on CCC test set.)

No #	Utterance	Actual	Predicted			
			[1]	[2]	[3]	[4]
1.	what do i do in the morning ?	qc	qc	qc	qc	qc
2.	they're all living ? yeah .	qc	qy^d	qy^d	qc	qc
3.	she does ?	qc	qc	qy	qc	qc
4.	gallstones ?	qc	qy	qy	qy	bh
5.	to georgia?	qc	qy	qc	qy	qc
6.	about my blood pressure ?	qc	br	qy^d	ng	qc
7.	what do you think she is laying down there for ?	qo	qw	qo	qo	qo
8.	why do you think you um , had high cholesterol ?	qo	qw	qy	qo	qo
9.	why you so scared like that ?	qo	qy	qy	qo	qo
10.	it just disappeared today ?	qy^d	qy	qy	qy^d	qy
11.	my teeth won't break on it ?	qy^d	^g	qy	qy^d	qy^d
12.	they all live in charlotte ?	qy^d	qy	qy	qy^d	qy
13.	no , i haven't seen any changes .	ng	other	ng	ng	ng
14.	no , i , no , i made , no , i got married , uh , -june twenty-first.	ng	sa	nn	ng	ng
15.	no , no i guess like opening in it .	ng	other	ng	ng	ng
16.	what that bird's thinking about .	qw^d	other	other	other	qw^d

Table 5.10: E
xample of utterances of confused pair and few more from **CCC**.[1] represents the model with one previous utterance context. [2] represents the model with 1 previous utterance and 1 previous DA, [3] represents model with all previous features and acoustic features, and [4] represents the model with additional features of similarity.

along with acoustic and similarity vectors have improved prediction for clarification requests and for a few cases of declarative wh-questions. The misclassification for the classes such as *qy^d*, *qc*, *qw^d* results due to the lack of training data and very few distinguishing words for the classifier to make an accurate judgment. More training data for these classes would help in increasing performance.

5.5.2 Effectiveness of BERT model as sentence encoder

Additionally, in light of recent successes in the use of large pre-trained language models for transfer learning on a range of NLP tasks and in the task of DA classification, such models are tested as sentence encoder and presented results in the table 5.11.BERT pre-trained model s used to obtain the utterance representation and compared it with fine-tuning the BERT model with our DA classification. A dense layer is simply used on top of the BERT model with fine-tuning when only the current utterance or 1 previous utterance is used. For longer context (more than 1 utterance), an LSTM layer is used

over the learned utterance representations to capture the context from BERT and then finally a dense layer with softmax is applied to get the final predictions.

Table 5.11 shows the results achieved by the different models on the two corpora. It is noteworthy that all models have demonstrated favorable classification results, especially on the CCC dataset. Model 1-7 shows the results of the BERT model fine-tuned for the task of rare class dialogue act recognition. BERT base models outperform the rest with the highest macro average F1 score of 0.48 and accuracy of 66% with 1 previous utterance and 1 previous DA as context on the CCC dataset. All these models outperform the BiLSTM models with Glove embeddings (see Figure 5.6). Previously, the highest macro average F1 score of 0.40 was obtained on the CCC dataset with the BiLSTM model with 1 previous utterance, 1 previous DA, acoustic features, and a similarity vector.

No.	Model	SwDA test set				CCC test set			
		Prec.	**Rec.**	**F1**	**Acc.**	**Prec.**	**Rec.**	**F1**	**Acc.**
0	Hierarchical-BiLSTM-LSTM (Table 5.7)	0.55	0.62	0.58	0.88	0.45	0.44	0.40	0.62
1.	FT-PRE-BERT-1 current utt	0.48	0.53	0.49	0.83	0.43	0.39	0.36	0.54
2.	FT-PRE-BERT-1 utt context	0.53	0.62	0.56	0.84	0.51	0.50	**0.46**	**0.64**
3.	FT-PRE-BERT-1 utt context+ 1 DA	0.51	0.60	0.55	0.84	0.56	0.48	**0.48**	**0.66**
4.	FT-PRE-BERT-1 utt context+ acoustic	0.55	0.57	0.55	0.83	0.47	0.44	0.41	0.60
5.	FT-PRE-BERT-1 utt context+ 1 DA + similarity	0.56	0.59	0.56	0.85	0.54	0.46	0.44	0.63
6.	FT-PRE-BERT-2-utt context	0.55	0.60	0.56	0.86	0.48	0.49	0.44	0.65
7.	FT-PRE-BERT-3-utt context	0.56	0.60	0.57	0.87	0.49	0.50	**0.46**	**0.65**
8.	FB-PRE-BERT 1 utt context	0.10	0.14	0.11	0.78	0.07	0.13	0.08	0.51
9.	FB-PRE-BERT 1 utt context +1 DA	0.15	0.20	0.16	0.78	0.09	0.16	0.11	0.50

Table 5.11: Comparison between different BERT-Based models on **SwDA** and **CCC**, in terms of accuracy, macro average precision, recall, and F1-measure.

On the other hand, the feature-based BERT model with pre-trained embeddings did not perform very well. The low macro average recall and F1 score in table 5.11 shows that models 8-9 are biased towards the most frequent classes. The model 8 and 9 are only able to identify classes such as ' *(b)*', 'other', '*yes-no (qy)*', and wh-questions. The classes of interest such as *qc,br* remain un-detected with 0.0 macro F1 score with these models.

Class-wise results To understand the results,the detailed classification report is examined for certain classes such as *signal non-understanding*, '*clarification request*', and a few question and answer tags. Table 5.12 shows the detailed classification report using the recall values for these classes.

With the FT-BERT model, the highest accuracy of 0.40 is obtained for *qc* tag with three utterance contexts, 0.38 with two utterances in context, and 0.22 with one utterance and one DA in context. This indicates that the BERT model with a longer context makes better predictions for clarification request class. In experiments with the BiLSTM model using Glove embedding, the model achieves an accuracy of 0.19 for qc class with 1 previous utterance, 1 DA, acoustic features with similarity vector as shown in Figure 5.7.

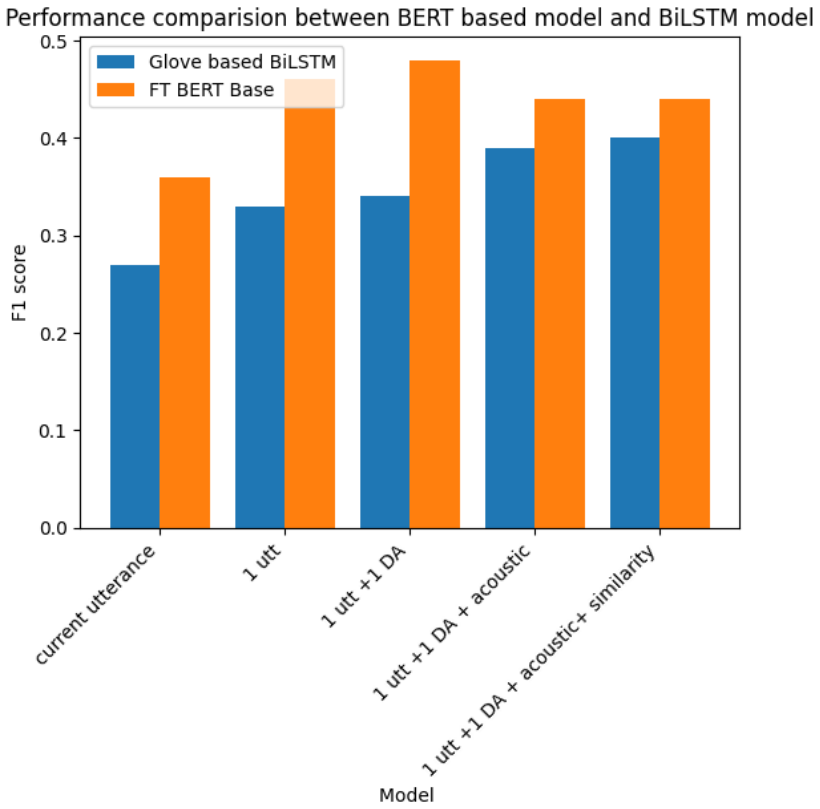


Figure 5.6: Comparison of *F1* score of difference between fine-tuned BERT-based and BiLSTM models with Glove embedding on **CCC** corpus.

Model	***br***	***qc***	***qw***	***qo***	***qw^d***	***qy^d***	***ny***	***na***	***sa***	***no***	***ng***
1 utt context	0.83	0.07	0.54	0.07	0.0	0.11	0.67	0.11	0.16	0.07	0.0
1 utt context + 1 DA	**0.89**	**0.22**	0.72	0.31	**0.13**	**0.33**	0.76	0.33	0.28	0.14	**0.32**
1 utt context + acoustic	**0.85**	0.13	0.54	0.41	0.09	0.24	0.64	0.14	0.19	0.20	0.21
1 utt context+ 1 DA + similarity	0.84	**0.28**	0.70	0.11	0.11	**0.36**	0.59	0.11	0.27	0.23	**0.26**
2 utt context	**0.85**	**0.40**	0.80	0.33	0.0	0.22	0.72	0.14	0.17	**0.33**	0.22
3 utt context	**0.85**	**0.38**	0.68	0.33	0.09	**0.43**	0.71	0.16	0.22	**0.53**	**0.25**
FB-BERT 1 utt context	0.00	0.00	0.11	0.0	0.0	0.00	0.00	0.0	0.00	0.00	0.00

Table 5.12: Accuracy of some of the rare classes on **CCC** test set with FT-PRE-BERT Model.

With BERT, 0.89 accuracy is achieved for the *signal non-understanding* class with one utterance and one dialogue act context. Longer contexts are also helpful resulting in an accuracy of 0.85 for *br* tag with 2 utterance and 3 utterance context. The accuracy of the '*no*' tag is greatly enhanced from 0.15 (from glove embedding) to 0.53 with the BERT model with three utterance contexts. In a similar line, the accuracy for declarative yes-no questions is also improved from 0.17 to 0.43 for BERT-based models.

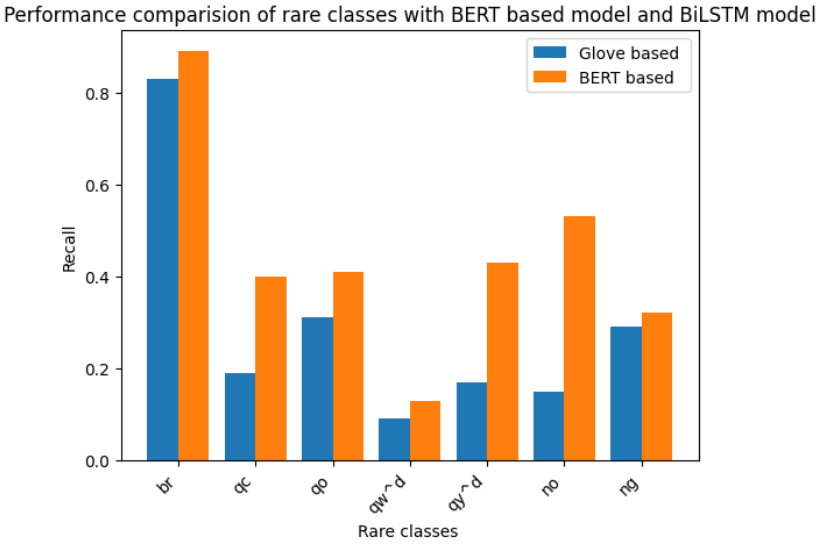


Figure 5.7: Comparison on the accuracy of rare classes for FT-PRE-BERT based and BiLSTM models with Glove embedding on **CCC** dataset.

5.5.3 Effectiveness of Conversational hierarchical BiLSTM-LSTM model with longer context

A conversational dialogue act tagger that takes into account the full conversation would use information from the entire conversation to make predictions about the speech acts in each turn. Here, instead of full conversation context, the context length is used, which is the number of utterances that will be processed in a conversation at a time.

Table 5.13 shows the results with our conversational level DA tagger. The first two rows show the results from table 5.7 to compare results with our conversational level dialogue act tagger. Here it can be seen that immediate previous utterance along with acoustic features and additional features gives the highest F1 score of 0.58 which is 0.4% higher than the highest macro F1 score obtained from the conversational rare class DA tagger. However, it is noticeable that the conversational tagger performs well with longer context (e.g up to 20 utterances) without considering any additional features such as previous DA and similarity vector with an F1 score of 0.54 with lexical features and 0.52 with both lexical and acoustic features.

Table 5.14 describes the results by adding a CRF layer on top of the conversational level DA tagger. This is very surprising that CRF does not help in improving the performance, in fact, it performs worse as compared to without using CRF. This is different from the other people's work who found that adding CRF is effective on standard DA tagging settings with more evenly distributed classes. This may be due to the fact that CRF does not suit very well for the task of rare class tagging where the tags are unevenly distributed. From the tagset of 20 classes, only five classes ('b', 'qy', 'qw', 'ny' and 'other')

Input	Context	SWDA			CCC		
		Acc.	F1.	Rec.	Acc.	F1.	Rec.
text	1	0.84	0.47	0.53	0.59	0.33	0.38
text+ speech+ additional features	1	0.88	0.58	0.62	0.62	0.40	0.44
text	3	0.84	0.45	0.45	0.55	0.30	0.33
text	5	0.85	0.48	0.49	0.56	0.34	0.35
text	8	0.86	0.52	0.52	0.54	0.34	0.35
text	10	0.87	0.51	0.52	0.55	0.31	0.35
text	15	0.87	0.52	0.53	0.56	0.33	0.36
text	20	0.88	**0.54**	**0.56**	0.55	**0.37**	**0.39**
text+speech	3	0.84	0.42	0.44	0.53	0.3	0.33
text+speech	5	0.84	0.43	0.45	0.56	0.31	0.34
text+speech	8	0.86	0.48	0.50	0.55	0.31	0.36
text+speech	10	0.86	0.50	0.53	0.56	**0.36**	**0.38**
text+speech	15	0.86	0.49	0.50	0.54	0.34	0.36
text+speech	20	0.87	**0.52**	**0.53**	0.56	**0.36**	**0.39**

Table 5.13: Conversational DA tagger performance with different context lengths on **SwDA** and **CCC**. The first two rows are results from the hierarchical BiLSTM model (see table 5.7) with the first row for 1 utterance only context and the second row for 1 utterance with acoustic features and additional features.

are frequent and the rest of them are less than 1% (see table 4.5).

Model	Input	Context	SWDA			CCC		
			Acc.	Rec.	F1.	Acc.	Rec.	F1.
Bi-LSTM-LSTM-CRF	text	3	0.85	0.40	0.43	0.53	0.31	0.29
Bi-LSTM-LSTM-CRF	text+speech	3	0.84	0.31	0.34	0.52	0.24	0.22
Bi-LSTM-LSTM-CRF	text	5	0.85	0.40	0.43	0.53	0.31	0.28
Bi-LSTM-LSTM-CRF	text+speech	5	0.85	0.30	0.35	0.52	0.23	0.22
Bi-LSTM-LSTM-CRF	text	8	0.86	0.42	0.46	0.53	0.32	0.28
Bi-LSTM-LSTM-CRF	text+speech	8	0.85	0.37	0.40	0.55	0.28	0.25
Bi-LSTM-LSTM-CRF	text	10	0.86	0.41	0.45	0.53	0.33	0.29
Bi-LSTM-LSTM-CRF	text+speech	10	0.85	0.35	0.39	0.53	0.29	0.25
Bi-LSTM-LSTM-CRF	text	15	0.86	0.42	**0.47**	0.53	0.33	**0.31**
Bi-LSTM-LSTM-CRF	text+speech	15	0.86	0.39	0.44	0.54	0.30	0.28

Table 5.14: Conversational DA tagger with **CRF** performance with different context lengths on **SwDA** and **CCC**.

5.6 Summary of research questions investigated

It is aimed to improve the performance of existing rare class DA tagger, particularly to improve the class-wise accuracy of certain classes such as clarification requests, signal non-understanding, questions types such as declarative wh-questions, open-ended questions, etc.

Q 1: **What kind of acoustic features have been used in literature that helped in the DA tagging task?**

Prosodic features are measures of the rhythm, intonation, and stress patterns of a speaker's voice. They have been shown to be useful in distinguishing between different dialogue acts, such as questions and commands. Shriberg et al. (1998) have utilized prosodic features that include duration, pause, F0, energy, and speaking rate and showed that prosody made significant contributions to the classification of DAs. Questions such as declarative questions have a similar order of words as statements and prosodic can help in distinguishing questions from statements. They have demonstrated the importance of each prosodic feature by a measure of 'feature usage' which is 'proportional to the number of times a feature was used in classifying the instances'(Shriberg et al., 1998). Duration-based features were used with a usage of 0.554, Fo-based features with a usage of 0.126 and pause features with a usage of 0.121. Rangarajan et al. (2007) presented a framework for DA tagging and integrated the prosodic cues with lexical information, resulting in a relative improvement of 11.8% over using lexical and syntactic features alone. They used RMS energy and pitch (f0) for each utterance as prosodic features. Arsikere et al. (2016) presented a set of 57 acoustic features for DA tagging that includes pitch & voicing, duration & pausing, intensity, and speaking rate & rhythm.

Overall, the features such as pitch (f0-based features), energy-based features, and duration features are found to be useful and studies showed a gain in performance. We were inspired by the comprehensive set of prosodic features by Shriberg et al. (1998) used for the task of DA recognition, aimed, to investigate how these acoustic features could be helpful for our chosen set of rare classes of DA's.

Q 2: **Which DA classes were performing better with the inclusion of acoustic features at utterance level?**

Our acoustic features improved the quality of recognition for a few classes, particularly for question types namely open-ended questions, declarative wh-questions, and some answer tags such as yes-answers, non-negative answers, and no-answers. Some instances of clarification requests (*qc*) are also predicted correctly by adding acoustic features (see table 5.10 for error analysis). Previously, Surendran and

Levow (2006), have shown the combination of lexical and acoustic features improved the recognition accuracy of certain classes such as clarify statements and reply-w tags. In chapter 4, a low accuracy of 0.09 was obtained on SwDA and 0.0 on the CCC dataset for the declarative-wh question. Here, by adding acoustic features, an improved accuracy of 0.22 is achieved on SwDA and 0.09 on the CCC dataset. This class has been reported in the literature with low accuracy (Duran et al., 2021; Chakravarty et al., 2019)

Q 3: **Do contextualized and pre-trained embedding is more helpful in DA recognition tasks than static pre-trained embeddings?**

Yes, contextualized pre-trained embeddings are found to be more helpful in the task of DA recognition than static pre-trained embeddings. BERT embeddings were used as contextualized embeddings and Glove embeddings as static pre-trained embedding. Contextualized embeddings take into account the context of the word or phrase being analyzed, which means that the embedding for each word would be different depending on the other words in the sentence which can lead to more accurate and nuanced results. pre-trained embeddings (static), on the other hand, are trained on large amounts of data and can be useful for a wide range of tasks. However, they do not take into account the specific context of the text being analyzed. Hierarchical BiLSTM Model with Glove embeddings and considering only current utterance gives an F1 score of 0.41 on the SWDA test set and 0.27 on the CCC test set. These scores are much more improved with BERT embedding and fine-tuning the model on our DA recognition task. With SWDA test data, we achieved an F1 score of 0.49 (vs 0.41 with Glove previously) and 0.36 for CCC data (0.27 with Glove embeddings). Adding one previous utterance context leads to a better F1 score (0.56 vs 0.47) on SwDA and (0.46 vs 0.33) on CCC test data. This also shows that the utterance representation that we got with contextual embeddings along with its previous utterance representation is more useful than static Glove embeddings.

It is also observed that with model utilizing longer context(e.g 2 utterances in context) with glove embeddings results in a decrease of 0.3 F1 score (from 0.33 to 0.30) while with BERT, longer context results in improving the F1 score (current utt: 0.36, 2 utt context: 0.44, 3 utt context: 0.46).

Q 4: **How well fine-tuning the BERT Model for the task of DA recognition improve the results over using pre-trained embeddings extracted as features?**

Fine-tuning the BERT model for the downstream task of DA recognition has led to significant improvements in results compared to using feature-based pre-trained embeddings. This is because BERT is a powerful language model that has been

trained on a large amount of data and can capture complex relationships between words and phrases. With fine-tuning BERT, we start with the pre-trained weights and then continue training the model on our smaller datasets (SwDA & CCC) that is specific to the downstream task. This allows the model to adapt to the specific nuances of the task and improve its performance. In contrast, using feature-based pre-trained embeddings during training means that the model is not able to adapt to the specific task and may not be able to capture all of the relevant information in the data. It can be seen in table 5.11, FT-PRE-BERT models performed very well with the highest F1 score 0f 0.57 on SwDA and 0.48 on CCC test set as compared to FB-PRE-BERT results in 0.16 F1 score for SwDA and 0.11 for CCC. With FZ-PRE-BERT also results in decreasing the F1-score to 0.29 on SwDA and 0.21 on CCC as compared to fine-tuned BERT model. FZ-PRE-BERT is heavily biased towards the most frequent classes, resulting in very poor performance on the least frequent classes of interest. In literature, Noble and Maraev (2021) also observed similar findings saying that fine-tuning the BERT model is more accurate on SwDA and the other corpus as compared to freezing weights during training. They got a macro F1 score of 36.75 on SwDA with the BERT-FT model. Devlin et al. (2018) demonstrated the results for fine-tuning the BERT and Feature-based pre-trained BERT model with Named Entity Recognition task and got an F1 score of 0.3 more than feature based BERT. However, they concluded that BERT is effective for both fine-tuning and feature-based approaches.

Here, the representation learned through a feature-based pre-trained approach is not performant and does not capture dialogical context information while fine-tuning the model helps in better understanding and capturing the dialogue phenomena. We particularly observed improved performance on the CCC dataset with fine-tuning the BERT on dialogue act tagging. In chapter 4, the best F1 score of 0.45 is obtained on CCC test set with glove embeddings with 1 utt and 1 previous DA as context. Here, in section 5.5.1, the highest F1 score of 0.40 was achieved with 1 utterance context, acoustic features, and similarity feature. With BERT fine-tuned model, we got an F1 score of 0.48 with 1 previous utterance and 1 DA as context.

Q 5: **Does building a conversational DA tagger that takes the full conversation in context is better than a DA tagger considering limited context length?**

We build a conversational hierarchical dialogue act tagger that takes the longer utterances in the context in the form of conversation. Here, only lexical information is considered along with acoustic features. Additional information such as similarity vectors and DA's are not considered as features. The hierarchical BiLSTM-LSTM model uses the neighboring sentences to learn the dependencies among consecu-

tive utterances. We got the highest F1 score of 0.54 with a context length of 20 utterances as a sequence in conversation, which is 0.4 lower than using 1 utterance context with additional information such as previous DA history, acoustic feature, and similarity vector. The conversational model with longer context with lexical information performs better than using both lexical and acoustic features (0.54 vs 0.52) with a context window of 20. When we increase the window more, to beyond 20, there is no further improvement.

Here, the Hierarchical BiLSTM model with limited context such as 1 previous utterance with additional information such as previous DA history, etc performs better than when using the conversational level hierarchical BiLSTM model which takes into account the longer context (e.g. 20).

Q 6: **Does adding a CRF layer to capture the contextual correlations between DAs help in predicting rare classes in a better way?**

This is quite surprising and contrary to our expectations that CRF does not perform well on the DA tagging task of rare classes. The conversational level DA tagger without using CRF performs better than using the CRF layer. This is different from the previous work with hierarchical conversational level models with CRF ((Raheja and Tetreault, 2019; Si et al., 2020b; Srivastava et al., 2019). This may be due to the fact that these models of dialogue act tagging worked with standard SwDA corpus where classes are more evenly distributed. In current case, CRF was not found useful for capturing DA dependency with data where most of the tags are less frequent (less than 1%). However, I left it as an open question and this will be further investigated in future.

5.7 Conclusion

In this section, we extend the set of experiments with the purpose of building an automatic DA tagger for the detection of rare classes from natural conversations. The overall goal was to improve the class-wise accuracy of certain classes of interest such as clarification requests, and signal non-understanding. The investigation focused on exploring the advantages of incorporating linguistic features, including pre-trained Glove embeddings and contextualized BERT embeddings, along with acoustic descriptors, to predict the DA tags for each utterance. The BERT model is also fine tuned for the downstream task of DA tagging with training on both the SwDA and CCC train sets. It is concluded that certain features were helpful for different types of DA's. Acoustic features are helpful for certain question classes such as open-ended questions, declarative wh-questions, and declarative yes-no questions. These along with similarity features are helpful for classes such as clarification requests. Fine-tuning the BERT model on the task of DA tagging

gives a performance boost particularly on the CCC dataset of 0.48 (macro F1 score) with the model utilizing 1 utterance and previous DA as context. While predicting rare class DA's is more challenging, our models show improvements using the different feature sets and fine-tuning the model to learn the relevant information within the utterance and across the utterances and capture dialogue-level information very well over the downstream task.

Future work will investigate the use of better prosodic features, and try combinations of different features to improve the class-wise performance. It is also aimed to use more conversational-styled BERT models such as BERT-base-cased-conversational [8] and Dialogue BERT (Gu et al., 2021) to capture discourse level coherence among utterances for the task of rare class DA tagging.

In the following chapters, an investigation is performed to check whether these rare class dialogue acts used as unigram and bigram sequences along with other general dialogue-level features are helpful for the AD classification.

[8] https://huggingface.co/DeepPavlov/bert-base-cased-conversational

CHAPTER 6

Are Interaction Patterns helpful in AD Diagnosis: An Experimental approach

This chapter explores different interactional aspects of communication particularly durational aspects of pauses, gaps, lapse, and attributable silence in the conversational corpus of Alzheimer's patients with a view to challenge claims of about different functionality of these pauses in the discourse of Alzheimer's patients. The annotation scheme for annotating the corpus with various pause types, along with the presentation of the distribution and duration of these pauses, will be discussed for both the AD patient group and the Non-AD group.

Here, it will also be investigated whether it is possible to combine the interactional dialogue features with disfluency features to improve the accuracy of detecting AD because the language has the same certain characteristics. In the context of AD, main focus will be on dialogue features for cognitive decline identification in spontaneous speech. A model will be designed that obtains prediction decisions based on these dialogue features and then combines them with disfluency features as language features as well with DA's based unigram, and bigram sequences obtained from experiments in chapter 4 and chapter 5 to get the final prediction score. Experimental results show that the proposed classification obtains very promising results on this conversational data set and suggests that AD can be successfully identified using interactional features of the spontaneous speech data in natural settings. This study advances our knowledge of how interaction patterns in natural conversation affect cognitive modeling across diverse activities, which has implications for the development of non-invasive, low-cost tools for widespread use in cognitive health monitoring.

6.1 Background

Much of the work to date in AD diagnosis has focused on properties of individual language, using various kinds of linguistic and acoustic features (Jarrold et al., 2014), or fluency, information content, and syntactic complexity (Fraser et al., 2016b,a; de Lira et al., 2011). However, this is often studied within particular individual language tasks, usually within specific domains including picture description (the commonly used Cookie Theft picture description task from the Boston Diagnostic Aphasia Examination (Goodglass et al., 2001)), story narration task (e.g. The Dog story (Le Boeuf, 1976)) and semi-structured interviews (e.g. Autobiographical Memory Interview (Kopelman et al., 1990)). Approaches to analysis and diagnosis therefore usually focus on aspects of individual language such as lexical, grammatical, and semantic features. Kavé and Dassa (2018), for example, examined Dementia via a picture description task in the Hebrew language, using ten linguistic features, and showed that relative to participants who were cognitively healthy, the AD group produced more frequent words, a lower type-token ratio, a higher number of pronouns in comparison to nouns and pronouns, and a smaller percentage of content words. Orimaye et al. (2017) build an automated diagnosis model using low-level linguistic features including lexical, syntactic, and semantic features (NGrams) from verbal utterances of Probable AD and control participants. In another line of research, Ahmed et al. (2013) argued that speech production, syntactic complexity, lexical content, semantic content, idea efficiency, and idea density are useful features of connected speech that are helpful to examine longitudinal profiles of deterioration in AD.

Fluency has also been shown to be indicative of AD. Patients with AD struggle with verbal fluency ,and object recognition, as well as tasks that require the use of semantic knowledge (Pasquier et al., 1995; López-de Ipiña et al., 2013). Patients with AD speak more slowly, pause for longer periods of time, and take more time to find the proper word, all of which add to speech disfluency (López-de Ipiña et al., 2013). Abel et al. (2009) used patient speech errors such as naming and repetition disorders and relate it to the problem of AD diagnosis. Rohanian et al. (2020a) used a deep multi-modal fusion model to show the predictive power of disfluency features in the identification of AD.

Speech contains pauses and is not continuous. A pause is an absence of speech. In human speech, pauses are an essential part and may possess different processes. Pause is needed to breathe, to plan or to think what to say next, and sometimes to see if somebody else wants to speak and to negotiate turn-taking. Pausing behaviour is often associated with a lack of fluency, and several research have recommended using different temporal speech analysis techniques to detect AD. Pauses in speech are frequently seen as a sign of lexical-semantic decline in patients, one of the earliest signs of AD (Pistono et al., 2019a). In a study, Davis and Maclagan (2010) examined the silent pauses in a story retelling task with an older woman on two different occasions and found changes in pauses function signaling difficulty in word finding to difficulty in finding key components

in the thread of a story.

(Forbes-McKay and Venneri, 2005) compared the word finding difficulties during the discourse in a picture description task among AD and healthy elderly subjects and stressed the fact that pauses, use of indefinite terms, and repetition are significantly more frequent in the AD group. According to Gayraud et al. (2011) AD patients produce more silence pauses than healthy control but they found no significant difference in the duration of pauses. This study was performed on spontaneous speech data of an autobiographical task of AD and healthy persons and also identified that silent pauses occur more often outside grammatical boundaries following more frequent words. Singh et al. (2001) have utilized different temporal measures including frequency of pauses, total pause time, mean duration of pause (MDP), *standardised pause rate* (SPR), *standardised phonation time* (SPT), and a few more to distinguish between AD and healthy control group by performing statistical analysis and discriminant analysis. Therefore, in order to better understand cognitive deficits during discourse processing, it may be helpful to look at pauses in the spontaneous speech of AD patients in connection to their cognitive impairment.

From a more linguistic perspective, silences in conversation have been analysed in terms of distinct categories, with several terms coined to distinguish these, especially pauses at speaker changes or turn changes. In one of the classic articles on pauses, (Sacks et al., 1978) distinguished three kinds of silences in speech; pause, Gap, and Lapse. This difference is based on the perceived length of silence, what preceded, and followed the silence in conversation. Pause is a simple silence that occurs within the same speaker's turn. This could be silence either at a transition relevance place (TRP) or silence at no TRP. The former is the situation when a speaker stopped speaking and continues to speak after the TRP. The latter other situation of pause could be when a speaker paused while speaking within the same utterance or sentence and it's not TRP. The reason for this pause could be breathing, planning, or others like word-finding difficulty, etc. Gaps referred to shorter silence at speaker change. A lapse is perceived as longer or extended silence between turns (speaker change). A lapse represents a discontinuity in the flow of conversation and is perceived as longer than the gap duration.

