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

Vassiliki Rentoumi1 George Paliouras1 Dimitra Arfani2 Katerina Fragkopoulou2 Spyridoula Varlokosta2
1NCSR ‘Demokritos’ 2National 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).

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
1. compare morphosyntactic complexity and lexical variation
2. confirm and explain differences in language produced by AD patients and normal controls (NC) using quantitative methods of evaluation.
3. 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).

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:
1. 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.
2. 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

<table>
<thead>
<tr>
<th>Feature/measure</th>
<th>AD vs. NC</th>
<th>Statistical Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lexical variety</td>
<td>AD &gt; NC</td>
<td>*p-value &lt; 0.05</td>
</tr>
<tr>
<td>Log TTR</td>
<td>AD &lt; NC</td>
<td>*p-value &lt; 0.05</td>
</tr>
<tr>
<td>Neum variation</td>
<td>AD &gt; NC</td>
<td>*p-value &lt; 0.05</td>
</tr>
<tr>
<td>Adjective variation</td>
<td>AD &gt; NC</td>
<td>*p-value &lt; 0.05</td>
</tr>
<tr>
<td>Adverb variation</td>
<td>AD &gt; NC</td>
<td>*p-value &lt; 0.05</td>
</tr>
<tr>
<td>Corrected variation</td>
<td>AD &gt; NC</td>
<td>*p-value &lt; 0.05</td>
</tr>
<tr>
<td>Number of noun phrases</td>
<td>AD &gt; NC</td>
<td>*p-value &lt; 0.05</td>
</tr>
<tr>
<td>Number of content words</td>
<td>AD &lt; NC</td>
<td>*p-value &lt; 0.05</td>
</tr>
</tbody>
</table>

System Description

Results

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

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


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: 145418).