Towards Automatic Detection of Abnormal Cognitive Decline and Dementia Through Linguistic Analysis of Writing Samples

Weissenbacher Davy
Department of Biomedical Informatics, ASU
Scottsdale, AZ, USA
davy.weissen-bacher@asu.edu

Johnson A. Travis
Department of Neurology
Mayo Clinic
Scottsdale, AZ, USA

Wojtulewicz Laura
Department of Biomedical Informatics, ASU
Scottsdale, AZ, USA

Dueck Amylou
Department of Biostatistics
Mayo Clinic
Scottsdale, AZ, USA

Locke Dona
Department of Psychiatry and Psychology, Mayo Clinic
Scottsdale, AZ, USA

Caselli Richard
Department of Neurology
Mayo Clinic
Scottsdale, AZ, USA

Gonzalez Graciela
Department of Biomedical Informatics, ASU
Scottsdale, AZ, USA
Graciela.Gon-zalez@asu.edu

Abstract

Given the limited success of medication in reversing the effects of Alzheimer's and other dementias, a lot of the neuroscience research has been focused on early detection, in order to slow the progress of the disease through different interventions. We propose a Natural Language Processing approach applied to descriptive writing to attempt to discriminate decline due to normal aging from decline due to pre-dementia conditions. Within the context of a longitudinal study on Alzheimer's disease, we created a unique corpus of 201 descriptions of a control image written by subjects of the study. Our classifier, computing linguistic features, was able to discriminate normal from cognitively impaired patients to an accuracy of 86.1% using lexical and semantic irregularities found in their writing. This is a promising result towards elucidating the existence of a general pattern in linguistic deterioration caused by dementia that might be detectable from a subject's written descriptive language.

1 Introduction

Alzheimer's disease is prevalent and becoming more so as the world's population ages (Prince et al., 2014). Since no cure is known, it is hoped that early detection and intervention might slow the onset of symptomatic cognitive decline and dementia. Clinical methods to detect Alzheimer's disease are typically applied well after symptoms have progressed to a troubling degree, and may be costly. Families, however, often report earlier signs of the disease through their language interactions with their elders. This has led clinical researchers to study linguistic differences to detect the disease in conversational speech (Asp and de Villiers, 2010). One approach is to search for non-informative phrases or semantic incoherences, which was confirmed to distinguish patients with Alzheimer's disease from controls (Nicholas et al., 1985). A strong limitation for its automatic application is the need of a trained expert to annotate the incoherences and scoring by hand.

We propose in this study to use Natural Language Processing (NLP) to evaluate samples of a patient's descriptive writing in order to attempt to

Proceedings of NAACL-HLT 2016, pages 1198–1207,
San Diego, California, June 12-17, 2016.

discriminate decline due to normal aging from decline due to pre-demented conditions. The Arizona Alzheimer's Disease Center (ADC) is a longitudinal study of patients with Alzheimer's disease and normal control subjects, who receive an annual battery of clinical and neuropsychological exams, to which we added the following brief and a simple task. Participants are asked to describe, in writing, a picture typically used within the speech-based Boston battery (Nicholas and Brookshire, 1993). We collected 201 descriptions written by ADC participants by hand, which were scanned, transcribed, and later analyzed. We describe here a statistical machine learning method relying on lexical, syntactical and semantical features to discriminate evidence of abnormal deterioration in the writings of the patients. Our results confirm a correlation between linguistic decline on this writing task and the cognitive decline revealed by the more time consuming neuropsychological test battery.

2 Background

Alzheimer's disease (AD) is a highly prevalent neurodegenerative dementia that increases exponentially with age. It is the most common form of dementia in the United States. AD is characterized by a severe memory deficit and at least one of the following: aphasia (an impairment of language, affecting the production or comprehension of speech and the ability to read or write), apraxia (loss of the ability to execute or carry out skilled movements and gestures), agnosia (inability to recognize and identify objects or persons), and a disturbance in the internal control of cognitive processes (such as reflection, planning, working memory, etc.) (American Psychiatric Association, 1994). While clinical testing often leads to an accurate diagnosis during its middle and late stages, several signs may alert a patient's family to much earlier stages of the disease even in the absence of frank aphasia (Obler and de Santi, 2000).