An alternative view to Sacks et al. (1978) categorization was advanced by Heldner and Edlund (2010) with few modifications including; overlaps, between speaker gaps, covered both gaps and lapse, and a within speaker silence refers to a pause. Levinson (1983) categorized silence into three categories: within-turn silence (*pause*), inter-turn silence (*lapse* or *gap*), and turn silence (*attributable silence*) using a turn-taking system that integrated its forms and functions. A body of research looked into turn silences within the context of CA and relevance theory studies, taking into consideration the psychological factors of the communicators, i.e., why they choose silence over other forms of communication to avoid providing a dispreferred response (Wang, 2019).

It is evident from the research that researchers have used different terms for these

silences, some authors only analyze silent pauses as silence while few take filled pauses e.g. ('um','uhhh').Using these concepts to analyze Alzheimer's discourse, Davis and Maclagan (2009) demonstrated that both filled and silent pauses are related to functions in narrative and conversations. They demonstrated that filled pauses (e.g. 'uh' and 'um') serve as placeholders and hesitation markers while silent pauses serve as a function for word finding, and planning at the word, and narrative level as well indicators of a decrease in other interactional and narrative skills. These pauses may also be defined by different thresholds. The most common threshold in literature used as a silence threshold is 200ms (Heldner and Edlund, 2010). The majority of pauses and gaps are shorter than 1000ms (Fors, 2011), and average gap durations of 345–456ms are reported in literature (Brady, 1968). Levinson and Torreira (2015) suggested that *gaps* of 700 ms or longer are associated with dispreferred responses, with 300 ms as the normal threshold. (Davis and Maclagan, 2010) utilized the convention of Crystal and Davy (2016) to distinguish between micro-pause (less than a second), average pause (less than two seconds), and long pause (longer than 2 seconds) with elderly people (speech rate decreases with age).

Because each addition to the conversation builds upon and responds to the partner's prior contribution, CA's focus on communication as a collaborative achievement shows that investigating interaction might offer more insight than an individual analysis of the contributions of the two halves. Perkins et al. (1998) explored turn-taking phenomena, repairs, and topic management in conversations with people having dementia and showed how failure to maintain topics frequently results in topic changes by the conversing partner and cognitive deficits can affect the ability to secure the conversational floor. Jones et al. (2016) explored interactions in dialogues between patients and clinicians during clinic visits, while Elsey et al. (2015) highlighted the role of carer, looking at interactions among a clinician, a patient, and a carer. They establish differential conversational profiles which distinguish between non-progressive functional memory disorder (FMD) and progressive neuro-degenerative Disorder (ND), based on the interactional behaviour of patients responding to neurologists' questions about their memory problems. Davis et al. (2014a) examined how effective communication can be with the usage of strategies such as quilting, go ahead, and indirect questions between residents with Dementia and their conversation partners, exploring various aspects including the impact of different types of questions, delayed responses, and the number of ideas in response using idea density.

Conversational clues were missed in traditional approaches like picture description tasks or narrative tasks or while analyzing individual speech. Interactional features, therefore, promise one way to help alleviate the above-discussed problems, by contributing to general, non-invasive methods of diagnosis that can be applied in natural everyday conversation, and some recent work has therefore investigated computational models using machine learning techniques. In a recent study, Mirheidari et al. (2019) performed an automated analysis for Dementia detection with CA-inspired features, together with

some language and acoustic features, achieving classification accuracy of 90%. Luz et al. (2018) build a predictive model based on content-free features extracted from dialogue interactions from spontaneous speech in more natural settings using the CCC corpus of patient interview dialogues (Pope and Davis, 2011). They got promising results with an accuracy of 86% with only dialogue interaction-based features with less reliance on the content of task/ dialogue. In a study building on the PREVENT Dementia project, de la Fuente Garcia et al. (2019) design a protocol for a study that uses conversational analysis to see if it may identify early behavioral indications of AD through dialog interactions. Interactional patterns are considered among the current challenges to be addressed to make the spoken dialogue systems usable by older adults or frail patients (Addlesee et al., 2019).

Dialog act-based conversation analysis through an initial corpus study in chapter 3 was introduced by us for the first time (Nasreen et al., 2019). This study is based on fine-grained analysis of questions and answers as several research with promising results on dementia detection have focused on these (Varela Suárez, 2018; Hamilton, 2005). Later on, Farzana et al. (2020) conducted a conversational analysis study based on DA tagging to capture the interaction patterns from the semi-structured picture description task at DementiaBank in terms of various DAs from the interviewer and the subject. In chapter 4, a DA tagger was developed to computationally model the conversations in terms of DA's that could be used as interaction sequences between patients and their conversational partners. Farzana and Parde (2022) also build a DA tagger based on the DA annotation scheme, and following the findings of our own DA tagger (see chapter 4) ((Nasreen et al., 2021c) and employing a collection of non-content interaction features for task-agnostic dementia detection and demonstrating their great utility in differentiating between dementia and healthy controls across tasks.

Here, our purpose is to investigate a new set of interactional features particularly the role of specific kinds of silences in AD conversations, and dialogue act based features and evaluate their use in a computational model for AD classification.

6.2 Research questions

It is hypothesized that using high-level interaction patterns as dialogue features could be beneficial when building automatic systems for predicting the diagnosis of Alzheimer's disease based on these interactional features. Therefore, in this chapter, it will be focussed on features derived from dialogue interaction, with a particular interest in specific types of silences and classes of features based on dialogue acts. In-depth, this study is conducted to answer the following research questions:

Q 1: What kind of dialogue features turn out to be the most prominent features that can aid in the prediction of Alzheimer's?

Q 2: Do any of the dialogue features fit in with the observations found in literature such as attributable silences, turn lengths, turn switches, pause rate, and speech rates?

Q 3: Does the functional division of silences into short pause (SP) and long pause (LP) within the same speaker and silences at turn changes like gaps and lapses contributes to improving the accuracy of predicting AD from Non-AD?

Q 4: Does the more specific dialogue acts feature with unigram and bigram sequences hold the predictive power to identify AD symptoms?

Q 5: Was the combination of dialogue features with disfluency features helpful in improving the accuracy of classification among AD and Non-AD?

6.3 Methodology

6.3.1 Dataset and participants

Our aim is to investigate the behaviour of AD patients based on the interaction patterns, including repairs and pauses within utterances and between turns, observed in a corpus of dialogue. This is a post hoc study based on an existing dataset, the Carolinas Conversation Collection (CCC) corpus (Pope and Davis, 2011), already discussed in Chapter 3 section 3.2.1. In chapter 3 and chapter 4, experiments were performed on a set of twenty patients including 10 AD and 10 Non-AD patients.

	AD (N=15)	**Non-AD (N=15)**
Age range	60-89	60-79
Years of Education	9-16	8-16
Gender	M:4	M:4
	F:11	F:11
Total duration of dialogues	152	179.7
Average dialogue duration	10.13	11.97

Table 6.1: Demographic data for AD and Non-AD patients, with dialogue duration in minutes.

In this chapter,the dataset is expanded by looking at a larger sample size of 30 patients. For this particular study, we use the transcript and audio recording from one dialogue conversation chosen randomly from each of a total of 30 patients: 15 AD-diagnosed patients (4 Males, 11 Females) and 15 patients (4 Males, 11 Females) with other chronic diseases including diabetes, heart problems, arthritis, high cholesterol, cancer, leukemia but not AD; no patients were diagnosed as having breathing problems.

These groups are selected to match the age range, to compare the different patterns of interaction, and to avoid bias. The demographic data of the participant is given in Table 6.1.

6.3.2 Annotation Scheme

In this section, the functional division of silences into different categories is considered, detailing the annotation procedure with examples and inter-rater agreement. Any silence lasting at least 0.5 seconds is considered for this specific study. To categorize the silences, Levinson (1983)'s definitions are employed: *pauses* (silences within a single speaker's turn), *gaps* and *lapses* (silences between speaker turns), and *attributable silences* (silences where speaker changes were expected but did not occur). This study further categorized pauses into *short pause* (***SP***) and *long pause* (***LP***). A *SP* is a silence that takes place within a single speaker turn, which are advised in the annotation protocol for average speech rates greater than 0.5 seconds and less than 1.5 seconds; a *LP* is a longer pause within a single speaker turn, normally at least 1.5 seconds. Guidelines were employed for these thresholds rather than strict rules, allowing for variations in speech rates. Annotators were given the discretion to determine the category of the pause based on their perception. Both SPs and LPs may occur either at a *transition relevance place* (TRP) or not at a TRP, but no speaker change occurred. TRPs are junctures at which the turn could pass from one speaker to another.

For inter-turn silences and attributable silences, explicit time thresholds are not used- annotators used their judgment when listening to the silences in the context of the conversation closely and categorized them according to the following definitions. A *gap (**GA**)* is defined as silence at a speaker change (i.e. turn boundary, with speaker change from I-P or vice versa P-I) which is not perceived as unusually long. Following Sacks et al. (1978), a *lapse (**LA**)* is then distinguished from a gap by not only being longer by "rounds of possible self-selection", but also involving a discontinuity in the flow of conversation. More precisely, annotators were told to annotate a silence as a lapse for unusually long silences in communication between two individuals, at TRPs, after which one participant (usually the interviewer in this dataset) initiates a new topic (topic shift). The final category, *attributable silence*, occurs when the current speaker selects another next speaker (by asking a question, by naming or by looking at them), thereby putting the selected speaker under the obligation to speak next, but for one reason or another, that selected speaker does not respond; after the silence, the current speaker, therefore, continues the conversation (Elouakili, 2017). Attributable silence is defined as a long silence after a question is asked from one party, no response from the other, and the first party then continues (example in the sample below at (2) with 4.2 seconds of silence in response to wh-question is an attributable silence).

1. *I*:" What other animals were on the farm?"
2. *P*: (4.2 seconds) ***[AS]***
3. *I*: "Like some pigs, hogs and chickens?"
4. *P*: "Um hmm."

Examples of these pause types with conversation samples are given in the **Appendix B**.We also differentiated between speakers (patient *P* and interviewer *I*) by assigning speaker ID *(SP_ID)* to each labelled pause.

The CCC conversations are recorded in a community center with background noise present, so, these silences after carefully listening to the audio together with the transcript with the help of the ELAN software are manually annotated (Sloetjes and Wittenburg, 2008).[1]

To check the inter-rater agreement, two annotators annotated the silences of at least 0.5 seconds in one randomly selected AD patient dialogue; both had a good knowledge of linguistics and were familiar with the annotation rules. A multi-rater version of Cohen's κ (Cohen, 1960) is used as described by Siegel and Castellan (1988) to establish the agreement of annotators in terms of the overall agreement on all pause types, and also in terms of each pause type individually – see Table 6.2. An overall substantial agreement of κ=0.66 is obtained for all categories of pauses. Lower, though still moderately strong, κ values for *LP* and *SP* are obtained as these are pauses within the same speaker utterances and patients are older people with lower speech rates, making it more difficult to decide whether there is a relatively shorter or longer pause at certain lengths around the recommended boundary of 1.5 seconds.

Feature name	**Acronym**	κ	A_o
Short Pause	*SP*	0.55	0.83
Long Pause	*LP*	0.46	0.79
Gap	*GA*	0.88	0.94
Lapse	*LA*	0.75	0.96
Attributable Silence	*AS*	0.66	0.98
Overall		0.66	0.75

Table 6.2: Inter-annotator agreement: Cohen's kappa (κ) and observed agreement *(A_o)*

6.3.3 Interactional features in AD speech

6.3.3.1 Temporal measures of dialogue interactions

Table 6.3 shows the extracted collection of high-level dialogue features to measure the interactions between *P* and *I*. There are 14 features for *P* and 12 features for *I* within the

[1] https://archive.mpi.nl/tla/elan

conversation and 6 features for overall conversation. This results in a set of 32 features representing the interaction within the natural dialogue conversations. The number of pauses within *P* or *I* were normalized by the number of words spoken by each respectively instead of normalising by the number of utterances because it may be possible that *P* speak less number of words per utterance.

Feature	**Description**
# *LA*	Total number of *LA* is sum of normalized no. of *LA* from *P–I* and *I-I*
Dur_LA	Sum of average *LA* duration from *P–I* and *I–I*
# *GA*	Total number of *GA* is the sum of normalized no. of *GA* from *P–I* and *I–P*
Dur_GA	Sum of average *GA* duration from *P–I* and *I–P*
# *overlaps*	No. of segments spoken simultaneously by both *P* and *I*. This feature indicates the frequency of occurrence that may be attributed to speech initiation difficulties. (Young et al., 2016)
#*Turn_switches per Minute*	This is calculated by the number of turns per 60 seconds.
Patient features	
# *SP*	Number of *SP* within *P* utterances normalized by the total # of words spoken by *P*.
Dur_SP	Total duration of *SP* normalized by the total duration of speech by *P* without pauses.
# *LP*	Number of *LP* within *P* utterances normalized by the total number of words spoken by *P*.
Dur_LP	Total duration of *LP* normalized by the total duration of speech by *P* without pauses.
# *GA(P–I)*	Number of *GA* at turn transition from *P–I* normalized by the total number of turns in the conversation
Dur_GA(P–I)	Average duration by considering the total duration of *GA (P–I)* divided by # *GA(P–I)*.
# *AS*	Normalised number of Attributable silence *AS* after posing the question from *I–P*.
Dur_ AS	Average Duration of *AS* from *I–P* with no response.
Standardized pause rate *(SPR)*	*SPR* is obtained by the total number of words spoken by *P* divided by the sum of *SP* and *LP*. (average words spoken per pause.)
Standardized Phonation time *(SPT)*	*SPT* is the total number of words spoken by *P* to the total speech time of the patient excluding *SP* and *LP*.

Feature	Description
Transformed Phonation rate *TPR*	"The arcsine of the square root of the phonation Rate (PR)" (Beltrami et al., 2018). *PR* is the speech time of *P* to the total speech time of *P* including *SP* and *LP*
Floor control ratio	This function quantifies dominance by measuring the relative length of time, the *P* spends speaking to the total speech time of the conversation (Aldeneh et al., 2019).
turn_length	This feature measures the number of words per turn spoken by *P*.
speech_rate	The number of syllables *P* produces each minute is known as 'speech rate'. It is derived by dividing *P*'s total number of syllables by the length of his or her speech (in minutes).
Interviewers features	
# SP	Number of *SP* within *I* utterances normalized by the total # of words spoken by *I*.
Dur_SP	Total duration of *SP* normalized by the total duration of speech by *I* without pauses.
# LP	Number of *LP* within *I* utterances normalized by the # of words spoken by *I*.
Dur_LP	Normalized duration of *LP*
# GA(I-P)	Number of *GA* at turn transition from *I–P* normalized by the total number of turns.
Dur_GA(I–P)	Average duration of *GA (P–I)* .
# LA(I–I)	Total # of *LA* is sum of all *LA (I–I)* normalized by # of turns.
Dur_LA(I–I)	Average *LA* duration from *I–I* with the *topic shift*.
# LA(P–I)	Normalized # of *LA* from *P–I* with a topic shift.
Dur_LA(P–I)	Average *LA* duration from *P–I* with the *topic shift*.
turn_length	This feature measures the # of words per turn spoken by *I*.
speech_rate	This feature measures the number of syllables per minute during a speech by *I*.

Table 6.3: The proposed interactional feature set.

6.3.3.2 Dialogue act features

The aim is to investigate the benefits of using DA labels predicted by our DA tagger in chapter 5 in the eventually intended downstream task in AD identification. Here rare class DA classes are used both as unigrams (*f1*) and as bigrams (*f2*) to capture characteristic local DA sequences. For this experiment, unigram DA's, bigram sequences are used containing the meaning-coordination *qc* and *br* DAs in a patient *(P)* utterances, preceded by question DAs from the interviewer *(I)*. total of 462 unique bigrams are found

from AD conversations and 338 bigrams from Non-AD conversations. Our focus lies solely on bigram sequences characterized by interviewers posing questions, succeeded by patients providing answer types, ultimately leading to the identification of 54 specific bigram sequences. For example *(I_qw,P_br)* is a bigram sequence in which *I* is posing a *qw* question which is followed by a *br* signal from *P*. Table 6.4 displayed some of the unigram and bigram DA features along with confusion ratio-based features. Two

Features	**Type (Total)**	**Details**
f1	Unigrams (39)	unigram DAs such as: *P_qy, P_ny, P_br, P_na, P_sa, I_qo, I_qw, I_b, I_qy*
f2	Bigrams (54)	bigram DAs sequences such as: *I_qw–P_br, I_qo–P_sa, I_qy–P_ny I_qw–P_qc, I_qw^d–P_qc*
f3	Confusion (2)	question_ratio, confusion_ratio
f4	Others (32)	other features from dialogue (see table 6.3) includes: normalized turn duration, Avg number of words per minute, turn switches per minute, number of overlaps

Table 6.4: The proposed dialogue act feature set.

aggregate features are also computed from these DAs as proxies for levels of patient confusion (*f3*): ***question_ratio*** (how many questions asked by the patient (*P*) out of total utterances spoken by *P*) and ***confusion_ratio*** (ratio of total *br* & *qc* to the total questions asked by *P*). Question_ratios were previously used by Khodabakhsh et al. (2015) in AD identification, considering question words such as 'what', 'which' etc. as a mark of confusion or request for further details. Here, this is replicated as question_ratio and add the more specific use of *qc* and *br* tags as confusion_ratio. These features were devised to capture different facets of patterns of global interaction that other feature groups could have missed. Moreover, these are combined with interactional features (*f4*) already discussed in the previous section 6.3.3.1. such as normalized turn lengths, an average number of words per minute (as used by Luz et al. (2018) for AD prediction), turn switches per minute, and the number of overlaps. Overlaps represent the number of segments spoken simultaneously by both speakers, with the intuition that these may be attributed to speech initiation difficulties.

6.3.4 Disfluency features

Detailed language use research helps us to find the indications of language impairment in AD and is a step toward the design of future clinical diagnostic tools. Schegloff et al. (1977) stated that self-repairs, pauses, and fillers are frequently used in regular speech. Disfluencies are typically interpreted as signs of communication concerns brought on by problems with production or self-monitoring (Levelt, 1983). People with AD are likely to experience issues with their language and cognitive abilities. Patients with AD tend to talk more slowly, pause for longer periods of time, and spend more time looking for the right word, all of which can lead to disfluency (López-de Ipiña et al., 2013).

In addition, it is aimed to combine the dialogue features with these language disfluencies present in conversations of AD. The features extracted in a recent study, revealing the usefulness of disfluency features in a diagnostic task of Alzheimer's Disease within the ADReSS challenge, are employed (Rohanian et al., 2020b). A deep-learning-driven model of incremental detection of disfluency created by Hough and Schlangen (2017) automatically annotated self-repairs (Rohanian et al., 2020b). It consists of deep learning sequence models that predict disfluency tags on the DementiaBank dataset using left-to-right, word-by-word word representation of incoming words, part-of-speech tags, and other variables. The disfluency tags are **edit terms** and **repair**s (verbatim repeats, substitutions, and deletions). Normally, it is considered that disfluencies have a reparandum-interregnum-repair structure. A verbal error that the speaker eventually corrects is known as a reparandum; the resulting expression is known as a repair. An interregnum word is a filler between the repair words and reparandum as in (6.1):

$$\texttt{John}\ \underbrace{\texttt{[likes}}_{\texttt{reparandum}}\ \texttt{+}\ \underbrace{\texttt{\{ uh \}}}_{\texttt{interregnum}}\ \underbrace{\texttt{loves]}}_{\texttt{repair}}\ \texttt{Mary} \tag{6.1}$$

Without reparandum and repair, the disfluency is reduced to a single **edit term**. The usage of more phrasal language like "I mean" and "you know" as well as marked, lexicalized edit terms like a filled pause ("uh" or "um") may also occur. Disfluency identification then involves identifying these components and their organizational structure.

The disfluency detector in this case labels each word as either a repair onset tag (designating the first word of the repair phase), an edit term (*edit_terms)*, or a fluent word. In order to get the most information from different types of disfluency, repairs are splitted between the broad classes of *verbatim repeats* (***Rpt***), *substitutions* (***Sub***), and *deletes* (***Del***):

1. " So [he, + he] brings the fresh flowers... "
   Repeats

2. "[Someone said that, + I heard someone out here say] it is getting quite cool outside, is it? "
   Substitution

3. "...and I looked [at + { uh}] and answered her question. . ."
   Deletes

Self-repairs are annotated automatically using a model of incremental detection of disfluency developed by Rohanian and Hough (2020) and Hough and Schlangen (2017)[2] by the authors of the tool. Rohanian and Hough (2020) report the automatic disfluency detector achieves an F1-score accuracy on detecting the first word of the repair phase at 0.743 and an F1-score accuracy of 0.922 on detecting all edit term words on the Switchboard disfluency detection test data. Its accuracy is considered adequate for our purposes. Automatically deriving the types of interest from the tagger's output, 4 disfluency tags are used for patients (*P*) and 4 for interviewers *(I)* resulting in a total of 8 disfluency features (details in table 6.5).

Feature	Description
Patient features	
# *edit_terms*	Number of # *edit_terms* within *P* utterances normalized by the *total # of words* spoken by *P*.
# *Rpt*	Number of verbatim repeats within *P* utterances normalized by the *total # of words* spoken by *P*.
# *Sub*	Number of substitutions within *P* utterances normalized by the *total # of words* spoken by *P*.
# *Del*	Number of deletes within *P* utterances normalized by the total # of words spoken by *P*.
Interviewer features	
# *edit_terms*	Number of # *edit_terms* within *I* utterances normalized by the total # of words spoken by *I*.
# *Rpt*	Number of verbatim repeats within *I* utterances normalized by the *total # of words* spoken by *I*.
# *Sub*	Number of substitutions within *I* utterances normalized by the *total # of words* spoken by *I*.
# *Del*	Number of deletes within *I* utterances normalized by the *total # of words* spoken by *I*.

Table 6.5: The proposed disfluency feature set.

6.4 Analysis

6.4.1 Statistical analysis

To investigate the importance of each feature, the mean and standard deviation (SD) are calculated for each group (AD and Non-AD). A non-parametric independent sample test

[2]The python implementation used is at https://github.com/clp-research/deep_disfluency

Feature	**AD**	**Non-AD**	***Mann-Whitney U test***	
	Mean (SD)	Mean (SD)	***p***	**U**
#LA	0.051 (0.053)	0.011 (0.020)	**0.013** *	171.5
Dur_LA	3.195 (2.592)	1.041 (1.927)	**0.026** *	166.0
# GA	0.228 (0.121)	0.104 (0.071)	**0.010** *	174.0
Dur_GA	1.400 (0.464)	1.100 (0.245)	0.067 -	156.0
# overlaps	0.073 (0.029)	0.109 (0.082)	0.595	99.0
#Turn_switches per Minute	2.544 (0.835)	3.510 (1.447)	**0.026***	59.5
Patient features				
# SP	0.034 (0.013)	0.032 (0.018)	0.455	130.5
Dur_SP	0.064 (0.022)	0.082 (0.06)	0.254	85.0
# LP	0.022 (0.016)	0.012 (0.017)	**0.013** *	171.5
Dur_LP	0.106 (0.078)	0.054 (0.065)	**0.016** *	169.5
# GA(P–I)	0.103 (0.067)	0.052 (0.054)	**0.015** *	170.5
Dur_GA(P–I)	1.515 (0.820)	1.000 (0.368)	0.098-	152.5
# AS	0.010 (0.013)	0.002 (0.002)	0.067-	157.0
Dur_ AS	2.468 (3.243)	0.414 (0.724)	**0.037** *	163.0
(SPR)	22.158 (12.54)	36.40 (28.19)	0.137	76.0
(SPT)	2.113 (0.531)	2.839 (0.060)	**0.002** **	41.0
TPR	1.041 (0.115)	1.114 (0.157)	0.081 -	70.0
Floor control ratio	0.596 (0.172)	0.712 (0.183)	0.098 -	72.5
turn_length	12.142 (6.59)	22.52 (20.34)	**0.007** **	168.5
speech_rate	164.91 (35.74)	180.1 (37.82)	0.345	89.0
Interviewers features				
# SP	0.013 (0.009)	0.017 (0.02)	0.935	110.0
Dur_SP	0.029 (0.020)	0.034 (0.036)	0.902	109.0
# LP	0.006 (0.006)	0.005 (0.007)	0.126	149.5
Dur_LP	0.033 (0.023)	0.021 (0.037)	0.061 -	157.5
# GA(I–P)	0.125 (0.068)	0.052 (0.033)	**0.002** **	184.5
Dur_GA(I–P)	1.363 (0.365)	1.011 (0.301)	**0.041** *	161.5
# LA(I–I)	0.020 (0.023)	0.027 (0.068)	0.305	137.5
Dur_LA(I–I)	3.291 (3.696)	1.316 (1.951)	0.106	151.5
# LA(P–I)	0.031 (0.037)	0.002 (0.003)	**0.009** **	175.0
Dur_LA(P–I)	2.552 (2.161)	1.163 (2.317)	0.081 -	155.0
turn_length	9.155 (4.320)	23.31 (22.31)	**0.001** *	34.0
speech_rate	195.49 (32.89)	183.05 (43.09)	0.325	137.0

Table 6.6: Descriptive statistics (Mean, SD) and statistical significance for the dialogue feature set. ** denotes highly significant at $p < 0.01$; * denotes significance at $p < 0.05$; **-** shows a trend toward significance at $p < 0.1$

(*Mann-Whitney U*) is selected on disfluency and interactional features due to the small sample size. The non-parametric test is applied as a two-tailed test for unpaired samples and unequal variances. The value $p < 0.05$ was chosen for statistical significance. IBM SPSS version 26.0 was used for the statistical analysis.

6.4.1.1 Interactional features analysis

Table 6.6 presents the mean, SD, *test statistic U* (for *Mann-Whitney U* test), and the p-values for each of the interactional features reported in Table 6.3. Significant differences between the AD and Non-AD groups are marked in bold. Compared to Non-AD populations, individuals with cognitive impairment or communication problems seem to express themselves differently. Overall, the total number of *GA* ($U = 174.0, p = 0.010$) and the total number of *LA* ($U = 171.5, p = 0.013$) are found to be significantly higher in the AD group. There were fewer turn switches in AD dialogues with a mean of 2.544 as compared to Non-AD dialogues with a higher mean of 3.510 ($U = 59.5, p = 0.026$). The mean duration of *LA* was also significantly higher in AD (3.195 ± 2.592) vs (1.041 ± 1.927) in Non-AD. The distribution of overlaps was not very eloquent among the two groups($U = 99, p = 0.595$).

Figure 6.1 shows distributions of three significant features with Figure 1(A-C) and Figure 1(D) represents the distribution of a non-significant feature ı.e Average duration of *LA (P–I)* between AD and Non-AD groups. There are more numbers of *AS* as shown in Figure 6.1(A) with longer silences in the AD group than in Non-AD. Y-axis shows the normalized duration while X-axis shows the frequency of durations of the *AS* in each group.

Patient Features: Our analysis found that patient's long pause, duration of long pause, number of gaps from *P–I* and duration of *AS* exhibit significant differences between AD and Non-AD patient groups. The more symptoms expressed by an individual, the more frequent and longer pauses are expressed($U = 169.5, p = 0.016$). These longer pauses within the patient's utterances signal the difficulty in word finding, to the problems of finding key components related to events, places, etc. Patients with cognitive impairment tend to pause longer at turn changes with more number of gaps (0.103 ± 0.067) vs (0.052 ± 0.054). Standardized phonation time of patients is significantly lower for AD patients, with a mean of 2.113 and variability of 0.531 for AD patients and a mean of 2.839 for Non-AD patients. Turn length goes up by an average of 20.34 for Non-AD patients and a lower average value of 12.142 for the AD group. This suggests that the more symptoms an individual develops, the less they express with shorter turn lengths ($U = 168.5, p = 0.007$). These results suggest AD patients produce a greater number of pauses with a longer duration (>1.5 seconds), with slower speech rates than Non-AD patients.

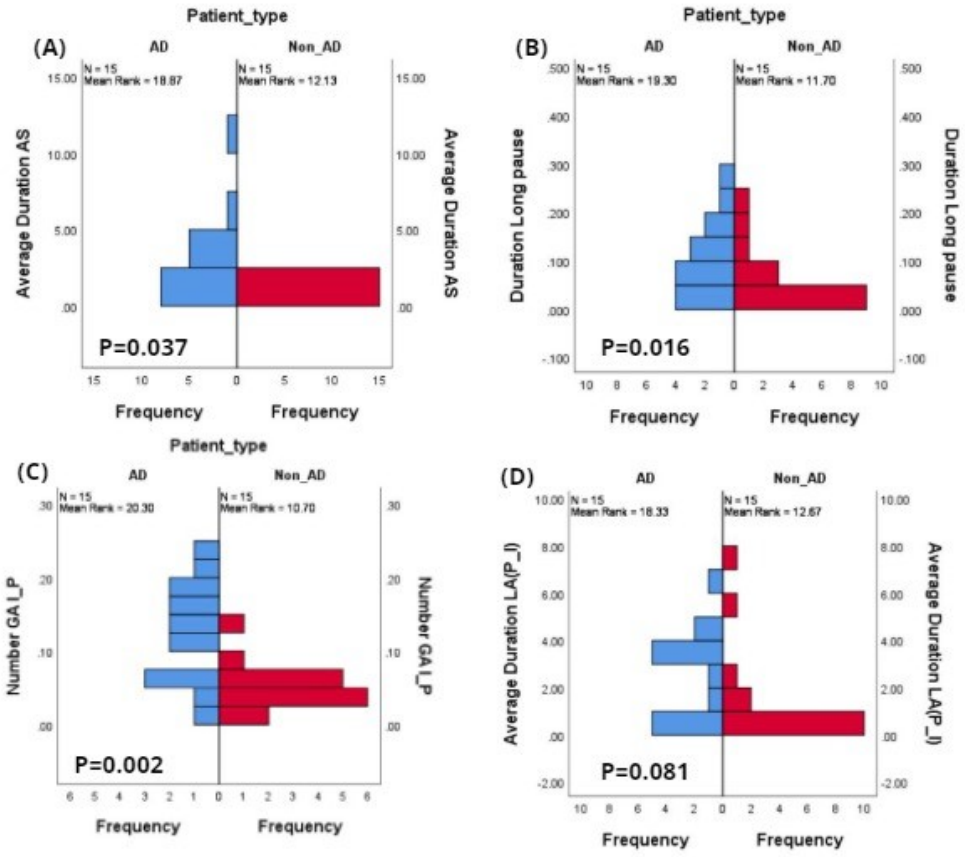


Figure 6.1: Feature value histograms for a selection of different pause types, showing differences in distributions between AD and Non-AD dialogues. **(A)** average duration of patient attributable silences *AS*; **(B)** duration of patient long pauses *LP*; **(C)** frequency distribution of interviewer-to-patient gaps *GA(I–P)*; **(D)** duration of patient-to-interviewer lapses *LA(P–I)*. (A),(B),(C) show distributions that are significant at $p < 0.05$, while for (D) $0.05 < p < 0.1$.

