LING 575 - Winter 2017
Project 2 Report

Analysis of Language for Dementia Diagnosis with Focus on Alzheimer's Disease

1 Introduction

Dementia "is characterized by a decline in memory, language, problem-solving and other cognitive skills that affects a person's ability to perform everyday activities."[1] These signs of decline are reflected in both cognitive (difficulty with planning, organizing, coordination, and motor functions) and psychological (paranoia, agitations, hallucinations) symptoms. It is possible to experience a single aspect of dementia (e.g. memory loss) while other cognitive and psychological aspects remain intact. The variations in which aspects are most affected contribute to the categorization of dementia types. For example, in Alzheimer's disease (AD), the most common type of dementia, patients suffer from cognitive decline due to the destruction of neurons in critical parts of the brain involved in cognitive functions. Symptoms of AD range from difficulty remembering recent information to impaired communication.

AD is estimated to be the 6th leading cause of death in the US, while there is no known cure and limited treatment options so far [2]. It is estimated that every 66 seconds, someone in the the US develops the disease, more than 5 million Americans are living with AD, and by 2050 this number could rise to as high as 16 million. The costs of health care for AD are therefore also high: in 2016, the caregivers provided approximately 18.2 billion hours, valued at over $230 billion [1]. Clearly, assessment of dementia is an urgent, important, as well as interesting research problem. This is crucial especially in earlier stages, potentially helping to understand the main symptoms and to provide preventive measures for the affected population.

Within AD, some clinicians further differentiate between subtypes of AD based on stages of severity: early mild cognitive impairment (EMCI), late mild cognitive impairment (LMCI), and finally AD [2]. Other common forms of dementia include[1]: vascular dementia (impaired judgement or ability to make decisions), dementia with Lewy bodies (memory loss and thinking problems, but more likely than people with AD to have early symptoms such as hallucinations), Parkinson's disease (problems with movement), frontotemporal dementia (changes in personality and behavior, difficulty with language), and others. For this project, I chose to focus on AD because of the available data (see Section 3) and the intriguing

[1] http://www.alz.org/dementia/types-of-dementia.asp

question of language in AD. In particular, as mentioned above, it is not necessarily the case that all AD patients experience obvious language difficulty; I was hoping to explore subtle language cues extractable using NLP tools.

Evaluation of AD (as well as evaluation of language disorders) are still done manually for the most part [3]. The most widely used test for dementia diagnosis is the Mini-Mental State Examination (MMSE) [4]. The MMSE tests a wide range of cognitive functions: orientation, registration, short-term memory, attention, calculation, visuo-spatial skills, and praxis [5], and is administered by clinicians. However, Roark et al. [6] as well as Sitek et al. [7] have pointed out that the MMSE is a coarse measure, providing insufficient sensitivity to subtler manifestations of cognitive decline such as Young Onset Dementia (YOD) and MCI. In addition, though MMSE is widely used, there is little agreement regarding the boundary points for classification of dementia vs. healthy. A variety of cut-off points have been proposed, ranging from 21-27 on a scale of 30 maximum points [5, 8]. Yet even more problematic is the relatively large variance in the assessment among clinicians on the same subject; Molloy et al. [9] reported an intra-rater standard deviation of as high as 4.8 per subject. In other words, there is a lack of ground truth for AD categorization.

From the language assessment point of view, Roark et al. [10] specifically emphasized the use of spontaneous speech elicitation, whereas the MMSE consists of only a small portion that is language specific. Many other researchers, such as Sajjadi et al. [11], also agree that narrative or conversational speech is important in assessment of language deficit, especially since such language more closely reflects communication in everyday life. Consequentially, the most commonly used task is the Cookie Theft (Figure 1) picture description task: participants are shown the picture and are asked to describe everything they see in the image. The dataset (DementiaBank) I am using also has the majority of data available in this specific task, and therefore will be my focus corpus for this study.

2 Related Work

Speech and NLP have become more relatively mature research areas over the last few decades, yet application to the clinical domain *on a large scale* [6] is still quite sparse. Much like research in the biomedical domain, there is the challenge of data availability and accessibility, due to privacy as well as ethical reasons. Fortunately, there seems to have been a push in application of NLP to the clinical domain, as evidenced by the increase in participation of NLP workshops in clinical psychology.