Given the repeated failures of experimental therapies targeting dementia stage AD, current strategies are targeting early intervention at preclinical and early symptomatic stages thereby necessitating more accurate methods for earlier detection of AD. Mild Cognitive Impairment (MCI), is defined as abnormal cognitive decline relative to age-matched peers that does not impair normal activities of daily living (Gauthier et al., 2006). AD is a frequent but not invariant cause. Some MCI patients may even recover, but all AD patients transition through the MCI stage before developing frank dementia (Petersen et al., 2001). As a result, an increasing number of clinical studies are trying to define and predict each stage in the life of an AD patient: normal, MCI and Alzheimer (Drummond et al., 2015).

2.1 Predicting Cognitive Decline with Language

Test batteries commonly used to measure cognitive decline include tests to evaluate the language production of patients, but they are criticized for their simplicity. For example, the Mini-Mental State Examination (MMSE), a widely used screening tool, asks to name 2 objects, to repeat a phrase, write a sentence and obey a 3-step instruction. Bucks et al., 2000, citing Sabat, 1994, assert that these structured tests break down language into artificial components that fail to represent the psychological and sociological context involved in daily conversations. As a consequence, such tests may be insensitive to early linguistic decline, when anomalies are already detectable by patients' families (Key-DeLyria, 2013).

More sophisticated exercises have been proposed to complement the existing linguistic test batteries (Asp and de Villiers, 2010). These exercises are centered around conversation and narration abilities of patients. Conversation and narration abilities are developed in the early age of children (around 2-3 years for conversation and around 4 years for narration). Since they play a fundamental role in cognitive and social development, they are intensively studied. Cognitive tests addressing narration capabilities can probe memory, spontaneity and the quality of interactions with the interlocutor. Tests can be complex, like narrating through informal conversation a habitual task, a memorable day of their life, or an event they participated in during the last week or month. Typically, the exact utterances are not captured, but rather the examiner notes if the narrative was coherent, or if the expected events were mentioned. Simpler tests ask patients to comment on an image, or a sequence of related images or to narrate a movie previously displayed. The patients participating in our study are receiving an extensive battery of tests annually

to which we added a linguistic task. We therefore opted for a simple exercise of image description to avoid exhausting our participants. While the majority of the exercises testing the narration abilities are spoken, with the exception of (Hayashi et al., 2015) and (Hirst and Feng, 2012), all studies work with a corpus of transcribed oral narratives. We opted for a written version for a direct analysis of written language, a form that remains relatively unexplored (Hayashi et al., 2015).

2.2 Clinical Studies for Linguistic Decline Prediction

A seminal longitudinal study (Snowdon et al., 1996) demonstrated that writing performance in young women correlated with development of AD in old age. Since then, clinical studies of cognitive decline have been scrutinizing all linguistic levels (Reilly et al., 2011), lexical, syntactical, semantical and pragmatic (Bolshakov and Gelbukh, 2004), in order to detect elements deteriorating with normal aging, those commonly observed degraded in the MCI stage, and finally their disintegration during the continuous phases of dementia. Various properties of language are studied, *e.g.* number of words, size of sentences, number and correctness of anaphoric references, number of propositions per sentence, number of relevant facts and the structure of the narration (Hier et al., 1985; Drummond et al., 2015). These properties are most often computed manually on samples of small size (usually around 50 patients) and appropriate statistical tests are used to determine the properties which can discriminate controls, MCI and AD patients.

From these studies has emerged a general pattern of pathological language decline observed during the MCI and the early stage of dementia (Obler and de Santi, 2000). Phonology and morphology are conserved. Syntax is also mostly spared even if it tends to be simplified. Degradations are mainly found at the lexical and semantical levels (Hier et al., 1985). At the lexical level, the vocabulary is reduced with fewer words and fewer occurrences. It becomes more abstract and vague with multiple phrasal repetitions (Xuan et al., 2011). At the semantic level, complex questions are reduced and, early in the dementia phase, patients have difficulty making exact and pertinent remarks (Nicholas and Brookshire, 1993). Empty words and incomplete sentences are often observed in oral exercises.

These alterations of the language seem to allow caregivers and researchers to distinguish decline due to normal aging from pathological decline but, further studies with larger patient numbers are needed to confirm these initial results. A significant limitation in clinical environment has been the need for a trained language pathologist to annotate and evaluate all linguistic productions of each patient examined. More recently, however, some efforts have been made to automate the annotation process using NLP techniques. The next section reviews the progress made.

