Automatic detection of linguistic indicators as a means of early prediction of Alzheimer's Disease

Vassiliki Rentoumi[1] George Paliouras[1] Dimitra Arfani[2] Katerina Fragkopoulou[2] Spyridoula Varlokosta[2]

[1]NCSR 'Demokritos' [2]National and Kapodistrian University of Athens

7th Language Disorders in Greek, 2-3 June 2018, Athens, Greece

Introduction

Although language changes in Alzheimer's Disease (hence, AD) have been well documented, there are several limitations in the language investigation of these groups. **A primary limitation** is that the language analysis of these groups is manual, a process which is **time consuming** and in most of the cases **subjective.**

The current study adopts a **computational approach** based on **machine learning** (hence, ML) to characterize **language samples** from people with AD in terms of linguistically defined criteria.

State of the art

ML methods have been used successfully to distinguish between patients with AD and AD patients with vascular load (Rentoumi et al. 2014) based on syntactic complexity and lexical variation.

ML methods have been used in order to identify primary progressive aphasia (Fraser et al. 2014).

ML methods have been used to identify various linguistic features in AD narrative speech (Fraser et al. 2016).

Recent linguistic analysis in AD has shown **more lexical errors** and **less syntactically complex** sentences than the control group (de Lira et al. 2011, Rentoumi et al. 2014).

Aims of this study

To adopt a computational approach in the analysis of written samples obtained from native speakers of Greek, diagnosed with mild and moderate AD, in order to

✓ compare morphosyntactic complexity and lexical variation

✓ confirm and explain differences in language produced by AD patients and normal controls (NC) using quantitative methods of evaluation.

✓ introduce a new framework in order to automatically detect early indicators of AD and facilitate the diagnostic process of AD.

Materials and Methods

Samples obtained with the Cookie Theft Description Task (Boston Diagnostic Aphasia Examination, Goodglass et al. 2001).

Participants

Demographics	AD (n=30)	Controls (n=30)	Statistical Significance
	Mean	Mean	*=p-value<0.05, n.s = not significant
Age (years)	66.48	68,03	n.s
Education (years)	12	13,93	n.s
Gender (male: female)	13:17	16:14	n.s
MMSE scores	22.68	28.26	*

Analytical Approach

Machine learning (ML) algorithms can learn from data. In our case what is learned is the **syntactic complexity** and **lexical variation** (vocabulary variation and information characteristics (features)) that the language data sets exhibit. Our ultimate aim is to employ a ML algorithm and features that will correctly classify every sample into its correct group.

The proposed methodology is articulated in **two consecutive stages**:

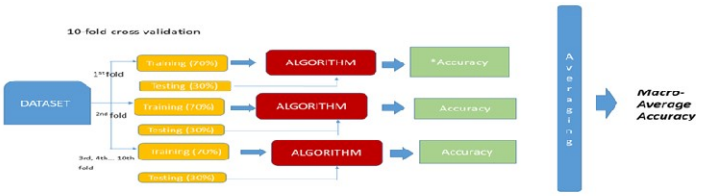
•**Feature extraction**: Automatic extraction of features with the use of a) a ***Part of Speech* (PoS) *tagger*** (Petasis 2014); b) the ***NP chunker for Greek*** (Petasis 2000); c) the ***Alzheimer's detector*** to create a set of values representing a number of distinct characteristics of each text.

•**Classification**: Automatic classification of written samples to their categories (AD vs. NC) with the use of ***Naive Bayes*** and ***SMO algorithms*** to assign a sample to the most likely class in the Waikato Environment for Knowledge Analysis (Hall et al. 2009).