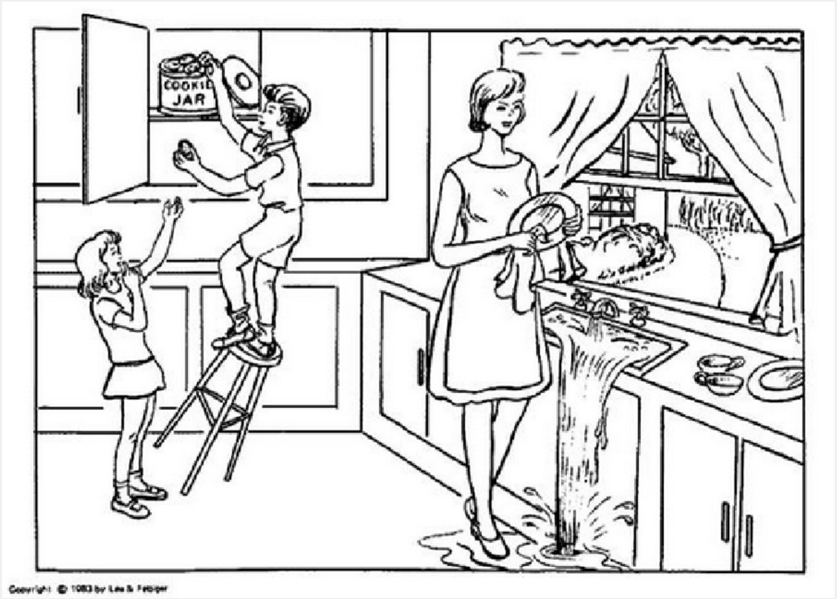

Figure 1: The "Cookie Theft" picture often used to assess language conditions in clinical settings

2.1 Diagnosis with Conversational Speech

Among the broader dementia space, Roark et al. investigated both linguistic [6] and acoustic [10] features of speech to detect MCI subjects from control. The authors recognized the utility of both text and speech features; in particular, noun and verb counts, syntactic complexity (as measured by Yngve score [12]), as well as pause durations and rates all proved to be useful. For Primary Progressive Aphasia (PPA) detection and subtype classification, Fraser et al. [13, 14] also found that syntactic complexity features were among the most useful. In addition, while acoustic features were not as useful in differentiating PPA from control, they were important in classification of PPA's subtypes.

Within the AD-specific domain, many researchers [15, 16] have studied a large feature set, both lexical and acoustic, to identify AD patients from healthy controls. Orimaye et al. [15] found most useful features to be number of predicates (raw counts and average), number of utterances, repetitions, and revisions; while Fraser et al. [16] found pronoun-noun ratio, NP → PRP production rule, adverb counts, verb counts, noun counts, and Honore's statistics to be among the most useful. Both studies used relatively standard machine learning methods: experimenting with logistic regression and SVM classifiers. Fraser et al. achieved a best-case accuracy of 82% on DementiaBank.

A slight deviation from AD detection includes Yancheva et al.'s work [17] , where the authors used a large feature set (477 features) to predict MMSE scores in DementiaBank. The authors obtained a mean absolute error of 3.83 in predicting MMSE, and found that focusing on longitudinal inter-subject conditions improved the error to 2.91.

Most recently, Fraser et al. [18] investigated the problem of differentiating late-life AD patients from depression patients, and vice versa. The researchers found that features used for AD detection did not result in false positives in depression patients who are otherwise healthy; however, detecting depression in AD was much more challenging. Similar to their previous works, Fraser et al. [18] also used a large feature set including both text and acoustics features, with the conclusion that, this case, acoustic features were highly important in distinguishing people with both AD and depression from people with only AD.

In an investigation to push towards a fully automated diagnosis pipeline, Zhou et al. [19] compared using hand-transcribed speech conversations vs. ASR outputs to detect AD in participants. Not surprisingly, they found that accuracy is higher using perfect transcripts, but also identified key features that have distinguishing power in both gold and ASR transcripts: word length and frequency, for example. In addition, the authors observed that accuracies can vary within a narrow band of WER; in other words, ASR transcripts with same low WER can contain drastically different information useful in detecting AD.