2.3 Automatic prediction of Linguistic Decline

A first hypothesis to detect the cognitive decline in an older person is to compare his/her writing at a young age with his/her writing at an old age. In (Hirst and Feng, 2012) sophisticated stylometric measures were tested to identify the differences caused by the disease in the style of three well-known authors (2 probable ADs and 1 healthy). However, not only were results not decisive given the small number of subjects, but this approach required a large amount of writings from the same person in order to establish the shift in the style of that person, conditions rarely met with common subjects. A variant of this approach is to compute two distinct profiles by modeling separately normal subjects and aphasic subjects from their writings. The results reported in (Holmes and Singh, 1996) report 88% of subjects correctly predicted from a corpus of 100 conversations. Few features were used and the computation of some of them still required a human intervention.

Bigger set of features can be explored with the use of NLP and machine learning. A first attempt in (Thomas et al., 2005) was to combine stylometric features (Stamatatos, 2009) and language model within a classifier. Their classifier obtained reasonable performances with 70% accuracy when distinguishing cognitively impaired from normal subjects in 95 oral interviews. In (Jarrold et al., 2010), the authors evaluated 80 features from various categories computed using dictionaries and predefined rules: positive sentiments words, socially related words, use of the first person, among

others. The performance reported an accuracy of 82.6% in the prediction task in 45 interviews.

The most efficient features for discrimination are semantic features which capture the abilities of a subject to understand and convey a set of pertinent information (Nicholas and Brookshire, 1993). Automatic computation of such features are still challenging for automatic systems. Therefore, several publications integrated heuristics for computing such features. A prototype to approximate the density of idea has been released by (Brown et al., 2008). Idea density can be thought of as the total number of assertions or claims whether true or false, in a proposition. The number of claims is estimated from the number of verbs, adjectives/adverbs and conjunctions given certain conditions. The integration of the idea density proved to be significant to separate AD subjects from controls in (Jarrold et al., 2010).

3 Methods

3.1 Corpus Description and Preprocessing

In the context of the ADC study we created a corpus for our experiments. At the day of writing, the total number of subjects participating in the ADC study was roughly 500 corresponding to about 200 normal controls, 100 with MCI and 200 with AD or other form of degenerative dementia. In the beginning of the year 2015, in collaboration with the five institutes participating on the ADC study, a cognitive test was added to the protocol of the study. Subjects were asked to describe an image at the end of their annual visit. This control image is the same for all subjects (Nicholas and Brookshire, 1993). The image (Figure 1) represents a family having a picnic near a lake. Subjects were asked to write (by hand) a detailed description of the scene in the picture. No time limit is imposed, and the time it takes them to write their description is noted. The test giver is asked to read the description when the subject completes it, asking the subjects to clarify any unreadable words and to write them in the descriptions. We collected 201 descriptions for this study, 154 from healthy subjects and 47 from subjects in decline. The collection process is ongoing[1].

Figure 1: The picnic scene described by the ADC cohort of patients.

We developed a web site to centralize the collection of the scans of the descriptions from the different institutions. The web site offers a basic interface to display the scans and to transcribe their contents. We trained a student (native English speaker) to transcribe the scans, preserving, as much as possible, the original presentation of the description (*i.e.* punctuations, uppercase, indents and new lines) as well as misspellings and crossed words.

Alzheimer Patient
Jane and Joe went out to blow But the weather was windy in the Oposit Direction, so they decided To blow the joint rather place and go home and have a bond fire in Their backyard and enjoy all the cooked things they could
Normal Patient
A family outing at a lake shore showed people doing several things. Mom and Dad sat on a blanket while dad read a book. Dad was over comfortable without his shoes, while mom listened to the radio and poured herself a cup of coffee. Junior was having fun flying his kite, and the family dog was interested in what all was going on. Another of the family was spending quiet time and fisherman, and another was playing in the shallow water. Other friends waved to them as they sailed by. It was a perfect day with just enough wind to move the flag and provide lift for the kite. It must have been comfortable sitting under the shade tree.

Table 1: Example of writings AD vs Normal Patient.

The descriptions are processed through an NLP pipeline composed of several off-the-shelf NLP modules. First, a homemade tokenizer and the Stanford Lemmatizer[2] are applied. Part of Speech as well as chunks are computed thanks to Genia

[1] The corpus is fully de-identified and will be publicly released at the end of the study.

