APPLYING NATURAL LANGUAGE PROCESSING AND MACHINE LEARNING TECHNIQUES TO AID IN THE DIAGNOSIS OF MILD COGNITIVE IMPAIRMENT AND EARLY DEMENTIA - A SYSTEMATIC REVIEW

A PREPRINT

Jomar Alcantara
Department of Computer Science
School of Engineering and Applied Sciences
Aston University
alcantaj@aston.ac.uk

Peter Sawyer
Department of Computer Science
School of Engineering and Applied Sciences
Aston University
p.sawyer@aston.ac.uk

George Vogiatzis
Department of Computer Science
School of Engineering and Applied Sciences
Aston University
g.vogiatzis@aston.ac.uk

Cristina Romani
Department of Psychology
School of Life and Health Sciences
Aston University
c.romani@aston.ac.uk

August 13, 2019

***K*eywords** MCI · Alzheimer's Disease · Machine Learning · Diagnosis

ABSTRACT

This is the paper's abstract . . .

1 Introduction

Dementia has been identified as one of those fast growing difficulties facing the world. A recent report suggests that in 2015 there were 46 million people with a diagnosis of dementia and that number is expected to hit 131.5 million by 2050 [1]. The report also states that the worldwide cost of dementia in 2018 is estimated to be in the region of one trillion US dollars.

A lot of work has gone into trying to find ways of improving the early diagnosis of Alzheimer's Disease (AD) and Mild Cognitive Impairment (MCI) with research focused on two areas - identifying biological markers and analyzing the cognitive decline of those who are suspected to have the disease [2]. As described above, the numbers of those suffering from AD and MCI are going to increase as the population ages [1] and thus it is important that we utilize technology wherever possible to aid clinicians in the detection of MCI and AD. At the present time diagnosis is typically conducted at memory clinics by trained clinicians [3]. I theorize that we may be able to enable an earlier diagnosis of those with MCI and AD using samples of spontaneous speech, natural language processing (NLP) and machine learning (ML).

There is a large body of research that looks at the decline in language in those with MCI and AD [2, 3]. However there is conflicting evidence in these studies about which declining language factors are associated of MCI and AD [2, 3]. Research therefore should look at these features in more detail and a clarification of this currently disorganised picture should go some way to helping researchers further understand the disease and it's progression. Another area of focus for research of this nature is the process of collecting appropriate language samples. Whilst collecting samples of language is comparatively unintrusive, researchers recognise that these samples require a rich sample of language that potentially cannot be generated by tasks such as the picture description task. Therefore, it would be useful to explore whether spontaneous discourse such a semi-structured interview, has the ability to put pressure on both the cognitive and linguistic systems in the same way as traditional cognitive tests such that it might be able to distinguish between healthy controls, those with MCI and those with AD. There is some evidence to support this. Berisha et al [4], has shown through a longitudinal language analysis of spontaneous speech that there are marked differences in this process between those who would go on to have a diagnosis of AD and a healthy control.

The question to be addressed in this systematic review is how has the field of machine learning and natural language processing addresses language deterioration in the diagnosis of Mild Cognitive Impairment and Early Alzheimer's Disease. The potential impact of this research in this area is immense. Research has shown that early diagnosis of people with AD or MCI improves sufferers quality of life and can, in some cases, slow the progress of the disease however the absence of a single test and the complexity of AD can create significant delays in diagnosis. Early diagnosis can increase the number of research opportunities for understanding the early stages of dementia and how the disease progresses so that more research can be conducted which may, in the future, lead to new treatments and other interventions.

The remainder of this article is organized as follows. Section 2 gives an account of the process of this Systematic Review. Our results are described in Section 3. We discuss the results and implications in Section 4. Finally, Section 5 gives the conclusions.

2 Methodology

A systematic literature review (SLR) describes a process which aims to identify, evaluate and interpret the research and literature in a given area. They are designed to provide a complete and exhaustive summary of the current evidence relevant to an identified research question. SLR's conduct a thorough search of all literature following a pre-defined protocol that specifies focused research questions, identifies criteria for the selection of studies and assessment of their quality, and forms to execute the data extraction and synthesis of results.

Common motivations for conducting an SLR are:

1. to summary all the evidence about a topic.
2. find gaps in the research.
3. to provide a ground for a fundament to new research.
4. and to examine how the current research supports a hypothesis.

Performing an SLR comprises the following steps:

1. identify the need for performing the SLR.
2. formulate research questions.
3. execute a comprehensive search and selection of primary studies.
4. assess the quality and extract data from the studies.
5. interpret the results.
6. report the SLR.