Features Extracted

Lexical variation measures	1. Lexical word variation	LV	Number of lexical word types/ number of lexical words
	2. Bi Logarithmic type token ratio	Log TTR	Log_e word types/ Log_e total words (tokens)
	3. Noun variation	NV	Number of noun types/ number of lexical words
	4.Adjective variation	ADJV	Number of adjective types/ number of lexical words
	5.Modifier variation	MODV	Number of adjective types + adverb types/ number of lexical words
	6.Adverb variation	ADVV	Number of adverb types/ number of lexical words
	7.Corrected verb variation	CVV	Number of lexical verb types/ number of verbs x 2
	8.Verb variation	VV	Number of lexical verb types/number of lexical words
	9. Brunet	W	$N^{V^{-0.165}}$ (N = number of word tokens ; V = vocabulary)
Syntactic complexity measures	10.Mean length of sentence	MLS	Number of words/ number of sentences
	11.Mean number of noun phrases	MNP	Number of noun phrases / Number of all sentences of each text

System Description



$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

$$MacroAverage_{Accuracy} = \frac{Acc1+Acc2+Acc3..+Acc10}{10}$$

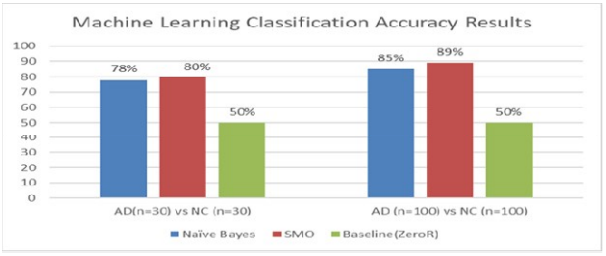
Results

Feature extraction Results	
LV	AD < NC
LogTTR	AD< NC
NV	AD < NC
ADJV	AD < NC
MODV	AD < NC
ADVV	AD < NC
CVV	AD < NC
VV	AD < NC
Brunet W	AD > NC
MLS	AD < NC
MNP	AD < NC

•**Lesser extent of Syntactic Complexity and Lexical Variation** in the language of AD vs. NC group

-Content words (NC > AD)

-Pronouns (AD > NC)



AD vs NC using NB, SMO > Baseline (p-values < 0,005)

- **1st classification task: 30 real** samples for each category,
- **2nd classification task: 70 synthetic samples**, employing the SMOTE algorithm.

For both comparisons A and B, in both classification tasks, NB and SMO significantly outperformed the baseline condition.

Classifiers (NB) efficiency is measured in terms of accuracy. The baseline condition was implemented using the ZeroR classifier provided by WEKA, which predicts the majority category.

Conclusions

The language of AD is **distinguishable** from the language of the NC group. Lexical variation and syntactic complexity are very good **discriminating factors** when it comes to distinguish the language of AD and NC groups. The current approach verifies our primary research hypothesis that cognitive deficits of AD patients can be reflected in their written language and these cognitive deficits are evidenced in both lexical variety and syntactic complexity domains.

Selected references

P. Garrard, V. Rentoumi, B. Gesierich, B. Miller, and M.L., Gorno-Tempini, "Machine learning approaches to diagnosis and laterality effects in semantic dementia discourse", Cortex Vol. 55, pp. 122-129, June 2014.

K.C. Fraser, JA. Meltzer, NL Graham, C. Leonard, G. Hirst, SE Black, and E. Rochon, "Automated classification of primary progressive aphasia subtypes from narrative speech transcripts", Cortex, Vol. 55, pp. 43-60, June 2014.

J.O. de Lira, K.Z. Ortiz, A.C. Campanha, P.H.F. Bertolucci, and T.S.C., Minett, "Microlinguistic aspects of the oral narrative in patients with Alzheimer's disease", International Psychogeriatrics Vol. 23 no. 3, pp. 404412, April 2011.

B. Roark, M. Mitchell, J.P Hossom, K. Hollingshead, and J. Kaye. "Spoken language derived measures for detecting mild cognitive impairment", IEE Transactions on audio, speech and language processing Vol. 19 no. 7, pp. 20812090, 2011.

Rentoumi, Vassiliki, Ladan Raoufian, Samrah Ahmed, Celeste A. de Jager, and Peter Garrard. "Features and machine learning classification of connected speech samples from patients with autopsy proven Alzheimer's disease with and without additional vascular pathology." Journal of Alzheimer's Disease 42, no. s3 (2014): S3-S17.

«This research has been financially supported by General Secretariat for Research and Technology (GSRT) and the Hellenic Foundation for Research and Innovation (HFRI) (Scholarship Code: 14541)".