[2] Available at http://stanfordnlp.github.io/CoreNLP/

tagger[3]. The descriptions are split into phrases by the sentence splitter found in the ANNIE tools suites of the Gate pipeline[4]. To compute the language models we have integrated the character Ngrams module provided by LingPipe[5] as well as a specific Perl module Text::NGrams (Keselj et al., 2003) for computing character Ngram frequencies. Finally, for computing the semantic features describe below (section 3.2.3), we compute vectors of words which are semantically close to a selected set of words that correspond to a model description. To generate these vectors we have selected the tool Word2Vec[6]. We used the vectors trained on part of Google News dataset (about 100 billion words).

For each sample writing, we have access to all information acquired during the ADCC study about the subjects enrolled. This includes personal information (*e.g.* gender, sex or education), social and medical information (*e.g.* social status, smoking habit, depression) as well as the subjects' tests administered during the visits. For our experiments, we used the primary diagnostic made during the last visit of a subject. If the subject was diagnosed with any form of dementia, including possible or probable Alzheimer's, or with MCI, the subject was labeled as *Declined*. If the subject was not diagnosed with dementia we checked the score measuring the cognitive status. This score is assigned by a neuropsychologist and it summarizes the performance of the subject during the cognitive exams. If the neuropsychologist diagnosed the subject as cognitively impaired or as demented, the subject is labeled as *Declined*. Finally, we checked the Clinical Dementia Rating (CDR) global score (Morris et al., 1997). The CDR is assigned using a semi-structured standardized interview completed with the subject's caregiver and the subject independently. The CDR score is used to help diagnose dementia, indicating: Normal, MCI, Early Dementia, Moderate Dementia, and Severe Dementia, depending on its value. Administrators of the CDR are trained in a standardized fashion. If the score of the CDR indicated the subject as MCI or Dementia, then we labeled the subject as *Declined*, otherwise the subject was *NotInDecline*. These labels were used as gold standard during our experiments.

[3] Available at http://www.nactem.ac.uk/GENIA/tagger/
[4] Available at https://gate.ac.uk/
[5] Available at http://alias-i.com/lingpipe/
[6] The tool and its documentation are available at https://code.google.com/p/word2vec/

3.2 A Classifier for Detecting Linguistic Decline

In order to automate the analysis of the descriptions of our 201 subjects we created a classifier to discriminate subjects in abnormal decline from subjects with normal aging decline. Our classifier incorporates various features proposed by us or found in the literature. The following sections details the features and the motivations for their use.

3.2.1 Lexical Features

Adjective/Noun/verb/Pronoun ratios (Thomas et al., 2005). Given an abnormal decline we expected an important impoverishment of the vocabulary. Our initial hypothesis was a sensitive diminution of the number of adjective and pronouns since they are indicative of a precise description and complex syntactic structures. These ratios were computed by taking the number of adjectives/nouns/verbs/pronouns divided by the total number of tokens contains in a description. We relied on the POS tags to determine if a word was a noun, adjective or verb. To find the pronouns we matched a list of 73 pronouns.

Type Token Ratio (Thomas et al., 2005). The use of this ratio was supported by the idea that a subject presenting an abnormal decline will see his/her vocabulary reduced and would tend to repeat general words. This ratio was computed by taking the size of the vocabulary of a description over the total number of tokens. The vocabulary was found by adding up the lemmas occurring in the description.

Documents, Sentences and Tokens length (Hirst and Feng, 2012). The length of the different components of a document are often a good indicator of the quality of the writing and the ability to produce long and complex descriptions. We expressed several statistics which describe the description. The description length is expressed in number of tokens and punctuations. The size of the longest and shortest sentences, *min-max sentence length*, were used as features as well as the average of the length of all sentences occurring in the description.

The average length of the tokens occurring in the description was also added as feature.

Misspelling Ratio (Proposed). For this ratio we considered only orthographic errors present in a description. Since longer descriptions are more likely to have more misspellings we normalized the metric by dividing the number of errors with the total number of tokens in the description. To discover automatically the misspellings we used the rule-based spell checker languagetool-3.0[7]. As for the previous ratios we assumed that a higher percentage of misspellings would reflect an underlying lexical problems.

3.2.2 Stylometric Features

Functional Words Ratio (Hirst and Feng, 2012). Functional words are known to be good indicators of a personal style (Stamatatos, 2009). We matched an extended dictionary of 337 entries to retrieve the functional words in our descriptions. The ratio was given by the number of functional words over the total number of tokens in a description.

Brunét's Index and Honoré's Statistic (Thomas et al., 2005). Both metrics are length insensitive versions of the *type token ratio* and often reported as useful features for discriminating abnormal decline in the literature. They were computed by the following equations:

Brunét's Index = $N^{V-0.165}$ and Honoré's Statistic = $100logN/1-\frac{V_1}{V}$ where V is the total vocabulary, N the total number of tokens and V_1 the total number of hapax.