Additionally, the findings imply that AD patients show higher variability in the time they either take to answer clinicians' questions(resulting in high values for the number of gaps from *I–P* with larger delays) or they preferred attributable silences (mean duration of 2.468 for AD patients as compared to 0.414 for Non-AD patients) instead of response. This *AS* is found significant among the groups ($U = 163, p = 0.037$). Notably, the floor control ratio is higher for Non-AD patients, suggesting that AD patients hold the floor for less time compared to Non-AD patients ($U = 72.5, p = 0.098$). There was a negative correlation between the AD severity and Standardised phonation time(number of words per minute excluding pauses) ($U = 41, p = 0.002$). On the contrary, the number of short pauses and duration of short pauses are not found to be significant between AD and Non-AD patients ($U = 130.5, p = 0.455$), suggesting that short pauses are present naturally for breathing and for planning at the word or phrase level. Differences in

terms of standardised pause rate and transformed phonation time are not statistically significant.

Another interesting fact can be visualized in Figure 6.2: the number of gaps is higher in the AD group as compared to the lapse in Figure 6.2 (A) but the durations of gaps are much shorter than the duration of an average lapse within the AD group in Figure 6.2 (C). Across the groups, the number of lapse and the number of gaps in the Non-AD group is lesser than the AD group (Figure 6.2 (A & B)).

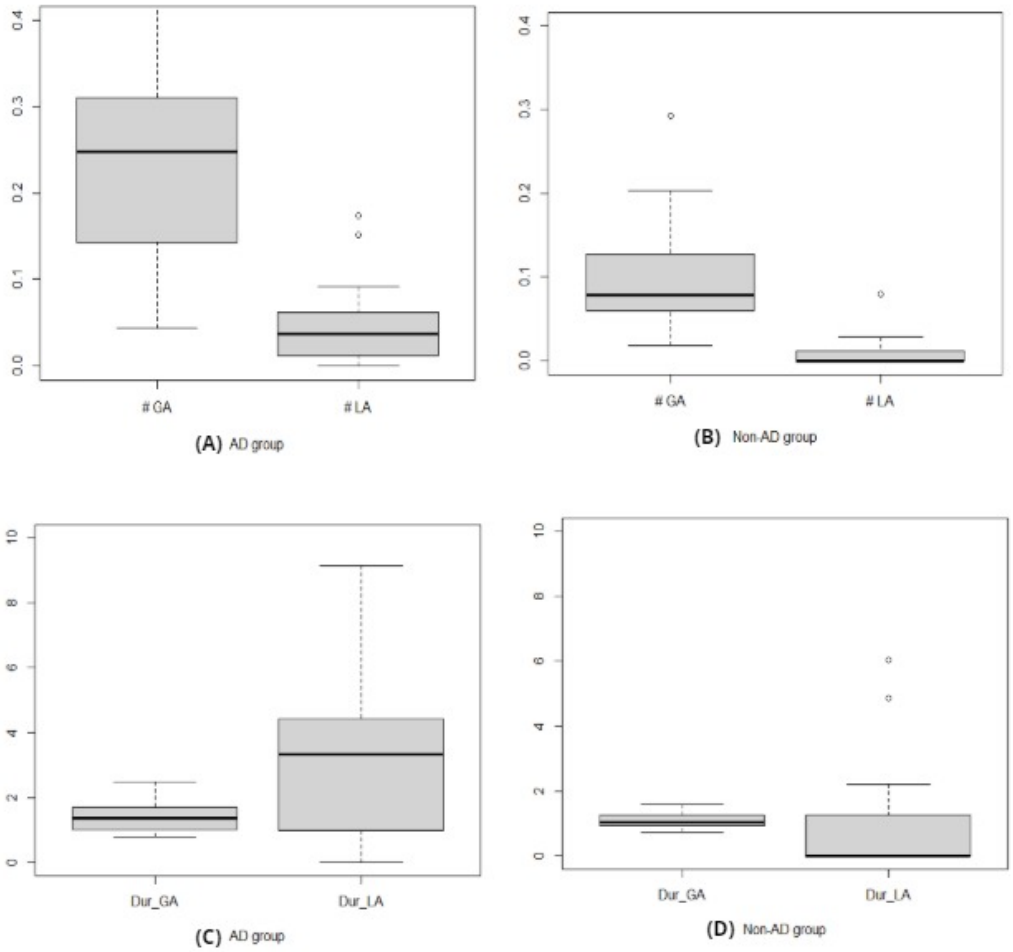


Figure 6.2: Boxplot showing distributions of gaps and lapse.

Interviewer Features: The duration of *LP* is approaching significance with the mean 0.033 (SD = 0.023) for interviewers with an AD patient being higher than 0.021 (SD = 0.037) for those with Non-AD patients. While only a tendency, It can tentatively be concluded that interviewers tend to insert longer silences while interacting with AD

patients. The number of *GA* at *I–P* turn changes is significantly greater at turn exchanges with AD patients, with an average of 0.103 with a longer duration of 1.515 compared to the mean of 0.052 with a relatively shorter duration on average of 1.011 at turn exchanges with Non-AD patients. The number of *LA* is also highly predictive among the two groups in the *P–I* turn changes. It means that the frequency of initiating a new topic from *I* after a considerable amount of silence after the patient has stopped speaking is higher in the AD group with a mean goes up by 0.031 for AD and substantially low with a mean of 0.002 for Non-AD patients. Finally, it is discovered the average turn length of interviewers with the AD patients was 9.155s (SD = 4.320), while it was 23.31s (SD = 22.31) with non-AD patients, the mirror image of the case with patient turn length, where AD patients have far longer turns. This reveals that although the interviewers paused for longer periods within their turns while interacting with AD patients they also tend to speak for a shorter period of time.

Our study provides strong evidence that these interactional features including pause duration, gaps, lapse duration, presence of attributable silences, phonation time, and turn length seem to be sensitive markers of cognitive decline and also distinguish the AD group from the Non-AD group. Here, for our classification task, we use these features as input to different machine learning algorithms.

6.4.1.2 Disfluency features analysis

Patient Features: Table 6.7 shows the results of our analysis indicating a substantial difference between Non-AD and AD patient groups in terms of the rate of patient *edit terms*, *repeats*, and *substitution* per word. The rate of edit terms is significantly higher (p=0.001) for AD patients with a mean of 0.029 (SD = 0.009) compared to 0.017 (SD = 0.006) for Non-AD patients. Furthermore, the rate of verbatim repeat disfluencies is significant (p=0.011) with a higher mean value for AD patients than non-AD patients (0.027 vs. 0.011). The results also show a strong relationship (p=0.045) between conditions and substitution disfluencies, again with higher rates for AD patients vs. non-AD patients (0.012 vs. 0.008). Disfluencies are seen as a sign of communication problems. It makes sense that higher disfluencies in the language would be noticeable because AD patients frequently have weak conversation flow and other communication issues. The rate of delete disfluencies is, however, not found to be significantly different between AD and Non-AD patients, possibly due to the lack of data as they are very rare.

Interviewer Features: As with patient features, it is found that there is a significantly greater rate of edit terms in conversations with AD patients (p=0.013) with a mean value of 0.009 (SD = 0.011) compared to 0.004 (SD = 0.004) for those with Non-AD patients. The rate of repeat disfluencies (p=0.048) is also significantly greater with a mean value of 0.010 (SD = 0.008) in interviewer speech with AD patient and a mean

Feature	**AD**	**Non-AD**	***Mann-Whitney U test***	
	Mean (SD)	Mean (SD)	***p***	**U**
Patient features				
# *edit_terms*	0.029 (0.009)	0.017 (0.006)	**0.001****	183.5
# *Rpt*	0.027 (0.015)	0.011 (0.13)	**0.011***	172.0
# *Sub*	0.012 (0.007)	0.008 (0.008)	**0.045***	161.0
# *Del*	0.005 (0.005)	0.003 (0.005)	0.256	137.0
Interviewer features				
# *edit_terms*	0.009 (0.011)	0.004 (0.004)	**0.013***	170.5
# *Rpt*	0.01 (0.008)	0.007 (0.006)	**0.048***	157.0
# *Sub*	0.05 (0.006)	0.004 (0.004)	0.743	145.0
# *Del*	0.002 (0.003)	0.001 (0.001)	0.154	153.0

Table 6.7: Descriptive statistics (mean, SD) and statistical significance of the disfluency feature set. ** denotes highly significant at $p < 0.01$; * denotes significance at $p < 0.05$

.

value of 0.007 (SD = 0.006) in interviewer speech with Non-AD individuals. The rate of delete and substitution disfluencies are not found to be significantly different in interviewer speech with AD and Non-AD patients. The fact that there are more disfluencies in the interviewer's speech suggests that trouble with communication is shared between both participants, in line with the Conversation Analytic emphasis on collaborative achievement.

6.4.1.3 Dialogue act features analysis

and bigram features among the two groups. Table 6.8 shows the statistical results for DA features mentioned in section 6.3.3.2. Our observation is that the AD group is producing more clarification request to interviewers in response to either questions or statements. it is suggested that a higher number of significant clarification request features could help distinguish the patients with AD from the Non-AD group. Although the signal non-understanding was higher within the AD patients group with a mean of 2.67 vs 1.93 for Non-AD but due to high standard deviation, it was not found significant. Our analysis also showed that *I_qy* are higher with a mean of 6.67 for AD patients than Non-AD patients (3.47). It is suggested that the effective use of simple yes-no questions, choice questions, and reduced structures could be helpful with fewer communication breakdowns. On the other hand, *I_qo & I_qr* shows a trend towards significance with a higher mean of 2.20 with open-ended questions asked from the Non-AD group as compared to the AD group (mean=1.0). On the other hand, more choice questions were asked from AD patients with a higher mean of 3.33 vs 1.53 for Non-AD patients. This result makes sense that interviewers are giving more directed questions to AD patients

Feature	**AD**	**Non-AD**	***Mann-Whitney U test***	
	Mean (SD)	Mean (SD)	***p***	**U**
Unigram features *(f1)*				
# *P_br*	2.67 (4.203)	1.93(4.284)	0.512	128.5
# *P_qc*	2.60 (2.131)	1.13 (1.356)	**0.041***	161.5
# *P_qy*	1.53 (1.457)	1 (1.690)	0.161	146.5
# *P_ny*	9.47 (8.476)	4.13 (4.612)	**.0165****	169.5
# *P_no*	2.33 (2.795)	0.67 (1.047)	**0.045***	161
# *P_nn*	1.73 (1.944)	0.67 (0.976)	0.160*	152
# *P_sa*	20.87 (12.188)	40.67(27.807)	**0.05** *	65.5
# *P_b*	11.53 (7.680)	5.40 (4.222)	0.08 -	176
# *P_na*	9.47 (11.993)	9.07 (10.559)	0.902	116
# *I_br*	1.80 (2.808)	2 (4.276)	0.653	123.5
# *I_qc*	3.07 (4.877)	2.40 (1.765)	0.305	87.5
# *I_qy*	6.67 (5.260)	3.47 (3.523)	**0.033***	163.5
# *I_qo*	1 (1.069)	2.20 (1.781)	0.061-	67
# *I_qr*	3.33 (2.895)	1.53 (2.295)	**0.041***	161.5
Bigram features *(f2)*				
I_qy–P_ny	2.67 (3.352)	0.73 (1.280)	0.061-	158
I_qy^d–P_ny	3.73((3.936)	2.80 (3.448)	0.294	137.5
I_qo–P_sa	0.47 (0.743)	1.33 (1.234)	.0560-	66
I_qy–P_nn	0.47 (0.640)	0.27 (0.594)	0.398	133.5
I_qy^d–P_qc	0.07 (0.258)	0.53 (0.834)	0.202	81
I_qy^d–P_ng	0.07 (0.258)	1.00 (1.363)	0.050-	65.5
I_qw–P_sa	3.80 (4.395)	4.47 (3.270)	0.305	87
I_qw–P_qc	0.27 (0.458)	0.13 (0.352)	0.539	127.5
Confusion features *(f3)*				
question_ratio	0.065 (0.050)	0.042 (0.047))	**.041***	161.5
confusion_ratio	1.077 (1.774)	0.305 (0.419)	**.045***	161

Table 6.8: Descriptive statistics (mean, SD) and statistical significance of the DA feature set. Statistical values of *f4* group features can be found in table 6.6. ** denotes highly significant at $p < 0.01$; * denotes significance at $p < 0.05$; **-** shows a trend toward significance at $p < 0.1$

.

by giving them choices. This means that questions like 'What do you want to do?' are very open and it's difficult for people with cognitive deficits to answer and that's why people often choose choice questions by reformulating questions ('what do you want to do?') as 'Do you want to go the cinema or do you want to go to the beach?' by giving two possible questions and that are easier to answer. It was also observed that statement answers (*sa*) are negatively correlated with the AD group with a lower mean of 20.87 as compared to a higher mean of 40.67 for the Non-AD group. Similarly, simple yes answers are more common in AD patients than Non-AD patients while dispreferred answers (*ng*)

are, however, not found to be significantly different between AD and Non-AD patients, possibly due to lack of data as they are very rare. Other-answers ('I don't know') are also significant in the AD group while Negative no-answer (*nn*) shows a trend towards significance among the two groups.

A total of 462 unique bigrams are found from AD conversations and 338 bigrams from Non-AD conversations. We are only interested in bigram sequences that are questions from interviewers followed by answer types from patients that result in selecting 54 bigram sequences. No bigram sequences of interest are found to be significant possibly due to the lack of these combinations of bigram tags, however fewer bigram sequences such as *I_qy_P_ny*, *I_qo_P_sa* and *I_qy^d–P_ng* shows trends towards significance. AD patients simply reply with more 'Yes' answers to simple Yes-No questions with a mean of 2.67 than Non-AD patients(mean: 0.73). In a similar line, Non-AD patients gave more explanation with 'yes' (*ng*) to declarative yes-no questions than AD patients (mean: 1.00 vs .07). This means AD patients tend to reply simpler with simply 'Yes' answers. While in contrast, few statement answers were given in response to open-ended questions from AD patients with a mean of 0.47 as compared to Non-AD patients (mean: 1.33). For sanity purposes, only presented a few interesting unigrams (*f1*) features in the result tables below. Detailed results could be found in Appendix D Table D.1 which shows the distribution of each unigram.

Figure 6.3 shows distributions of one significant unigram feature Figure 6.3 (A) and 6.3 (B-D) represents the distribution of three non-significant feature showing a trends towards significance.

6.5 Experiments

6.5.1 Experimental Setup

Our final goal is to perform a classification task to asses whether AD prediction can be enhanced by combining these interactional features with disfluency features and DA features.Three machine learning classifiers are used to examine the impact of these features, specifically: Logistic regression (LR), support vector machines (SVM), and multilayer perceptron (MLP). Each classifier is trained using disfluency features, interactional features, and DA features, and then by combining different combinations from these feature sets. As the dataset is fairly small, separate splits splits of data are not used for train and test. To provide a more accurate estimate of generalization accuracy, leave-one-out cross-validation (LOOCV) approach is used. In this procedure, one participant is chosen as the test, the classifier is trained on the remaining cases, and the process is repeated until all instances have been chosen for testing. In the end, resulting accuracies are aggregated into a final score. We build our models using Scikit-Learn Library (Pedregosa et al., 2011). The model is optimized with following

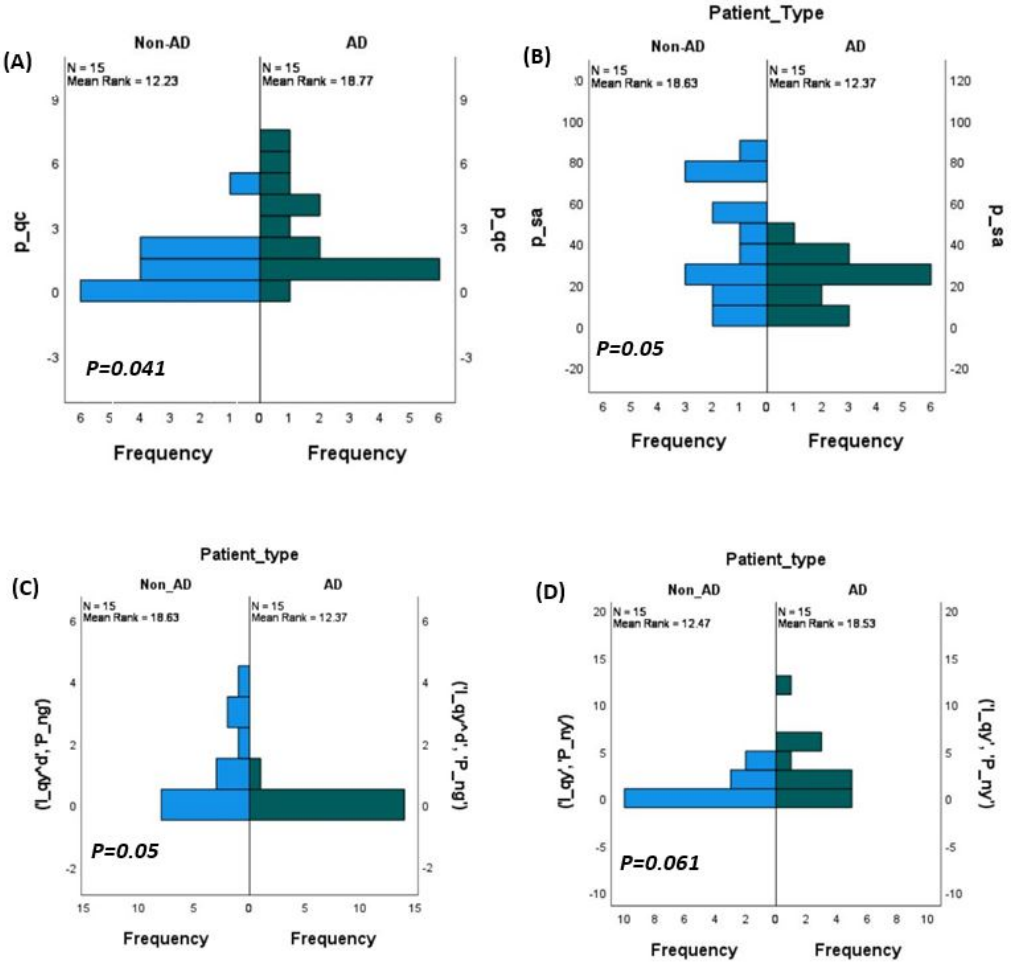


Figure 6.3: Feature value histograms for a selection of different unigrams, bigrams, showing differences in distributions between AD and Non-AD dialogues. **(A)** distribution of clarification request (*P_qc*); **(B)** Statement answers (*p_sa*); **(C)** Declarative yes-no question followed by yes plus explanation answer (*I_qy^d_P_ng*).; **(D)** Declarative yes-no question followed by simple yes answer (*I_qy^d_P_ny*).

hyper-parameters; logistic regression with $C \in \{0.001, 0.01, 0.1, 1, 10, 100, 1000\}$ using the 'liblinear' solver; SVM with $C \in \{0.1, 1, 10, 100, 1000\}$, $\gamma \in \{1, 0.1, 0.01, 0.001, 0.0001\}$, using the kernels 'rbf' and 'poly'; and MLP with the 'relu' activation function, hidden layer sizes of (2,3), and (3,4) and an initial learning rate of 0.01.

6.5.2 Feature Selection algorithm

A recursive feature elimination (RFE) method is selected on both interactional and disfluency feature sets to eliminate the weakest features with the purpose to remove

any dependencies and co-linearity. RFE is a feature selection method that removes a certain number of weak features per iteration and fits the model with the remaining features (sci, 2024). Initially, it trains the machine learning model using the entire set of features. The model assigns weights or ranks to each feature based on their importance in predicting the AD/Non-AD patient. Subsequently, the algorithm identifies and eliminates the least significant features according to these rankings. The model is then retrained on the reduced set of features, and the process iterates until a predetermined number of features is reached or until the model performance converges to an optimal level.

The key advantage of RFE lies in its ability to adapt to the characteristics of different datasets and models. By iteratively selecting and excluding features, RFE systematically navigates through the feature space, allowing it to capture the most informative features while discarding those that contribute less to the model's predictive power.Each classifier is trained with the top 15 ranked features based on RFE ranking.

6.5.3 Evaluation Metrics

Because our dataset is fairly balanced, results are reported in terms of *accuracy*, *precision*, *recall*, *F1* score, and area under the ROC curve (AUC) as evaluation metrics. Precision (true positive divided by true positive and false positive) quantifies the proportion of AD predictions that match actual cases of AD. The recall is a measure of the percentage of the actual AD occurrences that were detected (i.e. true positives divided by false negatives plus true positives). F1 is the harmonic mean of precision and recall. AUC is frequently used to assess how well clinical diagnostic and predicative models perform. (Zou et al., 2007)The trade-off between the true positive rate (TPR, recall of the AD class) and the false positive rate (FPR, 1- recall of the Non-AD class) is displayed using the ROC curve. From the very positive threshold, where every case is categorized as positive, to the very negative threshold, when every instance is classified as negative, that graph depicts the increase of the classification threshold. An AUC of 0.5 for a random classifier exists when a diagonal line is drawn from the origin (0, 0) to the target point (1, 1). Different clinical diagnostic scenarios may call for different TPR/FPR trade-offs, so the area under the curve (AUC) is used to express the overall level of diagnostic power; AUC greater than 0.75 is typically advised for clinical applications (Orimaye et al., 2017).

The AUC is calculated using the leave-pair-out cross-validation (LPOCV) to provide an informed comparison with the results from our baselines, which is a reliable technique used for unbiased AUC for clinical studies with small datasets (Airola et al., 2011; Smith et al., 2014).Every positive and negative class pair is assessed on a model trained on the remaining data, unlike previous cross-validation methods. For example for each iteration, one pair of AD and Non-AD is chosen as a test set for evaluating the model on the rest of the training data.

6.6 Experiments with interactional features

6.6.1 Baseline Model

It is aimed to measure how well our models perform in comparison to the random classifier baseline using a group of interactional features discussed in section 6.3.3.1. (Luz et al., 2018)'s work is also used and their results as the baseline with the dialogue interactions as features on the CCC dataset. Although it's difficult to compare the results directly due to the choice of different conversation sample chosen (38 dialogues vs. 30), our accuracy figures situated within a similar range, due to the same nature of the dataset and interactional features, it should give us a reasonable comparison. It is also difficult to compare these results directly to related work (Mirheidari et al. (2019)) on a different dataset (similar number of samples: 30 conversations between neurologist and patients) as they employed CA-inspired features with a combination of disfluency and acoustic features in more specific settings with a predefined set of questionnaires. While relying on features that can be more robustly derived from spontaneous speech, our accuracy scores are comparable (0.90), with a lower discrepancy between the classifications of the two groups.

6.6.2 Classification Results

Table 6.9 provides the classification accuracy measures obtained using all the extracted features. The performance is compared against all three classifier algorithms – LR, SVM, and MLP - using all dialogue features and with the top fifteen RFE feature set. It can be seen that SVM outperformed both LR and MLP for all dialogue features and with REF (top 15). Our dialogue features produced promising results in distinguishing AD from Non-AD with overall accuracy reaching 83% with the SVM classifier, showing that interactional patterns can provide salient cues to the detection of AD in dialogues. The results are further enhanced when adding with disfluency feature with an accuracy of (Acc 0.90) and F1 score of 89% suggesting that these different pause behaviours not only indicate word-finding difficulties as AD progresses but also mark disfluency and in certain situations were used to sustain social interaction as part of the compensatory language (e.g in case of attributable silences).

MLP performed in a similar manner with LR for disfluency features with the same accuracy and F1 score, however, it performs slightly worse with the dialogue features with F1 score of 76% when compared to LR and SVM. However, combing both features show an increase in accuracy of 80%. From the overall accuracy results with MLP, it can be concluded that MLP is a feed-forward neural network that is more parameters and data-hungry algorithm. Its performance is lower with a small number of samples and small feature space.

Model	Feature set	Acc.	Prec.	Rec.	F1 Score	AUC
Random Baseline	All	0.53	0.53	0.54	0.52	0.54
LR (Luz et al., 2018)	interaction	0.76	-	-	-	-
SVM (Luz et al., 2018)	interaction	0.84	-	-	-	0.89
SVM (Mirheidari et al., 2019)	CA inspired	0.96	-	-	-	-
SVM (Mirheidari et al., 2019)	All	0.90	-	-	-	-
LR	All	0.80	0.81	0.80	0.80	0.80
SVM	All	**0.83**	**0.83**	**0.83**	**0.83**	**0.87**
MLP	All	0.80	0.77	0.76	0.76	0.79
SVM	RFE (15)	**0.90**	**0.90**	**0.90**	**0.90**	**0.96**

Table 6.9: Comparison of results for the AD classification with three classifiers with LOOCV will all dialogue features.

Model	Method	Accuracy	AUC
LR	LOOCV	0.80	0.80
	LPOCV	0.80	0.80
SVM	LOOCV	0.83	0.87
	LPOCV	0.83	0.83

Table 6.10: Comparison between LOOCV and LPOCV

The corresponding ROC curve (AUC 0.87) is shown in Figure 6.4. It can be seen that these interactional features over dialogues had the effect of improving classification trade-offs between true positive and false positive rates, peculiarly reducing the false positives while increasing the true positives.

At the end, most discriminated features are ranked using RFE and this way, from the 34 features only 15 were retained (figure 6.5). It can be clearly seen from the figure that most of the interactional features including duration of attributable silence, duration of lapse either from *P–I* or *I-I*, number of gaps, number of laps, total duration of lapse, *I* turn length, *P* turns lengths are among the top 15 ranked features.

The intuition behind the *AS* confirms the findings of Levinson (1983) about attributable silences and aligned with conversational analysis studies that in order to avoid making a dispreferred response, people with cognitive impairment prefer silence over alternative forms of communication. Among the other useful features, not only number of gaps and lapse are found important but also the durations of gaps and lapse are

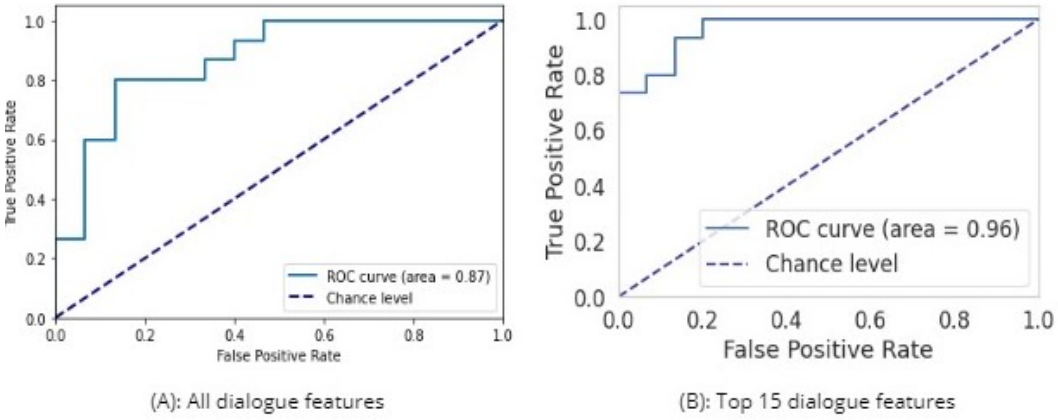


Figure 6.4: ROC curve for SVM classifier.

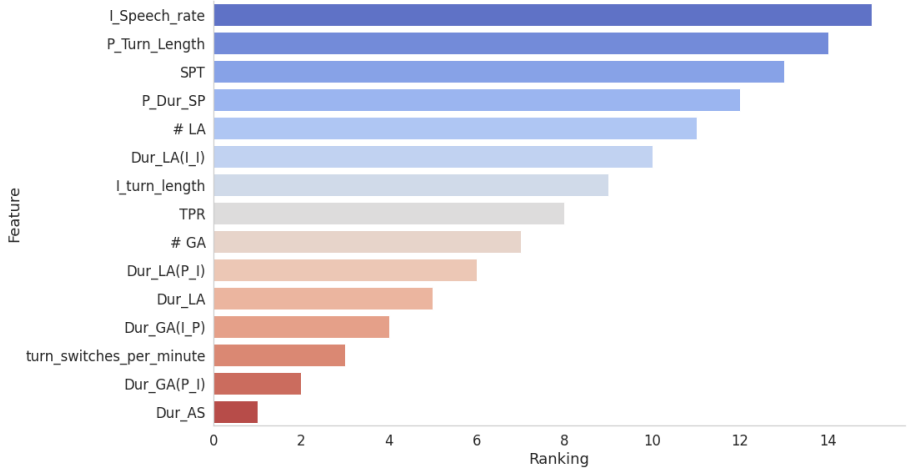


Figure 6.5: Feature ranking.

observed differently in both groups. Interestingly turn switches per minute, patient turn lengths, and standardised phonation time are negatively correlated with AD patients with higher mean values for Non-AD. That means Non-AD patients turn switches more frequently in conversations and have longer turn lengths as compared to AD individuals.

6.6.3 Error analysis

The results in Table 6.11 show that the SVM model with dialogue features attained the highest scores for both Non-AD and AD classes with all the feature sets included and also with considering only significant features ranked by RFE. The model obtains F1 scores of 0.90 for AD and 0.90 for Non-AD with only the top 15 ranked dialogue features compared to the model which attains an F1 score of 0.83 vs 0.84 for AD and Non-AD respectively with all features included. Although a higher recall value is obtained for

LR and MLP with the AD class, however, we got a fairly balance F1 score for both AD and Non-AD with all three classifiers. An F1 score (0.81 vs 0.79) is achieved with LR, (0.80 vs 0.72) with dialogue features and MLP, and a much higher F1 score (0.83 vs 0.84) for AD and Non-AD class respectively. False positive or false negative AD detection will depend on the application the model is utilized for, however as it stands, our dialogue feature set significantly decreases the false negatives of diagnosis while only slightly reducing the false positives.

Model	**Class**	***Pre.***	***Rec.***	***F1 Score***	***Accuracy***
Baseline	AD	0.67	0.53	0.59	0.53
(all)	Non-AD	0.40	0.55	0.46	
LR	AD	0.76	0.87	0.81	0.80
(all)	Non-AD	0.85	0.73	0.79	
SVM	AD	0.86	0.80	0.83	**0.83**
(all)	Non-AD	0.81	0.87	0.84	
MLP	AD	0.70	0.93	0.80	0.77
(all)	Non-AD	0.90	0.60	0.72	
SVM	AD	0.93	0.87	0.90	**0.90**
RFE(15)	Non-AD	0.88	0.93	0.90	

Table 6.11: Results of AD classification task with dialogue features only.

6.7 Experiments Combining dialogue features with disfluency features

Our next goal is to perform a classification task to asses whether AD prediction can be improved by combining these interactional features with disfluency features. The classification task is performed with the classifiers discussed in section 6.5.1 and with the same parameter for each classifier. The results are presented first for each feature set individually and then by the combined feature set.

