Detection of Alzheimer's disease based on automatic analysis of common objects descriptions

Laura Hernández-Domínguez
ÉTS, Quebec University
1100 Notre-Dame
Montreal, QC H3C1K3
laudobla@gmail.com

Edgar García-Cano
ÉTS, Quebec Univ.
1100 Notre-Dame
Montreal, H3C1K3
eegkno@gmail.com

Sylvie Ratté
ÉTS, Quebec Univ.
1100 Notre-Dame
Montreal, H3C1K3
Sylvie.Ratte@etsmtl.ca

Gerardo Sierra-Martínez
IINGEN, UNAM
Ciudad Universitaria
Mexico City, 04510
GSierraM@iingen.unam.mx

Abstract

Many studies have been made on the language alterations that take place over the course of Alzheimer's disease (AD). As a consequence, it is now admitted that it is possible to discriminate between healthy and ailing patients solely based on the analysis of language production. Most of these studies, however, were made on very small samples—30 participants per study, on an average—, or involved a great deal of manual work in their analysis. In this paper, we present an automatic analysis of transcripts of elderly participants describing six common objects. We used part-of-speech and lexical richness as linguistic features to train an SVM classifier to automatically discriminate between healthy and AD patients in the early and moderate stages. The participants, in the corpus used for this study, were 63 Spanish adults over 55 years old (29 controls and 34 AD patients). With an accuracy of 88%, our experimental results compare favorably to those relying on the manual extraction of attributes, providing evidence that the need for manual analysis can be overcome without sacrificing in performance.

1 Introduction

As life expectancy increases, age-related disorders increase as well, bringing great social, health and economic challenges for governments and societies in general. Researchers across the world are trying to find methods for detecting and treating these disorders in effective, non-invasive and cost-efficient ways.

AD affects one in ten adults over 65 years old in the United States (Alzheimer's Association, 2015). Interventions may be more effective in the early stages of dementia. Nevertheless, it is highly common, especially in low and middle income countries, to diagnose AD several years after the disease begins, leading to a treatment gap for early dementia sufferers (Alzheimer's Disease International, 2011). This gap could reduce the effectiveness of treatments, prolonging the patients' state of reduced independence. Alzheimer's Disease International (2015) identifies early diagnosis and treatment as a means of attenuating care costs and reducing this gap. Furthermore, an early diagnosis would allow the sufferers and their families to get their affairs in order by foreseeing the future better and preparing accordingly.

Many researchers have studied the early detection of AD. These studies usually follow two main approaches: the analysis of biomarkers and the examination of patients' decreasing cognitive abilities. The first approach yields reliable results in the detection of AD in its moderate and advanced stages, albeit still performing insufficiently in the early stages of the disease (Alzheimer's Association, 2015). The second approach has gained more attention in recent years, due to the fact that, in clinical practice, it has shown promise in the early detection of AD (Taler and Phillips, 2008; Schröder et al., 2010). Furthermore, when compared to the first approach, the analysis of the decline of cognitive abilities represents an inexpensive and noninvasive alternative.

Language skills are among the first cognitive abilities to diminish during the course of AD, with alterations appearing even before any symptom is experienced. Clinicians have designed many standard tests to evaluate language in elderly patients (Taler and Phillips, 2008), such as asking them to retrieve words from certain categories, to think of words that start with the same letter, to

Proceedings of the 7th Workshop on Cognitive Aspects of Computational Language Learning, pages 10–15,
Berlin, Germany, August 11, 2016. ©2016 Association for Computational Linguistics

name objects in pictures, etc. These tests, although sufficient to give a reasonably accurate diagnosis, present some problems in the clinical practice (Smith and Bondi, 2013). Such problems include production of nervousness and discomfort in elderly patients, as well as a "practice effect". Also, these tests do not necessarily describe patients' real performance in language production. This, apart from aiding in early detection of the disease, could help further our understanding of the disease, its progression and the parts of the brain affected in early stages (before the damage can be visible on MRI images).

In this article, we introduce our first experimental approach for automatic analysis of transcripts from elderly Spanish speakers. We aim to discriminate cognitively-healthy participants from (early and moderate) AD sufferers.