Character NGrams and Character NGram Frequencies (Thomas et al., 2005). Ngrams of words capture lexical regularities hidden in the writing style of an author as well as its syntactic complexity. They also help to highlight syntactic errors. Since sparsity problems raise quickly when Ngrams of words are created from a small size corpus, we preferred to use Ngrams of characters. By taking the most frequent Ngrams for both profiles Normal subjects and subjects in decline, we expected to capture the set of words which are the most indicative of each profile. We set the size of the Ngrams to 5 for the character NGrams and to 10 for the Character NGram Frequencies. We limited to the 2000 most frequent Ngrams. Those parameters were set manually and can be optimized in future experiments.

3.2.3 Semantic Features

Idea Density (Brown et al., 2008) To compute the idea density detailed in section 2.3, a heuristic to estimate the quantity of information convey in the description, we integrated the software CPIDR 3.2[8]. No change has been made in the set of rules used by the software.

Word2Vec Distance (Proposed). A characteristic of subjects in abnormal decline is their inability to convey pertinent information and to digress from the initial subject. To model this characteristic we propose a new feature which takes advantage of the specificity of our corpus: all subjects, normal and subjects in decline, are describing the same image. By taking only descriptions written by normal subjects we obtained a set of words describing correctly the image. We named this set *generative words*. All functional words were removed from this set. Our hypothesis was that subjects in decline would use less words from *generative words* and add more inappropriate words (given the context of the image). Since the size of our corpus is small, not all relevant words were present in *generative words*. We extended *generative words* into a set called *Word2Vec clusters* by adding for each word of *generative words*, the corresponding vector returned by Word2Vec. These vectors are composed by words semantically close to the generative words. This includes synonyms, meronyms, hyperonyms but also correlated words. At run time, when an unknown description was submitted to the system, we created a subset of *Word2Vec clusters*, called *Filtered Word2Vec clusters*, by taking all vectors V_i in *Word2Vec clusters* related to the words W_i occurring in the unknown description. We added V_i in *Filtered Word2Vec clusters* if W_i was the generating word of V_i or if W_i was a word occurring in V_i with W_i belonging to the set *generative words*. If W_i was found in a vector $V_j \in$ *Word2Vec clusters* but V_j was generated by a word w_j not occurring in the unknown description,

[7] Available at http://wiki.languagetool.org/java-api

[8] The software and its documentation are freely available at http://ai1.ai.uga.edu/caspr/

V_j was not added in *Filtered Word2Vec clusters*. This filtering step is crucial to guarantee good performances when using this feature. Additional tests were performed without filtering *Word2Vec clusters* and a significant drop of performances was observed due to noise or ambiguity in the vector generated by Word2Vec, for example vectors generated by *go*, *be* etc. The filtering step insures that the vectors of *Filtered Word2Vec clusters* contain only words semantically related with the content of the unknown description. Given the set of words in *Filtered Word2Vec clusters* the distance is the ratio of words W_i in *Filtered Word2Vec clusters* and total number of words in W_i.

3.2.4 Subject Features

All clinical information about the subjects participating in the ADC study were available during our experiments. We retained only criteria known to affect linguistic competences or known to contribute to the development of the disease. *Age* and *gender* are important factors for the Alzheimer's disease as well as the version of the *APOE* gene of a subject. The presence of an e4 allele increases significantly the risk of the disease. *Education* and *primary language* (native English speaker or not) are obvious attributes to consider to measure the linguistic abilities as well as the *social status* of the subject. A subject living alone, with relatives or spouse will not have the same opportunities to speak.

4 Results

We evaluated our classifier on the data mining platform *Weka*. This platform implements state-of-the art machine learning algorithms (Witten et al., 2011). The size of our corpus being small we opted for a leave-one-out cross validation. We chose the framework of a Bayesian Network (BN) (Koller and Friedman, 2009) to perform the evaluation of our classifier. For all following experiments we learned the structure of the network and its conditional probabilities automatically from our data. No Naive Bayes structure were *a priori* imposed during the training and the number of possible parents for a node were manually set to 20. We selected this machine learning algorithm because it learns complex decision functions, its decisions are interpretable by medical experts, it has very few global parameters to set up and it was fast to train on our problem.