It can be seen in table 6.12 that, SVM outperformed both LR and MLP for disfluency features, dialogue features, and a combination of both. Comparing the two feature sets, the best accuracy scores attained (with SVM) are equivalent with an accuracy of 83% and an F1 score of 83%. However, combining the two feature sets, the model got the highest accuracy of 90% with an F1 score of 0.89 with the SVM classifier. With LR, the model achieved an accuracy of 76% with disfluency features, 80% with dialogue features, and an increase in accuracy of about 6% when combing both feature sets with an accuracy value of 86%.

MLP performed in a similar manner with LR for disfluency features with the same accuracy and F1 score, however, it performs slightly worse with the dialogue features

Model	Feature set	*Accuracy*	*Pre.*	*Rec.*	*F1 Score*	*AUC*
LR	disfluency	0.76	0.77	0.76	0.76	0.74
	dialogue	0.80	0.81	0.80	0.80	0.80
	both	0.86	0.87	0.86	0.86	0.84
SVM	disfluency	0.83	0.82	0.83	0.83	0.85
	dialogue	0.83	0.83	0.83	0.83	0.87
	both	**0.90**	**0.90**	**0.89**	**0.89**	**0.89**
MLP	disfluency	0.76	0.77	0.76	0.76	0.75
	dialogue	0.80	0.77	0.76	0.76	0.79
	both	0.80	0.81	0.80	0.80	0.81

Table 6.12: Comparison of results for the AD classification with three classifiers with LOOCV.

with an F1 score of 76% when compared to LR and SVM. However, combing both features show an increase in accuracy of 80%. From the overall accuracy results with MLP, we can draw a conclusion as MLP is a feed-forward neural network that is more parameters and data-hungry algorithm. Its performance is lower with a small number of samples and small feature space.Due to the small quantity of training data we have, overfitting is very likely.

It is also worth examining the ROC AUC as it evaluates the different classifiers at different true positive rates and false positive rates. Figure 6.6(A) shows the ROC curve for the disfluency features with the SVM, with AUC 0.85, and with TPR 0.87 and FPR 0.20 at the chosen trade-off point. This trade-off point is chosen as it gives maximum accuracy.

The results are further enhanced when adding these disfluency language features with dialogue features reaching an accuracy of 90% and an *F1 score* of 0.90. These results suggest that different pauses behaviour not only indicate word-finding difficulties as AD progresses but also mark disfluency– in certain situations showing these were used to sustain social interaction as part of the compensatory language (e.g. in the case of attributable silences). The corresponding ROC curve is shown in Figure 6.6(B) with AUC 0.87, and the chosen trade-off between TPR and FPR (0.80 vs 0.13) with dialogue-only features. It can also be seen in Figure 6.6(C) that the overall classification performance was improved by merging these interactional features with disfluency features over dialogues to AUC=0.89, and improving trade-offs between true positive (0.93) and false positive rates (0.13), minimising the false positives while increasing the true positives.

Table 6.13 reports the top 15 discriminative features from the combined feature set using REF.

Luz et al. (2018) use a probabilistic graphical model to classify AD patients in the CCC, using a slightly bigger dataset but with shorter dialogue conversations. They use only interactional features, and achieve comparable accuracies of 0.757 with LR and 0.837 with SVM classifiers, but do not investigate the role of different pause types, or

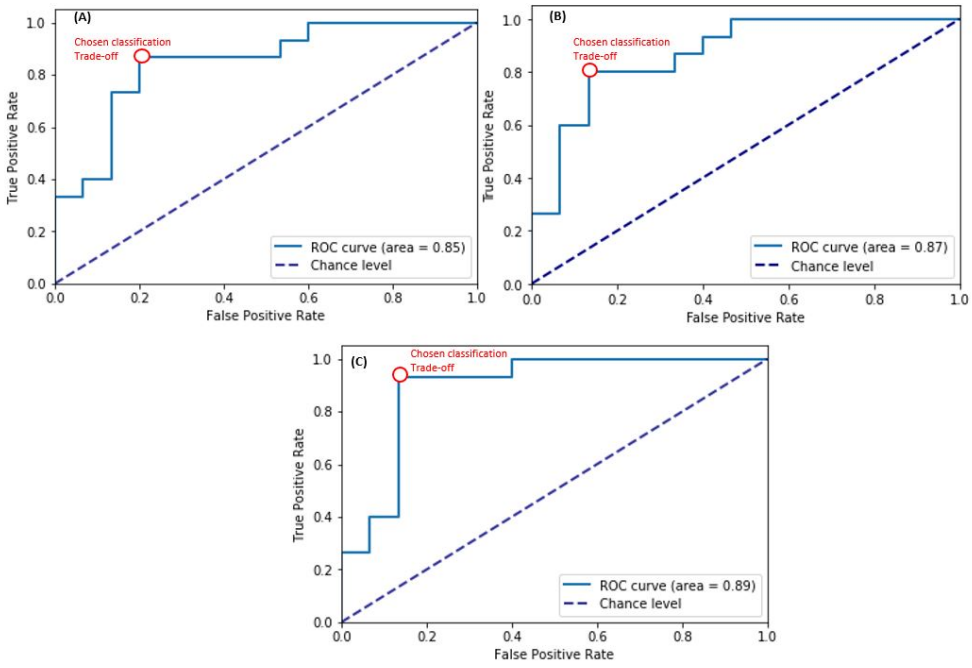


Figure 6.6: ROC curves for SVM classification experiments with (A): disfluency features, (B): interactional features, (C) combined feature set. The red bubble shows the chosen trade-off point for classification experiment results in Table 6.12.

the combination with fluency. Interestingly, they found that AD patients produce longer turns with more words and a higher speech rate; this contrasts with our results, in which AD patients produce fewer words than Non-AD patients, with lower speech rates (See Table 6.14). It is note that our findings align better with other research (Pistono et al., 2019b; Kavé and Dassa, 2018; Themistocleous et al., 2020; Martínez-Sánchez et al., 2013). Themistocleous et al. (2020) reported a higher speech rate in the healthy control group than in the MCI group while Martínez-Sánchez et al. (2013) have shown a lower speech rate in AD than as compared to the control group. Lira et al. (2014) performed a cross-sectional study and showed that less words were said by AD patients than by healthy people, although there was no difference between groups with mild and moderate AD in this regard. Mirheidari et al. (2019) go a step further, combining CA-inspired interaction features including turn-taking behaviour with some acoustic and language features, to achieve a classification accuracy of 90% similar to this study. However their approach is based on structured interviews with chosen topics and question types, in more clinical settings, and the use of features that directly target particular aspects of this structure (e.g. responses to particular setting-specific questions).

Features	Type	Ranking
Dur_AS	Interactional	1
turn_switches_per_minute	Interactional	2
Dur_LA	Interactional	3
Dur_LA (P-I)	Interactional	4
#GA	Interactional	5
TPR	Interactional	6
P_RPT	disfluency	7
I_turn_length	Interactional	8
Dur_LA (I-I)	Interactional	9
# LA	Interactional	10
I_edit_terms	disfluency	11
P_edit_terms	disfluency	12
SPT	Interactional	13
P_Turn_Length	Interactional	14
I_Speech_rate	Interactional	15

Table 6.13: Top 15 ranked features including disfluency and interactional features by RFE.

Measure	Luz et al. (2018)		Nasreen et al. (2021c)	
	AD	Non-AD	AD	Non-AD
Speech rate (sylb per min)	168 (35.6)	180.8 (28.4)	164.9 (35.7)	180.1(37.8)
avg. words per min	166.5	155.9	150.7	176.25
Norm. turn duration	4.1	3.0	5.91	4.01
turn duration	255.8	97.3	433	428
Avg. no. of words	742.5	314.6	895.6	1186.6

Table 6.14: Comparison of our approach with Luz et al. (2018)'s work based on certain measures.

6.7.1 Error Analysis

The results in Table 6.15 show that the SVM model with disfluency and interactional features attained the highest F1 score, recall, and precis on for both Non-AD and AD classes; both classes are shown in order to provide a measure of both sensitivity (recall of the positive AD class) and specificity (recall of the non-AD class), standard measures for diagnostic tests. Note that due to the small dataset, differences between modes are indicative rather than statistically significant - see confidence intervals in Table 6.15. The model achieves F1 scores of 0.90 for both the AD and the Non-AD classes. Combining the disfluency features with interactional features particularly improves the recall of the AD class (i.e. improves the sensitivity of the classifier): the SVM model with both feature sets has a recall of 0.93, improving over using disfluency features alone at 0.87 and over the 0.80 achieved with interactional features. The specificity (recall for the

Model	Class	Pre.	Recall	F1 Score	Accuracy	95% CI
SVM (disfluency)	AD	0.81	0.86	0.83	0.83	0.70-0.96
	Non-AD	0.85	0.80	0.82		
SVM (dialogue)	AD	0.86	0.80	0.83	0.83	0.70-0.96
	Non-AD	0.81	0.87	0.84		
SVM (both)	AD	0.93	0.87	**0.90**	0.90	0.79-0.99
	Non-AD	0.92	0.86	0.89		

Table 6.15: Results of AD predictions with both disfluency and dialogue features with SVM classifier.

non-AD class) was lowest when using disfluency features only at 0.80, significantly lower than the 0.87 achieved by both using dialogue features alone and combining both feature sets. A balanced F1 score for both the AD and Non-AD classes with all three combinations was achieved overall with our chosen threshold (0.84 vs 0.83 for disfluency features, 0.83 vs 0.84 with interactional features, and 0.90 for the combined feature sets). Higher sensitivity or greater specificity for AD detection will be more or less beneficial depending on the application the model is utilized for and this can be achieved in line with the AUC results shown in Fig. 6.6, but as it stands using the combined feature set considerably increases the sensitivity of AD diagnosis over the most sensitive single feature set classifier (disfluency features) whilst maintaining a high specificity on par with that achieved using dialogue features.

It can be observed the confusion matrices of predictions of the SVM Model with disfluency, interactional, and combining both in Figure 6.7 which show the influence of (A) and (B) on (C).

6.8 Experiments with DA features

The Linear SVM classifier model is chosen to classify the instances of AD and Non-AD. The z-score normalization is applied to of all features including unigram, bigram, and confusion ratio features. Here, we opt for the cross-validated, recursive feature elimination (RFECV) feature selection method. By employing recursive feature elimination to remove 0 to N features (where N is the number of features), REFCV chooses the optimal subset of features for the chosen estimator. The best subset is then chosen based on the model's cross-validation score. Recursive feature elimination removes n features from a model by repeatedly fitting the model and deleting the weakest features at each iteration.

A recent study (Li et al., 2022) has been using Dialogue act tagger along with other modalities including speech, language and interaction however they have not published the results yet. As a baseline, a random classifier is used with unigram features only.

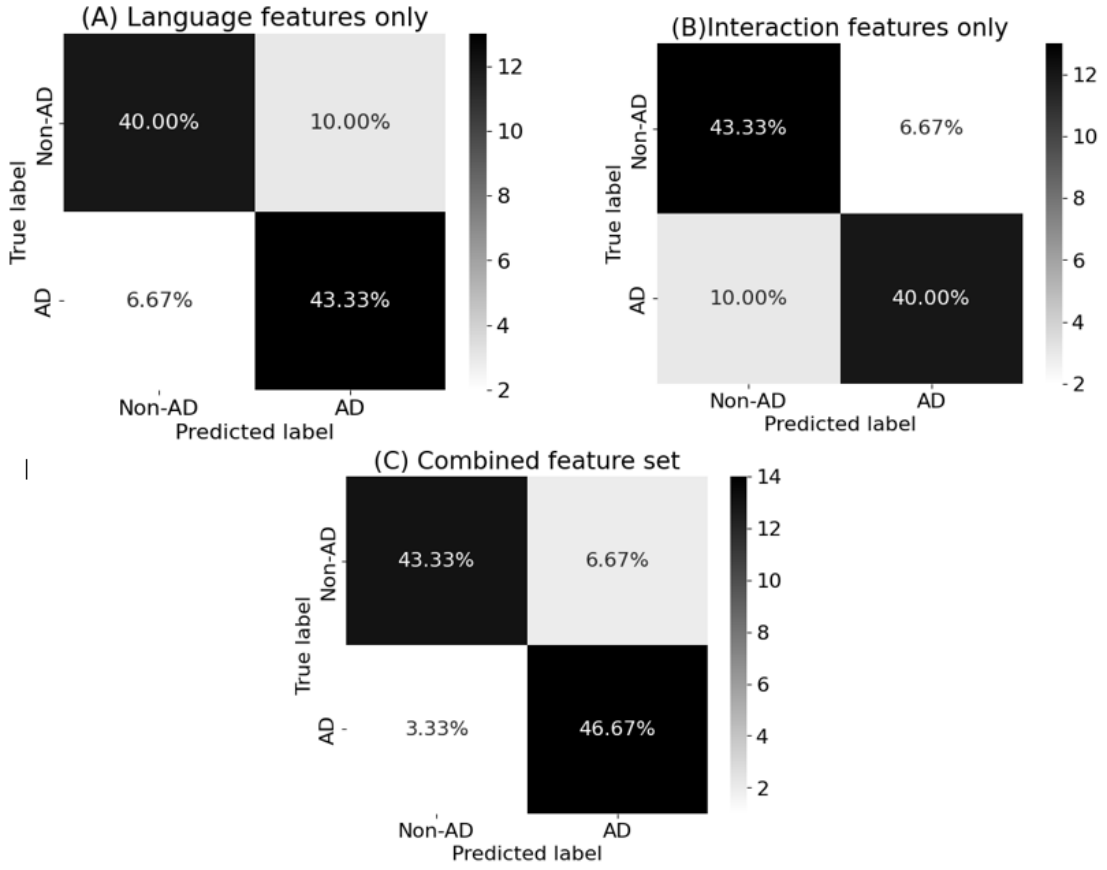


Figure 6.7: Confusion matrices for AD classification task with different feature sets.

The model got an accuracy of 73% with all unigram features with the same recall of 73% for both AD and Non-AD classes. However, using RFE-based chosen features with the pipeline, the model got 87% accuracy with 31 unigrams selected during REFCV. The recall for the AD class is much higher than Non-AD class (0.933 vs 0.800). The accuracy is dropped when using only the bigram sequence with a recall of 0.733 for AD and 0.467 for Non-AD class. Combining unigram with Bigram and confusion ratio improve the accuracy by 16% with a higher recall of 0.800 for the AD class than Non-AD class (0.7333). A performance increase is obtained by combining unigram features and confusion ratio features (*f3*) over when including bigram features. It is found that unigram DA tags along with confusion ratios help as features in AD detection. Adding confusions features to the unigram features led to improvement (ACC 0.933 vs 0.867) for AD and Non-AD class. Unigrams with confusion ratios and unigrams alone outperformed all other combinations. The results with Bigram sequences are not very high contrary to our expectations possibly due to the reason that the sequences in which our interest lies, based on CA studies are very rare in this dataset.

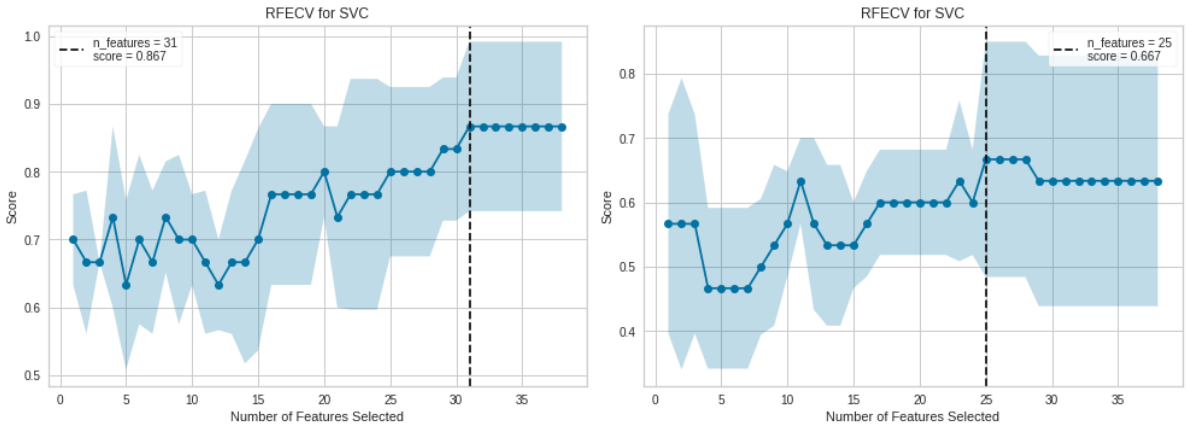
Model	Class	Precision	Recall	F1 Score	Accuracy	95% CI
Random (baseline)	AD	0.500	0.467	0.483	0.500	0.194 - 0.539
	Non-AD	0.500	0.533	0.516		
SVM (All *f1*)	AD	0.824	0.733	0.733	0.733	0.575 - 0.892
	Non-AD	0.733	0.733	0.733		
SVM *f1 (31)*	AD	0.824	0.933	0.875	**0.867**	0.745 - 0.988
	Non-AD	0.923	0.800	0.857		
SVM *f2 (25)*	AD	0.579	0.733	0.647	0.600	0.425 - 0.775
	Non-AD	0.636	0.467	0.538		
SVM *f1+f3*	AD	0.824	0.933	0.875	**0.867**	0.745 - 0.988
	Non-AD	0.923	0.800	0.857		
SVM *f1+f2+f3*	AD	0.750	0.800	0.774	0.767	0.615 - 0.918
	Non-AD	0.786	0.733	0.759		
SVM Significant *f1+f2+f3*	AD	0.800	0.800	0.800	**0.800**	0.657 - 0.943
	Non-AD	0.800	0.800	0.800		
SVM *f1+f2+f3+f4*	AD	0.684	0.867	0.765	0.733	0.575 - 0.892
	Non-AD	0.818	0.600	0.692		
SVM Significant *f1+f2+f3+f4*	AD	0.750	0.800	0.774	0.767	0.615 - 0.918
	Non-AD	0.786	0.733	0.759		

Table 6.16: Results of AD classification task with SVM classifiers with unigram (*f1*), bigram (*f2*), confusion ratios (*f3*) and interactional feature (*f4*) feature sets, using LOOCV, with 95% confidence intervals (CI).

Using only the unigram DA set yields a better AD prediction than using bigrams or combining with confusions and bigrams (Acc scores 0.867 vs. 0.733 vs. 0.767), making it the most informative set of features for AD classification (selected based on feature selection).

The model with significant features from unigram, bigram DA's, and interactional feature set combined with confusion ratios achieves a fair accuracy score of 73% with a higher recall of 0.867 for the AD class than Non-AD class (0.600). At a per-class level, performance for the AD class exceeded that of the Non-AD class. However, using RFECV within a pipeline and using significant features improves the overall accuracy to 77% with a decrease in recall of AD (0.80) and an increase in Non-AD class (0.73).

Effect of Unigram DA tags it is found that unigram tags help as features in AD detection. Among the unigrams, patient's yes-answers, no-answers, backchannels, patient's 'I don't know answers and clarification requests are among the most important ones. our model got an accuracy of 73% using all unigram DA's while an increase in performance of 13% with 31 unigrams chosen using feature selection (See Figure 6.8).



((a)) Unigrams **((b))** Bigrams

Figure 6.8: REFCV feature selection with Linear SVM with (a) unigram and (b) bigram feature sets.

Effect of Bigram DA tags Only DA bigrams with a between-class training set frequency difference of more than 0 were kept after filtering, deleting those bigram sequences for which there is not a single instance for any of the patients and only 38 bigram DA sequences are left. Among the chosen 25 bigrams DA's *('I_qy', 'P_ny'), ('I_qo', 'P_sa'), ('I_qw^d', 'P_br'), ('I_qw^d', 'P_sa'), ('I_qw', 'P_qc')* and *('I_qw', 'P_br')* are some of the top ranked bigram sequences. This seems to include some sequences which are identified in our earlier corpus study work in chapter 3 such as the way questions are responded with signal non-understanding and clarification requests from the AD group than other group. Here, it can be seen that open questions such as *qw* and *qw^d* are responded to with signal non-understanding indicating that the patient responds with kind of non-understanding signal or seeking some sort of clarification using *qc* requests. It turns out that our classifier does not give us very good accuracy with these bigram sequences but it might be because these sequences are rare. But it is interesting that these findings fit with our corpus study findings and findings from literature (CA studies).

Effect of combining DA's with Interactional features It is a bit surprising that adding more features does not really seems to be helping the classifier's performance. It may be due to the small dataset with 30 occurrences of patients. There is a point when adding more feature did not really helps in improving the performance.

Effect of feature selection Overall, using RFECV with a pipeline and using an optimal number of features as a minimum number of features gives higher performance

than using all features. Using the optimal number of features through RFECV with unigram gives an accuracy of 87% over using all unigrams with an accuracy of 73%. Choosing optimal features with unigram (*f1*), bigrams (*f2*), and confusion ratios (*f3*) gave 87% and without using feature selection, the classifier performance is lower (77%). With all four types of feature and feature selection, the accuracy is though low than the previous combination but still higher than using all features from these four feature sets.

Features	**Type**	**Importance**
P_nn	Unigram	0.324
P_ny	Unigram	0.218
P_qw^d	Unigram	0.216
Total_No_GA	Interactional	0.160
I_P_Avg_GA_Dur	Interactional	0.146
P_b	Unigram	0.141
I_qr	Unigram	0.130
I_qy	Unigram	0.124
I_P_Avg_Dur_AS	Interactional	0.123
P_no	Unigram	0.122
P_qw	Unigram	0.105
I_^g	Unigram	0.099
I_Speech_rate	Interactional	0.095
I_qw^d	Unigram	0.081
P_I_Avg_Dur_LA	Interactional	0.079

Table 6.17: Top 15 important features including unigrams and interactional features.

6.9 Summary of the research questions investigated:

Some of the research questions that this study aimed to investigate in this chapter are:

Q 1: **What kind of dialogue features turns out to be the most prominent features that can aid in the prediction of Alzheimer's?**

We study features of interaction in clinical interviews in relation to AD. A conversational CCC dataset is used that includes interactions between the patient and interviewer. It is found that among the 34 dialogue features investigated, 14 are statistically found significant among the AD and Non-AD. It is also confirmed this with a classification experiment and got an overall 83% accuracy with all dialogue features and 90% with the top 15 ranked dialogue features. It is found that the more an individual pause within the utterances and hold the floor for less time with smaller turn lengths, the higher the chances of AD. Our results also showed

that AD patients used attributable silence as a kind of dispreferred action. AD individuals tend to produce more repeats and repairs with substitutions than non-AD individuals. Lapses either from *I–I* or *P–I* indicate the minimal response from the patient and initiation of a new topic. Lapse durations are higher as compared to gaps. Other than these general dialogue features, more specific DA unigram, confusion ratios, and bigram sequences containing the meaning-coordination *qc* and *br* also show fairly good accuracy. All these findings suggest that the combination of these features holds the predictive power to identify AD.

Q 2: **Do any of the dialogue features fit in with the observations found in literature such as attributable silences, turn lengths, turn switches, pause rate, and speech rates?**

Attributable silences are considered a sign of dispreferred action in communication (Levinson, 1983; Wang, 2019). We found many situations where interviewers faced dispreferred responses from patients in response to questions. The mean normalized duration of *AS* in the case of the AD group is 2.468 seconds as compared to 0.414 seconds in the Non-AD group. Turn switches and turn lengths are higher in the Non-AD group ((Mirheidari et al., 2019; Aldeneh et al., 2019). AD individuals have slower speech rates as compared to other groups (Boschi et al., 2017; Tóth et al., 2018; Pistono et al., 2019a). Standardised phonation time and transformed phonation rate were also found significant among the individuals of two groups and also confirm the findings of literature (Tóth et al., 2018; Beltrami et al., 2018).

Q 3: **Does the functional division of silences into short pause (SP) and long pause (LP) within the same speaker and silences at turn changes like gaps and lapses contribute in improving the accuracy of predicting AD from Non-AD?**

This categorization of silences into short and long pauses and in turn changes into gaps and lapse gives us an insight that they all present different functionality within the conversations. It is observed different patterns of silence particularly with long pauses and lapse and gaps. AD patients produce more longer pauses within the sentence boundary showing lexical retrieval difficulty. Similarly, the duration of lapses in AD conversations is longer than in Non-AD conversations. This means there are many situations where the conversational partner initiates a new topic after receiving a continuous minimal response. Some conversation analysis studies highlighted the importance of silences Mirheidari et al. (2019); Gayraud et al. (2011); Perkins et al. (1998) and the short and long pauses and silences at turn changes (Davis and Maclagan, 2010; Ramanathan, 2013). Distribution of gaps and lapses were also found different among the AD and Non-AD patients (figure 6.1). Total number of gaps, the average duration of gaps, the number of lapse, the

average duration of laps, and gaps from turn transition from *I–P* and *P–I* were positively correlated with AD (Table 6.6). The number of gaps, number of lapse, and duration of lapse are also ranked in the most discriminating features among the entire set of features (figure 6.5).

Q 4: **Does the more specific dialogue acts feature with unigram and bigram sequences hold the predictive power to identify AD symptoms?**

The unigram DAs and a few instances of bigram DA sequences are examined as dialogue features and found these sequences useful in the prediction task. These features were created to capture different facets of patterns of global interaction that other feature groups could have missed. Among the unigram DA's, This study finds patient's simple *yes-answers*, *no-answers*, *I don't know* answers' with interviewer's choice questions, yes-no questions, and tag questions among the most effective ones. Confusion features including question ratios and confusion ratios were also found helpful when combined with unigram features. Moreover, we have also tested these DA sequences with the more general dialogue features discussed earlier in this chapter and found the combination is useful in the task. Farzana et al. (2020) used ISO standard DA tagset to annotate the conversations from the DementiaBank dataset and used as features to predict AD from healthy control (Farzana and Parde, 2022). Li et al. (2022) proposed a novel AD detection architecture consisting of two major modules: an ensemble AD detector with speech, language, interaction, and DA tagging and a proactive listener based on DA tagging. These recent studies indicate that DA tags are being used as a measure of interaction and could be useful and helpful in AD prediction tasks. Upstream diagnostic or assessment applications may be able to benefit from these more generic properties rather than retraining models for new tasks, which has fascinating implications.

Q 5: **Was the combination of dialogue features with disfluency features helpful in improving the accuracy of classification among AD and Non-AD?**

The choice of combining interactional aspects of communication with disfluency features was inspired by research in literature that combines various aspects of communication, either lexical, semantic, and syntactic aspects or combining features from different modalities. Disfluency features at its own produce fairly good accuracy results. Rohanian et al. (2020b) got 72% using disfluency features, Fraser et al. (2016a) got 81% with top ranked language features.The disfluency features and the dialogue features are combined. Promising results (90% accuracy) are achieved by combining the language aspects of communication with interactional aspects of communication suggesting that disfluency features combined with inter-

actional features can serve as strong predictors for AD identification and could be integrated into clinical assessments through natural dialogues/conversations.

6.10 Conclusion

This chapter focused on features based on these DA tags and other interaction traits to explore the usefulness of interaction patterns as content-free features in less structured dialogue conversations. Our study have presented different NLP techniques on conversational dialogue features in Carolina's conversation collection. Disfluency features and interactional features are utilized with DA features from the dialogues of AD and Non-AD patients.A statistical analysis of dialogue features is performed and find out most of the interactional features including attributable silences, gaps, lapses, turn lengths, turn switches per minute, etc as sensitive cues in discriminating AD patients from Non-AD patients. It is also observed that in natural conversation not only the patient's conversation characteristics are affected but also several patterns can be observed while interacting with interviewers or carers in real settings. These interactional features are less dependent on the content of a conversation due to not relying on specific tasks like picture description task or particular questions. Our experiments showed that these interactional features have identifiable patterns that can be a strong predictor of AD detection. This study also attempted to classify AD using machine learning classifiers combing disfluency, dialogue interaction features, and DA-based features.The highest accuracy is achieved with disfluency-only features (Acc 83%) with the SVM classifier, with dialogue features it attains 83% accuracy with SVM and with significant features and accuracy of 90%. The Method obtains an overall accuracy of 90% with an F1 score of 89% with both disfluency features and interactional features with the SVM classifier. However, promising results are achieved by combining the language aspects of communication with interactional aspects of communication suggesting that disfluency features combined with interactional features can serve as strong predictors for AD identification and could be integrated into clinical assessments through natural dialogues/conversations. These interpretable interaction features including DA-based features in cognitive health screening tasks show promising performance in AD detection.

It is also intended to include more linguistic markers related to the severity of AD in this study. We want to use idea density and the number of ideas stated to take a more principled approach. At the dialogue feature level, it is further planned to include dialogue act level tags that will provide interpretation of a speaker's intent at the utterance level, different tags for questions, answers types, clarification requests, and signal misunderstanding, sequence of DA tags with language modeling to predict the disrupted communication patterns in natural conversations with AD.

CHAPTER 7

Shifting towards Multimodal Alzheimer's Disease Detection with feature Exploration: Linguistic, Acoustic and Interactional behavior based descriptors

In this chapter, techniques are developed for automatically detecting Alzheimer's disease (AD) in conversational interactions, encompassing speech and transcripts In chapter 6, Analysis was conducted on interactional features by integrating them with disfluency features and dialogue act-based features. Specific behaviors, such as seeking clarification, reiterating ideas, pausing at different conversation levels, and prompting conversational partners through clarification requests, back-channeling, or topic shifts, are examined. In this connection, dialog act sequences can be used to identify dementia through their potential to capture significant interaction patterns. In this chapter, it is aimed to add acoustic features from the speech of AD/Non-AD patients with lexical information from dialogues and combine this combination with previously explored interactional features and DA sequences. Two models are proposed: one that immediately incorporates features by concatenating them after feature extraction, providing the model with the concatenated feature vector as input, and the second that gathers unimodal judgments from various standard classifiers, one for each of the text and audio modalities, and then merges the results using a late fusion (LF) method for the final result prediction. Additionally, a multimodal fusion-based deep learning model is presented that uses simultaneous use of speech transcriptions and acoustic features to determine whether a speaker in a natural conversation has AD.s

7.1 Background

The most common type of dementia, Alzheimer's disease (AD) is an irreversible brain disorder associated with a progressive loss in people's cognitive abilities. The use of spontaneous speech to derive pathologically appropriate biomarkers for AD detection has therefore become a focus of research. State-of-the-art studies have proven that AD is identifiable from spontaneous speech (Luz et al., 2018), using both speech and transcripts (Li et al., 2022). Language production's quantitative and qualitative components are described using linguistic variables, for example via the decline in lexical-semantic abilities, word comprehension, verbal fluency, and syntactic processing for particular kinds of tasks such as picture description Boschi et al. (2017); Fraser et al. (2016a). AD-related changes can also affect acoustic features of speech, suggesting that speech analysis could provide measures of early disease progression (Lin et al., 2020). In a study of Lin et al. (2020), jitterDDP and shimmerLocal were found to be most significant towards incident Dementia. Several studies have used language-independent acoustic features only, achieving comparable accuracy to linguistic approaches (Weiner et al., 2018; Chakraborty et al., 2020); AD patients can show patterns of frequent hesitation, longer pauses, lower articulation and speech rates, and lower floor control ratio (Beltrami et al., 2018). Other acoustic features including prosodic, energy-based, spectral, and spectral aspects (jitter, shimmer, harmonics-to-noise ratio, Mel-frequency cepstral coefficients (MFCCs) can also correlate with AD (Ning and Luo, 2020). Ambrosini et al. (2019) revealed a 73% rate of accuracy in identifying MCI from spontaneous speech when using specific acoustic features such as *pitch*, *voice breaks*, *shimmer*, speech pace, and *syllable duration*. HNR was discovered to be a more sensitive gauge of vocal function, with a considerable reduction of HNR evident in older speakers. This research was done to differentiate between vocal alterations that occur with normal aging and those that are linked with disease (Ferrand, 2002).