2.2 Diagnosis with Written Text

So far, the language aspects of cognitive decline have been mostly evaluated through spontaneous speech and their resulting transcripts. Written text also is a potentially useful source of language information to diagnose dementia, especially since text would be much easier to process, de-identify, and available in larger quantities than speech. Early works have eluded to written text's potential utility: Snowdon's famous Nun Study [20], started in 1986, recruited 678 Catholic sisters, each was asked to write a short biographical sketch of her life. Snowdon [21] later analyzed these texts, which were written in each sister's early life, and found strong, consistent correlations between idea density and severity of Alzheimer's disease pathology in the neocortex.

Several more recent works that analyzed texts for signs of dementia include Garrard et al.'s work [22], which compared three Iris Murdoch's novels, to find signs of cognitive decline in her last work, *Jackson's Dilemma.* Iris Murdoch was of particular interest because her AD diagnosis was confirmed post mortem, and *Jackson's Dilemma* was widely criticized by the literary community at the time for its inconsistency and relatively lower quality. Pakhomov et al. [3] also examined longitudinal changes in syntactic complexity of Iris Murdoch's writings and found clear patterns of decreasing grammatical complexity. Le et al. [23, 24] expanded on this line of work, including writings of Agatha Christie, who was suspected of experiencing AD, and P.D. James, a confirmed healthy control. The authors found correlations of decreased complexity with age of the authors such as a decline in vocabulary size and use of passives.

Most recently, Weissenbacher et al. [25] are creating a corpus of written narrative picture descriptions from participants with AD, MCI, and control. For their initial study, the authors performed classification using a variety of lexical, stylometric, semantic features, and subject meta features. This is among the very few (maybe even first) works to use modern word embeddings as one of the features. Weissenbacher et al. achieved an accuracy as high as 86% with carefully selected features, with the most important features being word2vec and ngrams found from their ablation studies.

3 Data

The data for my experiments and analyses are from the DementiaBank[2] corpus [26], and specifically the Cookie Theft description task subset. The amount of data available in this set is listed in Table 1. Participants are presented with the Cookie Theft picture (Figure 1), and are asked to describe what they see. This subset includes both manual transcripts of the clinical sessions, as well as audio recordings. For this project, however, I am only studying language features extracted from transcriptions.

Table 1: Statistics of the DementiaBank corpus. (*) There was actually **one** subject (subject 172) that had a label changed from "Control" (in their first visit) to "Dementia" (in their remaining 3 visits). The statistics here merged this subject to the "Dementia" group because of the majority of the labels (3 out of 4) for this subject were in "Dementia".

	Control	Dementia*
# subjects	98	194
# total visits (recordings)	241	307
those w/ 1 visit	25	117
those w/ 2 visits	28	53
those w/ 3 visits	28	12
those w/ 4 visits	9	9
those w/ 5 visits	8	3

The transcriptions were all manually annotated with the CHAT guideline [27], which are quite informative. Figures 2 and 3 show an examples of transcription excerpts in this data set. `*PAR*` denotes the participant's turn, and `*INV` denotes the interviewer's turn. The format `[: text]` denotes assimilation (as in `[: going to]`), `[x N]` stands for N repetitions, `[/]` or `[//]` denote interruption point in disfluencies - which I am using as a marker for

[2] https://talkbank.org/access/DementiaBank/English/Pitt.html

```
*INV: tell me all of the things you see going on .
*PAR: alright . [+ exc]
*PAR: a little girl is reaching for the <cookie> [/] cookie that the boy’s
reaching for the cookie to give to her while <the> [/] the &uh stool is
being tipped . [+ gram]
*PAR: the mother is drying the dish while the water is running out
of the spigot .
*PAR: &um let’s see . [+ exc]
*PAR: action [x 4] . [+ exc]
*PAR: no birds, geese . [+ gram]
*PAR: (..) it’s all I can see . [+ exc]
*INV: okay .
```

Figure 2: Excerpt transcription from a healthy control.

pauses, and [+ `gram`] denotes grammatical error. Details of other markers can also be found in the CHAT manual [27].

As evident from these excerpts, the amount of data is still relatively small in NLP standards: the average narration length of each participant in each visit is only 100 words. However, at a quick glance, one can see there are certain intuitive differences between a healthy control’s narration and a potential dementia patient’s, more disfluencies and interruption points stand out as one observation in this example.

4 Feature Extraction and Analysis

I extracted common bag-of-words (BOW) and syntax features, as well as other less common features suggested by Fraser et al. in [16]. In particular, my features are grouped in 3 sets, all computed only for the participant’s sides: BOW features, syntax features, and length features. The complete feature set description can be found in Table 2.