Our first experiment evaluated the performances of our classifier when all features were used (Table 2). We confirmed the quality of our classifier by comparing its performances with a baseline classifier. The baseline classifier predicted the majority class label

Classifier	Accuracy (%)	FN	FP
Baseline	76.6	47	0
Bayesian Network			
- All Features	83.1	25	9
- Selected Features	86.1	21	7

Table 2: Performances of the classifiers for decline detection. Considering Decline as the targeted class, False Positive are Normal patients labeled as patients in decline and False Negative are patients in decline labeled as Normal patients.

NotInDecline for all instances. The baseline system obtained 76.6% of accuracy (Acc). With this setting, our classifier obtained a better score with 80.6% Acc. and thus demonstrated its abilities to learn the difference between normal subjects from subjects in abnormal decline using linguistic features.

We proceeded to an ablation study to assess the benefits of each feature. We removed one at a time each feature, or complementary features such as *min-max length of sentence*, and rerun the training/testing of our classifier. The results are detailed in Table 3. For brevity we did not report in the table the features which did not change the score of our classifier once removed.

Feature removed	Accuracy (%)
None	83.1
Misspelling Ratio	85.1
Word2Vec Distance	81.9
Brunét's Index	82.1
Average Sentence Length + Min-Max Sentence Length	83.6
Ngram Frequencies	85.1
Ngrams	81.6
Patient APOE	84.1
Patient Age	82.6

Table 3: Performances of the Bayesian Network during the ablation study.

In the light of the ablation study we performed a second experiment to determine the optimal per-

formances of our classifier. We run several feature selection/reduction algorithms implemented in the *Weka* platform. The Correlation-based Feature Selection algorithm (CFS) (Hall, 1999) found a set of features which maximized the performances of the classifier. Under this setting our classifier outperformed the baseline system with a score of 86.1 Acc. against 76.6 Acc (Table 2). Inspection of the confusion matrix shown that the classifier correctly recognized 24 patients in abnormal decline and 149 normal patients. Considering *Decline* as the targeted class, our classifier mistakenly predicted 7 False Positives (FP) and 21 False Negatives (FN). We reproduced comparable performances with other machine learning algorithms using this set of features. A multilayer perceptron got a score of 84.6% Acc., a random forest 81.1% Acc. and a bagging algorithm 83.6% Acc. Five features only were selected by the CFS algorithm: Ngrams, Honoré's Statistic, Misspelling Ratio, Age and the Word2Vec Distance. This set of features differs from the set indicated by the ablation study but obtained better performances on our task. When trained and tested using only the four features which improved the classification during the ablation study, the score of the classifier reached 85.6 Acc. with 4 FP and 25 FN.

From these experiments we can conclude that our system showed promising performances when learning to discriminate subjects in abnormal cognitive decline from their writings. The ablation study and the set of optimal features found by the CFS algorithm seem to confirm the existence of the general pattern postulated in the clinical literature where lexical and semantical capacities are damaged during the cognitive decline. The most important features were the semantic features, Ngrams and Word2Vec, with a total drop of 2.7 points when they were removed. Both features capture the tendency of the subjects in decline to describe few topics of the image, resulting in a low Word2Vec distance, and to digress from the description task by mentioning several facts or statements that could not be inferred from the image or were not plausible with its content. These digressions caused the system to compute a higher probability for the description written by a subject in decline to be generated by the profile of the abnormal subjects and a low probability for being generated by the profile of the normal subjects. The profile of abnormal subjects contained more words than the profile of normal subjects, this latter containing only words related to the image.
The decline of the lexical capacities are suggested by the higher number of misspellings made by subjects in decline as well as the positive role of the Brunét's Index or Honoré's Statistic Brunet during the classification.

4.1 Analysis of Errors

The prediction of abnormal decline is a hard learning problem. Since it is still difficult to clinically diagnose the cognitive decline and potentially the following dementia, the labels of the target class in our corpus remains uncertain. Patients labeled normal can quickly show sign of decline and MCIs can recover over time. Therefore, for our analysis, we focused more on the capacity of our classifier to detect good descriptions rather than to strictly predict the target class. Additional analysis of our errors will be carried out by pathologists specialized in aphasia.