Usually, there exist studies within language tasks in specific domains or in conversational dialogue (Mirheidari et al., 2019; Luz et al., 2018). Luz et al. (2020) presented the first criterion in the task-specific context using speech recordings and transcripts of participant descriptions of spoken pictures prompted by the Boston Diagnostic Aphasia Exam's Cookie Theft picture (Goodglass et al., 2001). In this ADReSS challenge, they used different feature sets based on energy, Mel-Frequency Cepstral Coefficients (MFCC), fundamental frequency (F0), and so on, as well as their statistical functionals achieving an accuracy of 0.625 with decision trees and 0.563 with SVM. Modeling multimodal input for AD detection has also been studied, Campbell et al. (2020) examined two fusion strategies with linguistic features and acoustic features, achieving 75% accuracy. Shah et al. (2021) used a weighted majority-vote ensemble technique for classification and selected the three top-performing acoustic models along with the language model that performed the best, resulting in a 83% prediction accuracy. Rohanian et al. (2020a)

used a deep learning multimodal model with fusion strategy with gating using lexical, acoustic, and disfluency features achieving an accuracy of 0.792 on AD classification.

Some other work focuses less on the individual and more on the properties of their interaction with others. Conversation Analysis (CA) studies show that dialogue with Dementia has characteristic features that would be missed if analyzing only individual speech (Elsey et al., 2015; Varela Suárez, 2018), but these studies are generally qualitative and/or small-scale. Some computational work on dementia is starting to fill this gap, focusing on interaction patterns such as turn-taking behavior, disfluency, repair, repetition, and topic management. Luz et al. (2018) use dialogue interaction features from the speech in a predictive model, with an impressive accuracy of 86%. . Mirheidari et al. (2019) go a step further, combining CA-inspired interaction features including turn-taking behavior with some acoustic and language features, to achieve a classification accuracy of 90%. In a recent study, Li et al. (2022) proposed a novel diagnosis architecture consisting of an ensemble AD detection module (including language, disfluency, acoustic and interaction features) and a proactive listener module usable in the dialogue system of conversational robots for healthcare. The proactive listener module uses dialogue act tagger along word extractor to generate responses based on DA tags ('statement', 'question', 'answer') and generate responses accordingly. A corpus analysis was performed by Farzana et al. (2020) using DA tagging on the DementiaBank dataset. In a similar study, Farzana and Parde (2022) combined DA features with some interactive features on two tasks including a picture description task and verbal fluency from Pitt Corpus. They achieved an accuracy of 0.79 with all interaction, N-gram DA features, and ratios features.

This use of interaction cues has the potential to be more versatile in AD prediction and monitoring in more daily life settings than individual language tasks (Addlesee et al., 2019). However, work so far either looks only at interaction rather than combining it with other modalities (e.g. (Luz et al., 2018)) or relies on particular interactional settings such as interviews with chosen topics or question types (e.g. (Mirheidari et al., 2019)) or more task-specific settings (e.g picture description task (Farzana et al., 2020)).

Building on all of this, in this thesis so far we did: In chapter 3, an initial investigation was performed on CCC corpus vs a more generic SwDA corpus to see the differences and how the different occurrences of DA are different in AD and Non-AD patients. Based on findings, we build a DA tagger in chapter 4 to automatically tag the utterances in conversations with DA's based on our tagset. In chapter 5, The emphasis was specifically placed on enhancing the class-wise accuracy of these rare class Dialogue Acts (DAs) that are deemed noteworthy and distinctive in the majority of Conversation Analysis (CA) studies. In chapter 6, temporal aspects of communication are explored with dialogue interactions and in combination with disfluencies found in natural language. A comprehensive analysis is performed using interactional features (including different pauses types, speech rate, floor control ratio, pause rate, turn lengths, etc.) with disfluency features and got an

accuracy of 90%. In a follow-up study, performed dialog act-based conversation analysis along with confusion and interactional features for the AD classification with an accuracy score of 0.80 (Nasreen et al., 2021a). This work was based on our initial investigation in a corpus study on semi-structured conversations on the CCC dataset.

In this study, the issues of combining features from different modalities is discussed, using a combination of dialogue interaction, lexical and acoustic features, and analysing semi-structured interviews obtained in more natural settings. Both early fusion (feature-based) and late fusion (decision-based) fusion techniques will be used. Early fusion involves concatenating features to combine them as soon as they are extracted, then feeding the resulting feature vector to the classifier. One classifier for each modality is used by the late fusion classifier to create unimodal decision values, which are then combined for the final prediction score using a weighting approach. In order to capture the interaction between various modalities in detecting symptoms and maximize the usage and combination of each modality, a model-based fusion using deep learning is constructed, inspired by recent endeavors in multimodal fusion for diagnosing Alzheimer's disease. A comparison will be made with both early fusion and late fusion approaches.

7.2 Motivation & Research Questions

The research questions in this chapter can be summed up as follows:

Q 1: What are the acoustic features associated with AD? Do they align with the state-of-the-art methods in literature?

Q 2: Do we get benefits by integrating different modalities with different combinations of features?

Q 3: Does building a model that takes a weighted average decision based on a decision from each modality (LF) better than the one which first combines the features and made the decision (EF)?

Q 4: Does building a deep learning multi-model helps better in learning AD cues than with traditional models with these multimodality-based features?

7.3 Methodology

Our approach is to build a model based on interaction cues from dialogue conversations, and lexical cues and combined these with acoustic features, to predict whether an individual has AD or not. This is the same task (i.e. diagnosis) with the same data set (i.e. CCC) as used in the previous chapter 6 with the inclusion of new features (i.e lexical features and acoustic features), combining with previously explored features

using different fusion approaches. It consists of four main parts: feature engineering, feature selection, learning algorithms, and multimodal fusion strategies. Using features from the audio and text data, we ran the experiments mentioned below to predict AD:

1. An experiment with different lexical features including TF-IDF and Glove embedding with traditional ML classifiers.
2. A Convolutional neural network (CNN) model on sentence level BERT embeddings that learn local patterns between current utterances and surrounding utterances.
3. Traditional ML models utilising unimodal audio and text features that includes lexical representations, DA Ngrams, confusion ratios, acoustic features, and dialogue interaction features.
4. Traditional ML models with EF and LF to test the effect of the combination of each modality.
5. A multimodal CNN model using lexical (BERT embeddings), acoustic, and interaction information to classify AD.

7.3.1 Feature Engineering

7.3.1.1 Interactional Features

The similar set of pause-based features mentioned in Chapter 6 is employed. Just as a recap: five categories of pauses are used: *short pauses (SP)*, *long pauses (LP)*, *gaps (GA)*, and *lapses (LA)* and *Attributable silence* (*AS*). *SPs* and *LPs* are silences within one individual's speech, with *SPs* less than 1.5 seconds and *LPs* greater than 1.5 seconds.

Other features encode general characteristics of the interaction, includes the number of overlaps, *turn length* (number of words per turn), *floor control ratio* (amount of time during which *P* speaks, relative to the total speech time of the conversation), *standardized pause rate* (ratio of total words spoken by *P* to the total pauses (including *SP & LP*)), *phonation rate* (total time spoken by *P* to total spoken time including *SP and LP* by *P*), and *speech rate* (number of words per minute). The annotation protocol for these interactional features is described in (Nasreen et al., 2021c) (Chapter 6: section 6.3.2).

7.3.1.2 DA Features

Unigram DA and bigram DA sequence feature from chapter 6 are used as a local interaction pattern within the dialogue here in this experiment. Confusion ratios are also used and to produce the input for DA based classifier, the unigram, and bigram DA sequences are concatenated with confusion ratios. Here, in this chapter ***Ngram DA*** will represent features from unigram, bigram, and confusion ratios. In an experiment

with the deep learning model, the DA one hot encoded vector are concatenated with utterance representation to represent as one input. This addition of DA with utterance will help to find useful local interaction patterns within the local context. Ngram DA's represent a combined representation of unigram, bigrams, and confusion ratios. In other experiments, counts of Ngrams features as used in the previous chapter on per patient basis.

Now, it is aimed to add lexical and acoustic features as a set of new features in next sections (section 7.3.1.3 and 7.3.1.4).

7.3.1.3 Lexical features from text

Different lexical representations, comprising TF-IDF feature vectors, GloVe embeddings, and contextualized BERT embeddings, have been utilized as features. The language impairments produced by AD patients include non-coherent repetitions. The significance of each word in a document is represented by TF-IDF features, which average out each word's importance throughout the entire dataset (Guerrero-Cristancho et al., 2020). By including TF-IDF features, it is possible to model patients' vocabularies and the significance of their spoken words in the transcripts. The lexical feature representations were extracted from the transcripts using a pre-trained GloVe model (Pennington et al., 2014) with an embedding size of 100 dimensions space. To obtain the TF-IDF feature vector and glove embedding features, only patient utterances are utilized.

TF-Hub BERT was used to load an official, pre-trained BERT model trained on MEDLINE/PubMed [1] and fine-tuned on SQuAD 2.0 [2]. The reason to choose this version of BERT is that it is pre-trained on MEDLINE/PubMed that contains medical domain literature and we find there are conversations in Carolina's collection dataset in which patient talks about their medical history or the medical treatment they are on. For each utterance in the transcript of each speaker, this model was used to extract embeddings of shape (u,768), where u is dependent on the length of the input. The largest embedding had an u value of 387 after embeddings for each conversation had been retrieved. As a result, the remaining conversations were padded to have the same shape, producing an embedding of (387,768) for each conversation. Given the transcript and DA information (based on chapter 4 & chapter 5), only question-type and answer-type utterances are extracted from the conversations. These question-answer sequences are then fed into the BERT tokenizer and BERT model to get the utterance level embeddings. A question or answer utterance embeddings are obtained from a CLS token. This corresponds to the first token of the output (after batch dimension) [3].

[1] MEDLINE/PubMed dataset

[2] Stanford Queston Answering (SQuAD 2.0) dataset

[3] The idea of inclusion of question utterances from the interviewer is inspired by Williamson et al. (2016), who found the interviewer's questions (or avatar's text) to contain highly predictive features

7.3.1.4 Acoustic Features

OpenSMILE v2.1 (Eyben et al., 2010) was used to extract acoustic features from the audio recordings, for a total of 30 dialogue conversations. OpenSMILE is open-source software that has been previously used for AD classification using audio features (Lin et al., 2020; Warnita et al., 2018). Recently OpenSMILE was also used to develop machine learning models and create a benchmark speech dataset for AD speech classification and regression task (Luz et al., 2020).

A set of 64 audio features was extracted and higher-order statistics (mean, min, max, standard deviation) were computed (See detail in appendix B.2). The output data comprises comma-separated files with each column indicating a different acoustic feature and each row having the acoustic value inside a 10 ms frame that was taken from each segment with 10 ms. Using the utterance timing information provided in the transcripts, the participant's utterances are extracted (either *P* or *I*) and calculated average values of the features per utterance basis. To create a single vector for each utterance duration, the data for each 10 ms frame are averaged. A standard zero mean and variance normalization was applied to each feature. The detail of acoustic features is given in Table 7.1. Jitter and shimmer, measure cycle-to-cycle variations of the fundamental frequency

Type	**Feature names**
Frequency related	Fundamental frequency *(f_0)*, jitter, voicing probability
Energy, amplitude related	RMS energy, log RMS energy, *shimmer*, *loudness*, *Harmonic to noise ratio (HNR)*
Spectral parameters	4 Mel-frequency Cepstral coefficients *(MFCCs)* [1-4], *delta MFCCs* [1-4], *delta-delta MFCCs* [1-4]

Table 7.1: Acoustic feature set

(*f0*) and amplitude, respectively. Shimmer and jitter have been reported significant in patients of neurodegenerative diseases (Gayraud et al., 2011; Lin et al., 2020). The significant decrease of HNR in elderly people (Ferrand, 2002) may be attributable to AD.

7.3.2 Feature Selection

Feature selection (FS) reduces the dimensionality of the feature set by choosing a subset of relevant features. FS algorithms are classified as Filter, wrapper, and embedded methods. The recursive feature elimination (RFE) method most widely used from the wrapper method is an iterative process that removes a specific number of features, is the model with the remaining features, and examines the effect on classification accuracy (Pedregosa et al., 2011). Those features making the least contribution are removed

recursively until the desired number of features are left. A variant of RFE is used here, recursive feature elimination with cross-validation (RFECV). To find the optimal number of features cross-validation is used with RFE and maintain scores for different feature subsets and select the best scoring subset of features. The RFECV is used to make sure that FS only used training data and not test data.

7.3.3 Learning Algorithm

Due to the low number of samples, compared to the dimensionality of the feature space, traditional machine learning classifiers are used rather than more complex neural networks, as the former has the potential to provide a rational trade-off between classification performance, run-time complexity, and the risk of overfitting (Taschwer et al., 2018). In this study, traditional ML classifiers were used: Logistic Regression (LR), Support Vector Machines (SVM), and Random Forests (RF). However, to find more local interaction within the utterances and surrounding utterances instead of high dimensional (dialogue level) feature space, a CNN model is used that took BERT embeddings/ Glove embedding as input for each utterance, produced a learned representation at utterance level. At the higher level, CNN will learn useful information with the utterance and its surroundings to give more insightful information. This on the whole will produce a dialogue-level feature representation of features per patient. At the end, a fully connected layer with a sigmoid function is used to produce binary predictions. A multimodal CNN is used to experiment with lexical representations learned from CNN using BERT embedding, concatenated with acoustic and interactional features, and fed into a simple dense layer with a sigmoid activation function to make decisions.

The LR model is with a range of regularization parameters (0.1,10,100); SVM with RBF and polynomial kernels, cost C (0.1,100) and gamma (0.001,0.1); and RF with 60 trees of maximum depth of 5. The same hyperparameters were used for all experiments. We experimented with CNN with filter sizes (32,64,128) and kernel sizes (2,3,4,5) and choose a filter of 32 units and a kernel size of 4. The model is trained using ADAM optimizer and for the loss function, Binary Cross-Entropy is used to model binary outcomes. Embedding size was set to 100 dimensions GloVe pre-trained embeddings, with 768 dimensions for BERT embeddings. Early stopping was used to avoid over-fitting the network.

7.3.4 Fusion Strategy

Two different fusion strategies were employed in this study. In the *early fusion (EF)* method, the values of each feature for both acoustic and interactional features are normalized using the standard scalar feature of scikit-learn (Géron, 2019) and then concatenated directly (Figure 7.1(a)). The *late fusion (LF)* or decision-level strategy utilized

the same normalization for each feature set but with predictions made individually for each feature set (See Figure 7.1(b)). The prediction scores of each classifier are then combined using either a hard voting ensemble or a standard soft voting ensemble method (Sagi and Rokach, 2018): hard voting predicts the final output by considering the individual labels (hard level); soft voting computes based on scores/probabilities (soft level) of the involved classifiers. There are two approaches of soft voting that are employed: Averaging and Blending. Soft voting (Averaging) computes the average probability for each class over each component classifier and eventually bases the final prediction on maximum average probability. In the case of the *blending* approach, a meta-classifier is used on the predictions of individual classifiers. The meta-model uses this set of predictions as a set of features (input). Final predictions will therefore be produced by the meta-classifier. Among our classifiers, LR and RF provide prediction probabilities directly, while the SVM outputs were transformed into prediction probabilities using Praat scaling (Taschwer et al., 2018). [4] LF allows the use of different models for each of the modalities providing more flexibility as different predictors can model different individual modality better. However, it misses the important cues from the low-level interaction between the modalities.

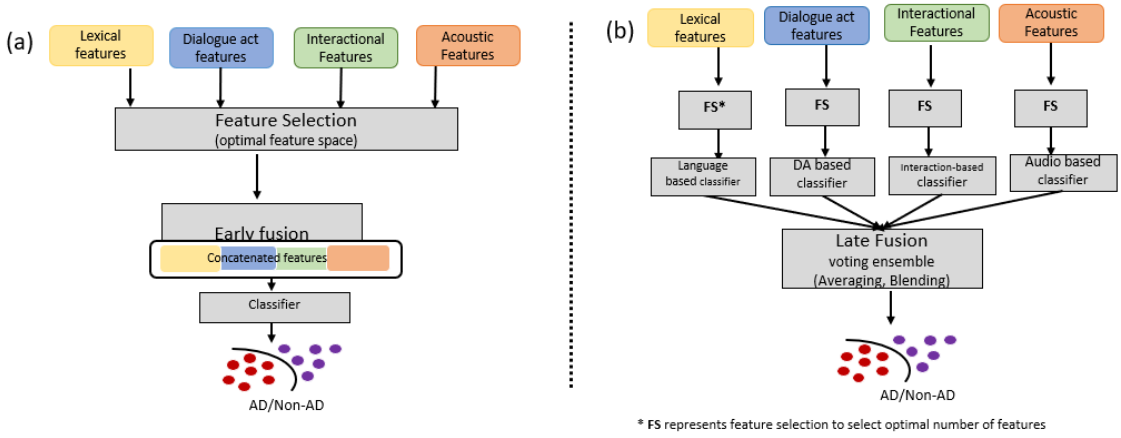


Figure 7.1: Early fusion (a) and late fusion (b) based model architectures.

In the experiment with multimodal CNN, a model-based fusion strategy is implemented. A CNN model used to fuse the results from modality-specific transcriptions (text) based embeddings with low-level descriptors of acoustics along with one hot encoded vector representation of DA's. The model-based fusion allows end-to-end training of multimodality feature representation and fusion components. However, the results are

[4]Note that Praat scaling is only implemented for kernel SVM in scikit-learn framework for Python and only the kernel version of SVM is used and not the linear SVM for late fusion.

less interpretable as its hard to tell on what the predictions relying and which features plays an important role in each of the modalities.

In addition to testing the effect of each feature set separately on the classification task, the combinations of two modalities is compared with different feature sets using EF and LF. The model-based fusion technique is also compared to the early fusion and late fusion strategies.

7.4 Experimental Setup

7.4.1 Dataset

The dataset selected for this study is the Carolina's Conversation Collection (CCC) Pope and Davis (2011), which includes a comparable number of dialogues and the same set of patients as examined in earlier chapters. Both transcripts and audio data were utilized in this study. Specifically, 38 dialogues from individuals with Alzheimer's disease (AD) were accessible, and for this study, we randomly selected 15 dialogues from AD patients and 15 from non-AD individuals.

7.4.2 Evaluation Metrics

We set up our experiments to investigate which acoustic features and dialogue interaction features are most effective for predicting AD. Due to the fairly small dataset, we used leave-one-(patient)-out cross-validation (LOOCV) to get a better estimation of generalization accuracy. The dataset is balanced in terms of classes; we choose precision, recall, F1-score, and accuracy as evaluation metrics as we used in chapter 6 for classification of AD and non-AD.

7.4.3 Baseline Model

The performance of our model is compared with Luz et al. (2018)'s work on the same CCC corpus with dialogue interaction features. Luz et al.'s dataset is slightly bigger than ours (38 dialogues vs. 30). Although the features set are not directly comparable, they utilized only interactional aspects of conversation including dialogue duration, average turn duration, normalized duration, average number of words, and average words per minute from the spontaneous speech.

7.5 Results and Discussion

In table 7.2, the results of the classification task are reported with lexical features that include: TF-IDF features, Glove embeddings, and BERT embedding features, with LR,

Classifier	Features	Class	Prec.	Rec.	F1	Acc.
LR	TF-IDF	Non-AD	0.72	0.87	0.79	0.77
		AD	0.83	0.67	0.74	
	Glove	Non-AD	0.75	0.60	0.67	0.70
		AD	0.67	**0.80**	0.73	
SVM	TF-IDF	Non-AD	0.72	0.87	0.79	0.77
		AD	0.83	0.67	0.74	
	Glove	Non-AD	0.75	0.80	0.77	**0.77**
		AD	0.79	**0.73**	0.77	
RF	TF-IDF	Non-AD	0.75	0.80	0.77	**0.77**
		AD	0.79	**0.73**	0.76	
	Glove	Non-AD	0.79	0.73	0.76	**0.77**
		AD	0.75	**0.80**	0.77	
CNN	Glove	Non-AD	0.83	0.67	0.74	**0.77**
		AD	0.72	**0.87**	0.74	
	BERT	Non-AD	0.66	0.69	0.67	0.67
		AD	0.67	0.64	0.66	
	BERT + DA's	Non-AD	0.75	0.60	0.67	0.70
		AD	0.67	0.80	0.73	
Multimodal CNN	BERT+DA's +Acoustic+ Interaction	AD	0.71	0.67	0.69	0.70
		AD	0.69	0.73	0.71	

Table 7.2: Individual classifiers with lexical features, with all and top-ranked features with FS.

SVM, RF, and deep learning model i.e CNN. It can be seen from the results that the model got a higher recall for the AD class with Glove embeddings as compared to TF-IDF features for all three machine learning based classifiers. We got an accuracy of 0.77 with LR with recall for AD class (0.80 vs 0.67) with Glove vs TF-IDF. An even higher accuracy is obtained with an SVM of 0.77 with slightly lower recall for AD class (0.73 vs 0.67) with Glove vs TF-IDF features. With RF, the accuracy of 0.77 is achieved with recall for AD class (0.80 vs 0.73) for Glove vs TF-IDF features. Moving towards deep learning models, with CNN, the accuracy is dropped to 0.67 with BERT embeddings and 0.70 with combining BERT embeddings at utterance level with DA's of each utterance. Surprisingly, it did not get an improvement using BERT over GLove, as people usually find better results with BERT. Therefore, to check, DA's and other acoustic features are added. However, adding acoustic and interaction features does not improve the accuracy further. This may be because deep learning models are more data-hungry models and perform well if the dataset size is larger, and with this number of data points, the accuracy has deteriorated. Given the limited sample size, the model is executed in 10 repeated runs, and the average performance across these 10 iterations is reported to mitigate variability. Another contributing factor may be that BERT is pre-trained on text data that may not align well with our dialogue dataset.

From the results from table 7.2, further experiment will be performed with only Glove embeddings as lexical features and used them in experiments with other feature sets

using fusion strategies.

Classifier	Features	Modality	Acc. [all]	Acc. [FS]	# features
Baseline Luz et al. (Luz et al., 2018)					
LR	Interaction	Speech	0.75	-	-
SVM	Interaction	Speech	0.83	-	-
RF	Interaction	Speech	0.81	-	-
Our Models					
LR	Lexical	Text	0.77	-	-
	Acoustic	Speech	0.70	0.80	25
	Ngram DA's	Text	0.77	0.80	48
	Interaction	Text + Speech	0.80	**0.83**	12
SVM	Lexical	Text	0.70	-	-
	Acoustic	Speech	0.70	0.80	25
	Ngram DA's	Text	0.77	0.80	48
	Interaction	Text + Speech	0.77	**0.83**	12
RF	Lexical	Text	0.77	-	-
	Acoustic	Speech	0.80	0.77	25
	Ngram DA's	Text	0.63	0.70	48
	Interaction	Text + Speech	0.73	0.77	15

Table 7.3: Individual classifiers with different feature sets, with all and top-ranked features with FS.

Results with individual feature set In Table 7.3, we present our model's performance in a cross-validation setting with each feature set individually against that of the baseline models that only utilize interaction features on AD/Non-AD classification.

Among all features, 25 acoustic features, 48 Ngram DA's features, and 15 interactional features are selected after RFECV feature selection method. For AD classification, our LR model with top ranked Ngram DA's achieved an accuracy of 0.80 and 0.83 with top ranked interaction features outperforming the baseline (LR: 0.75). SVM also achieves similar accuracy with top ranked interaction features with an accuracy of 0.83 over the baseline (0.83). On the other hand with RF, an accuracy of 0.77 is achieved with optimal FS. Among All three classifiers, LR and SVM performed best with top ranked interaction features (0.83). With Ngram DA's features LR and SVM achieves the highest accuracy of 0.80 against RF with 0.77 with the interaction feature set. However, SVM achieves the highest accuracy with our top ranked interaction features (0.87) and 0.80 with all interaction features over LR and RF. Random forest did not perform very well with all three feature sets with feature selection based features, however, it did best with acoustic features with 0.80 with all acoustic features and 0.77 with FS based acoustic features.

With the fusion of different combinations of features, the model got higher accuracies as compared to using individual feature sets. The overall results support our assumptions that the errors and noise of the individual modalities can be reduced more effectively by a model that combines data from many modalities and feature sets.

Effect of feature selection (FS) RFECV is performed[5] with pipeline in a cross-validation setting on acoustic, interactional, and Ngram DA's features. 25 features are found from the acoustic feature set as the minimum optimal number of features, 12 from the interactional feature set, and 48 from Ngram DA's set securing good classification results. The most significant acoustic feature was LogHNR_SD (r=0.60) positively correlated with AD, known to be important in acoustic analysis for the diagnosis of pathological voices. Among others loudness_SD (r=0.56), raw fundamental frequency (r= 0.44), variation in jitter _DDP (r=0.45), intensity (r=0.44) are all positively correlate with AD[6]. MFCC[2]_mean and MFCC[3] double delta SD are found to be negatively correlated with AD class. Among interactional features, duration of AS (r=0.41), duration of lapses (r=0.43), and duration of gaps (*P_I*)(r=0.38) are all positively correlated with AD class while patient turn lengths (r= -0.40), standardized phonation time (r= -0.54) were among the negatively correlated with AD class.

It can be seen that in Figure 7.2, all models (LR models) with FS based features perform better than using all the features from each set. Similarly, with the fusion of different features based on FS gave better accuracy results than using all features from each set.

7.5.1 Fusion Analysis

From Table 7.4, It can be viewed that the performance of our three models across different feature sets. Our models, integrating features with early fusion give better AD prediction than using these feature sets individually in all three cases. LR and SVM achieve the highest accuracy of 0.93 and 0.90 with FS from acoustic, interaction, and Ngram DA's features. With acoustic features, LR got an accuracy of 0.80, adding interaction features gives a performance boost of 0.3 which is further improved by adding Ngrams with acoustic and interaction feature with overall accuracy of 0.93. On the other hand, the interaction features set alone gives promising results (0.80 with LR and 0.77 with SVM) for all interactional features and (0.83 with both LR and SVM) with top ranked features. Adding acoustic features with interaction using early fusion with LR has increased the

[5]The results based on RFECV are different from the results that are published in INTERSPEECH, 2021 (Nasreen et al., 2021b) as previously standard feature selection (RFE) was used for selecting top-ranked features. Here it is made sure that the model only trained on training data and did not see a test set during each fold.

[6]Detail can be found here: https://osf.io/3fd8x/?view_only=8d864851fbd74be5b53c0ef86335a25a

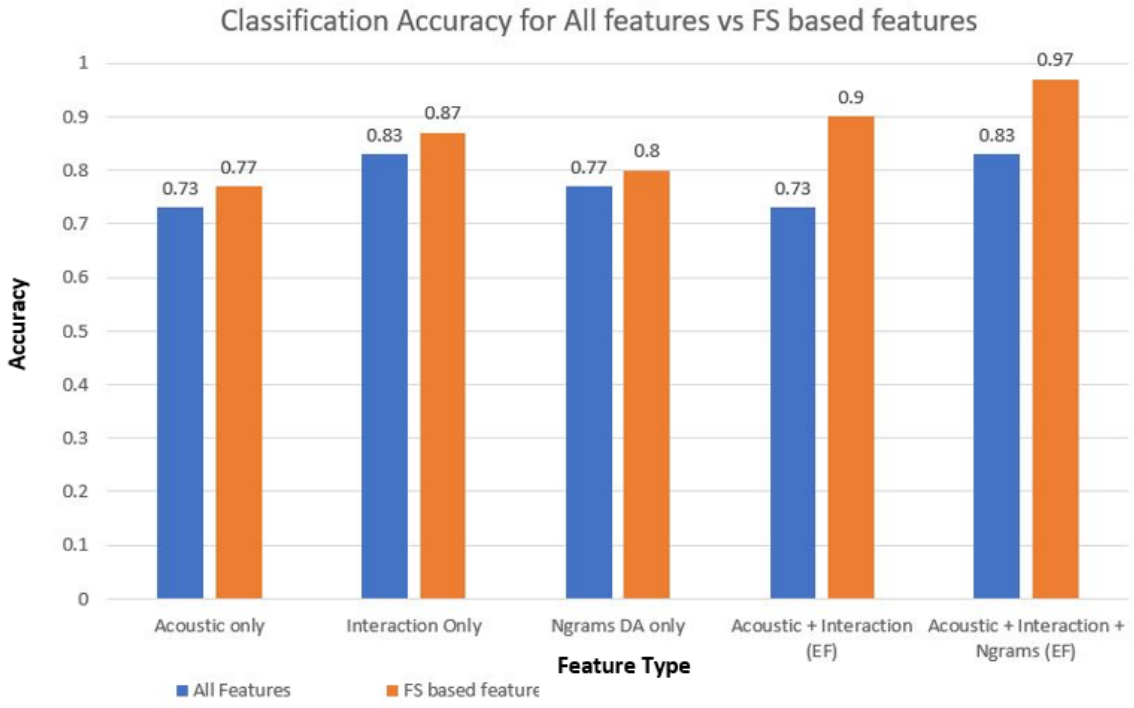


Figure 7.2: Comparison of performance (accuracy) for all features from each feature set vs FS based features.

performance from 0.83 to 0.87 with FS and from 0.80 to 0.90 with all features (both acoustic and interactional). Furthermore, adding Ngrams to acoustic and interaction has increased accuracy from 0.87 to 0.93 with FS and 0.90 from 0.83 with all features. However, no performance gain is observed with SVM by adding acoustic features and Ngrams when considering all features (acc.:0.87). But with top ranked features, SVM model with acoustic and interaction gives an overall accuracy of 0.87 which is further increased by adding Ngram DA's to 0.90. This indicates that adding the features at early stages before classification helps in better learning the interaction cues between different modalities (text vs speech).