2 Related Work

Relatively few authors (Bucks et al., 2000; Jarrold et al., 2010; Guinn and Habash, 2012; Guinn et al., 2014; Jarrold et al., 2014; Alegria et al., 2013) have researched the automatic discrimination of AD patients using language analyses of transcripts, although there is a growing interest in recent years. In most studies, researchers examined the free discourse of elderly English-speakers. The most often-used features are part-of-speech rates, lexical richness measures, pauses, and incomplete words. Overall accuracy ranges from 73% to 95%, but between authors there is a disagreement on the features used. Some works, like Khodabakhsh et al. (2015) even minimize the usefulness of these types of features.

Most of these studies used very small samples (8-32 AD patients and 16-51 controls) taken in different settings (phone/face-to-face conversations, hospital/familiar environment, inconsistent thematics, etc.). These differences make it difficult comparing their findings. Given the small size of the samples, it would be helpful to use corpora with constrained settings, like restricted discourse and controlled environments, in order to discard differences attributable to factors unrelated to language. Moreover, further studies with non-English speakers would help us to enrich our understanding of language alterations due to AD.

In a different approach, Guerrero et al. (2016) trained a Bayesian Network using manually extracted conceptual components along with age, gender, and educational level as prior probabilities to detect AD. Their corpus consisted of transcripts of Spanish elderly participants orally describing six objects (Peraita and Grasso, 2010). The authors reported the following performance—accuracy, precision, recall (sensitivity), F_1-score, false positive rate, and false negative rate—:

Acc	**Pre**	**Rec**	**F_1**	**FPR**	**FNR**
0.91	0.94	0.87	0.90	0.05	0.01

Table 1: Results by Guerrero et al. (2015).

For this work, we studied the restricted-discourse corpus used in Guerrero et al. (2016), and trained an SVM using some of the linguistic features used by previous authors in the analysis of free conversations. We additionally incorporated two scarcely explored part-of-speech-based features—*conjunction rate* and *secondary verb rate*—. We compared our automatic analysis results to those obtained by Guerrero et al. using manually extracted conceptual components.

3 Methods

3.1 Corpus

Peraita and Grasso (2010) created a dataset[1] of oral descriptions in Spanish to study linguistic pathologies related to dementia, particularly AD. All recollections were obtained with the written informed consent of the participants (Grasso et al., 2011). The authors granted us permission to use their corpus for this study. We choose this corpus because its availability, restricted discourse and homogeneous recollections facilitate the comparison of our results with those of other researchers. Likewise, the size of this sample is comparable to the largest samples used in related works.

The cohort used by Guerrero et al. (2016) in their study includes a total of 69 participants (30 controls and 39 AD patients previously diagnosed by neurologists) aged between 55 and 95 years old. For each participant, Peraita and Grasso recorded free oral descriptions of six common objects (referred in their work as "semantic categories"): *dog*, *apple*, *pine* (living things), *car*, *trousers* and *chair* (non-living things). These descriptions were manually transcribed. Any interactions with or interventions by the interviewer

[1] http://www.uned.es/investigacion-corpuslinguistico/

Participant's description of *dog*:

"It is a loving animal. They are loving. They love the owner. They obey him. You leave him home alone and he cries. He misses you and when you arrive home he is very glad to see you. He guards the house pretty well. When he hears a noise in the stairs, he barks, meaning, he guards the house. They like you to pet them, to love them. They make great company."

Conceptual components:

Taxonomic:
- it's an animal.

Functional:
- he guards the house;
- they make great company.

Evaluative:
- they are loving; [...]

Figure 1: Translated sample from the corpus.

were excluded from the transcription. In addition, the authors of the corpus noted if a participant went "off-topic", but did not include these utterances in the transcript. They annotated this corpus to show marks of interruptions, off-topic, and unintelligible words.

In their study, Peraita and Grasso (2010) manually analyzed and extracted attributes from the description of each object. These attributes were divided into eleven categories: *taxonomic*, *types*, *parts*, *functional*, *evaluative*, *places/habitat*, *behavior*, *cause/generate*, *procedural*, *life cycle*, and *others*. In Figure 1, we provide a translated version of a sample taken from the corpus.

From the sample, we removed all participants with no utterances in the description of one or more objects. Additionally, since our objective was to evaluate the performance of a classifier for early detection of AD, we proceeded like Guerrero et al. (2016) and only considered the controls and patients in the early and moderate stages of AD. Our final sample consisted of a total of 63 participants (29 controls and 34 AD patients).