The 7 FPs where all primary diagnosed normal during their last visit. Their ages varied from 69 to 86 year old. Our manual inspection of their writings revealed that 4 descriptions presented strong irregularities which may explain the decision of our classifier. In the first case we found short descriptions containing misspellings, repeated phrases, ungrammatical sentences and descriptions focused on small details of the image. In the second case, descriptions were longer but they all contained digressions such as *"The turtle is shuffling back to be with the water."* (no turtle is drawn in the image), or *"Mom is torn between the playtime there and being being with her friends back home"* (the woman seems perfectly relaxed). Additional analysis of such digressions on our corpus are needed to know how strongly they are correlated with the decline. The reasons the classifier tagged the last 3 descriptions as *Decline* remained unclear. The Bayesian Networks learned for these instances are currently analyzed to understand which features deceived the classifier. The BN classifiers learned are probabilistic directed acyclic graphs which represent causal relations between variables. They can be displayed in a dedicated Graphical User Interface where values for different variables observed can be manually imposed to see the

changes on the likelihood of the others unseen variables.

The 21 FNs can be separated in 3 groups: 2 patients whose primary diagnosis were AD, 11 whose primary diagnosis were MCIs and 8 normal patients but whose cognitive exams results (3 patients) or global CDR (5 patients) showed signs of decline.

Our corpus contains in total 7 cases of patients diagnosed with AD, 5 cases were correctly classified by the system and 2 incorrectly, making it fairly sensitives to strong signs of decline. The majority of the classifier's errors were made on light and mild impairments. In order to understand these errors we randomly selected 10 descriptions written by these patients and proceeded to a manual examination. A clear difference with the descriptions of the FPs is the absence of digressions. Only one description mentioned some implausible facts, others strictly described the image with most of its topics commented. 6 descriptions presented anomalies like misspellings, phrases repeated, verbs/auxiliaries missing, incomplete sentences or wrong choices of pronouns and, for 2 of them, a simplified syntax with unnatural constructions (*e.g. "A coulle having a picnic, the man with a book the girl pouring a soda."*). The 4 remaining descriptions exhibit a good quality and would be difficult to discriminate with linguistic features only.

5 Conclusion and Perspectives

With the general aging of the population more attention has been given to Alzheimer's disease. In this study we presented a NLP system to predict early signs of cognitive decline, which precedes the disease, based on the analysis of written descriptions of an image. To perform our experiments we created a corpus which is, to the best of our knowledge, unique by its nature and its size. With a final score of 86.1% Accuracy our system outperformed our baseline system and showed state-of-the-art performances with existing classifiers working on oral interviews. Our results suggest a correlation between abnormal cognitive decline and the dislocation of the language ability. Our ablation study revealed that our system discriminates patients with abnormal decline using lexical and semantical irregularities found in their writings, consolidating the hypothesis of a general pattern in the linguistic impairment already postulated in the literature. The analysis of its classification errors showed the limitation of our approach: the presence of linguistic irregularities are not always sufficient to diagnose abnormal decline and may not always be observed in writings of patients already diagnosed in abnormal decline. To overcome this limitation we are currently designing a classifier based on Conditional Random Fields. This classifier will integrate all information available about our patients (i.e. medical, cognitive, linguistic, and imaging information) and will allow the representation of the performances of the patients over the time.

Acknowledgement

Research reported in this publication was partially supported by the NIH/NIA under grant P30 AG019610.

References

American Psychiatric Association, 1994. Diagnostic and statistical manual of mental disorders (4th ed.).

Elissa Asp and Jessica de Villiers. 2010. When Language Breaks Down: Analysing Discourse in Clinical Contexts. Cambridge University Press.

Igor A. Bolshakov and Alexander Gelbukh, editors. 2004. Computational Linguistics: Models, Resources, Applications. Igor A. Bolshakov and Alexander Gelbukh.

C. Brown, T. Snodgrass, S.J. Kemper, R. Herman, and M.A. Covington. 2008. Automatic measurement of propositional idea density from part-of-speech tagging. Behavior Research Methods, 40(2):540–545.

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.

T. Hayashi, H. Nomura, R. Mochizuki, A. Ohnuma, T. Kimpara, K. Suzuki, E. Mor. 2015. Writing Impairments in Japanese Patients with Mild Cognitive Impairment and with Mild Alzheimer's Disease. Dementia Geriatric Cognitive Disorders Extra, 5:309-319

C. Drummond, G. Coutinho, R. Paz Fonseca, N. Assuno, A. Teldeschi, R. de Oliveira-Souza, J. Moll, F. Tovar-Moll, and P. Mattos. 2015. Deficits in narrative discourse elicited by visual stimuli are already present in patients with mild cognitive impairment. Frontiers in Aging Neuroscience, 7(96).

