INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Tackling the ADReSS challenge: a multimodal approach to the automated recognition of Alzheimer's dementia

Matej Martinc, Senja Pollak

Jozef Stefan Institute, Ljubljana, Slovenia
matej.martinc@ijs.si, senja.pollak@ijs.si

Abstract

The paper describes a multimodal approach to the automated recognition of Alzheimer's dementia in order to solve the ADReSS (Alzheimer's Dementia Recognition through Spontaneous Speech) challenge at INTERSPEECH 2020. The proposed method exploits available audio and textual data from the benchmark speech dataset to address challenge's two subtasks, a classification task that deals with classifying speech as dementia or healthy control speech and the regression task of determining the mini-mental state examination scores (MMSE) for each speech segment. Our approach is based on evaluating the predictive power of different types of features and on an exhaustive grid search across several feature combinations and different classification algorithms. Results suggest that even though TF-IDF based textual features generally lead to better classification and regression results, specific types of audio and readability features can boost the overall performance of the classification and regression models.

Index Terms: Cognitive Decline Detection, Computational Linguistics, Natural Language Processing, Speech Processing

1. Introduction

Alzheimer's Disease (AD) is the most common underlying cause of dementia, a neurodegenerative disease that leads to behavior and personality changes, such as decline in cognitive abilities and memory loss. AD is age-related and due to recent population trends suggesting large increases in elderly population [1], development of efficient methods for AD early detection and management has become of utmost importance.

The ADReSS (Alzheimer's Dementia Recognition through Spontaneous Speech) challenge [2] at INTERSPEECH 2020 [3] deals with automatic detection of AD from audio recordings and corresponding transcripts of subjects participating in a picture description task. The challenge defines two subtasks: Subtask 1 is a binary classification, i.e., to determine whether a patient has dementia or not, and SubTask 2 aims to determine the mini-mental state examination scores (MMSE) for each patient, i.e., a regression task.

The related work on AD classification reports accuracies of up to around 80% when best features are selected from a large set of linguistic and audio features [4, 5], or just linguistic features [6]. The accuracy in most cases decreases to below 70% in studies that consider only audio features [7], an exception being a study by Haider et al. [8], where the best accuracy of 78.7% is reported when an active data representation (ADR) feature extraction method is employed. When it comes to the regression task of determining the MMSE, we are aware of just one study that tackled it, reporting a mean absolute error (MAE) of 3.83 [5].

Due to findings from the related work and a relatively small size of the training set (108 training examples), our approach to both tasks was based on an extensive grid search over all possible feature combinations for each of the several pre-chosen classifiers and regressors[1]. These feature sets include several audio features (e.g., MFCC, ADR...) and a diversity of text features, covering different aspects of text transcripts (e.g., semantic features such as unigrams, syntactic features based on universal dependencies, which are in recent natural language processing research replacing the traditional part-of-speech tags and language dependant parsers, and statistical features indicating the readability of the text). The main contributions of this paper are as follows:

- Systematic evaluation of 16 distinct feature sets engineered from the audio signals and text transcripts and an insight into how they can be combined in the most efficient way.
- Deployment of novel universal dependency based features, and additional readability features for automated AD detection (i.e. ARI [9], GFI [10] and SMOG [11]).
- Development of a number of dementia AD classification and regression models with good performance and an available code for all experiments.

2. Methodology

Our core methodology consists of three parts, feature engineering (Section 2.1), choosing the learning algorithms (Section 2.2) and selection of the best feature combinations (Section 2.3).

2.1. Feature engineering

Features employed in the conducted experiments can be roughly divided into four distinct types, **audio features**, **TF-IDF features**, **readability features** and **embeddings**.

2.1.1. Audio features

All audio features were generated from the normalised audio-chunks, i.e., the .wav files extracted from the audio recordings of the AD and non-AD patient's speech after applying voice activity detection [2]. The following feature sets were constructed:

- **Mean MFCC**: means of first 13 mel-frequency cepstral coefficient features averaged across all audio recordings of each patient's speech. Window width of 25 ms and a stride of 10 ms were used in the extraction.
- **ADR**: an active data representation cluster based method for feature extraction [8] employed on Geneva minimalistic acoustic parameter set (**eGeMAPS**) and **MFCC**

[1]Code for the experiments is available under the MIT license at https://github.com/matejMartinc/ADReSSchallenge.

http://dx.doi.org/10.21437/Interspeech.2020-2202

features. Note that in our implementation, the self-organising maps (SOM) [12] clustering was replaced by a more widely used k-means clustering, with k=30.