2.1 Search strategy

The main research question this SLR aims to address is: "How has the field of machine learning and natural language processing addressed language deterioration in the diagnosis of Mild Cognitive Impairment and Early Alzheimer's Disease?".

This SLR builds upon these questions and additionally presents the results of the other two additional research questions. Further, the key terms related to MCI and the other names by which it is known were included in the search string to ensure that relevant studies about machine learning and the diagnosis of a disease were also retrieved, even if not specifically mentioned in the paper's title or abstract.

To address the research questions, a search string was defined using the PICO approach, which decomposes the main research question into four parts:

1. Population - Studies that present research on mild cognitive impairment and dementia. Mild Cognitive Impairment and Dementia keywords were selected from the Systematized Nomenclature of Medicine–Clinical Terms (SNOMED-CT).
2. Intervention - Intervention: ML or MS techniques. The ML keywords were selected from the branch "Machine Learning Approaches" of the "2012 ACM Computing Classification System". The MS keywords were selected by A2.
3. Comparison
4. Outcome - Outcome: Prognosis on dementia and comorbidies. The prognosis keywords were provided by A4.

Using this process, the main research question was decomposed into four research questions:

- RQ1: Which NLP and ML techniques are being used in dementia research?
- RQ2: What features / data characteristics of text (variables, determinants and indicators) that are considered when applying the ML techniques (n-grams, PoS Tagging etc)?
- RQ3: What are the goals of the studies that employ NLP / ML techniques for prognosis of dementia?
- RQ4: Do the studies focus on time as factor?

The automated searches were performed in the Pubmed, Web of Science, Scopus and IEEE databases. Table 1 shows the search string used for the Pubmed automated search, but note that this search string was adapted to each of the other databases' search context.

(dementia OR MCI OR Mild Cognitive Impairment OR Alzheimer's OR Mild Neurocognitive Disorder OR AD) AND TOPIC: (machine learning OR Data Mining OR Decision Support System OR NLP OR Natural Language Processing) AND TOPIC: (prognosis OR prognostic estimate OR predictor OR prediction OR model OR patterns OR diagnosis OR diagnostic OR forecasting OR projection OR Deep Language Model OR Deep Neural Network) AND TOPIC: (classification OR regression OR kernel OR support vector machines OR Gaussian Process OR Bayesian Network OR Factor Analysis OR Deep Learning OR Neural Networks OR Maximum Likelihood OR Principal Component Analysis OR Markov OR Linear Model OR Mixture Model OR Perceptron Algorithm OR Logical Learning OR relational learning OR Supervised Learning OR Unsupervised Learning OR clustering OR Decision Tree) AND TOPIC: (Language OR Cognitive OR Speech OR Conversation OR Connected Speech OR Picture Description OR Discourse Analysis OR Verbal Fluency)
Searched - 4th April 2019 - Generated 1257 Articles

Table 1: Example of Search Terms for Web of Science database

Scopus	Web of Science	PubMed	IEEE Xplore
1002	991	376	230

Table 2: Number of articles found from each database - complete search terms appear in Appendix B

2.2 Study selection

A total of 1490 unique papers were identified through the searches conducted above. Each paper was initially reviewed by just the title and the abstract. This yielded a total of 25 potential papers. Papers were then evaluated by the JA based on the inclusion and exclusion criteria (see Table 3). Where JA could not reach a decision, PS and GV were consulted and majority vote on inclusion was conducted.

Inclusion Criteria	Exclusion Criteria
Be a primary study in English; AND address research on dementia and comorbidities; AND address at least one ML or MS technique; AND address a prognosis related to dementia and comorbidities; AND use cognition or language decline as a factor for analysis.	Be a secondary or tertiary study; OR be written in another language other than English; OR do not address a research on dementia and comorbidities; OR do not address at least one ML or MS technique; OR do not address a prognosis related to dementia and comorbidities.

Table 3: Inclusion and Exclusion Criteria

After the identification of these papers, a one-iteration backward snowballing process was carried out using the reference lists of the original set of papers looking for studies that were missed in the original searches. This resulted in 41 additionally identified papers. Throughout the whole selection process, PS and GV acted as additional assessors in the case where there was uncertainty about whether a paper should be included.