Serge Gauthier, Barry Reisberg, Michael Zaudig, Ronald C Petersen, Karen Ritchie, Karl Broich, Sylvie Belleville, Henry Brodaty, David Bennett, Howard Chertkow, Jeffrey L Cummings, Mony de Leon, Howard Feldman, Mary Ganguli, Harald Hampel, Philip Scheltens, Mary C Tierney, Peter Whitehouse, and Bengt Winblad. 2006. Mild cognitive impairment. The Lancet, 367(9518):1262 – 1270.

Daniel B. Hier, Karen Hagenlocker, and Andrea G. Shindler. 1985. Language disintegration in dementia: Effects of etiology and severity. Brain and Language, 25(1):117–133.

Graeme Hirst and Vanessa Wei Feng. 2012. Changes in style in authors with alzheimer's disease. English Studies, 93(3):357–370.

D. Holmes and S. Singh. 1996. A stylometric analysis of conversational speech of aphasic patients. Literary and Linguistic Computing, 11:45–60.

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, Brain Informatics, volume 6334 of Lecture Notes in Computer Science, pages 299–307. Springer Berlin Heidelberg.

Vlado Keselj, Fuchun Peng, Nick Cercone, and Calvin Thomas. 2003. N-gram-based author profiles for authorship attribution. In Pacific Association for Computational Linguistics.

Daphne Koller and Nir Friedman. 2009. Probabilistic Graphical Models: Principles and Techniques. MIT press.

Linda E. Nicholas and Robert H. Brookshire. 1993. A system for quantifying the informativeness and efficiency of the connected speech of adults with aphasia. Journal of Speech, Language, and Hearing Research, 36(2):338–350.

M. Nicholas, L. Obler, M. Albert, and N. HelmEstabrooks. 1985. Empty speech in alzheimer's disease and fluent aphasia. Journal of Speech and Hearing Research, 28(3):405–410.

Loraine K. Obler and Susan de Santi, 2000. Methods for Studying Language Production, chapter 20. Lise Menn and Nan Bernstein Ratner, lawrence erlbaum associates edition.

Ronald C. Petersen, Rachelle Doody, Alexander Kurz,

Richard C. Mohs, John C. Morris, Peter V. Rabins, Karen Ritchie, Martin Rossor, Leon Thal, and Bengt Winblad. 2001. Current concepts in mild cognitive impairment. Archives of Neurology, 58(12):1985–1992.

Martin Prince, Emiliano Albanese, Maelenn Guerchet,¨ and Matthew Prina. 2014. World alzheimer report 2014. Alzheimer's Disease International (ADI).

Jamie Reilly, Joshua Troche, and Murray Grossman, 2011. Language Processing in Dementia, chapter 12, pages 336–368. Wiley-Blackwell.

S. R. Sabat. 1994. Language function in alzheimer's disease: a critical review of selected literature. Language and Communication, 14:331–351.

David A. Snowdon, Susan J. Kemper, James A. Mortimer, Lydia H. Greiner, David R. Wekstein, and William R. Markesbery. 1996. Linguistic ability in early life and cognitive function and alzheimer's disease in late life: Findings from the nun study. JAMA, 275(7):528–532.

Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. JASIST, 60(3):538–556.

Calvin Thomas, Vlado Keselj, Nick Cercone, Kenneth Rockwood, and Elissa Asp. 2005. Automatic detection and rating of dementia of alzheimer type through lexical analysis of spontaneous speech. In Proceedings of the IEEE International Conference on Mechatronics and Automation.

Ian H. Witten, Eibe Frank, and Mark A. Hall. 2011. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann.

Le Xuan, Ian Lancashire, Graeme Hirst, and Regina Jokel. 2011. Longitudinal detection of dementia through lexical and syntactic changes in writing: A case study of three british novelists. Literary and Linguistic Computing, 26(4):435–461.

JC. Morris and C. Ernesto and K. Schafer and M. Coats and S. Leon and M. Sano and LJ Thal and P. Woodbury. 1997. Clinical dementia rating training and reliability in multicenter studies: the Alzheimer's Disease Cooperative Study experience. Neurology 48(6):1508-1510.

S. Key-DeLyria. 2013. What are the methods for diagnosing MCI? Neurophysiology and Neurogenic Speech and Language Disorders, 23: 14-22.

Mark A. Hall. 1999. Correlation-based Feature Selection for Machine Learning. Ph.D. thesis, University of Waikato.