For language modeling, I used the SRILM toolkit [28]; POS tags and syntactic trees were parsed by the Berkeley parser [29]. I am also noting a few clarifications here:

- V denotes number of words types; N denotes number of raw tokens; clean tokens are raw tokens excluding disfluencies and pauses
- All features are computed per visit, i.e. per recording
- Rate features were computed by normalizing their counts by the number of raw tokens

Table 2: Full feature set explored and used in this study

Group	Feature	Description
BOW	num tokens raw (N)	number of raw tokens
	num tokens clean	number of clean tokens
	content rate	rate of content words
	disfluent rate	rate of disfluent words
	V1	number of words appearing only once
	honore	Honore's statistics, computed as logN/(1-V1/V)
	ttr	type:token ratio
	COOKIE	number of key words COOKIE
	COUNTER	number of key words COUNTER
	CURTAIN	number of key words CURTAIN
	SINK	number of key words SINK
	STOOL	number of key words STOOL
	WINDOW	number of key words WINDOW
	lm prob	language model log-probability
length	conv length	conversation length: total number of turn-utterances
	visit length	length of visit in ms
	mean length clean	number of clean tokens / conv length
	mean length raw	number of raw tokens / conv length
	verbal rate clean	number of clean tokens / visit length
	verbal rate raw	number of raw tokens / visit length
syntax	RB counts, rates	adverb counts and rates
	NN counts, rates	noun counts and rates
	PRP counts, rates	pronoun counts and rates
	VB* counts, rates	verb counts and rates
	NP counts, rates	NP constituent counts and rates
	VP counts, rates	VP constituent counts and rates
	constituent counts, rates	total constituent counts and rates
	ADVP → RB counts, rates	production rule ADVP → RB counts and rates
	NP → PRP counts, rates	production rule NP → PRP counts and rates
	NP → DT NN count, rates	production rule NP → DT NN counts and rates

```
*PAR: (..) well I see the mother .
*PAR: what's she doin(g) ? [+ exc]
*PAR: pourin(g) the [/] (.) the faucet down on the floor . [+ gram]
*PAR: <she got> [//] yeah she got that runnin(g) on the floor . [+ es]
*PAR: she's wipin(g) the dishes .
*PAR: and the little boy's gonna [: going to] break his neck (.)
tryin(g) to get into the cookie jar
*PAR: and the little girl's yellin(g) for more
*PAR: (..) they're headin(g) into a disaster &=laughs
*INV:  +< &=laughs okay .
*INV:  (...) anything else ?
*PAR: well the little boy's gonna [: going to] &br break his neck
there if he doesn't watch out .
```

Figure 3: Excerpt transcription from a dementia patient (diagnosed as ProbableAD).

- The language model used for computing lm-prob was trained on Switchboard data, since Switchboard consists of natural speech more closely matching DementiaBank's data
- The Berkeley parser was also retrained on Switchboard data before applying it to DementiaBank data

Tables 3 through 6 show the feature analysis in terms of their potential distinguishing power between the AD and healthy control groups. Among the length features, mean length, both raw and clean, as well as conversation length are not statistically significant ($p > 0.01$). This is not surprising, since we could relatively see the length similarity from the excerpt examples, where both AD and control subjects spent about the same amount of words to describe the picture.

Table 3: Feature comparison between subject groups: **Length features**

	Control		Dementia			
Feature	Mean	Std. dev	Mean	Std. dev	tval	pval
conv length	12.86	5.28	12.47	6.62	0.78	4.34E-01
visit length	57766	24348	68318	34836	-4.17	3.53E-05
verbal rate raw	2.18E-03	5.15E-04	1.84E-03	6.44E-04	6.90	1.46E-11
verbal rate clean	2.04E-03	4.96E-04	1.66E-03	6.11E-04	7.89	1.65E-14
mean length raw	9.82	2.72	9.77	2.76	0.22	8.27E-01
mean length clean	9.22	2.70	8.85	2.68	1.61	1.08E-01

Regarding BOW features, almost all are statistically significant ($p < 0.01$) except for lm-prob and ttr. What was a bit surprising was the fact that control subjects tend to have a slightly lower content word rate than dementia patients. However, studies such as [30] have also found that dementia patients (progressive non-fluent aphasics (PNFA), not AD though) exhibit a decrease in the use of function words, which might explain the higher content-word rates in this case. Other BOW differences between the groups are what one would expect: healthy controls have less pausing and fewer disfluencies, as well as more key words that are "obligatory" to be used in this particular picture description such as "cookie", "stool", and "window". The potential explanation for this key-word difference is that a healthy control is much more likely to use these specific words, while a subject with dementia might substitute "cake" for "cookie", and "chair" for stool, exhibiting signs of decline in vocabulary specificity.