A better AD prediction is obtained when only the text modality is used than when the speech modality is used. However, integrating multimodal features based on both text and speech with early and late fusion has significantly improved AD predictions. In terms of late fusion strategy, soft voting is utilized with two simple fusion approaches: averaging and blending. Under the averaging criterion, an optimal performance of 0.90 of AD detection accuracy is obtained with LR and 0.87 with SVM with acoustic, interaction, and Ngram DA's while the blending approach produced 0.87 of AD detection accuracy with both LR and SVM with FS from acoustic, interaction and Ngram DA sets. It is clear that the proposed approach of fusing various feature sets from the text and speech

Classifier	Features	Acc. [all]	Acc. [FS]
	Early Fusion (EF)		
LR	Acoustic+Interaction	0.83	**0.87**
	Acoustic+interaction+ Ngram DA's	**0.90**	**0.93**
	Acoustic+interaction+ Ngram DA's+ lexical	**0.90**	-
SVM	Acoustic+ Interaction	**0.87**	**0.87**
	Acoustic+ Interaction+ Ngram DA's	**0.87**	**0.90**
	Acoustic+interaction+ Ngram DA's+ lexical	0.83	-
RF	Acoustic+ Interaction	0.83	0.83
	Acoustic+interaction+Ngram DA's	0.87	0.87
	Acoustic+interaction+ Ngram DA's+ lexical	0.80	-
	Late Fusion (LF) with Averaging (*Avg)* and Blending (*blend*)		
LR	Acoustic+Interaction (avg)	0.80	0.83
	Acoustic+Interaction (blend)	0.80	0.87
	Acoustic+interaction+ Ngram DA's (Avg)	**0.87**	**0.90**
	Acoustic+interaction+ Ngram DA's (Blend)	0.80	**0.87**
	All (avg)	**0.87**	-
	All (blend)	**0.83**	-
SVM	Acoustic+Interaction (avg)	0.83	0.80
	Acoustic+Interaction (blend)	0.80	0.83
	Acoustic+interaction+ Ngram DA's (Avg)	**0.87**	**0.87**
	Acoustic+interaction+ Ngram DA's (Blend)	**0.87**	**0.87**
	All (avg)	0.83	-
	All (blend)	0.80	-
RF	Acoustic+Interaction (avg)	0.83	0.80
	Acoustic+Interaction (blend)	0.77	0.73
	Acoustic+interaction+ Ngram DA's (Avg)	0.80	0.77
	Acoustic+interaction+ Ngram DA's (Blend)	0.77	0.70
	All (avg)	0.80	-
	All (blend)	0.77	-

Table 7.4: Early and Late fusion of different feature set, with all and top-ranked features with FS.

modalities improves AD detection accuracy when compared to the best results provided by optimal unimodal (single feature set from Table 7.3) data by 10% for both EF and LF with LR and SVM classifiers. Although the proposed approach uses the simplest fusion approaches, it still enhances the performance against classification results from a single feature set. The major factor producing better outcomes is the fact that the proposed method's base classifiers are powered by various kinds of optimal feature sets and data modalities. As a result, the best outcomes from unimodal data are combined, and the resulting outcomes are better than the best outcomes from unimodal data (or from a

single feature set).

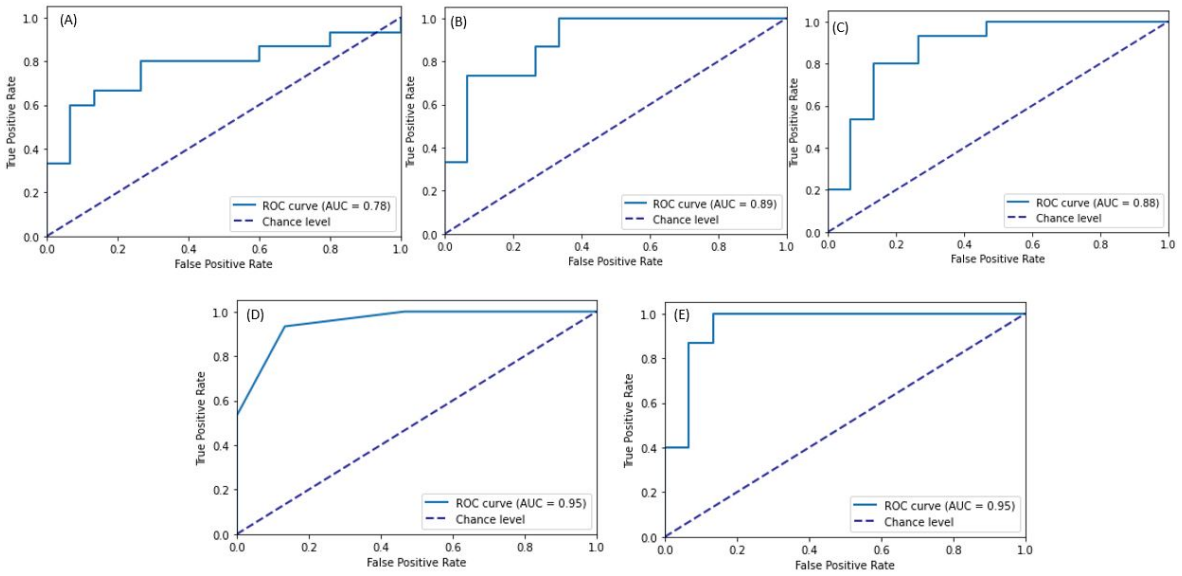


Figure 7.3: Receiver operating characteristics (ROC) curves: (A) LR model with all acoustic features, (B) LR model with FS based acoustic features, (C) LR model with all Ngram DA features, (D) LR model with acoustic, interaction and Ngram DA' with *LF* (*Averaging*), (E) LR Model with FS based acoustic, interaction and Ngrams with *EF*.

When comparing the robustness of a developed model to baseline models, the ROC curve is a more trustworthy evaluation tool in machine learning. In comparison to models with ROC curves that have less AUC, a model is deemed robust if it has a higher AUC value. Hence, to confirm the efficacy of the proposed model utilizing the fusions approach, the ROC curves of the three single feature set based classifiers and two from the fusion of feature sets with EF and LF (averaging) approaches are plotted (see Figure 7.3). The ROC curves show that the LR classifier has the highest AUC with FS-based early fusion of acoustic, interaction, and Ngram features, which is 0.95 (Figure 7.3 E) with an accuracy of 0.93. On the other hand, an AUC of 0.95 is also produced by averaging the prediction probabilities from all three feature sets, i.e., acoustic + interaction + Ngram DA's with an accuracy of 0.90. Therefore, both accuracy and AUC, which measure the effectiveness of the proposed model's fusion of various feature sets, are validated.

Experiment without manually annotated interaction features To test the robustness of an automated system for AD diagnosis, an experiment is performed with excluding the interaction features that were extracted from the manually annotation of pauses in natural dialogues. For the said purpose, we only considered *#overlaps, #turn_switches_per_minute, Floor_control_ratio, speech_rate,* and *turn_length*. These features are combined with acoustic features and Ngram DA features using early fusion

with LR and SVM classifiers and results are presented in table 7.5. By comparing these results with results in table 7.4, it can be clearly seen that by excluding the manual annotation based pauses features, the performance of LR model is reduced from 0.93 to 0.87 with feature selection, while with all features, it is reduced from 0.90 to 0.83. While pause-based features play a significant role, the system is still able to achieve accuracy comparable to state-of-the-art methods (e.g (Luz et al., 2018)) even without these features.

Classifier	**Features**	**Acc. [all]**	**Acc. [FS]**
Manually extracted feature based experiment(from Table 7.4 for comparison)			
LR	Acoustic+interaction (manual)+ Ngram DA's	**0.90**	**0.93**
Without manually extracted interaction features			
LR	Acoustic+Ngram	0.83	0.80
	Acoustic+interaction (auto)+ Ngram DA's	**0.83**	**0.87**
SVM	Acoustic+Ngram	0.87	0.83
	Acoustic+interaction (auto)+ Ngram DA's	0.87	0.83
RF	Acoustic+Ngram	0.80	0.73
	Acoustic+interaction (auto)+ Ngram DA's	0.80	0.83

Table 7.5: Early fusion of Ngram DA features with acoustic features and automated (auto) interaction features with all and top-ranked features with FS.

7.5.2 Error Analysis

The results in Table 7.6 show that the EF strategy with top-ranked acoustic, interactional, and Ngram DA's features obtains the highest F1 score for Non-AD and AD group with LR with a precision of 1.0 for Non-AD and recall of 1.0 for AD class with an F1 score of 0.93 for Non-AD and 0.94 for AD class. With SVM, and with top-ranked features, the model got an F1 score of 0.90 for both AD and Non-AD groups. With LF, similar accuracy of 0.90 is obtained by averaging with the same feature sets with logistic regression. Detailed class-wise results with only the LR classifier in Table 7.6 are reported.

Combining interactional and acoustic features particularly improves recall (0.93) of the AD class: acoustic features alone (with LR) give recall 0.67 for the AD class, increasing to 0.87 with top-ranked features, while interactional features alone give 0.87 with all features and 0.87 with the optimal 12 features. Adding Ngrams along with FS-based acoustic and interaction, improves the recall of the AD class from 0.93 to 1.0, while for Non-AD class, it improved from 0.80 to 0.87. It can be clearly seen that the fusion of features either EF or LF (averaging or blending) has significantly improved the recall of the AD class as well as Non-AD class. Figure 7.4 also shows the accuracy of AD and Non-AD classes with a confusion matrix. It can be observed that the confusion matrices of predictions of the LR Model with (A) all acoustic features, (B) FS based

Feature set	Fusion	No.	Class	Prec.	Rec.	F1	Acc.
Acoustic	-	All	Non-AD AD	0.69 0.71	0.73 0.67	0.71 0.69	0.70
Acoustic	-	25	Non-AD AD	0.85 0.77	0.73 0.87	0.79 0.81	0.80
Interactional	-	All	Non-AD AD	0.85 0.77	0.73 0.87	0.79 0.81	0.80
Interactional	-	12	Non-AD AD	0.86 0.81	0.80 0.87	0.83 0.84	0.83
Ngram DA's	-	All	Non-AD AD	0.72 0.83	0.87 0.67	0.79 0.74	0.77
Ngram DA's	-	48	Non-AD AD	0.77 0.85	0.87 0.73	0.81 0.79	0.80
Acoustic + interaction		EF significant	Non-AD AD	0.93 0.82	0.80 0.93	0.86 0.87	0.87
Acoustic + interaction (blend)		LF significant	Non-AD AD	0.93 0.82	0.80 0.93	0.86 0.87	0.87
Acoustic + interaction+ Ngram DA's		EF significant	Non-AD AD	1.00 0.88	0.87 1.00	**0.93** **0.94**	**0.93**
Acoustic + interaction+ Ngram DA's (Avg)		LF significant	Non-AD AD	0.93 0.88	0.87 0.93	**0.90** **0.90**	**0.90**

Table 7.6: Comparison of results for the AD classification, shown as precision, recall, F1, and accuracy per class with **LR Model**.

acoustic features, and (C) False negatives or false positives for AD detection will depend on the application the model is used for, however as it stands, integrating the most relevant features significantly reduces the false negatives of diagnosis while still only slightly reducing the false positives.

7.6 Summary of the research questions investigated:

Some of the research questions that are aimed to investigate in this chapter are:

Q 1: **What are the acoustic features associated with AD? Do they align with the state-of-the-art methods in literature ?**

Through Pearson correlational analysis, we looked for the characteristics that set AD speech distinct from Non-AD speech.There was a significant positive association between the logarithmic power of the Harmonic-to-noise ratio and AD (r = 0.60). This finding supports previous research that found elderly individuals with a

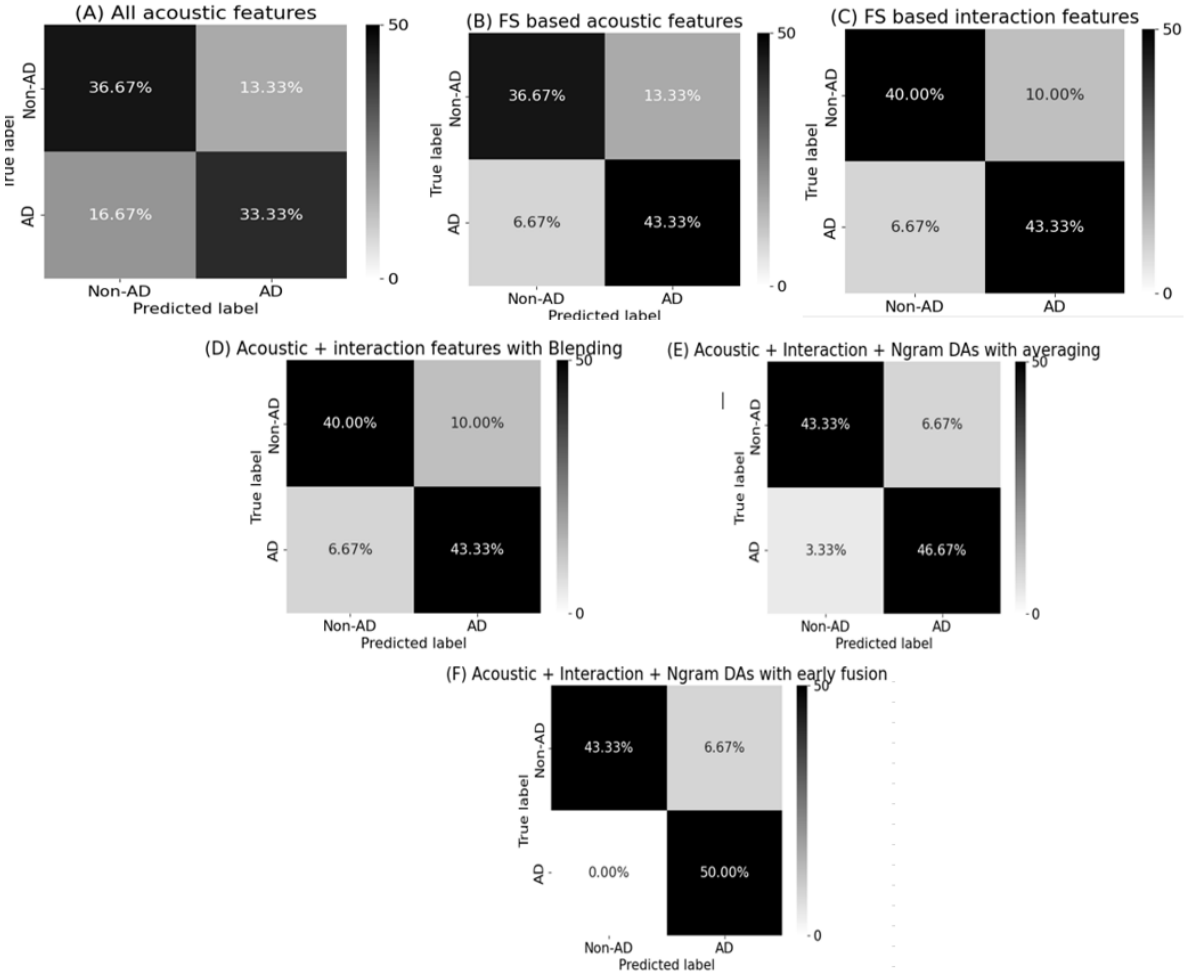


Figure 7.4: Confusion matrices for AD classification task with different feature sets and with the fusion of different feature sets.

decrease in HNR (Ferrand, 2002) (mAD= 1.49 vs mNon-AD= 1.52). jitterDDP_mean was also positively correlated (r= 0.32) with AD with higher values (mAD= 0.00047 vs Non-AD= 0.00041) indicating towards AD (Lin et al., 2020). Teixeira et al. (2013) says that the primary cause of jitter is an uncontrolled vibration of the vocal cords, and patients with pathologies frequently have voices with greater jitter values. Intensity_sma_SD also has a strong positive correlation with AD (r= 0.44) with higher values for AD patients (mean= $5.75 * 10^{-}6$) than Non-AD ($1.88 * 10^{-}6$).

Q 2: **Do this study get benefits by integrating different modalities with different combinations of features?**

The choice of combining interactional aspects of communication with lexical aspect and with acoustic features have shown significant improvements over the results when using these feature set alone. When using acoustic features only from speech

data, an accuracy of 70% is obtained with LR and 73% with SVM. Adding Ngram DA's and Interaction features further improved the accuracy of the system to 90% with LR and 87% with SVM with an early fusion strategy with all features. Selecting features based on FS and combing feature sets further give a performance boost to 93% & 90% with LR and SVM. Adding lexical features did not improve the results(90% vs 80%) with LR and SVM. On the other hand, with the LF strategy, averaging the predictions of individual classifier produce better results than the blending approach with LR. An accuracy of 90% is achieved with LR and averaging fusion strategy with acoustic, interaction, and Ngram DA's and 87% with the same feature set with blending fusion strategy. These findings are aligned with studies of literature that have used fusion strategies for combining text and speech modality features for AD and MCI prediction tasks. Campbell et al. (2020) utilized both linguistic and speech-based features and applied two fusion strategies with an accuracy of 0.80 for fusion I and 0.82 for fusion II strategy for AD classification showing that AD detection is greatly improved with fusion strategies. Shah et al. (2021) have used a majority voting ensemble on three best-performing acoustic and best-performing language models computing the final predictions using a weighted combination of the individual model predictions. Chakraborty et al. (2020) experimented with only audio biomarkers and found that using early fusion achieves an improved performance of up to 77% as compared to when using each set of audio biomarkers separately (66%) to distinguish between AD, MCI, and healthy controls. These results are aligned with our findings as well: Chakraborty et al. (2020) reported best results up to 77% with feature selection before early fusion of features. They achieved an F1 score of 64.7% with all four feature sets, improving to 74.7 with feature selection and 77% with feature selection before fusion. They achieved an even higher F1 score up to 81% with late fusion when using a different classifier (not the random forest) as the decision classifier.

Our results showed that the fusion of both text-based and speech-based modalities with different combinations of features and feature selection improves the detection of AD.

Q 3: **Does building a model that takes average decision based on a decision from each modality is better than the one which first combines the features and made decision?**

Early fusion of FS and all features gave overall better results with all combinations over decision-based fusion (both *average* and *blend* approach). We got 93% accuracy with FS of acoustic, interaction, and Ngrams with LR model with EF while with SVM, it achieves 90% with *averaging* and 87% with *blending*. This may be due to

the fact that feature selection and the classifier itself benefit more from cross-model relations than using each modality-based classifier separately.

Q 4: **Does building a deep learning multi-model helps better in learning AD cues than with traditional models with these multimodality-based features?**
The EF/LF models often require handcrafted features. Deep models, on the other hand, may identify the optimal set of features during training. Additionally, unsupervised feature generation can be carried out using deep models like LSTM and CNNs, which can subsequently be combined with a more complex decision layer. When the dataset is very small (small sample size), deep models do not perform well. Lack of training data for networks may be the root of the decreased performance. In our case, deep learning models (i.e CNN) do not perform very well with the small number of patient instances available. In contrast, ML classifiers perform well with both early fusion and late fusion of features at the feature level and decision level. In the future, it is aimed to add more patient data in our experiments to get generalizable accuracy results.

7.7 Conclusion

In this section, it is aimed to look at the advantages of combining Ngram-based DAs with linguistic, acoustic, and interactional behavior descriptors to find evidence of AD. Three ML models are presented with these feature sets individually and with early fusion and late fusion methods which consume transcripts and speech data to classify whether a speaker in a natural setting/ environment in the form of natural conversation has Alzheimer's Disease. Our best models were an LR (both EF and LF with averaging) and SVM with EF with FS based acoustic features, interactional features, and Ngram DA's features. It has been shown that different combinations of features with fusion strategies and feature selection methods have drastically improved the performance of the models using these features individually.

CHAPTER 8

CONCLUSIONS AND FUTURE WORK

The current study adopts the strategy of analyzing the interactions within language as it pertains to a wide range of AD symptoms and more thoroughly characterizes the progression of the condition. At the beginning of the thesis, the main objectives of this research, with the research questions that seek solution were listed and then an extensive review of the up-to-date literature work being studied were described in chapter 2.

Chapter 3 investigates through a corpus study what are the most significant interactional patterns that are effective in AD diagnosis. Results of our initial corpus study were presented on the conversational dataset with the objective of looking into different distributions of questions being asked, response behaviors, signals of non-understanding, and clarification requests and showed that AD patients exhibit different patterns during conversations. It was found out that more yes-no questions and fewer wh-questions were asked from AD group than Non-AD group. More signal non-understanding and clarification requests were produced in response to wh-question and simple yes-no questions from the AD patients. This answers **Research Question 1** from chapter 1 section 1.3.

In chapter 4, I experimented with building a rare class DA tagger for identifying conversational phenomena such as clarification requests, etc. The model learns the utterance representations with few previous utterances and previous DA's as features It includes evaluation of our models with several setups of utilizing previous DA's history either a gold standard or predicted ones during both training and testing of the model. Although performance is low for few rare-class DA's, adding contextual information and previous DA tags boosts performance in several cases. The purpose of building this automatic DA's tagging model as an auxiliary task is to produce DA's that can help in identifying indicators of AD. This experimentation, answered research **Research**

Question 2.

Chapter 5, extends the experiments with rare class dialogue act tagging by including different lexical representations of utterances as well as acoustic features from speech. The purpose is to improve the class-wise accuracy of certain classes. I compared the performance against both static utterance representation and contextualized utterance representation. It was observed that different features are appropriate for different DA classes. Adding acoustic features had improved the performance for the question categories such as the declarative yes-no question, open-ended question, and wh-question while additional features along with acoustic helped in improving class wise accuracy of clarification request. This is further improved with contextualized utterance representation. **Research Question 3** is investigated in this chapter.

In chapter 6, the benefits of using interactional patterns are investigated in conversations to identify indicators of AD. The results showed that by combining interactional features with disfluency features helps more in improving the accuracy of AD predictions. Edit terms with repeats and substitutions were more frequent in the AD group than the Non-AD group. AD patients produce more longer pauses within the sentence boundary showing lexical retrieval difficulty. Similarly, the duration of lapses in AD conversations is longer than in Non-AD conversations. DA's as unigram and bigram features are also included for AD classification experiments and found these sequences useful for AD. The set of experiments in this chapter answers our **Research Questions 4 and 5**.

Later in chapter 7, multimodal fusion-based models were presented using early fusion, late fusion, and multimodal deep learning models to classify whether a speaker involved in the semi-structured conversation has AD. Our best model used interaction features including pauses, Ngram DA sequences, and a set of acoustic features with early fusion and feature selection. This chapter answers the **Research question 6 and 7**.

8.1 Summary of contributions

Following is a list of contributions that are achieved in this thesis:

1. A rare class DA annotation scheme is developed that is suitable for conversational dialogues to capture the dialogue phenomena and present a comparative analysis of the subject's interaction patterns through a corpus study.

2. A subset of Carolina's collection corpus is annotated with the proposed rare class DA annotation scheme. These annotations are available for the research community for further follow-up work and can be used after getting access to the CCC dataset[1].

3. A subset of the CCC corpus now includes an additional collection of conversational pauses with annotations.

[1]Annotations: https://osf.io/8w9z2/?view_only=ee08242870f24ae7ab6754ddf9a0176a

4. To capture the dialogue interaction, a rare class dialogue act tagger is developed with deep-learning models that leverage utterance representation with word embeddings, speaker change information, a few previous utterances, and DA's as context to predict the DA for the current utterance. The model utilised both static utterance representation and pre-trained contextualized utterance representation along with acoustic features from speech data. I also fine-tune the pre-trained BERT model on the downstream task on both corpora and achieve the most robust performance (macro F1: 0.58 0n SwDA and 0.48 on CCC). Later it was shown that the challenging setup of rare-class DA labeling for better recognizing rare classes in the CCC data set really helped to detect Alzheimer's disease better.

5. To capture the full conversation context, a conversational level DA tagger is developed using a CRF to model the sequence of DA tags , following an approach used successfully in general DA tagging work (e.g. (Kumar et al., 2018; Srivastava et al., 2019)). However, this CRF approach did not perform very well with our rare DA classes, suggesting that it is not well suited for our task with its highly uneven class distribution.

6. A set of interactional features is proposed including conversation pauses, for the identification of AD and it showed that they yield high utility in differentiating between AD and Non-AD. The disfluency features are also combined with interactional features and it was shown that the combination of the interaction features with the disfluency features is helpful for the AD classification task in a natural setting. These models were able to achieve performance comparable to the work that is based on content-driven task-specific settings.

7. Both unimodal and multimodal AD detection methods were tested and the performance is compared and contrasted over several combinations from different feature sets including interactional features, acoustic features, and DA's based features e.g. unigram, bigram, and confusion ratios. These models learn AD markers using speech and text visual modalities. A comprehensive research of fusion strategies for including early fusion, late fusion, and multimodal fusion is also presented. Lastly, it is shown that feature selection along with fusion strategy outperforms when the features are alone used for AD identification.

8.2 Limitations and Future Work

Stage of disease: The experimental work presented in this thesis focuses on dialogue interaction using a conversational dataset. One of the limitations of this study is that the dataset includes only older patients with diagnosed dementia, it cannot directly tell us whether these patterns extend to early-stage diagnosis; it can only allow us to

observe patterns linked with AD at relatively advanced stage. The detailed cognitive scores are not available outside the original study. It does, however, have the advantage of containing relatively free conversational interaction, than the more formulaic tasks in e.g. DementiaBank. In a longitudinal study for monitoring and diagnosis of Dementia, a dataset is currently being collected that contains transcribed and recorded speech in several sessions for both control and Dementia patients (though contains less number of participants) (Gkoumas et al., 2021). It would be interesting in the future to explore this dataset in terms of interaction, interactional features, conversational pauses, and DA-based features and see how the progression can be captured through these features. In addition, their is also a need for the collection of new dataset that is conversational in nature and it would contain more number of AD patients.

Size of dataset: In addition to this, the other limitation of the study is that 30 participants were used for the experiments though balanced in terms of AD and Non-AD groups and results may not be generalizable. Although some recent studies focusing on interaction have used kind of similar number of participants (Mirheidari et al. (2019) used 30 conversations for neurologist-patient dataset) or a slightly bigger set (Luz et al. (2018) used 38 dialogues from the same CCC corpus). This study may be further extended by including more samples from the corpus.

Automatic pause detection: Identifying pauses and their function in speech is key to analyzing conversations and also can be useful for the automatic diagnosis of dementia like AD. Our findings from conversational pause analysis and experimental work in chapter 6 show that these pauses are very strong predictors and could be used for conversational agents in the clinical domain. In the future, we will try to automate pauses detection as an auxiliary task for AD classification. For this purpose, our aim to utilize lexical input, acoustic information, and DA-based information as well. As DA per utterance basis reveals important information such as the previous utterance is a question and the noticeable silence shows no response in terms of attributable silence.

More interactional aspects: Capturing interaction in terms of question-answer and finding whether the response is relevant or not needs further consideration and will be investigated in the future. I may also explore and further categorize question categories in terms of questions related to semantic and episodic memory could provide useful insights into the cognitive processes. It would be interesting to explore advanced acoustic analysis techniques, such as leveraging models like wav2vec (Yuan et al., 2017) and Whisper for enhanced word embedding, could provide valuable insights and contribute to the continual refinement of the study's methodologies.

Applicability to other Mental diseases In future, it will be investigated whether the proposed interactional patterns and features, initially designed for the analysis of Alzheimer's disease, can be extended and applied to other conditions such as bipolar disorder or depression. Individuals with depression may exhibit reduced cognitive processing speed, may ask fewer clarification questions, and answer questions slowly with low speech rates, reflecting a commonality with Alzheimer's disease. Examining the generalizability of the identified patterns may provide a broader understanding of their applicability across various mental health contexts.

State of the art NLP modeling: In-depth research has been done on contextualized embeddings in NLP. There are numerous domain-specific variations of contextualized text embeddings that have been pre-trained. To improve the performance of a particular domain task, representation learning in particular domains seeks to encode domain-specific information. In the future, more experiments will be performed with domain-specific representation learning methods that are helpful to capture behavioral tendencies. For the said purpose, DialogueBERT (Gu et al., 2021), MentalBERT (Ji et al., 2021), and similar alternatives may be investigated.

BIBLIOGRAPHY

Recursive feature elimination (rfe). https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.RFE.html, 2024. Accessed: March 9, 2024.

Stefanie Abel, Walter Huber, and Gary S Dell. Connectionist diagnosis of lexical disorders in Aphasia. *Aphasiology*, 23(11):1353–1378, 2009.

Angus Addlesee, Arash Eshghi, and Ioannis Konstas. Current challenges in spoken dialogue systems and why they are critical for those living with Dementia. *arXiv preprint arXiv:1909.06644*, 2019.

Samrah Ahmed, Anne Marie F. Haigh, Celeste A. de Jager, and Peter Garrard. Connected speech as a marker of disease progression in autopsy-proven Alzheimer's disease. *Brain : a Journal of Neurology*, 136(Pt 12):3727–3737, 2013. ISSN 14602156. doi: 10.1093/brain/awt269.

Antti Airola, Tapio Pahikkala, Willem Waegeman, Bernard De Baets, and Tapio Salakoski. An experimental comparison of cross-validation techniques for estimating the area under the ROC curve. *Computational Statistics & Data Analysis*, 55(4): 1828–1844, 2011.

Zakaria Aldeneh, Mimansa Jaiswal, Michael Picheny, Melvin McInnis, and Emily Mower Provost. Identifying Mood Episodes using dialogue features from clinical interviews. *arXiv preprint arXiv:1910.05115*, 2019.

Emilia Ambrosini, Matteo Caielli, Marios Milis, Christos Loizou, Domenico Azzolino, Sarah Damanti, Laura Bertagnoli, Matteo Cesari, Sara Moccia, Manuel Cid, et al. Automatic speech analysis to early detect functional cognitive decline in elderly population. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 212–216. IEEE, 2019.

Anne H. Anderson, Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, Catherine Sotillo, Henry S. Thompson, and Regina Weinert. The hcrc Map Task Corpus. *Language and Speech*, 34(4):351–366, 1991. doi: 10.1177/002383099103400404. URL https://doi.org/10.1177/002383099103400404.

Jeremy Ang, Yang Liu, and Elizabeth Shriberg. Automatic dialog act segmentation and classification in multiparty meetings. *Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, 1:I/1061–I/1064 Vol. 1, 2005.

Harish Arsikere, Arunasish Sen, AP Prathosh, and Vivek Tyagi. Novel acoustic features for automatic dialog-act tagging. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6105–6109. IEEE, 2016.

Meysam Asgari, Jeffrey Kaye, and Hiroko Dodge. Predicting mild cognitive impairment from spontaneous spoken utterances. *Alzheimer's & Dementia: Translational Research & Clinical Interventions*, 3(2):219–228, 2017.

Alzheimer's Association. Alzheimer's disease facts and figures. *Alzheimer's & Dementia*, 13(4):325–373, 2017.

Alzheimer's Association. 2023 alzheimer's disease facts and figures. *Alzheimer's & Dementia*, 19(4):1598–1695, 2023.

Daniela Beltrami, Gloria Gagliardi, Rema Rossini Favretti, Enrico Ghidoni, Fabio Tamburini, and Laura Calzà. Speech analysis by natural language processing techniques: a possible tool for very early detection of cognitive decline? *Frontiers in aging Neuroscience*, 10:369, 2018.

Veronica Boschi, Eleonora Catricala, Monica Consonni, Cristiano Chesi, Andrea Moro, and Stefano F Cappa. Connected speech in neurodegenerative language disorders: a review. *Frontiers in Psychology*, 8:269, 2017.

Chandrakant Bothe, Cornelius Weber, Sven Magg, and Stefan Wermter. A context-based approach for dialogue act recognition using simple Recurrent neural networks. *arXiv preprint arXiv:1805.06280*, 2018.

Paul T Brady. A statistical analysis of on-off patterns in 16 conversations. *Bell System Technical Journal*, 47(1):73–91, 1968.