3.2 Linguistic features

For this work, we used a combination of 5 features that most authors have found suitable: verb, noun and preposition rates, Brunet's *W* index, and Honoré's *R* Statistics. Additionally, in a previous non-automatic study regarding the preservation of syntax in AD, Kemper et al. (1993) found that sentences produced by cognitively healthy adults usually contain more secondary verbs and conjunctions. We incorporated these findings, resulting in a total of 7 features.

Part-Of-Speech features:

- *Verb*, *noun*, *preposition* and ***conjunction rates***: the number of verbs, nouns, prepositions and conjunctions per 100 words, respectively.
- ***Secondary verb rate***: number of secondary verbs divided by the total number of verbs.

Lexical richness features:

- We used *Brunet's W index* (Brunet, 1978) to determine the richness of speakers' vocabularies:

$$W = N^{(V-1.165)} \quad (1)$$

  Where N is the total number of words used and V is the vocabulary size (number of different words used).

- *Honoré's R Statistics* (Honoré, 1979) measures lexical richness based on the number of a speakers once-mentioned words:

$$R = (100logN)/(1 - V_1/V) \quad (2)$$

  Where N is the total number of words used, V is the vocabulary size, and V_1 is the number of words mentioned only once.

3.3 Implementation

To perform the binary classification, we used a Support Vector Machine (SVM) implementation of the Python library scikit-learn (Pedregosa et al., 2011). For the automatic tokenization, lemmatization, and part-of-speech extraction, we used FreeLing 3.0 (Padró and Stanilovsky, 2012), an open source language analysis tool suite. We selected this package for its good performance in Spanish (although it also supports other languages), and for the way it encapsulates multiple text analysis services in a single application.

In their experiments, Guerrero et al. (2016) trained their Bayesian Network without directly linking risk factor variables (such as age, gender, or education) to the rest of the model. Instead, they used these *a priori* probabilities as deterministic inputs. In our experiments, we did not consider

these variables. We performed our classification based solely on linguistic features.

Using the above-mentioned risk factor variables and their correlation to AD could be useful in the improvement of the overall accuracy of these types of experiments. However, these correlations vary significantly depending on factors such as country, race, quality of life, diet, pollution, environment, etc. Moreover, in most countries, there are no reliable statistics about these correlations (Alzheimer's Disease International, 2015). Furthermore, most AD datasets have very few participants, and their distributions are not usually an accurate representation of the population. In practice, training an algorithm with the socio-demographic information presented in these datasets would lead to biased results.

In the core of their Bayesian Network, Guerrero et al. (2016) calculated the probability of a person having a lexical-semantic-conceptual deficit (LSCD)—which is considered by the authors as a major sign of cognitive impairment—in two main categories: "living things" and "non-living things". The authors obtained these probabilities based on the number of attributes present on each of the 11 categories; first individually for each object, and then jointly for the main category to which they belonged (living / non-living things). The reason behind this categorical division is that previous researchers have found an important difference in the number of attributes of living and non-living things in the descriptions given by AD patients in early stages and those given by healthy individuals. The authors used the k-means++ algorithm to discretize the presence of LSCD given the number of living and non-living things' attributes mentioned by a participant.

We designed two different experiments. In the first experiment, we followed the lead of Guerrero et al. (2016) and divided each human subject's descriptions into living and non-living things. From this, we extracted a total of 14 linguistic features (*set1*): 7 features (verb, noun, preposition, secondary verb and conjunction rates, Brunet's *W* index, and Honoré's *R* Statistics) from their descriptions of living things, and (the same) 7 features from their descriptions of non-living things. In the second experiment, we considered all the descriptions from each human subject as a unit and extracted the 7 linguistic features (*set2*).

Calibration: We tested two SVM kernels for both experiments, linear and Radial Basis Function (RBF). We used a 5-fold cross validation to calibrate the value of their respective hyperparameters. Cross validation used 80% of the data for training and 20% for testing. We shuffled the training and testing samples and selected them at random. For *set1*, the best model (**accuracy=86%**) used an RBF kernel (C=1.0, gamma=0.0001). The best model (**accuracy=88%**) for *set2* used a linear kernel (C=0.1). The best models' accuracies reflect the performance of the classifiers when dealing with completely unseen data.