In total, 66 papers were selected to be fully read. A quality assessment questionnaire (see Table 4) was developed based on Kitchenham's guidelines and was used to minimize the chance of bias in the selection process. Studies were graded on a 12 point scale and any studies which scored less than 8 points were excluded for quality reasons. The ones that successfully passed the filtering criteria described earlier had their relevant data extracted.

Variable	Definition
Conditions Studied	For which dementia disorder is the study deriving a prognosis.
Database used in the study	Name and origin of the data source used to derive the prognosis of the studied dementia.
Dataset Categories	Classes in which the data units were divided into.
Follow-up period	Period of time, which the data units were followed.
Techniques used	Natural Language techniques AND/OR ML techniques that were used to build the diagnostic models.
Features generated	NLP generated features used in building the diagnostic models.
Aim of the Study	The goal of the built diagnostic models.

Table 4: Quality Assessment Questionnaire

In this phase, a paper could also be rejected due to inclusion and exclusion criteria because the selection process adopted an inclusive approach. This means that during the reading of the titles and abstracts, in the case where the information provided was incomplete or too general it was selected to be fully read in the posterior phase. A common example is the case when the data analysis technique specified in the abstract was merely "classification", so it was not possible to know if any machine learning occurred.

In total, 37 studies composed the final set of included primary studies and had their relevant data extracted, 7 papers were rejected due quality reasons, and 34 papers were rejected due to failing the inclusion and exclusion criteria. One reason for the high number of the latter was the decision to exclude the papers that used solely statistical methods as data analysis techniques to build the prognostic models. The selected studies were also assessed for the risk of

cumulative evidence bias. This was done by checking, in the case of the same research group with different studies in the final set of included primary studies, if it was justified having both studies (i.e different samples).

2.3 Data collection

For the data collection, a base extraction form was defined in the protocol, but later in the study it was evolved based on the research group discussions. Table 4 lists and defines the collected variables.

In addition to these variables other basic data about the studies was collected, these were: title, authors, journal/source, year and type of publication. No summary measures were used. Summary tables were used for the synthesis of results and no additional analyses were carried out

3 Results

3.1 Machine Learning methods

Intro - Addresses Research Question 1

Machine Learning technique used	Paper Number
Logistic Regression	Orimaye
Support Vector Machines	Orimaye
Linear Discriminant Analysis	Orimaye
Decision Trees	Orimaye

3.2 Traditional Machine Learning methods

3.2.1 Logistic Regression

Logistic Regressions is a probabilistic approach to classification and is dependant on the assumption that your input space can be separated into two distinct regions, one for each class, by a linear boundary. In 2D space, this boundary is a line and in 3D space, this boundary is defined as a plane.

In the results of this survey.

3.2.2 Support Vector Machines

Support Vector Machine (SVM) was originally proposed as an algorithm for classification problems; it is a relatively new technique compared to the other ML approaches. The classification process consists of mapping the data points (usually the study subjects) into a feature space composed of the variables that characterize these data points, except for the outcome variable. Then, the algorithm finds patterns in this feature space by defining the maximum separation between two or more classes, depending on the problem to be solved. Contrary to some regression techniques, SVMs are not dependent on a pre-determined model for data fitting, although there are still algorithm specifications to be considered (e.g. choice of a kernel function); instead, it is a data-driven algorithm that can work relatively well in a scenario where sample sizes are small compared to the number of variables, reason why it has been widely employed by diagnostic studies in tasks related to the automated classification of diseases.

Regarding the SLR results, SVMs were present in ??/41 selected studies, in 38 proposed models, and being by far the most used machine learning technique in the dementia diagnosis research. These numbers account for the traditional SVM and variations. In all of the ?? selected studies the SVMs focused at binary classifications where the task was to discriminate mild cognitively impaired (MCI) patients that will or will not develop Alzheimer's Disease (AD). In the general case, the problem is posed as either MCI converters versus MCI non-converters, or progressive MCI versus stable MCI classification. This outlines a situation in which a regression problem (when will the MCI patients convert to AD?) is formulated as a classification problem (which MCI patients will convert to AD in X months?) to be solved. Reasons for this could be due to limitations in the data used, i.e. the limited follow-up periods of the subjects included in the studies.