- **Average duration** of audio recordings of each patient.

In addition, we also tested predictive power of mean root-mean-square, zero-crossing rate, spectral bandwidth, rolloff and centroid of audio samples, and the ADR feature extraction method on the emobase, ComParE 2013 and Multi-Resolution Cochleagram (MRCG) feature sets, as in [8], but did not use them in further experiments due to bad performance.[2]

2.1.2. TF-IDF features

TF-IDF features, which have been used in previous AD detection studies [13], were generated from the transcriptions of audio recordings[3] by generation of word and character n-gram tokens and employing bag-of-words vectorization and term frequency - inverse document frequency (TF-IDF) weighting on the derived tokens. The following tokens were used in vectorization and TF-IDF weighting:

- **Unigram** tokens, i.e., single words
- **Bigram** tokens, i.e., sequences of two adjacent words
- **Char4gram** tokens, i.e., sequences of four adjacent characters
- **Suffix** tokens, i.e., word suffixes of length 3
- **POS tag** bigrams, i.e., sequences of two adjacent part-of-speech tags
- **Grammmatical dependency (GRA)** features modelling grammatical relations between words in the input text, generated by the organizers of the challenge [2].
- **Universal dependency (UD)** features, i.e., a sequential representations of grammatical relations generated using the Stanford universal dependency parser [14]. For each word in the text, a tuple containing the type of grammatical relation (e.g., a determiner, nominal subject...) and the distance between the word at hand and its related word is generated. Unigrams, bigrams and trigrams of these tuples are used in our experiments.

2.1.3. Embeddings

Since related work reports promising results when word embeddings are used for AD detection [6, 15], we test several doc2vec embedding representations [16], namely *doc2vec text* representations generated from transcript texts, *doc2vec POS tags* representations generated from transcript POS tag sequences, *doc2vec GRA* representations generated from GRA features and *doc2vec UD* representations generated from UD feature sequences. We only use **doc2vec UD** features in further experiments, others were discarded due to bad performance.

2.1.4. Readability features

We experiment with several readability features. The hypothesis is that readability measures capture the complexity of language, which can be related to AD (AD patients display a decrease in the syntactic complexity of language [17] and have trouble in understanding the meaning of more complex words [18]):

- **Gunning fog index (GFI) [10]** was designed to estimate the years of formal education a person needs to understand the text on the first reading. It is calculated as $\text{GFI} = 0.4(\frac{\text{totalWords}}{\text{totalSentences}} + 100\frac{\text{longWords}}{\text{totalSentences}})$, where longWords are words longer than 7 characters.
- **Automated readability index [9] (ARI)** was also designed to return values corresponding to the years of education required to understand the text and is calculated as $\text{ARI} = 4.71(\frac{\text{totalCharacters}}{\text{totalWords}}) + 0.5(\frac{\text{totalWords}}{\text{totalSentences}}) - 21.43$
- **The SMOG grade (Simple Measure of Gobbledygook) [11]** is a readability formula mostly used for checking health messages and is calculated as $\text{SMOG} = 1.0430\sqrt{\text{num3Syllables}\frac{30}{\text{totalSentences}}}3.1291$, where the num3Syllables is the number of words with three or more syllables.
- **Number of unique words (NUW)**, normalized with the number of all words in the transcript.

Besides the readability features above, we also experimented with Flesch reading ease [19], Flesch-Kincaid grade level [19] and Dale-Chall [20] readability formulas, which were not used in further experiments due to bad performance.

2.2. Learning algorithms

Classification experiments were conducted by using four distinct classification algorithms from the Scikit library [21], namely Xgboost [22] (with 50 gradient boosted trees with max depth of 10), Random forest (with 50 trees of max depth of 5), SVM (with linear kernel and 2 box constraint configurations, 10 and 100) and Logistic regression (LogR) (with 2 distinct regularization configurations, 10 and 100). Regression experiments were conducted by using four distinct regression algorithms, namely Xgboost, SVM, Random forest and Linear regression (LinR). For Xgboost, SVM and Random forest same hyperparameters were used as for classification, while for LinR we used default parameters.