Table 4: Feature comparison between subject groups: **BOW features**

	Control		Dementia			
Feature	Mean	Std. dev	Mean	Std. dev	tval	pval
content rate	0.68	0.14	0.73	0.19	-3.60	3.49E-04
pause rate	0.01	0.01	0.03	0.03	-8.42	3.96E-16
disfluent rate	0.18	0.10	0.24	0.15	-5.67	2.31E-08
honore	19.67	5.62	18.29	5.80	2.81	5.14E-03
lm-prob	-24.09	6.17	-22.76	6.33	-2.47	1.40E-02
ttr	0.60	0.07	0.60	0.09	0.23	8.21E-01
v1	51.33	16.43	47.00	17.06	3.01	2.73E-03
COOKIE	3.17	2.18	2.01	1.92	6.52	1.80E-10
CURTAIN	0.14	0.45	0.05	0.26	3.07	2.30E-03
COUNTER	0.37	0.78	0.10	0.34	5.09	6.22E-07
SINK	2.22	1.82	1.36	1.68	5.69	2.16E-08
STOOL	2.18	1.64	1.26	1.34	7.04	6.92E-12
WINDOW	1.47	1.66	0.58	1.15	7.10	5.42E-12

Among the syntax features, raw counts don't seem as useful as rates. Constituent counts in particular, raw or rate, are not significantly different between the two groups, which is consistent with the fact that conversation and utterance lengths are also similar among the two groups. Among those most potentially useful, in terms of statistical significance, are adverb rates, verb rates, noun phrase rates, and production rule NP $\rightarrow$ DT NN. This observation is also consistent with previous feature analysis [16] and the fact that primary progressive aphasics (another form of dementia) have verb-finding difficulties [31, 7].

Table 5: Feature comparison between subject groups: **Syntax features; raw counts**

	Control		Dementia			
Feature	Mean	Std. dev	Mean	Std. dev	tval	pval
RB counts	9.24	7.51	13.28	8.93	-5.76	1.40E-08
NN counts	32.21	15.81	33.65	16.72	-1.03	3.02E-01
PRP counts	9.16	6.49	10.17	7.58	-1.67	9.54E-02
VB* counts	21.45	10.53	17.38	11.36	4.34	1.67E-05
NP counts	46.30	24.41	48.62	25.45	-1.08	2.79E-01
VP counts	33.04	16.24	29.06	17.30	2.77	5.82E-03
constituent counts	249.41	125.30	240.17	127.94	0.85	3.96E-01
ADVP → RB	3.12	2.79	3.64	3.16	-2.04	4.15E-02
NP → PRP	7.40	5.36	8.71	6.73	-2.54	1.15E-02
NP → DT NN	12.29	6.21	9.20	5.55	6.06	2.70E-09

Table 6: Feature comparison between subject groups: **Syntax features; rates over tokens**

	Control		Dementia			
Feature	Mean	Std. dev	Mean	Std. dev	tval	pval
RB counts	0.07	0.05	0.12	0.07	-8.34	6.19E-16
NN counts	0.27	0.07	0.30	0.09	-4.76	2.47E-06
PRP counts	0.07	0.02	0.08	0.03	-4.64	4.34E-06
VB* counts	0.18	0.04	0.14	0.04	8.94	6.48E-18
NP counts	0.37	0.08	0.42	0.09	-6.22	9.71E-10
VP counts	0.27	0.05	0.24	0.05	6.02	3.20E-09
constituent counts	2.01	0.14	2.03	0.16	-2.01	4.44E-02
ADVP → RB	0.02	0.02	0.03	0.02	-4.09	4.94E-05
NP → PRP	0.06	0.02	0.07	0.03	-5.72	1.76E-08
NP → DT NN	0.10	0.03	0.08	0.03	7.22	1.79E-12

5 Classification Experiments and Results

Similar to previous works, I set up classification experiments to detect AD patients from healthy controls. Because of the small data set, all experiments are done with 10-fold cross validation, and I report mean F1 (dementia = class 1; control = class 0), as well as mean accuracy scores.