3.2.3 Linear Discriminant Analysis

3.2.4 Decision Trees

A Decision Tree (DT) is a classification algorithm in which the learned knowledge is represented in a tree structure that can be translated to if-then rules. DT's learning process is recursive and starts by testing each input variable as to how well each of them, alone, can classify the labeled examples. The best one is selected as a root node for the tree and its descendant nodes are defined as the possible values (or relevant ratios) of the selected input variable. The training set is then classified between the descendant nodes according to the values of the selected input variable. This process is repeated recursively until no more splits in the tree are possible. Like SVMs, DTs do not depend on a pre-defined model and are mostly used to find important interactions between variables. Being intuitive and easy to interpret, DTs have been used in prognostic studies as a tool for determining prognostic subgroups. In this SLR, DTs were the second most frequently used ML technique, present in 6/37 selected studies and proposed in 7 models. It was employed for the same reason as SVM; except for one study that investigated the evolution of patients diagnosed with cognitive impairment no dementia (CIND) to AD.

Curry, Singer and Habash - Alz (66.7 - Pres, 66.7 - Recall) and Control (67.9 - Pres, 67.9 Recall)

3.2.5 Naive Bayes

Curry, Singer and Habash - Alz(80.8 - Pres, 0.75 - Recall) and Control (79.3 - Pres, 82.1 Recall) - In the naive bayes classifier, they identified that pauses, go-ahead, fillers and incomplete words as the most significant features.

3.2.6 Artificial Neural Networks

An Artificial Neural Network (ANN) is a methodology that performs multifactorial analyses, which is desirable in the health area as medical decision-making problems are usually dependent of many factors. An ANN is composed of nodes connected by weighted edges in a multi-layer architecture that comprises: an input layer, one or more hidden layers and an output layer. In the training process, inputs and outputs values are known to the network, while the weights are incrementally adjusted so that the outputs of the network are approximate to the known outputs. Despite being a powerful predictor, ANNs are 'black boxes', which means that they are not able to explain their predictions in an intuitive way, contrary to DTs or BNs. Also, they require the specification of the architecture to be used beforehand (i.e. the number of hidden layers).

3.2.7 K-nearest neighbours

K Nearest Neighbors (KNN) is a classification algorithm that takes a data point from an unknown class and assigns it as an input vector in the feature space. Then, the classification process follows by assigning the unknown class data point to the class in which the majority of the K nearest data points belong to. The distance between data points is usually measured by Euclidean distance, but it is possible to employ other measures. KNN is one of the simplest ML classification algorithms and have been used in a wide range of applications; however, it can be computationally expensive in a highly dimensional scenario. Further, it considers all features to be equally weighted, which can be a problem if the data has superfluous attributes.

3.3 Deep Learning methods

3.3.1 Deep language space neural network for classifying mild cognitive impairment and Alzheimer-type dementia - Orimaye, Wong and Wong (2018)

In this paper, Orimaye et al use deep-deep neural networks language models (D2NNLM) to learn linguistic changes that distinguish the language of patients with MCI and AD-type dementia from the healthy controls using higher order n-grams. An ordinary DNNLM uses lower order n-gram N-dimensional sparse vectors as descrete feature representations to train the neural network with multiple hidden layers.

3.4 Features of Language

Intro - Addresses Research Question 2

Asgari, Kaye and Dodge (2017) [**?**] used another form of word frequency measurement. Using recordings of unstructured conversations (with standardized preselected topics across subjects) between interviewers and interviewees they grouped spoken words using Linguistic Inquiry and Word Count (LIWC) which is a technique used to categorize words into

features such as negative and positive words [**?**]. They were able to successfully used machine learning algorithms to distinguish between these two groups with an accuracy of 84%.

3.4.1 Measures of Semantic Complexity

Intro!

Type token ratio (TTR) is the ratio obtained by dividing the types (The total number of different words) by the tokens (the total number of words in an utterance).

$$TTR = numberOfUniqueWords/totalNumberOfWords. \tag{1}$$

Brunet's Index (W) differentiates itself for TTR, as it is not impacted by the length of the text itself. Brunet's Index is defined by the following equation:

$$W = N^{V(-0.165)} \tag{2}$$

where N is the total length of the utterance being measured and V is equal to the total vocabulary being used by the subject. Brunet's Index usually has a score of between 10 and 20, with high numbers indicating a more rich vocabulary compared to low numbers.

Honore's Statistic is based on the idea that vocabulary richness is implied when a speaker uses a greater amount of unique words. This is indicated by the following equation:

$$R = (100 \log N)/(1 - V1/V) \tag{3}$$

where v1 is equal to the number of unique words, V is the total vocabulary used and N is the total number of words in the utterance being measured.