2.3. Exploration of feature space and model selection

Our approach is based on the early future-level fusion between different types of audio and textual features and relies on identification of feature combinations with the best synergy effect (see Figure 1). In order to do that, an extensive grid search across 65,535 combinations of 16 different feature sets (i.e., 4 audio, 7 TF-IDF, 1 embeddings and 4 readability feature sets) for each of the learning algorithms was conducted on the train set in a 10-fold cross-validation (CV) setting. For classification, accuracy is used for the performance evaluation, and for regression, root mean square error (RMSE) is used, same as for the official challenge evaluation [2].

The ADReSS challenge allows for submission of 5 distinct test set prediction tries. Therefore we identify 5 best performing classification models with non-identical predictions on the test set according to the grid search results. Their predictions on the test set are used for a majority vote ensemble, the output of which is used as one of the submissions. The other four submissions are test set predictions of the four best performing classification models. 5 submissions for regression are generated by first identifying 4 best performing regression models that do not

[2]The Logistic regression classifiers leveraging each of these feature sets did not outperform the majority baseline in the 10-fold cross-validation setting on the train set.

[3]Parts of the transcriptions that refer to the interviewer, and not the patient, were not used.

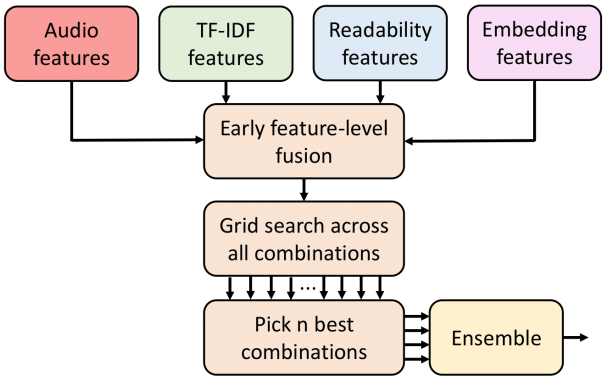

Figure 1: *Exploration of the feature space. Four types of features are combined by a concatenation of feature vectors (i.e., early feature-level fusion) and a grid search across all feature combinations is conducted. The best performing models employing best feature combinations are used for generating predictions on the test set, which are finally used for the ensembling (late prediction-level fusion).*

produce identical predictions on the test set and then calculating the mean of the predicted MMSE scores of these four best performing models in order to produce the fifth submission.

3. Experimental setting

In this Section we quickly overview the dataset and present the experiments conducted and results achieved in the scope of the ADReSS challenge. The Section is divided into three parts, Dataset (Section 3.1) Feature evaluation (Section 3.2) and Experimental results (Section 3.3).

3.1. Dataset

The dataset consists of recordings and transcripts of Cookie Theft picture descriptions by 78 AD and 78 non-AD participants of the Boston Diagnostic Aphasia Exam [23] and is balanced in terms of gender and age. Altogether the dataset contains 4,076 normalized speech segments, on average 24.86 per participant, and one transcript per each participant. It is split into a train set containing 108 examples and the test set containing 48 examples. For details, see [2].

3.2. Feature evaluation

In this experiment we explore the classification and regression performance of distinct feature sets in the 10-fold CV setting on the train set. SVM with box constraint of 10 was used in the feature evaluation experiments. Results for classification are presented in Figure 2. In general, TF-IDF features outperform all other feature types and among them, the best features are Char4grams that by themselves achieve the accuracy of 86.4%. While all TF-IDF feature sets lead to accuracy of about 70% or more, other types of features generally achieve accuracies between 50% and 60%, the only exception being ARI, which achieves accuracy just slightly above 60%. The worst performing feature is another readability measure, GFI, achieving accuracy just slightly above the chance level (51.8%). Among the audio features, the best performing are MFCC features (accuracy of 57.6%) and the worst are ADR features generated on the eGeMAPs (accuracy of 54.7%).