Jason Brandt. The hopkins verbal learning test: Development of a new memory test with six equivalent forms. *The Clinical Neuropsychologist*, 5(2):125–142, 1991.

H. Bunt, V. Petukhova, A. Malchanau, K. Wijnhoven, and A. Fang. The DialogBank. In *LREC*, 2016.

Harry Bunt, Jan Alexandersson, Jean Carletta, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Kiyong Lee, Volha Petukhova, Andrei Popescu-Belis, Laurent Romary, et al. Towards an ISO standard for dialogue act annotation. 2010.

Edward L Campbell, Laura Docío-Fernández, Javier Jiménez Raboso, and Carmen García-Mateo. Alzheimer's Dementia detection from audio and text modalities. *arXiv preprint arXiv:2008.04617*, 2020.

Rupayan Chakraborty, Meghna Pandharipande, Chitralekha Bhat, and Sunil Kumar Kopparapu. Identification of Dementia using audio biomarkers. *arXiv preprint arXiv:2002.12788*, 2020.

Saurabh Chakravarty, Raja Venkata Satya Phanindra Chava, and Edward A Fox. Dialog acts classification for question-answer corpora. In *ASAIL@ ICAIL*, 2019.

Sandra Bond Chapman, Amy Peterson Highley, and Jennifer L Thompson. Discourse in fluent Aphasia and Alzheimer's disease: Linguistic and pragmatic considerations. *Journal of Neurolinguistics*, 11(1-2):55–78, 1998.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. SMOTE: Synthetic Minority Over-Sampling Technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

Nitesh V Chawla, Nathalie Japkowicz, and Aleksander Kotcz. Special issue on learning from imbalanced data sets. *ACM SIGKDD explorations newsletter*, 6(1):1–6, 2004.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

Jacob Cohen. *Statistical power analysis for the behavioral sciences*. Routledge, 2013.

Mark G Core and James Allen. Coding dialogs with the DAMSL annotation scheme. In *AAAI fall symposium on communicative action in humans and machines*, volume 56. Boston, MA, 1997.

David Crystal and Derek Davy. *Investigating english style*. Routledge, 2016.

Boyd Davis and Margaret Maclagan. Pauses, fillers, placeholders and formulaicity in Alzheimer's discourse. *Fillers, pauses and placeholders*, 93:189, 2010.

Boyd Davis, M Maclagan, and Shenk D. Exploring interactions between visitors and residents with Dementia, with a focus on questions and the responses they evoke. In *The Routledge handbook of language and health communication*, pages 344–360. Routledge, The city, 2014a.

Boyd Davis, Margaret Maclagan, and Dena Shenk. *Exploring communicative interactions between visitors and assisted-living residents with Dementia*, chapter 21. Routledge, 2014b. doi: 10.4324/9781315856971.ch21. URL https://www.routledgehandbooks.com/doi/10.4324/9781315856971.ch21.

Boyd H Davis and Margaret Maclagan. Examining pauses in Alzheimer's discourse. *American Journal of Alzheimer's Disease & Other Dementias®*, 24(2):141–154, 2009.

Sofia de la Fuente Garcia, Craig W Ritchie, and Saturnino Luz. Protocol for a conversation-based analysis study: Prevent-ed investigates dialogue features that may help predict Dementia onset in later life. *BMJ open*, 9(3):e026254, 2019.

Juliana Onofre de Lira, Karin Zazo Ortiz, Aline Carvalho Campanha, Paulo Henrique Ferreira Bertolucci, and Thaís Soares Cianciarullo Minett. Microlinguistic aspects of the oral narrative in patients with Alzheimer's disease. *International Psychogeriatrics*, 23 (3):404–412, 2011.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR*, abs/1810.04805, 2018. URL http://arxiv.org/abs/1810.04805.

Mark Dingemanse, Seán G Roberts, Julija Baranova, Joe Blythe, Paul Drew, Simeon Floyd, Rosa S Gisladottir, Kobin H Kendrick, Stephen C Levinson, Elizabeth Manrique, et al. Universal principles in the repair of communication problems. *PloS one*, 10(9): e0136100, 2015.

Cláudia Drummond, Gabriel Coutinho, Rochele Paz Fonseca, Naima Assunção, Alina Teldeschi, Ricardo de Oliveira-Souza, Jorge Moll, Fernanda Tovar-Moll, and Paulo Mattos. Deficits in narrative discourse elicited by visual stimuli are already present in patients with mild cognitive impairment. *Frontiers in Aging Neuroscience*, 7:96, 2015.

Nathan Duran and Steve Battle. Probabilistic word association for dialogue act classification with Recurrent neural networks. In *International Conference on Engineering Applications of Neural Networks*, pages 229–239. Springer, 2018.

Nathan Duran, Steve Battle, and Jim Smith. Sentence encoding for Dialogue Act classification. *Natural Language Engineering*, page 1–30, 2021. doi: 10.1017/S1351324921000310.

Samira Elouakili. A conversation analysis approach to attributable silence in Moroccan conversation. *International Research in Education*, 5(2):1–21, 2017.

Christopher Elsey, Paul Drew, Danielle Jones, Daniel Blackburn, Sarah Wakefield, Kirsty Harkness, Annalena Venneri, and Markus Reuber. Towards diagnostic conversational profiles of patients presenting with Dementia or functional memory disorders to memory clinics. *Patient Education and Counseling*, 98(9):1071–1077, 2015.

Florian Eyben, Martin Wöllmer, and Björn Schuller. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1459–1462, 2010.

Shahla Farzana and Natalie Parde. Are interaction patterns helpful for task-agnostic Dementia detection? an empirical exploration. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 172–182, 2022.

Shahla Farzana, Mina Valizadeh, and Natalie Parde. Modeling dialogue in conversational cognitive health screening interviews. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1167–1177, 2020.

Carole T Ferrand. Harmonics-to-noise ratio: an index of vocal aging. *Journal of voice*, 16 (4):480–487, 2002.

Steven H Ferris and Martin Farlow. Language impairment in alzheimer's disease and benefits of acetylcholinesterase inhibitors. *Clinical interventions in aging*, 8:1007, 2013.

M F Folstein, S E Folstein, and P R McHugh. Mini-Mental Status. a practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, 12(3):189–198, 1975.

Katrina E Forbes-McKay and Annalena Venneri. Detecting subtle spontaneous language decline in early Alzheimer's disease with a picture description task. *Neurological Sciences*, 26(4):243–254, 2005.

Kristina Lundholm Fors. An investigation of intra-turn pauses in spontaneous speech, 2011.

Kathleen C Fraser, Jed A Meltzer, Naida L Graham, Carol Leonard, Graeme Hirst, Sandra E Black, and Elizabeth Rochon. Automated classification of primary progressive Aphasia sub types from narrative speech transcripts. *cortex*, 55:43–60, 2014.

Kathleen C Fraser, Jed A Meltzer, and Frank Rudzicz. Linguistic features identify Alzheimer's disease in narrative speech. *Journal of Alzheimer's Disease*, 49(2):407–422, 2015. ISSN 18758908.

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

Kathleen C. Fraser, Frank Rudzicz, and Graeme Hirst. Detecting late-life depression in Alzheimer's disease through analysis of speech and language. In *Proc. CLPsych*, pages 1–11, San Diego, CA, USA, June 2016b. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W16-0301.

Kathleen C Fraser, Frank Rudzicz, and Graeme Hirst. Detecting late-life depression in Alzheimer's disease through analysis of speech and language. In *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*, pages 1–11, 2016c.

Simone Fuscone, Benoit Favre, and Laurent Prévot. The contribution of dialogue act labels for convergence studies in natural conversations. In *Proceedings of the 24th Workshop on the Semantics and Pragmatics of Dialogue-Full Papers*, 2020.

Frederique Gayraud, Hye-Ran Lee, and Melissa Barkat-Defradas. Syntactic and lexical context of pauses and hesitations in the discourse of Alzheimer patients and healthy elderly subjects. *Clinical linguistics & phonetics*, 25(3):198–209, 2011.

Aurélien Géron. *Hands-on Machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*. O'Reilly Media, 2019.

Dimitris Gkoumas, Bo Wang, Adam Tsakalidis, Maria Wolters, Arkaitz Zubiaga, Matthew Purver, and Maria Liakata. A longitudinal multi-modal dataset for Dementia monitoring and diagnosis. *arXiv preprint arXiv:2109.01537*, 2021.

Harold Goodglass, Edith Kaplan, Sandra Weintraub, and Barbara Barresi. The Boston Diagnostic Aphasia Examination. 2001.

DALIA GOTTLIEB-TANAKA, JEFF SMALL, and ANNALEE YASSI. A programme of creative expression activities for seniors with Dementia. *Dementia*, 2(1):127–133, 2003.

Sergio Grau, Emilio Sanchis, Maria Jose Castro, and David Vilar. Dialogue act classification using a Bayesian approach. In *9th Conference Speech and Computer*, 2004.

Xiaodong Gu, Kang Min Yoo, and Jung-Woo Ha. Dialogbert: Discourse-aware response generation via learning to recover and rank utterances. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12911–12919, 2021.

Juan S Guerrero-Cristancho, Juan C Vásquez-Correa, and Juan R Orozco-Arroyave. Word-embeddings and grammar features to detect language disorders in Alzheimer's disease patients. *TecnoLógicas*, 23(47):63–75, 2020.

Heidi Ehernberger Hamilton. *Conversations with an Alzheimer's patient: An interactional sociolinguistic study*. Cambridge University Press, 2005.

Mattias Heldner and Jens Edlund. Pauses, gaps and overlaps in conversations. *Journal of Phonetics*, 38(4):555–568, 2010.

Julian Hough and David Schlangen. Joint, incremental disfluency detection and utterance segmentation from speech. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 326–336, 2017.

William Jarrold, Bart Peintner, David Wilkins, Dimitra Vergryi, Colleen Richey, Maria Luisa Gorno-Tempini, and Jennifer Ogar. Aided diagnosis of dementia type through computer-based analysis of spontaneous speech. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 27–37, 2014.

Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. Mentalbert: Publicly available pretrained language models for mental healthcare. *arXiv preprint arXiv:2110.15621*, 2021.

Danielle Jones. A family living with Alzheimer's disease: The communicative challenges. *Dementia*, 14(5):555–573, 2015.

Danielle Jones, Paul Drew, Christopher Elsey, Daniel Blackburn, Sarah Wakefield, Kirsty Harkness, and Markus Reuber. Conversational assessment in memory clinic encounters: interactional profiling for differentiating Dementia from functional memory disorders. *Aging & Mental Health*, 20(5):500–509, 2016.

Alaa Joukhadar, Nada Ghneim, and Ghaida Rebdawi. Impact of Using Bidirectional Encoder Representations from Transformers (BERT) Models for Arabic Dialogue Acts Identification. *Ingénierie des Systèmes d Inf.*, 26(5):469–475, 2021.

Daniel Jurafsky, Elizabeth Shriberg, and Debra Biasca. Switchboard SWBD-DAMSL Shallow-Discourse-Function Annotation Coders Manual. *URL http://web. stanford. edu˜ jurafsky/ws97/manual. august1. html*, 1997.

Elke Kalbe, Josef Kessler, Pasquale Calabrese, R Smith, AP Passmore, Met al Brand, and R Bullock. Demtect: a new, sensitive cognitive screening test to support the diagnosis of mild cognitive impairment and early Dementia. *International journal of geriatric psychiatry*, 19(2):136–143, 2004.

Nal Kalchbrenner and Phil Blunsom. Recurrent Convolutional neural networks for discourse compositionality. *arXiv preprint arXiv:1306.3584*, 2013.

Gitit Kavé and Ayelet Dassa. Severity of Alzheimer's disease and language features in picture descriptions. *Aphasiology*, 32(1):27–40, 2018. ISSN 14645041. doi: 10.1080/02687038.2017.1303441.

Gitit Kavé and Ayelet Dassa. Severity of Alzheimer's disease and language features in picture descriptions. *Aphasiology*, 32(1):27–40, 2018.

Gitit Kavé and Yonata Levy. Morphology in picture descriptions provided by persons with Alzheimer's disease. *Journal of speech, language, and hearing research*, 2003.

Ali Khodabakhsh, Fatih Yesil, Ekrem Guner, and Cenk Demiroglu. Evaluation of linguistic and prosodic features for detection of Alzheimer's disease in Turkish conversational speech. *EURASIP Journal on Audio, Speech, and Music Processing*, 2015(1):1–15, 2015.

Hae-Young Kim. Statistical notes for clinical researchers: chi-squared test and Fisher's Exact test. *Restorative dentistry & endodontics*, 42(2):152–155, 2017.

Diederik P Kingma and Jimmy Ba. Adam: A method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980*, 2014.

MD Kopelman, BA Wilson, and AD Baddeley. The autobiographical memory interview: a new assessment of autobiographical and personal semantic memory in Amnesic patients. *Journal of Clinical and Experimental Neuropsychology*, 11(5):724–744, 1989.

MD Kopelman, BA Wilson, and AD Baddeley. The autobiographical memory interview (manual). *Thames Valley Test Company, Bury St. Edmunds, England*, 1990.

Michael D Kopelman. Remote and autobiographical memory, temporal context memory and frontal atrophy in Korsakoff and Alzheimer patients. *Neuropsychologia*, 27(4): 437–460, 1989.

Harshit Kumar, Arvind Agarwal, Riddhiman Dasgupta, and Sachindra Joshi. Dialogue act sequence labeling using hierarchical encoder with CRF. In *Proceedings of the AAAI conference on artificial intelligence*, 2018.

John Lafferty, Andrew McCallum, Fernando Pereira, et al. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Icml*, volume 1, page 3. Williamstown, MA, 2001.

Christine Le Boeuf. *Raconte...: 55 historiettes en images*. L'école, 1976.

Ji Young Lee and Franck Dernoncourt. Sequential short-text classification with Recurrent and Convolutional neural networks. *arXiv preprint arXiv:1603.03827*, 2016.

Willem JM Levelt. Monitoring and self-repair in speech. *Cognition*, 14(1):41–104, 1983.

Stephen C Levinson. Pragmatics Cambridge University Press. *Cambridge UK*, 1983.

Stephen C Levinson and Francisco Torreira. Timing in turn-taking and its implications for processing models of language. *Frontiers in psychology*, 6:731, 2015.

Ruizhe Li, Chenghua Lin, Matthew Collinson, Xiao Li, and Guanyi Chen. A dual-attention hierarchical recurrent neural network for dialogue act classification. *arXiv preprint arXiv:1810.09154*, 2018.

Yuanchao Li, Catherine Lai, Divesh Lala, Koji Inoue, and Tatsuya Kawahara. Alzheimer's dementia detection through spontaneous dialogue with proactive robotic listeners. In *HRI*, pages 875–879, 2022.

Honghuang Lin, Cody Karjadi, Ting FA Ang, Joshi Prajakta, Chelsea McManus, Tuka W Alhanai, James Glass, and Rhoda Au. Identification of digital voice biomarkers for cognitive health. *Exploration of Medicine*, 1:406, 2020.

Juliana Onofre de Lira, Thaís Soares Cianciarullo Minett, Paulo Henrique Ferreira Bertolucci, and Karin Zazo Ortiz. Analysis of word number and content in discourse of patients with mild to moderate Alzheimer's disease. *Dementia & neuropsychologia*, 8: 260–265, 2014.

Yang Liu, Kun Han, Zhao Tan, and Yun Lei. Using context information for dialog act classification in dnn framework. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2170–2178, 2017.

Karmele López-de Ipiña, Jesus-Bernardino Alonso, Carlos Manuel Travieso, Jordi Solé-Casals, Harkaitz Egiraun, Marcos Faundez-Zanuy, Aitzol Ezeiza, Nora Barroso, Miriam Ecay-Torres, Pablo Martinez-Lage, et al. On the selection of non-invasive methods based on speech analysis oriented to automatic Alzheimer disease diagnosis. *Sensors*, 13(5):6730–6745, 2013.

Saturnino Luz, Sofia de la Fuente, and Pierre Albert. A method for analysis of patient speech in dialogue for Dementia detection. *arXiv preprint arXiv:1811.09919*, 2018.

Saturnino Luz, Fasih Haider, Sofia de la Fuente, Davida Fromm, and Brian MacWhinney. Alzheimer's Dementia recognition through spontaneous speech: the ADReSS challenge. *arXiv preprint arXiv:2004.06833*, 2020.

Saturnino Luz, Fasih Haider, Sofia de la Fuente, Davida Fromm, and Brian MacWhinney. Detecting cognitive decline using speech only: The adresso challenge. *arXiv preprint arXiv:2104.09356*, 2021.

Harry Maltby, Julie Wall, T Goodluck Constance, Mansour Moniri, Cornelius Glackin, Marvin Rajwadi, and Nigel Cannings. Short utterance dialogue act classification using a Transformer Ensemble. *UA-DIGITAL 2023: UA Digital Theme Research Twinning*, 2023.

F Martínez-Sánchez, JJG Meilán, J García-Sevilla, J Carro, and JM Arana. Oral reading fluency analysis in patients with Alzheimer disease and asymptomatic control subjects. *Neurología (English Edition)*, 28(6):325–331, 2013.

Guy Mckhann, David Drachman, and Marshall Folstein. views & reviews Clinical diagnosis of Alzheimer ' s disease :. *Neurology*, 34(7):939—-944, 1984. ISSN 0361-9230. doi: 10.1186/alzrt38.

Stefano Mezza, Alessandra Cervone, Giuliano Tortoreto, Evgeny A Stepanov, and Giuseppe Riccardi. ISO-standard domain-independent dialogue act tagging for conversational agents. *arXiv preprint arXiv:1806.04327*, 2018.

Bahman Mirheidari, DJ Blackburn, Kirsty Harkness, Traci Walker, Annalena Venneri, Markus Reuber, and Heidi Christensen. An avatar-based system for identifying individuals likely to develop Dementia. In *Interspeech 2017*, pages 3147–3151. ISCA, 2017.

Bahman Mirheidari, Daniel Blackburn, Traci Walker, Markus Reuber, and Heidi Christensen. Dementia detection using automatic analysis of conversations. *Computer Speech & Language*, 53:65–79, 2019.

Shamila Nasreen, Matthew Purver, and Julian Hough. A corpus study on questions, responses and misunderstanding signals in conversations with Alzheimer's patients. In *Proceedings of the 23rd Workshop on the Semantics and Pragmatics of Dialogue-Full Papers. SEMDIAL, London, United Kingdom (Sep 2019), http://semdial.org/anthology/Z19-Nasreen semdial*, volume 13, 2019.

Shamila Nasreen, Julian Hough, and Matthew Purver. Rare-class dialogue act tagging for Alzheimer's disease diagnosis. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 290–300, 2021a.

Shamila Nasreen, Julian Hough, Matthew Purver, et al. Detecting Alzheimer's disease using interactional and acoustic features from spontaneous speech. Interspeech, 2021b.

Shamila Nasreen, Morteza Rohanian, Julian Hough, and Matthew Purver. Alzheimer's Dementia recognition from spontaneous speech using disfluency and interactional features. *Frontiers in Computer Science*, page 49, 2021c.

Liu Ning and Kexue Luo. Using text and acoustic features to diagnose Mild Cognitive Impairment and Alzheimer's disease. 2020.

Bill Noble and Vladislav Maraev. Large-scale text pre-training helps with dialogue act recognition, but not without fine-tuning. In *Proceedings of the 14th International Conference on Computational Semantics (IWCS)*, pages 166–172, 2021.

Peter Noone. Addenbrooke's Cognitive Examination-III. *Occupational Medicine*, 65: 418–420, 2015.

Jun Ogata, Masataka Goto, and Katunobu Itou. The use of acoustically detected filled and silent pauses in spontaneous speech recognition. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4305–4308. IEEE, 2009.

Sylvester O Orimaye, Jojo SM Wong, Karen J Golden, Chee P Wong, and Ireneous N Soyiri. Predicting probable Alzheimer's disease using linguistic deficits and biomarkers. *BMC Bioinformatics*, 18(1):34, 2017.

Daniel Ortega and Ngoc Thang Vu. Neural-based context representation learning for dialog act classification. *arXiv preprint arXiv:1708.02561*, 2017.

Daniel Ortega and Ngoc Thang Vu. Lexico-Acoustic neural-based models for dialog act classification. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6194–6198. IEEE, 2018.

Daniel Ortega, Chia-Yu Li, Gisela Vallejo, Pavel Denisov, and Ngoc Thang Vu. Context-aware neural-based dialog act classification on automatically generated transcriptions. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7265–7269. IEEE, 2019.

Serguei VS Pakhomov, Susan E Marino, and Angela K Birnbaum. Quantification of speech disfluency as a marker of medication-induced cognitive impairment: An application of computerized speech analysis in neuropharmacology. *Computer Speech & Language*, 27(1):116–134, 2013.

Pinelopi Papalampidi, Elias Iosif, and Alexandros Potamianos. Dialogue act semantic representation and classification using Recurrent Neural Networks. *Proc. SEMDIAL*, pages 77–86, 2017.

Florence Pasquier, Florence Lebert, Laurence Grymonprez, and Henri Petit. Verbal fluency in Dementia of frontal lobe type and Dementia of Alzheimer type. *Journal of Neurology, Neurosurgery & Psychiatry*, 58(1):81–84, 1995.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

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

Lisa Perkins, Anne Whitworth, and Ruth Lesser. Conversing in Dementia: A conversation analytic approach. *Journal of Neurolinguistics*, 11(1-2):33–53, 1998.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*, 2018.

A Pistono, J Pariente, C Bézy, B Lemesle, J Le Men, and M Jucla. What happens when nothing happens? an investigation of pauses as a compensatory mechanism in early Alzheimer's disease. *Neuropsychologia*, 124:133–143, 2019a.

Aurélie Pistono, M Jucla, C Bézy, B Lemesle, J Le Men, and J Pariente. Discourse macrolinguistic impairment as a marker of linguistic and extralinguistic functions decline in early Alzheimer's disease. *International journal of language & communication disorders*, 54(3):390–400, 2019b.

Charlene Pope and Boyd H Davis. Finding a balance: The Carolina's Conversation Collection. *Corpus Linguistics and Linguistic Theory*, 7(1):143–161, 2011.

Matthew Purver, Jonathan Ginzburg, and Patrick Healey. On the means for clarification in dialogue. In *Current and new directions in discourse and dialogue*, pages 235–255. Springer, 2003.

Vipul Raheja and Joel Tetreault. Dialogue act classification with context-aware Self-Attention. *arXiv preprint arXiv:1904.02594*, 2019.

Nithin Ramacandran. Dialogue Act Detection from Human-Human Spoken Conversations. *International Journal of Computer Applications*, 67(5), 2013.

Vai Ramanathan. *Alzheimer discourse: Some sociolinguistic dimensions*. Routledge, 2013.

Vivek Rangarajan, Srinivas Bangalore, and Shrikanth Narayanan. Exploiting prosodic features for dialog act tagging in a discriminative modeling framework. *Proceedings of Interspeech, Antwerp, Belgium*, 2007.

Danielle N Ripich, Diane Vertes, Peter Whitehouse, Sarah Fulton, and Barbara Ekelman. Turn-taking and speech act patterns in the discourse of Senile Dementia of the Alzheimer's type patients. *Brain and Lsanguage*, 40(3):330–343, 1991.

Brian Roark, Margaret Mitchell, John-Paul Hosom, Kristy Hollingshead, and Jeffrey Kaye. Spoken language derived measures for detecting mild cognitive impairment. *IEEE Transactions on Audio, Speech, and Language processing*, 19(7):2081–2090, 2011.

Kepa Rodríguez and David Schlangen. Form, intonation and Function of Clarification Requests in German Task-Oriented Spoken Dialogues. In *Proceedings of the 8th Workshop on the Semantics and Pragmatics of Dialogue (Catalog)*, pages 101–108, Barcelona, July 2004.

Morteza Rohanian and Julian Hough. Re-framing Incremental deep language models for dialogue processing with multi-task learning. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 497–507, 2020.

Morteza Rohanian, Julian Hough, and Matthew Purver. Multi-Modal Fusion with Gating Using Audio, Lexical and Disfluency Features for Alzheimer's Dementia Recognition from Spontaneous Speech. In *Proc. Interspeech 2020*, pages 2187–2191, 2020a. doi: 10.21437/Interspeech.2020-2721. URL http://dx.doi.org/10.21437/Interspeech.2020-2721.

Morteza Rohanian, Julian Hough, and Matthew Purver. Multi-modal fusion with gating using audio, lexical and disfluency features for Alzheimer's Dementia recognition from spontaneous speech. *Proc. Interspeech 2020*, pages 2187–2191, 2020b.

Jonathan D Rohrer, Martin N Rossor, and Jason D Warren. Syndromes of nonfluent primary progressive Aphasia: a clinical and neurolinguistic analysis. *Neurology*, 75(7): 603–610, 2010.

Harvey Sacks, Emanuel A Schegloff, and Gail Jefferson. A simplest systematics for the organization of turn taking for conversation. In *Studies in the organization of conversational interaction*, pages 7–55. Elsevier, 1978.

Omer Sagi and Lior Rokach. Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):e1249, 2018.

Emanuel A Schegloff, Gail Jefferson, and Harvey Sacks. The preference for self-correction in the organization of repair in conversation. *Language*, 53(2):361–382, 1977.

Z Shah, J Sawalha, M Tasnim, S Qi, E Stroulia, and R Greiner. Learning Language and acoustic models for identifying Alzheimer's Dementia from speech. *Front. Comput. Sci. 3: 624659. doi: 10.3389/fcomp*, 2021.

Dena Shenk. Watching what you say : Walking and conversing in Dementia preliminary studies background : Relationship between. *Topics in Geriatric Rehabilitation*, 27(4): 268–277, 2011. doi: 10.1097/TGR.0b013e31821e58db.

Elizabeth Shriberg, Andreas Stolcke, Daniel Jurafsky, Noah Coccaro, Marie Meteer, Rebecca Bates, Paul Taylor, Klaus Ries, Rachel Martin, and Carol Van Ess-Dykema. Can prosody aid the automatic classification of dialog acts in conversational speech? *Language and speech*, 41(3-4):443–492, 1998.

Yuke Si, Longbiao Wang, Jianwu Dang, Mengfei Wu, and Aijun Li. A hierarchical model for dialog act recognition considering acoustic and lexical context information. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7994–7998, 2020a. doi: 10.1109/ICASSP40776.2020.9053061.

Yuke Si, Longbiao Wang, Jianwu Dang, Mengfei Wu, and Aijun Li. A hierarchical model for dialog act recognition considering acoustic and lexical context information. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7994–7998. IEEE, 2020b.

Sidney Siegel and NJ Castellan. Measures of association and their tests of significance. *Nonparametric statistics for the Behavioral Sciences*, pages 224–312, 1988.

Sameer Singh, Romola S Bucks, and Joanne M Cuerden. Evaluation of an objective technique for analysing temporal variables in dat spontaneous speech. *Aphasiology*, 15(6):571–583, 2001.

Han Sloetjes and Peter Wittenburg. Annotation by category-ELAN and ISO DCR. In *6th international Conference on Language Resources and Evaluation (LREC 2008)*, 2008.

Jeff A Small and JoAnn Perry. Do you remember? how caregivers question their spouses who have Alzheimer's disease and the impact on communication. *Journal of Speech, Language, and Hearing Research*, 48(1):125–136, 2005.

Gordon CS Smith, Shaun R Seaman, Angela M Wood, Patrick Royston, and Ian R White. Correcting for optimistic prediction in small data sets. *American journal of epidemiology*, 180(3):318–324, 2014.

Vivek Kumar Rangarajan Sridhar, Srinivas Bangalore, and Shrikanth Narayanan. Combining lexical, syntactic and prosodic cues for improved online dialog act tagging. *Computer Speech & Language*, 23(4):407–422, 2009.

Saurabh Srivastava, Puneet Agarwal, Gautam Shroff, and Lovekesh Vig. Hierarchical capsule based neural network architecture for sequence labeling. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2019.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3):339–373, 2000.

Dinoj Surendran and Gina-Anne Levow. Dialog act tagging with Support Vector Machines and Hidden Markov Models. In *Ninth International Conference on Spoken Language Processing*, 2006.

Mario Taschwer, Manfred Jürgen Primus, Klaus Schoeffmann, and Oge Marques. Early and late fusion of classifiers for the MediaEval Medico Task. In *MediaEval*, 2018.

João Paulo Teixeira, Carla Oliveira, and Carla Lopes. Vocal acoustic analysis–jitter, shimmer and HNR parameters. *Procedia Technology*, 9:1112–1122, 2013.

Charalambos Themistocleous, Marie Eckerström, and Dimitrios Kokkinakis. Voice quality and speech fluency distinguish individuals with Mild Cognitive Impairment from Healthy Controls. *Plos one*, 15(7):e0236009, 2020.

Calvin Thomas, Vlado Keselj, Nick Cercone, Kenneth Rockwood, and Elissa Asp. Automatic detection and rating of Dementia of Alzheimer type through lexical analysis of spontaneous speech. In *IEEE International Conference Mechatronics and Automation, 2005*, volume 3, pages 1569–1574. IEEE, 2005.

László Tóth, Ildikó Hoffmann, Gábor Gosztolya, Veronika Vincze, Gréta Szatlóczki, Zoltán Bánréti, Magdolna Pákáski, and János Kálmán. A speech recognition-based solution for the automatic detection of mild cognitive impairment from spontaneous speech. *Current Alzheimer Research*, 15(2):130–138, 2018.

Quan Hung Tran, Ingrid Zukerman, and Gholamreza Haffari. A hierarchical neural model for learning sequences of dialogue acts. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 428–437, 2017.

Kelvin K F Tsoi, Lingling Zhang, Nicholas B Chan, Felix C H Chan, Hoyee W Hirai, and Helen M L Meng. Social Media as a Tool to Look for People with Dementia Who Become Lost : Factors That Matter. *Proceedings of the 51st Hawaii International Conference on System Sciences*, 9:3355–3364, 2018.

Ana Varela Suárez. The question-answer adjacency pair in Dementia discourse. *International Journal of Applied Linguistics*, 28(1):86–101, 2018.

Chunrong Wang. A Relevance-theoretic approach to Turn Silence. In *4th International Conference on Contemporary Education, Social Sciences and Humanities (ICCESSH 2019)*. Atlantis Press, 2019.

Tifani Warnita, Nakamasa Inoue, and Koichi Shinoda. Detecting Alzheimer's disease using gated convolutional neural network from audio data. *arXiv preprint arXiv:1803.11344*, 2018.

Nick Webb, Mark Hepple, and Yorick Wilks. Dialogue act classification based on intra-utterance features. In *Proceedings of the AAAI Workshop on Spoken Language Understanding*, volume 4, page 5. Citeseer, 2005.

Jochen Weiner, Miguel Angrick, Srinivasan Umesh, and Tanja Schultz. Investigating the effect of audio duration on Dementia detection using acoustic features. In *INTERSPEECH*, pages 2324–2328, 2018.