I experimented with 4 types of classifiers: logistic regression, SVM with linear kernel, SVM with RBF kernel, and decision trees. All models, except decision trees, were tuned with regularization parameters $C = \{0.01, 0.1, 1, 10, 10\}$ and the best model was selected by mean accuracy score. The models were trained using the `scikit-learn` package [32].

For each model, I experimented with using all features in a certain group (BOW, length, syntax), as well as choosing only the subset in each group shown to have statistically significant differences between two classes. These results are shown in Tables 7 through 9.

Table 7: Classification results (mean score across 10 folds) using **length** features; numbers in parentheses are standard deviation across folds

	All features		Significant subset	
Classifier	F1	Accuracy	F1	Accuracy
Logistic regression	0.66 (0.07)	0.61 (0.09)	0.72 (0.003)	0.56 (0.004)
SVM, linear kernel	0.68 (0.05)	0.60 (0.06)	0.66 (0.06)	0.58 (0.06)
SVM, RBF kernel	0.72 (0.01)	0.56 (0.01)	0.72 (0.01)	0.56 (0.02)
Decision tree	0.61 (0.06)	0.57 (0.06)	0.63 (0.05)	0.61 (0.04)

Table 8: Classification results (mean score across 10 folds) using **BOW** features; numbers in parentheses are standard deviation across folds

	All features		Significant subset	
Classifier	F1	Accuracy	F1	Accuracy
Logistic regression	0.71 (0.06)	0.68 (0.05)	0.73 (0.05)	0.68 (0.03)
SVM, linear kernel	0.73 (0.06)	0.67 (0.06)	0.73 (0.06)	0.68 (0.05)
SVM, RBF kernel	0.72 (0.02)	0.58 (0.04)	0.71 (0.05)	0.60 (0.06)
Decision tree	0.64 (0.06)	0.62 (0.07)	0.62 (0.08)	0.60 (0.08)

Finally, I also trained the 4 models on the combined feature set, with and without features not statistically significant. Results for this experiment are shown in Table 10. Overall, using all features extracted, even though some showed little statistical significance, still resulted

Table 9: Classification results (mean score across 10 folds) using **syntax** features; numbers in parentheses are standard deviation across folds

	All features		Significant subset	
Classifier	F1	Accuracy	F1	Accuracy
Logistic regression	0.76 (0.07)	0.73 (0.06)	0.76 (0.06)	0.74 (0.05)
SVM, linear kernel	0.77 (0.06)	0.74 (0.06)	0.73 (0.08)	0.70 (0.08)
SVM, RBF kernel	0.73 (0.01)	0.59 (0.02)	0.71 (0.01)	0.56 (0.01)
Decision tree	0.70 (0.07)	0.67 (0.06)	0.68 (0.08)	0.66 (0.07)

in the best F1 score and classification accuracy. This was achieved by the logistic regression model, with an F1 score of 0.79 and an accuracy of 0.76. This accuracy score is around 6 points below state of the art systems on this dataset [16], which achieved 0.82 accuracy using both lexical and acoustic features. This result suggests that speech features are important in complementing text-only features in this classification task.

Table 10: Classification results (mean score across 10 folds) using **all** features; numbers in parentheses are standard deviation across folds

	All features		Significant subset	
Classifier	F1	Accuracy	F1	Accuracy
Logistic regression	**0.79** (0.04)	**0.76** (0.04)	0.76 (0.07)	0.73 (0.07)
SVM, linear kernel	0.78 (0.05)	0.74 (0.07)	0.77 (0.05)	0.74 (0.07)
SVM, RBF kernel	0.72 (0.004)	0.56 (0.004)	0.72 (0.004)	0.56 (0.004)
Decision tree	0.66 (0.09)	0.62 (0.08)	0.67 (0.08)	0.63 (0.06)

As error analysis, I applied the best learned model on the whole data set, and looked at cases of false positive and false negatives. There were 49 instances of false positives (healthy classified as AD), and 68 instances of false negatives (AD classified as healthy). Personally, the false negatives were more interesting to me, so I also looked at the ground truth diagnosis of such instances. As noted in Table 1, there was one subject whose diagnosis changed from control to MCI, which is the mild form of early dementia. Therefore, I suspected that most false negative samples would likely be MCI diagnosis. However, among the 68 false negatives, only 14 were diagnosed with MCI, while 26 were diagnosed with Probable AD, and the rest were a mix of memory decline and vascular dementia. This was quite surprising to me, since Probable AD is a diagnosis relatively severe into dementia. Since my text features could not detect these patients from healthy controls, here is where speech features might have been significantly useful.