The feature performance on the regression task is somewhat consistent with the performance on the classification task (See Figure 3). TF-IDF features outperform other feature types and Char4grams are again the best features (achieving RMSE of 5.32). Also, ARI is again the best readability feature. On the other hand, MFCC features, which showed the best performance among audio features in the classification setting, are the worst features in the regression setting (achieving RMSE of 8.66). The best performing audio feature is the mean duration of the audio clips.

3.3. Experimental results

Results of the five best performing classification and regression models are presented in Table 1. The best classification accuracy of 77.8% on the official test set was achieved when a LogR model with a regularization strength (C) of 10 was trained on GFI, NUW, Duration, Char4gram, Suffix, POS tag and UD features. The same model also achieved the best accuracy in the CV setting, a much higher accuracy of 92.7%. On the other hand, for regression, the best RMSE score of 4.4388 on the test set was achieved by the SVM model with the box constraint of 10 trained on NUW, Bigram, Char4gram, Suffix, POS tag and GRA features, which performed the worst out of the four best regression models in the CV setting. While the ensemble of models produced the worst classification result on the test set, it ranked as second best on the regression task, although its performance was still much worse than the performance of the best model.

4. Discussion

The large discrepancies between the CV and test set classification performances suggest all the models overfitted, since all the models performed worse on the official test set than in the CV setting. The same can be said for four out of five regression models. Overfitting could be to some extent explained with the small size of the train set and might be limited by reducing the number of features. The one exception to the overfitting is the best performing regression model, which achieved a better RMSE score on the test set than in the CV setting. A more thorough error analysis would be required to explain this deviation.

Logistic/linear regression and SVMs with linear kernels proved better than Xgboost and Random forest models for both tasks. Some previous studies [24] suggest that these models work especially well on textual features and this could also explain their good performance on the tasks at hand, where textual TF-IDF features are the best performing features.

Besides the best performing textual features (Char4grams) and POS tags, which appear in most of the best classification and regression feature combinations, GFI and NUW also appear in 5 out of 9 best combinations, which suggests that readability measures add some useful information to the models. Interestingly, UD features only appear in best configurations for classification. When it comes to audio features, the best performing feature for classification appears to be Duration (appearing in 3 out of 5 best combinations) and the best performing feature for regression is MFCC ADR, appearing in 3 out of 4 best combinations. The doc2vec UD embedding features did not appear in any of the best combinations, most likely due to a very small train set which prohibits the successful training of an efficient embedding model.

Overall, our results outperform the baseline by a large margin [2] and are slightly worse than the results reported in the

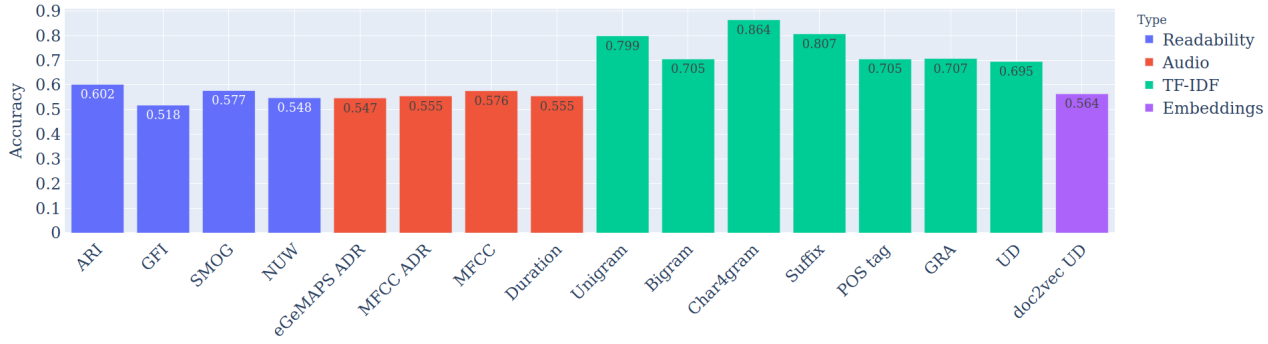

Figure 2: *SVM (with box constraint of 10) classification performance with different features.*