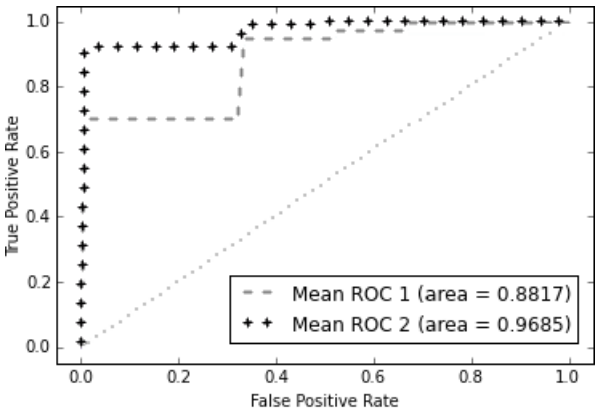

Figure 2: ROC curves and areas under the curves for the classifiers trained with *set1* and *set2*.

4 Results

We evaluated the two best models using 5-times-10-fold cross-validation over the dataset. In Table 2 we show the average of the most common performance metrics—accuracy, precision, recall (sensitivity), F_1-score, FPR (false positive rate), and FNR (false negative rate)—used for medical applications. Additionally, we obtained the ROC curves and areas under the curves (AUC) for both experiments (see Figure 2).

Set	Acc	Pre	Rec	F_1	FPR	FNR
1	0.87	**0.91**	0.88	0.88	**0.08**	0.12
2	**0.88**	0.89	**0.90**	0.88	0.10	**0.10**

Table 2: Performance metrics obtained with the classifier trained with *set1* and *set2*.

5 Discussion and future directions

As shown in Table 2, the differences in accuracy and F_1-score between the AD classifiers trained with *set1* and *set2* of features are not very per-

ceptible. The classifier of the second experiment has a slightly higher sensitivity (2% more), which means that it has a lower tendency of letting AD participants go unrecognized. When comparing the AUC of both classifiers, the difference is more noticeable; *set2* performed better than *set1*. From this, we concluded that for the linguistic features considered, there is no need to separate participants' descriptions into living and non-living categories.

Guerrero et al. (2016) reported an accuracy of 91% (see Table 1) and an AUC of 0.9636. They used a Bayesian Network fed with manually-extracted attributes and incorporated participants' socio-demographic information as *a priori* deterministic inputs. We obtained an accuracy of 88% and an AUC of 0.9685 by performing automatic language analysis, without taking into account any socio-demographic information. Although the manually extracted attributes' classifier performs slightly better, automatic language analysis reduces time and human effort and provides consistency and replicability.

There is another cohort of 143 speakers from Argentina in the corpus used in this work. The corpus is provided in a read-only application, and manually transforming the data into text format took a great amount of time. For this reason, we only analyzed the cohort of Spanish participants as Guerrero et al. (2016) did. To our knowledge, no experimental work has been done over it yet. Our next step will be to experiment with this cohort to explore intralanguage variations. We also intend to perform a study on less restrictive discourse contexts, like the work of Prud'hommeaux and Roark (2011) with story retellings.

For our first set of experiments, we selected some basic linguistic features commonly used in free spontaneous discourse analysis, but applied them to a particular restricted discourse context with very encouraging results for detecting AD in its early and moderate stages. In future experiments we will test more sophisticated linguistic features, and perform computational syntactic and semantic analysis. Furthermore, we will investigate performance of other classification algorithms. An in-depth analysis of features used and their relevance in this task is also planned.

Acknowledgments

We would like to thank Prof. Herminia Peraita from the National University of Distance Education in Spain for sharing her dataset. This research was partially supported by the FRQNT 177601 and 194703 files, the CONACYT 323619 scholarship, and the Ministère des Relations Internationales et de la Francophonie and CONACYT Mexico (XV Groupe de Travail Québec-Mexique 2015-2017).

References

Renne Alegria, Celia Gallo, Mirian Bolso, Bernardo dos Santos, Cleide Rosana Prisco, Cassio Bottino, and Nogueira Maria Ines. 2013. Comparative study of the uses of grammatical categories: Adjectives, adverbs, pronouns, interjections, conjunctions and prepositions in patients with Alzheimer's disease. *Alzheimer's & dementia: the journal of the Alzheimer's Association*, 9(4):P882, jul.

Alzheimer's Association. 2015. 2015 Alzheimer's Disease Facts and Figures. Technical report, Alzheimer's Association.

Alzheimer's Disease International. 2011. World Alzheimer Report 2011. The benefits of early diagnosis and intervention. Technical report, Alzheimer's Disease International.