James R Williamson, Elizabeth Godoy, Miriam Cha, Adrianne Schwarzentruber, Pooya Khorrami, Youngjune Gwon, Hsiang-Tsung Kung, Charlie Dagli, and Thomas F Quatieri. Detecting depression using vocal, facial and semantic communication cues. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, pages 11–18, 2016.

Raphael Wittenberg, Bo Hu, Luis Barraza-Araiza, and Amritpal Rehill. Projections of older people with Dementia and costs of Dementia care in the United Kingdom, 2019–2040. *London: London School of Economics*, 2019.

World Health Organization. First WHO ministerial conference on global action against Dementia: meeting report, who Headquarters, Geneva, Switzerland, 16-17 march 2015. *First WHO Ministerial Conference*, 2015.

Jessica A Young, Christopher Lind, and Willem van Steenbrugge. A conversation analytic study of patterns of overlapping talk in conversations between individuals with dementia and their frequent communication partners. *International journal of language & communication disorders*, 51(6):745–756, 2016.

Ye Yuan, Guangxu Xun, Qiuling Suo, Kebin Jia, and Aidong Zhang. Wave2vec: Learning deep representations for biosignals. In *2017 IEEE International Conference on Data Mining (ICDM)*, pages 1159–1164. IEEE, 2017.

Zining Zhu, Jekaterina Novikova, and Frank Rudzicz. Detecting cognitive impairments by agreeing on interpretations of linguistic features. *arXiv preprint arXiv:1808.06570*, 2018.

Matthias Zimmermann. Joint segmentation and classification of dialog acts using Conditional Random Fields. In *Tenth Annual Conference of the International Speech Communication Association*, 2009.

Kelly H Zou, A James O'Malley, and Laura Mauri. Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. *Circulation*, 115(5): 654–657, 2007.

APPENDIX A

MANUAL OF ANNOTATION FOR CCC CORPUS

A.1 Complete list of Dialogue acts

A complete list of DA tags that are used in annotation of Carolina's Corpora are listed in table A.1. Type represent category of tags , tags are symbols used to represent these dialogue acts and example of each dialogue act is given in last column. Section A.2 gives detailed description with example and context provided for all tags used to represent question's categories. Section A.3 covered all dialogue acts used for possible answer tags for all question types. Section A.4 gives description with examples of all other tags that are used to label the utterances in this corpus.

A.2 Question Types

Yes-No Questions(qy): Used with Yes-No questions only if it have pragmatic force as well as have syntactic and prosodic marking of a yes-no question like:

- Subject-inversion
- Question intonation
- Do-support

Few example from Carolina Conversation collection are given below in table A.2:

Wh-Questions and Declarative Wh-Questions wh interrogative questions with subject-auxiliary inversion are tagged with 'qw'. Some wh questions are in place wh question are tagged with 'qw^d'(see Examples in table A.3).

	Type	**Tag**	**Example**
1	Yes-No Question	*qy*	- did you go anywhere today?
2	Wh Questions	*qw*	When do you have any time to do your homework?
3	Declarative Yes-No Questions	*qy ^d*	You have two kids?
4	Declarative Wh Questions	*qw ^d*	Well what are your hours?
5	Or Question	*qr*	— did he um, keep him or did he throw him back?
6	Tag questions	*^g*	But they're pretty aren't they?
7	Open ended questions	*qo*	And uh -how do you think -that work helps you?
8	Clarification Question	*qc*	In Charlotte?
9	Non-understanding signal	*br*	-pardon?
10	Backchannel in question form	*bh*	Really?
11	Yes answer	*ny*	Yeah.
12	Yes- plus expansion	*ny^e*	yeah, but they're ~
13	Non-yes answer (affirmatve non yes answer)	*na*	- oh I think so. [laughs]
14	No answers	*nn*	No.
15	Negative non-no answers	*nn^e*	- no, I belonged to the Methodist church.
16	Other answer	*no*	- I, I don't know.
17	Statement answer	*sa*	– hell hot and heaven beautiful.
18	Backchannel/Acknowledge	*b*	um hmm.
19	Repeat Phrase	*b^m*	a grouper.
20	other	other	(everything else)
21	pause	*p*	pause , - (three seconds)
22	Non-verbal	*x*	[laughter],[cough] etc.

Table A.1: Complete List of proposed tagset.

Or Questions Or questions(*qr*) also called choice questions consists of multiple options in the question form which facilitate the respondent to choose any one of them to answer and lesson the burden of planning to response (see examples below in table A.4).

Tag Questions A confirmation question after a simple statement that often have auxiliary inversion at the end(don't you?) is represented by '^g' tag. Response to a tag question will be similar to yes-no questions (See table A.5).

Open ended questions open ended questions(*qo*) is a variant of wh-questions which put syntactic contraints on the answer and involves the thinking process like ' what do you think about it?'. (See example in A.6).

Conversation #	DA Type	Speaker:Utterance#:sub-utterance#	Example
Wakefield_brock_001_01	*qy*	*I*:3:2	- did you go anywhere today?
	qy	*I*:57:2	are you going to the Biltmore house?
Mason_Davis_001_01	*qy*	*I*:50:1	wh~, did you live on a farm?

Table A.2: Examples of *qy* questions.

Conversation #	DA Type	Speaker:Utterance#:sub-utterance#	Example
x	*qw*	*I*:35:1	- did you go anywhere today?
	qw	*I*:26:1	what about children?
	qw^d	*I*:86:3	So you were saying eh– that you like to do what?

Table A.3: Examples of *qw and qw^d* questions.

Conversation #	DA Type	Speaker:Utterance#:sub-utterance#	Example
Wakefield_brock_001_01	*qr*	*I*:63:3	when, was it like recent or
	+	*I*:63:4	- was it when they were little?
	qr	*I*:125:1	— did he um, keep him or did he throw him back?

Table A.4: Examples of Or (*qr*) question

Clarification Question Clarification question are more specific and are generated in response to a questions to clarify what was asked specifically(in utterance 86:1) or to the clarify the answer/statements in response to question as given in utterance:66:1 in table A.7.

Conversation #	DA Type	Speaker:Utterance#:sub-utterance#	Example
Tabor_Aver_001	^*g*	*I*:10:9	But they're pretty aren't they?
	^*g*	*I*: 36:1	Oh, li yeah, you think would like that, yeah?

Table A.5: Examples of tag (^*g*) questions.

Conversation #	DA Type	Speaker:Utterance#:sub-utterance#	Example
Zachary_Baile_001	*qo*	*I*:90:1	what's your greatest memory of your grandkids?
	sa	*P*: 91:1	oh, let me see.
	p	*P*: 91:2	(three seconds)
	sa	*P*: 91:3	most of them like to stay with us.

Table A.6: Examples of open-ended (*qo*) questions.

Conversation #	DA Type	Speaker:Utterance#:sub-utterance#	Example
Mason_Davis_001_01	*qw*	*I*:85:1	where is that church?
	qc	*P*:86:1	Fountain Hill?
	qw	*I*:64:1	what do you do?
	sa	*P*:65:1	- I'm a teacher.
	qc	*I*:66:1	Preacher?

Table A.7: Examples of clarification (*qc*) questions

Signal Non-Understanding Signal non-understanding is generated by a person in response to a question that they have not understood and are tagged with 'br'. They are used as marker of non-understanding, so a same question or statement is repeated in response to this type of signal/question. These are slightly different from *qc* as *br* are more generic in nature and usually represented with words like 'sorry?', 'Pardon?','Mam?', 'Sorry sir?' etc (Example given in table A.8).

Conversation #	DA Type	Speaker:Utterance#:sub-utterance#	Example
Wakefield_brock_001_01	*qy*	*I*83:2	— does Caroline go down to her?
	br	*P*:84:1	-pardon?
Mason_Davis_001_01	*qy*^d	*I*:22:3	so, he goes off preaching
	+	*I*:22:4	— and you stay here.?
	br	*P*:23:1	–m'am?

Table A.8: Examples of signal non-understanding(*br*) questions

Backchannel Question Backchannel is a continuer which takes the form of a question is represented by *bh*. Words like 'really? ', ' Right? ', ' Yeah? ', and ' Have you? ' are indicators of back-channel continuer and form very generic form of confirmation.

Repeat Questions Same tag is used when question is repeated but we have added an extra column to link to the utterance for which question is being repeated. Question can

Conversation #	DA Type	Speaker:Utterance#:sub-utterance#	Example
Wheadon_Lee_001	*other*	*P*:52:2	- huh, it used to be something special. it used to be my Mother's birthday.
	bh	*I*:53:1	Really?

Table A.9: Examples of back-channel (*bh*)

be repeated in following cases:

1. When the other participant gave signal of non understanding.

2. When there is no response or silence

Repetition will be considered for two cases:

1. When same question is repeated as in utterance 17:1.

2. When the speaker reformulate the question for other person to understand it better as in utterance 144:1.

For simple repeat we will just add utterance number of previous question but for reformulation we will add keyword 'reformulation' with the utterance number shown in table A.10.

Conversation #	DA Type	Speaker:Utt#: sub-utt#	Example	Repeat question
Wakefield_brock_001_01	*qw*	*I*:15:1	- where's she been?	
	br	*P*:16:1	-pardon?	
	qw	*I*:17:1	Where is she been?	**15**
Mason_Davis_001_01	*qy*	*I*:47:1	were you on a farm?	
	br	*P*:48:1	–m'am?	
	qy	*I*:49:1	wh , did you live on a farm?	**47-reformulation**
Wakeman_Rhyne_001_01	*qy*	*I*:142:3	well, are you, are you restricted from certain foods?	
	br	*P*:143:1	- what?	
	qy	*I*:144:1	like, do they, do they make you eat certain foods because your medication?	**47-reformulation**

Table A.10: Examples of repeat questions.

A.3 Answer Types

Yes Answer These are affirmative answers that are yes and its variant. Variant of a yes answer can be 'Yes', 'Yeah', 'uh-huh', 'yup', 'yep', 'oh yeah' etc. A yes answer is represented by *ny* example shown in table A.11. This could be generated in response to *qy*, *qy^d*, *^g*, *qc*, and*bh* question types.

Conversation #	**DA Type**	**Speaker:Utterance#:sub-utterance#**	**Example**
Mason_Davis_001_01	*qy*	*I*:50:1	wh , did you live on a farm?
	ny	*P*:51:1	Yeah.
Wakefield_brock_001_01	*qy*	*I*:7:1	did Tiffany do that?
	ny	*P*:8:1	Yes.

Table A.11: Examples of Yes Answers (*ny*)

Yes plus explanation 'ny^e' is used for yes answer that is answered with an explanation to simple yes-no questions, tag questions, clarification requests, declarative yes-no questions and backchannels.

Conversation #	**DA Type**	**Speaker:Utterance#:sub-utterance#**	**Example**
Mason_Davis_001_01	*qy*	*I*:Utt:28:1	– do you have children?
	ny^e	*P*:Utt:29:1	yeah, but they're .
	+	*P*:Utt:29:2	- big children now. grown.

Table A.12: Examples of Yes plus explanation Answers (*ny^e*)

Non-yes answer(affirmative non-yes answer) '*na*' is used for an affirmative yes answer that does not contain a 'Yes' and it's variant. But this is given as response to questions like yes-no, tag questions, clarification requests etc.

Negative answer and with Explanation '*nn*' is used for 'no' as negative answer and its variant like 'not', 'none' or 'no' while negative answers with an expansion are labelled with '*nn^e*'. These could be answers to simple yes-no questions, tag questions, clarification requests, declarative yes-no questions and backchannels.

Other Answer '*no*' is is used for other answers like 'I don't know, 'may be' in response to 'yes-no questions',tag questions etc.

Conversation #	DA Type	Speaker:Utterance#:sub-utterance#	Example
Taggart_Blac_001	*qy*	*I*:11:1	Did you mind picking cotton? unclear
	na	*P*:12:1	we, we just had to unclear do it.
Wakefield_brock_001_01	*qy*	*I*:71:2	do you think Carol's going to like your haircut?
	na	*P*:72:1	- oh I think so. [laughs]

Table A.13: Examples of Non-Yes answer (*na*).

Conversation #	DA Type	Speaker:Utterance#:sub-utterance#	Example
Tabor_Aver_001	*qy*	*I*:5:1	Did you go anywhere yummy?
	nn	*P*:6:1	No
Mason_Davis_001_01	*qy*	*I*:92:2	– were you Primitive Baptist?
	nn^e	*P*:93:1	- no, I belonged to the Methodist church.

Table A.14: Examples of negative answer and negative plus explanation (*nn and nn^e*).

Conversation #	DA Type	Speaker:Utterance#:sub-utterance#	Example
Wakefield_brock_001_01	*qy^d*	*I*:51:1	are you going to go with them
	+	*I*:51:2	– to see the Christmas lights?
	no	*P*:52:1	Oh
	+	*P*:52:2	I, I don't know

Table A.15: Examples of other answer (*no*).

Statement-Answer In response to a wh-question, declarative wh-question, and open ended question, '*sa*' tag will be used for the answers rather than simple declarative statement (*sd*).

A.4 Other Tags

Backchannel/Acknowledge 'b' is usually referred to as continuer. Most common form of continuer are: 'uh-huh', 'yeah', 'right', 'oh', 'yes', 'okay' , 'oh yeah', 'huh', 'sure',

Conversation #	DA Type	Speaker:Utterance#:sub-utterance#	Example
Mason_Davis_001_01	*qw*	*I*:31:3	- what does he preach about?
	sa	*P*:32:1	– hell hot and heaven beautiful.
Wakeman_Rhyne_001_01	*qw*	*I*:6:1	what types of food do you like the best?
	sa	*P*:7:1	– vegetables, meat.

Table A.16: Examples of declarative wh- answer (*sd-qw*)

'um', 'huh-uh', ' uh', 'mmm hmm' (See table A.17).

Conversation #	DA Type	Speaker:Utterance#:sub-utterance#	Example
Mason_Davis_001_01	*other*	*I*:37:1	it's beautiful out there.
	other	*P*:38:1	before I was married I joined there.
	b	*I*:39:1	yeah.

Table A.17: Examples of Continuer/acknowledge (*b*)

Repeat Phrase Repeat phrase is a combination of 'b' and '^m'. We code it as the Backwards function "Repeat-phrase" linking it with previous utterance. (See Example in Table A.18)

Conversation #	DA Type	Speaker:Utterance#:sub-utterance#	Example
Wakefield_brock_001_01	*sd*	*I*:119:1	it sort of looks like a blow fish a little bit. [laughs]
	+	*I*:119:2	it's really big
	b^m	*P*:120:1	it's big.

Table A.18: Examples of repeat phrase (*b^m*)

pause 'p' is used to represent pauses within the conversation. Transcripts are annotated with pauses duration like ' {silence } {three seconds}' or like '– hell hot and'. The former one is used as when the segment is only silence and we will use 'p' to tag those silence segments. The later one is silence within utterance that will be later used for intra-silences pauses in future.

Conversation #	DA Type	Speaker:Utterance#:sub-utterance#	Example
Mason_Davis_001_01	*other*	*P*:86:1	it's a nice church.
	b^m	*I*:87:1	a nice church.
	p	*I*:87:2	{silence} {six seconds}.

Table A.19: Examples of pauses (*p*).

Non-verbal expressions 'x' is used to represent non-verbal expressions within the conversations such as 'coughing', 'sneezing', 'laughing' etc.

Conversation #	DA Type	Speaker:Utterance#:sub-utterance#	Example
Tabor_Aver_001	*other*	*P*:136:6	Uh, I don't drink orange juice. [Chuckles]
	x		[Chuckles]

Table A.20: Examples of non-verbal expressions (*x*).

Declarative statement/ Other 'sd' is used for non-opinion statements possibly general statements. Later including declarative statements and all other possible statements are tagged as 'other'..

Conversation #	DA Type	Speaker:Utterance#:sub-utterance#	Example
Addison_McBain_04	*sd*	*I*:33:1	oh, my medicine is up here with my coffee time.
Wakefield_brock_001_01	*sd*	Ms. Brock:123:1	- that's a big fish!
Addison_McBain_04	*sd*	Ms. Addison:45:1	yeah, I take insulin. I take insulin and at night I take uh [telephone rings] NovoLogs with my meals.

Table A.21: Examples of declarative statement (*sd*).

A.5 Guidelines

Following guidelines are followed during the annotation process:

1. All utterances are arranged by turn and assigned unique number incrementally. That means that if at time t Interviewer says something, then the next utterance

at time t+1 is from the patient.You don't have to annotate empty turns like (*I*:) or (*P*:). Interviewer is denoted with *I*and *P* for Patient.

2. Within turn, there can be multiple sub-utterances spoken by same person e.g 10:3 mean turn 10 and sub utterance three.

3. Some utterances contain few words and are continuations of the previous utterance of the same speaker, or a preamble of the following utterance. Assign either same DA or '+' that you gave to that speaker's previous or subsequent utterance.

4. If more than one tag is applicable to an utterance, choose the tag corresponding to its main function keeping context of previous utterance. **only one tag will be assigned to each utterance**.

5. Assign 'Other' tag if an utterance does not fit within any of proposed tagset categories.

A.6 Ethical Considerations

A Research Ethics application was submitted and approved for the studies contained within this thesis. The letter confirming this is included here, together with the CITI course certificate required by MUSC for gaining access to CCC corpus.

For Office Use Only:
Rec Reference ……
Date received: ……

Queen Mary
University of London

Application form – Queen Mary Ethics of Research Committee

1 Name, department and email address of applicant
Matthew Purver Electronic Engineering and Computer Science m.purver@qmul.ac.uk
2 Title of study
Analysing Spoken Dialogue Structure with Alzheimer's Disease
3 Investigators
Matthew Purver (Reader), Julian Hough (Lecturer), Shamila Nasreen (PhD student), Morteza Rohanian (PhD student). Cognitive Science Research Group School of Electronic Engineering and Computer Science
4 Proposed timetable
24 months from receipt of data.
5 Other organisations involved
No other organisations involved in this analysis work. The data to be used has already been collected and transcribed for research purposes by Medical University of South Carolina (MUSC), USA.
6 Other REC approval
Our use of the data will be subject to MUSC's own REC approval and conditions, to be obtained after QMUL approval. See http://carolinaconversations.musc.edu/help/access/approval
7 Nature of project e.g. undergraduate, postgraduate
Postgraduate
8 Purpose of the research
To investigate analysis methods for interactional features in spontaneous speech of patients with Alzheimer's Disease (AD), and their potential for use in automatically recognising the presence and/or severity of AD. Specifically, the

CITI PROGRAM

Completion Date 10-Mar-2019
Expiration Date 09-Mar-2022
Record ID 30863000

This is to certify that:

Shamila Nasreen

Has completed the following CITI Program course:

GCP – Social and Behavioral Research Best Practices for Clinical Research (Curriculum Group)
GCP – Social and Behavioral Research Best Practices for Clinical Research (Course Learner Group)
1 - Basic Course (Stage)

Under requirements set by:

Medical University of South Carolina

CITI
Collaborative Institutional Training Initiative

Verify at www.citiprogram.org/verify/?wc16c64d8-87af-4158-af30-c4c92bcb79ac-30863000

APPENDIX B

ANNOTAIONS

B.1 Examples of Pauses Types

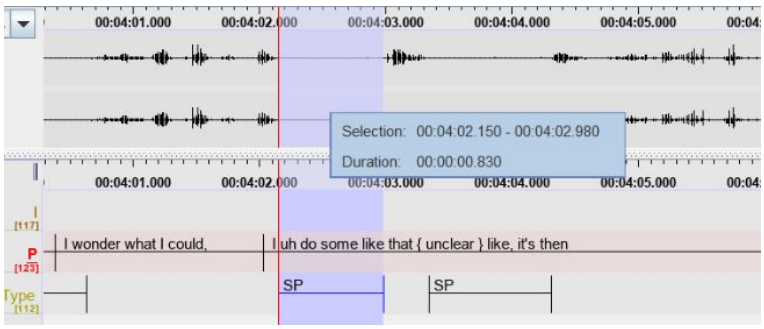


Figure B.1: Example of Short pause (SP)

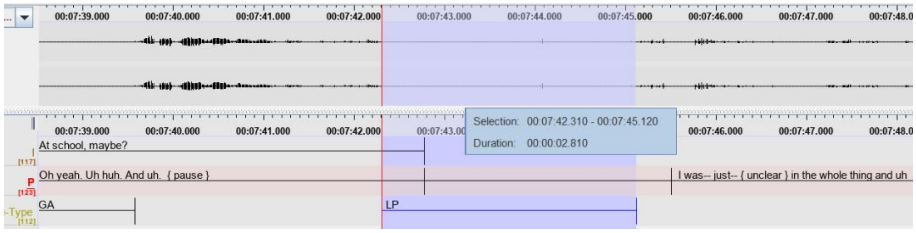


Figure B.2: Example of Long pause (LP)

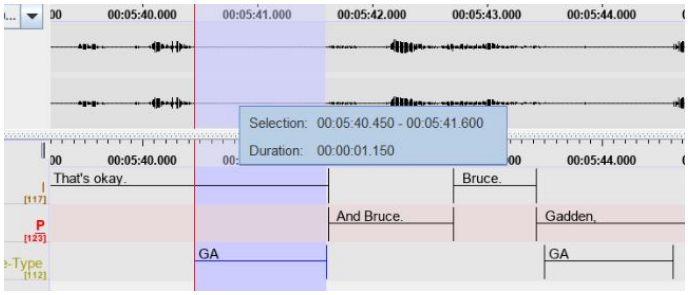


Figure B.3: Example of Gap (GA)

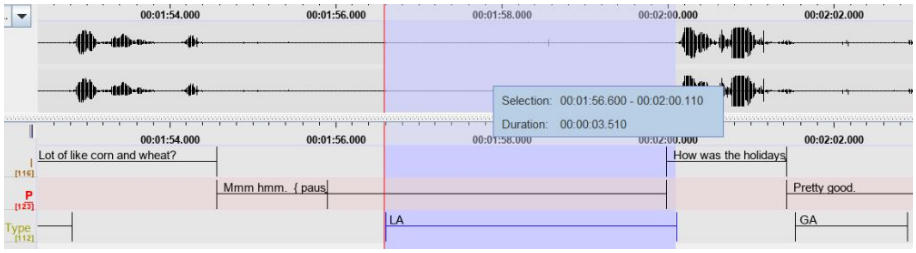


Figure B.4: Example of lapse(LA) with a topic shift by asking a question about holidays.

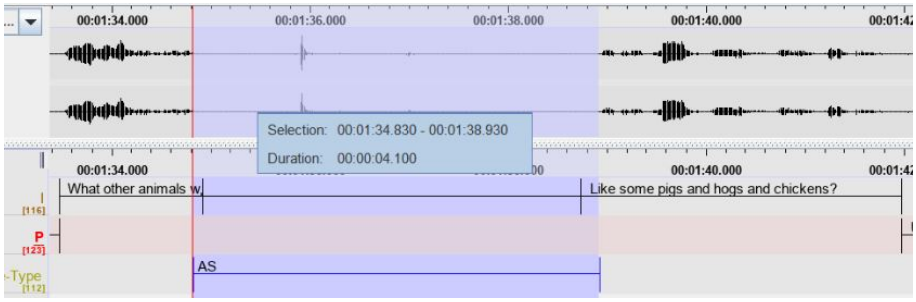


Figure B.5: Example of attributable silence (AS) of 4.1 seconds after a question from Interviewer(I) to patient (P)

.

B.2 Acoustic features annotation

No.	Feature	Annotation
1	F0final_sma	Fundamental frequency smoothed contour
2	pcm_RMSenergy_sma	Root mean square signal frame energy with smoothed contour
3	pcm_LoGenergy_sma	Logarithmioc power of energy

No.	Feature	Annotation
4	voicingFinalUnclipped_sma	voicing probability of the final fundamental frequency candidate. unclipped mean it was not set to zero when falls below the voicing hreshold.
5	F0_raw_sma	Raw fundamental frequency with smoothed contour
6	pcm_intensity_sma	
7	pcm_loudness_sma	The loudness as the normalised intensity raised to a power of 0.3
8	jitter_local	The local (frame-to-frame) Jitter (pitch period length deviations)
9	jitter_DDP_sma	The differential frame-to-frame Jitter (the 'Jitter of the Jitter')
10	shimmer_local	The local (frame-to-frame) Shimmer (amplitude deviations between pitch periods)
11	Linear_HNR	It quantifies the relative amount of additive noise in the voice signal.
12	logHNR_sma	Logarithmic power of Harmonic to noise ratio.
13	jitterLocal_sma_de	1st order delta coefficient of local (frame-to-frame) Jitter.
14	jitterDDP_sma_de	1st order delta coefficient of he differential frame-to-frame Jitter.
15	shimmerLocal_sma_de	1st order delta coefficient of the local (frame-to-frame) Shimmer.
16	Linear_HNR_de	1st order delta coefficient of the harmonic to noise ratio.
17	logHNR_sma_de	1st order delta coefficient of logarithmic Harmonic to noise ratio.
18	pcm_mfcc[1]	Mel-Frequency cepstral coefficients 1 (frame based).
19	pcm_mfcc[2]	Mel-Frequency cepstral coefficients 2
20	pcm_mfcc[3]	Mel-Frequency cepstral coefficients 3 .
21	pcm_mfcc[4]	Mel-Frequency cepstral coefficients 4
22	pcm_mfcc_de[1]	1st order delta coefficient of pcm_mfcc[1]
23	pcm_mfcc_de[2]	1st order delta coefficient of pcm_mfcc[2]
24	pcm_mfcc_de[3]	1st order delta coefficient of pcm_mfcc[3]
25	pcm_mfcc_de[4]	1st order delta coefficient of pcm_mfcc[4]
26	pcm_mfcc_de_[1]	2nd order delta coefficient of pcm_mfcc[1]

No.	Feature	Annotation
27	pcm_mfcc_de_[2]	2nd order delta coefficient of pcm_mfcc[2]
28	pcm_mfcc_de_[3]	2nd order delta coefficient of pcm_mfcc[3]
29	pcm_mfcc_de_[4]	2nd order delta coefficient of pcm_mfcc[3]

Association of acoustic features with annotations

APPENDIX C

RESPONSE DATA FOR DIFFERENT QUESTION TYPES

C.1 Responses

Table C.1 lists the responses against each question type separately for each response category

Question-type	Response-category	AD	Non-AD
qy	ny	0.25	0.04
	ny^e	0.18	0.18
	na	0.20	0.07
	no	0.04	0.18
	nn	0.06	0.14
	nn^e	0.17	0.21
	br	**0.11**	0.00
	qc	**0.03**	0.00
qy^d	ny	0.22	0.25
	ny^e	0.21	0.34
	na	0.25	0.25
	no	0.04	0.00
	nn	0.03	0.00
	nn^e	0.13	0.06
	br	**0.05**	0.00
	qc	**0.01**	0.00
^g	ny	0.33	0.20
	ny^e	0.19	0.20

Question-type	Response-category	AD	Non-AD
	na	0.29	0.40
	no	0.00	0.00
	nn	0.00	0.20
	nn^e	0.05	0.00
	br	**0.14**	0.00
	qc	0.00	0.00
qr	ny	0.09	0.00
	ny^e	0.35	0.00
	na	0.35	1.00
	no	0.04	0.00
	nn	0.00	0.00
	nn^e	0.13	0.00
	br	0.00	0.00
	qc	0.00	0.00
qc	ny	0.26	0.13
	ny^e	0.04	0.25
	na	0.52	0.25
	no	0.00	0.25
	nn	0.04	0.13
	nn^e	0.04	0.00
qw and qw^d	sd-qw	0.84	0.91
	br	**0.12**	**0.05**
	qc	**0.04**	**0.02**
	no	0.04	0.02
	ny	0.01	0.00

Table C.1: Frequency distribution of responses against each question type for AD group and Non-AD group

APPENDIX D

STATISTICAL ANALYSIS RESULT FOR DA UNIGRAM AND BIGRAM FEATURES

D.1 Unigram DA features statistical analysis results

Feature	**AD**	**Non-AD**	**Mann-Whitney U test**	
	Mean (SD)	Mean (SD)	***p***	**U**
Unigram features				
p_^g	0.60 (0.737)	0.53 (0.834)	0.653	122
p_qo	0 (0)	0.40 (1.121)	0.539	97.5
p_qr	0 (0)	0.20 (0.561)	0.150	97.5
p_qw	1.53 (1.767)	0.87 (2.066)	0.106	151.5
p_qw^d	0.47 (0.915)	0.07 (0.258)	0.345	136
p_qy	1.53 (1.457)	1 (1.690)	0.161	146.5
p_qy^d	1.27 (1.870)	1.33 (2.410)	0.655	122.5
p_b	11.53 (7.680)	5.40 (4.222)	0.080-	176
p_b^m	1.27 (1.335)	2.07 (2.251)	0.412	92
p_bh	0.60 (1.242)	.07 (0.258)	0.345	136
p_br	2.67 (4.203)	1.93 (4.284)	0.512	128.5
p_qc	2.60 (2.131)	1.13 (1.356)	**0.041***	161.5
p_other	29.73 (29.961)	31.33 (37.130)	0.967	113.5
p_na	9.47 (11.993)	9.07 (10.559)	0.902	116
p_ng	0.07 (0.258)	0.60 (1.242)	0.345	89
p_nn	1.73 (1.944)	0.67 (0.976)	0.160	152
p_no	2.33 (2.795)	0.67 (1.047)	**0.045***	161
p_ny	9.47 (8.476)	4.13 (4.612)	**0.016****	169.5
p_sa	20.87 (12.188)	40.67 (27.807)	**0.05***	65.5
I_^g	2.93 (3.127)	1.27 (1.486)	0.074-	156
I_qo	1 (1.069)	2.20 (1.781)	0.061-	67
I_qr	3.33 (2.895)	1.53 (2.295)	**0.041***	161.5
I_qw	7.40 (5.865)	5.80 (3.745)	0.713	121.5
I_qw^d	0.93 (0.961)	0.60 (1.121)	0.179	142.5
I_qy	6.67 (5.260)	3.47 (3.523)	**0.033***	163.5
I_qy^d	9.33 (8.608)	11.40 (8.253)	0.412	92.5
I_b	18.47 (12.889)	31.20 (22.951)	0.081-	70.5
I_b^m	2.40 (2.131)	2.33 (2.225)	0.881	116
I_bh	1.40 (1.765)	0.73 (1.580)	0.174	145.5
I_br	1.80 (2.808)	2 (4.276)	0.653	123.5
I_qc	3.07 (4.877)	2.40 (1.765)	0.305	87.5
I_other	50.07 (66.034)	30.07 (34.704)	0.624	125
I_na	3.33 (2.717)	8.67 (10.668)	0.217	82
I_ng	0.07 (0.258)	0.60 (1.242)	0.345	89
I_nn	1.60 (1.805)	1(1.309)	0.285	139
I_no	2.07 (2.865)	0.93 (1.223)	0.305	137.5
I_ny	5.07 (8.447)	2.87 (4.103)	0.106	151.5
I_sa	15.20 (16.967)	24.13 (23.600)	0.345	89.5

Table D.1: Descriptive statistics (mean, SD) and statistical significance of the DA feature set. ** denotes highly significant at $p < 0.01$; * denotes significance at $p < 0.05$; **-** shows a trend toward significance at $p < 0.1$