Finally, I did a brief analysis focusing on subjects that have longitudinal data available, i.e.

those who participated in at least 2 sessions. I computed the variance in all the features, and ranked them by mean of the variance across sessions. Comparing these averages between the control and AD groups, there was little to no difference regarding which features varied more for one group vs. the other. Specifically, both groups showed large average variance in visit length and conversation length, while both also showed small average variance in verbal rates (both raw and clean).

6 Conclusion

In this project, I analyzed a variety of text features to detect AD patients from healthy controls, guided by previous literature on this task using the DementiaBank dataset. In attempting to reproduce some aspects of previous works, my findings are consistent with other researchers in terms of feature utility and the optimistic potential of using NLP for dementia diagnosis. Several challenges also became clearer to me, specifically the existing data sparsity barrier and the much larger feature search space compared to the amount of available data. Nevertheless, there does seem to be an increasing interest and appreciation in the NLP community for clinical applications, and more share-able corpora are being developed. This growing community will hopefully help NLP for clinical applications soon catch up with more mature NLP domains.

References

[1] Alzheimer's Association. 2016 Alzheimer's Disease Facts and Figures. `http://www.alz.org/facts/`. Accessed March 13, 2017.

[2] Joseph Bullard, Cecilia Ovesdotter Alm, Xumin Liu, Qi Yu, and Rubén Proaño. Towards early dementia detection: Fusing linguistic and non-linguistic clinical data. In *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*, pages 12–22, San Diego, CA, USA, June 2016. Association for Computational Linguistics.

[3] Serguei V.S. Pakhomov, Dustin Chacon, Mark Wicklund, and Jeanette Gundel. Computerized assessment of syntactic complexity in alzheimer's disease: a case study of iris murdoch's writing. *Behavior Research Methods*, 43(1):136–144, 2011.

[4] Marshal F. Folstein, Susan E. Folstein, and Paul R. McHugh. "mini-mental state": A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, 12(3):189 – 198, 1975.

[5] C. Thomas, V. Keselj, N. Cercone, K. Rockwood, and E. Asp. Automatic detection and rating of dementia of alzheimer type through lexical analysis of spontaneous speech. In

Mechatronics and Automation, 2005 IEEE International Conference, volume 3, pages 1569–1574 Vol. 3, 2005.

[6] B. Roark, M. Mitchell, J. Hosom, and K. Hollingshead. Syntactic Complexity Measures for Detecting Mild Cognitive Impairment. In *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, BioNLP '07, pages 1–8, Stroudsburg, PA, USA, 2007. Association for Computational Linguistics.

[7] Emilia J. Sitek, Anna Barczak, and Michał Harciarek. Neuropsychological assessment and differential diagnosis in young-onset dementias. *Psychiatric Clinics of North America*, 38(2):265 – 279, 2015. Young-Onset Dementias.

[8] G. Tosto, M. Gasparini, AM. Brickman, F. Letteri, R. Renie, P. Piscopo, G. Talarico, M. Canevelli, A. Confaloni, and Bruno G. Neuropsychological predictors of rapidly progressive Alzheimer's disease. *Acta Neurol Scand*, 36(6), 2015.

[9] D William Molloy, Timothy IM Standish, and David L Lewis. Screening for mild cognitive impairment: Comparing the smmse and the abcs. *The Canadian Journal of Psychiatry*, 50(1):52–58, 2005. PMID: 15754666.

[10] B. Roark, M. Mitchell, J. Hosom, K. Hollingshead, and J. Kaye. Spoken language derived measures for detecting mild cognitive impairment. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19(7):2081–2090, Sept 2011.

[11] Seyed Ahmad Sajjadi, Karalyn Patterson, Michal Tomek, and Peter J Nestor. Abnormalities of connected speech in the non-semantic variants of primary progressive aphasia. *Aphasiology*, 26(10):1219–1237, 2012.

[12] Victor H. Yngve. A model and an hypothesis for language structure. *Proceedings of the American Philosophical Society*, 104:444–466, 1960.