Alzheimer's Disease International. 2015. World Alzheimer Report 2015. The Global Impact of Dementia. Technical report, Alzheimer's Disease International.

Etienne Brunet. 1978. *Vocabulaire de Jean Giraudoux: structure et évolution: statistique et informatique appliquées à l'étude des textes à partir des données du Trésor de la langue*. Slatkine (book).

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

Lina Grasso, M. Carmen Díaz-Mardomingo, and Herminia Peraita-Adrados. 2011. Deterioro de la memoria semántico-conceptual en pacientes con enfermedad de Alzheimer. Análisis cualitativo y cuantitativo de los rasgos semánticos. *Psicogeriatría*, 3(4):159–165.

José María Guerrero, Rafael Martínez-Tomás, Mariano Rincón, and Herminia Peraita-Adrados. 2016. Bayesian Network Model to Sup-port Diagnosis of Cognitive Impair-ment Compatible with an Early Diagnosis of Alzheimers Disease. *Methods of Information in Medicine*, 55:42–49.

Curry Guinn and Anthony Habash. 2012. Language Analysis of Speakers with Dementia of the Alzheimer's Type. In *Association for the Advancement of Artificial Intelligence Fall Symposia*, pages 8–13. AAAI.

Curry Guinn, Ben Singer, and Anthony Habash. 2014. A comparison of syntax, semantics, and pragmatics in spoken language among residents with Alzheimer's disease in managed-care facilities. In *2014 IEEE Symposium on Computational Intelligence in Healthcare and e-health (CICARE)*, pages 98–103. IEEE, dec.

A Honoré. 1979. Some simple measures of richness of vocabulary. *Association for literary and linguistic computing bulletin*, 7(2):172–177.

William L Jarrold, Bart Peintner, Eric Yeh, Ruth Krasnow, Harold S Javitz, and Gary E Swan. 2010. Language Analytics for Assessing Brain Health : Cognitive Impairment , Depression and Pre-symptomatic Alzheimer's Disease. In Yiyu Yao, Ron Sun, Tomaso Poggio, Jiming Liu, Ning Zhong, and Jimmy Huang, editors, *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, chapter Brain Info, pages 299–307. Springer Berlin Heidelberg, Berlin, Heidelberg.

William Jarrold, Bart Peintner, David Wilkins, Dimitra Vergryi, Colleen Richey, Maria Luisa Gorno-Tempini, and Jennifer Ogar. 2014. Aided diagnosis of dementia type through computer-based analysis of spontaneous speech. In *Proceedings of the ACL Workshop on Computational Linguistics and Clinical Psychology*, pages 27–36.

Susan Kemper, Emily LaBarge, Richard Ferraro, Hintat Cheung, Him Cheung, and Martha Storandt. 1993. On the preservation of syntax in Alzheimer's disease: Evidence from written sentences. *Archives of neurology*, 50(1):81–86.

Ali Khodabakhsh, Fatih Yesil, Ekrem Guner, and Cenk Demiroglu. 2015. 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):9, mar.

Lluís Padró and Evgeny Stanilovsky. 2012. Freeling 3.0: Towards wider multilinguality. In European Language Resources Association, editor, *Proceedings of the Language Resources and Evaluation Conference*, pages 2473–2479, Istanbul, Turkey.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, and J. Vanderplas. 2011. Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, 12:2825–2830.

Herminia Peraita and Lina Grasso. 2010. Corpus lingüístico de definiciones de categorías semánticas de personas mayores sanas y con la enfermedad del alzheimer. Technical report, Fundación BBVA.

Emily Tucker Prud'hommeaux and Brian Roark. 2011. Extraction of Narrative Recall Patterns for Neuropsychological Assessment. In *INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association*, pages 3021–3024.

Johannes Schröder, Britta Wendelstein, and Ekkehard Felder. 2010. Language in the Preclinical Stage of Alzheimer's Disease. Content and Complexity in Biographic Interviews of the ILSE Study. In *Klinische Neurophysiologie*, volume 41, page S360.

Glenn E Smith and Mark W Bondi. 2013. *Mild Cognitive Impairment and Dementia: Definitions, Diagnosis, and Treatment*. OUP USA, illustrate edition.

Vanessa Taler and Natalie A. Phillips. 2008. Language performance in Alzheimer's disease and mild cognitive impairment: a comparative review. *Journal of clinical and experimental neuropsychology*, 30(5):501–56, jul.