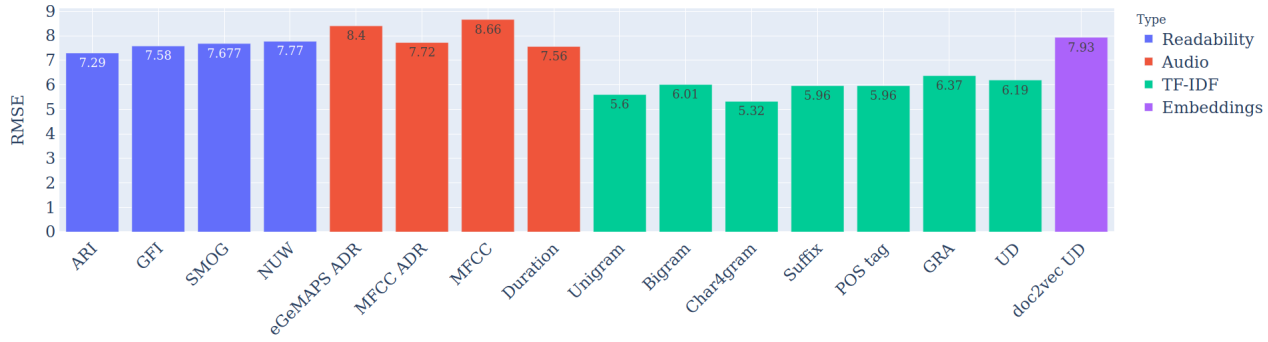

Figure 3: *SVM (with box constraint of 10) regression performance with different features.*

Table 1: *Results of the Cross validation (CV) and official test set experiments in terms of accuracy and RMSE.*

Classification			
Feature set	Model	CV score	Test set score
GFI,NUW,Duration,Character 4-grams,Suffixes,POS tag,UD	LogR (C=100)	**0.927**	**0.7708**
Duration,Character 4-grams,Suffixes,POS tag,UD	SVM (C=10)	0.918	0.7500
NUW,Duration,Unigram,Suffixes,POS tag,UD	LogR (C=10)	0.917	0.7500
GFI,Duration,Unigram,Bigram,Suffixes,POS tag,UD	SVM (C=10)	0.908	0.7500
duration,Unigram,Bigram,Suffixes,POS tag,UD	LogR (C=10)	0.907	/
Ensemble	/	/	0.7292
Regression			
Feature set	Model	CV score	Test set score
GFI,NUW,MFCC ADR,Bigram,Character 4-grams,Suffixes,POS tag	LinR	**5.008**	5.1878
GFI,NUW,MFCC ADR,Character,4-grams,Suffixes,POS tag	LinR	5.021	5.4312
GFI,MFCC ADR,Character 4-grams,Suffixes,POS tag	LinR	5.032	5.4483
NUW,Bigram,Character 4-grams,Suffixes,POS tag,GRA	SVM (C=10)	0.505	**4.4388**
Ensemble	/	/	5.0574

related work [4, 5], which have been achieved on a much larger and also unbalanced DementiaBank's Pitt corpus [25].

5. Conclusions

In this paper we have presented a multimodal approach to the ADReSS (Alzheimer's Dementia Recognition through Spontaneous Speech) challenge. The proposed method relies on a feature-level fusion between different feature types and an extensive grid search across all feature combinations, and exploits both audio and textual data for the automatic detection of Alzheimer's dementia.

The results suggest that a multimodal approach leads to better performance than unimodal approaches but also suggest caution about using many different features due to the overfitting risk. Besides testing new features (e.g., clinical features such as concept counts), our future work will therefore be focused on reducing the number of features in order to avoid overfitting, while still sustaining the predictive performance of the classification and regression models.

6. Acknowledgements

The authors acknowledge the financial support from the project SAAM – Supporting Active Ageing through Multimodal coaching (grant agreement no. 769661).

7. References

[1] W. H. Organization *et al.*, "Mental health action plan 2013-2020," 2013.

[2] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, "Alzheimer's dementia recognition through spontaneous speech: The ADReSS Challenge," in *Proceedings of INTERSPEECH 2020*, Shanghai, China, 2020. [Online]. Available: https://arxiv.org/abs/2004.06833

[3] *INTERSPEECH 2020 – 21th Annual Conference of the International Speech Communication Association, September 14-18, Shanghai, China, Proceedings*, 2020.