[13] Kathleen C. Fraser, Frank Rudzicz, and Elizabeth Rochon. Using text and acoustic features to diagnose progressive aphasia and its subtypes. In Frédéric Bimbot, Christophe Cerisara, Cécile Fougeron, Guillaume Gravier, Lori Lamel, François Pellegrino, and Pascal Perrier, editors, *INTERSPEECH*, pages 2177–2181. ISCA, 2013.

[14] 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 subtypes from narrative speech transcripts. *Cortex*, 55:43 – 60, 2014. Language, Computers and Cognitive Neuroscience.

[15] Sylvester Olubolu Orimaye, Jojo Sze-Meng Wong, and Karen Jennifer Golden. Learning predictive linguistic features for alzheimer?s disease and related dementias using verbal utterances. In *Proceedings of the Workshop on Computational Linguistics and*

Clinical Psychology: From Linguistic Signal to Clinical Reality, pages 78–87, Baltimore, Maryland, USA, June 2014. Association for Computational Linguistics.

[16] 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, October 2015.

[17] Maria Yancheva, Kathleen Fraser, and Frank Rudzicz. Using linguistic features longitudinally to predict clinical scores for Alzheimer's disease and related dementias. In *Proceedings of SLPAT 2015: 6th Workshop on Speech and Language Processing for Assistive Technologies*, pages 134–139, Dresden, Germany, September 2015. Association for Computational Linguistics.

[18] 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, San Diego, CA, USA, June 2016. Association for Computational Linguistics.

[19] Luke Zhou, Kathleen C. Fraser, and Frank Rudzicz. Speech recognition in Alzheimer's disease and in its assessment. *Interspeech 2016*, pages 1948–1952, 2016.

[20] D.A. Snowdon. Aging and Alzheimer's Disease: Lessons From the Nun Study. *The Gerontologist*, 37:150–156, 1997.

[21] D.A. Snowdon, L.H. Greiner, and W.R. Markesbery. Linguistic Ability in Early Life and the Neuropathology of Alzheimer's Disease and Cerebrovascular Disease: Findings from the Nun Study. *Annals of the New York Academy of Sciences*, 903:34–38, 2000.

[22] Peter Garrard, Lisa M. Maloney, John R. Hodges, and Karalyn Patterson. The effects of very early alzheimer's disease on the characteristics of writing by a renowned author. *Brain*, 128(2):250–260, 2005.

[23] Xuan Le. Longitudinal Detection of Dementia Through Lexical and Syntactic Changes in Writing. Master's thesis, University of Toronto, Computer Science, January 2010.

[24] Xuan Le, Ian Lancashire, Graeme Hirst, and Regina Jokel. Longitudinal detection of dementia through lexical and syntactic changes in writing: a case study of three british novelists. *LLC*, 26(4):435–461, 2011.

[25] Davy Weissenbacher, Travis A. Johnson, Laura Wojtulewicz, Amylou Dueck, Dona Locke, Richard Caselli, and Graciela Gonzalez. Automatic prediction of linguistic decline in writings of subjects with degenerative dementia. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1198–1207, San Diego, California, June 2016. Association for Computational Linguistics.

[26] James T. Becker, Francois Boiler, Oscar L. Lopez, Judith Saxton, and Karen L. McGonigle. The natural history of alzheimer's disease: Description of study cohort and accuracy of diagnosis. *Archives of Neurology*, 51(6):585–594, 1994.

[27] Brian MacWhinney. *The CHILDES Project: Tools for Analyzing Talk*. Lawrence Erlbaum Associates, 3rd edition, 2000.

[28] Andreas Stolcke. SRILM – an extensible language modeling toolkit. In *Proceedings of ICSLP*, volume 2, pages 901–904, Denver, USA, 2002.

[29] Slav Petrov and Dan Klein. Improved inference for unlexicalized parsing. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics, Proceedings of the Main Conference*, pages 404–411, 2007.

[30] M. Grossman and S. Ash. Primary Progressive Aphasia: A Review. *Neurocase*, 10(1):3–18, 2004.

[31] K.E. Forbes, A. Venneri, and M.F. Shanks. Distinct patterns of spontaneous speech deterioration: an early predictor of Alzheimer's disease. *Brain Cogn.*, 48(2-3):356–361, 2002.

[32] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.