[4] K. C. Fraser, J. A. Meltzer, and F. Rudzicz, "Linguistic features identify alzheimer's disease in narrative speech," *Journal of Alzheimer's Disease*, vol. 49, no. 2, pp. 407–422, 2016.

[5] M. Yancheva and F. Rudzicz, "Vector-space topic models for detecting alzheimer's disease," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2016, pp. 2337–2346.

[6] J. S. Guerrero-Cristancho, J. C. Vásquez-Correa, and J. R. Orozco-Arroyave, "Word-embeddings and grammar features to detect language disorders in alzheimer's disease patients," *TecnoLógicas*, vol. 23, no. 47, pp. 63–75, 2020.

[7] S. Luz, "Longitudinal monitoring and detection of alzheimer's type dementia from spontaneous speech data," in *2017 IEEE 30th International Symposium on Computer-Based Medical Systems (CBMS)*. IEEE, 2017, pp. 45–46.

[8] F. Haider, S. De La Fuente, and S. Luz, "An assessment of paralinguistic acoustic features for detection of alzheimer's dementia in spontaneous speech," *IEEE Journal of Selected Topics in Signal Processing*, 2019.

[9] E. A. Smith and R. Senter, "Automated readability index," *AMRL-TR. Aerospace Medical Research Laboratories (US)*, pp. 1–14, 1967.

[10] R. Gunning, *The technique of clear writing*. McGraw-Hill, New York, 1952.

[11] G. H. Mc Laughlin, "Smog grading - a new readability formula," *Journal of reading*, vol. 12, no. 8, pp. 639–646, 1969.

[12] T. Kohonen, "The self-organizing map," *Proceedings of the IEEE*, vol. 78, no. 9, pp. 1464–1480, 1990.

[13] J. Bullard, C. O. Alm, X. Liu, Q. Yu, and R. A. Proano, "Towards early dementia detection: fusing linguistic and non-linguistic clinical data," in *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*, 2016, pp. 12–22.

[14] M.-C. De Marneffe, T. Dozat, N. Silveira, K. Haverinen, F. Ginter, J. Nivre, and C. D. Manning, "Universal stanford dependencies: A cross-linguistic typology." in *LREC*, vol. 14, 2014, pp. 4585–4592.

[15] B. Mirheidari, D. Blackburn, T. Walker, A. Venneri, M. Reuber, and H. Christensen, "Detecting signs of dementia using word vector representations," in *Interspeech*, 2018, pp. 1893–1897.

[16] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *International conference on machine learning*, 2014, pp. 1188–1196.

[17] B. Croisile, B. Ska, M.-J. Brabant, A. Duchene, Y. Lepage, G. Aimard, and M. Trillet, "Comparative study of oral and written picture description in patients with alzheimer's disease," *Brain and language*, vol. 53, no. 1, pp. 1–19, 1996.

[18] P. Scheltens, "100 questions and answers about alzheimer's disease," 2004.

[19] J. P. Kincaid, R. P. Fishburne Jr, R. L. Rogers, and B. S. Chissom, *Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel*. Institute for Simulation and Training, University of Central Florida, 1975.

[20] E. Dale and J. S. Chall, "A formula for predicting readability: Instructions," *Educational research bulletin*, pp. 37–54, 1948.

[21] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.

[22] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.

[23] H. Goodglass, E. Kaplan, and B. Barresi, *BDAE-3: Boston Diagnostic Aphasia Examination–Third Edition*. Lippincott Williams & Wilkins Philadelphia, PA, 2001.

[24] F. Rangel, P. Rosso, M. Potthast, and B. Stein, "Overview of the 5th author profiling task at pan 2017: Gender and language variety identification in twitter," *Working Notes Papers of the CLEF*, pp. 1613–0073, 2017.

[25] J. T. Becker, F. Boiler, O. L. Lopez, J. Saxton, and K. L. McGonigle, "The natural history of alzheimer's disease: description of study cohort and accuracy of diagnosis," *Archives of Neurology*, vol. 51, no. 6, pp. 585–594, 1994.