Guinn, Singer and Habash [5], they found in their corpus that there was no significant difference between interviewers and those with dementia when applying these measures. However, when they compared these results with a control dataset, they did find a significant difference in Honore's statistic. They explained this interesting results by suggesting that the interviewer used was trying to match (intentionally or unintentionally) the lexical richness of the person they were interviewing. In comparing healthy older adults with those suffering from dementia, they found that there was an increase in lexical diversity which is contrary to other research. They did note that conversations involving those with dementia contained roughtly 50% less speech than those with controls, and that TTR in particular is not a suitable measure because it is more sensitive to length. They also note that Brunet's Index and Honore's Statistic are better statistics that control for the total length of the conversation and controls had statistically greater lexical diversity on those measures.

3.5 Quantity - Total number of words

Several studies report that adults with moderate AD produce fewer words than controls on picture description, however other studies found no differences in total words among groups of controls and patients with MCI or AD. Murray and Nicholas et al investigated normal controls, patients with AD and older adults with depression and found no group differences in total words. In contrast, Lira 2014 found that controls produced more total words than patients with AD but found no difference between mild and moderate groups.

3.6 Syntax and Morphology (Language Form)

Syntax can be defined as the rules that govern how words can be combined to form sentences, whilst Morphology is the system that governs the structure of words and the construction of word forms. Multiple studies of language decline in dementia included at least one measure of syntax or syntactic complexity. Common constructs included words per clause, grammatical form (measures of an appropriate use of syntactic conjections, tenses, conditionals, subordinate clauses and passive constructions), measures of phrase length and proportions of words in sentences. Some researchers have explored the use of formulaic language in those with dementia, the theory being that well practiced phrases are less effortful and therefore place low load on the cognitive abilities of those with AD. The general hypothesis motivating these studies is that either working memory limitations or semantic memory limitations in AD affect one's ability to use complex constructions.

3.7 N-grams and skip-grams

One of the first features discussed as a potential predictor of MCI or AD is the n-gram. An n-gram is a contiguous sequence of n items from a given sample of text or speech. The items can be phonemes, syllables, letters, words or base pairs according to the application. For example, given the sequence of words "to be or not to be", this extract is said to contain six 1-gram sequences (to, be, or, not, to, be), five 2-gram sequences (to be, be or, or not, not to, to be), four 3-gram sequences(to be or, be or not, or not to, not to be) and so on. This is useful as, given a large portion of text or speech, we can predict the probability of a word being close by to a given word. A number of researchers have used n-grams as features. One of the first attempts to use machine learning and natural language techniques to look was conducted by Thomas [**?**] who was able to successfully demonstrate the ability of machine learning algorithms to analyse n-grams as well as other features to outperform a naive rule-based classifier which always selects the modal class. Orimaye et al (2017) [**?**] investigated the use of machine learning algorithms to detect differences primarily in n-gram use to distinguish between those with a diagnosis of AD and healthy controls. Their main finding supported n-grams as the most significant predictor. One of the criticisms is the use of picture description tasks and n-grams. Because the language generated by this task is content specific the n-grams generated are only specific to the task given and cannot be generalised.

Skip-grams are a variant of n-grams in which word tokens are skipped intermittently while creating n-grams. For example, take the sentence 'I am going to London', there are four conventional bigrams: 'I am', 'am going', 'going to', and 'to London' - using skip-grams, we might skip a word to create additional bigrams such as: 'I going' and 'going London'. Orimaye defined k-skip-n-grams as a set of n-gram tokens with the following equation, where n is the specified n-gram (e.g. 2 for a bigram and 3 for a trigram), m is the number of tokens in a given sentence, k is the number of word skip between n-grams given that k < m and a = {1,..., m-n}

One problem with this approach is existing research currently uses language generated from picture description tasks. Given the nature of these tasks, the language generated is relatively constrained in comparison to language generated spontaneously.

$$T_{n-gram} = W_a, ..., W_{a+n-k}, ..., W_{a+n}, ..., W_{m-n}, ..., W_{(m-n)+n-k}, ..., W_m \tag{4}$$

Thus for the sentence 'I am going to London', 1-skip-2-grams will give {I going, am to, going London} and 1-skip-3-grams will give {I going to, I am to, am to London, am going London}

3.8 Mean length of utterance (MLU)

Murray found that MLU was not a distinguishing factor among health adults, adults with depression and adults with AD. Ripich et al found a decrease in MLU in adults with severe AD over time, and this was supported by findings of Le et al in their studies of authors [**?**]

3.9 Proportion of verbs to nouns plus verbs

Kave and Levy used a verb index to capture syntactic complexity and found that adults with AD expressed the same amount of verbs to nouns plus verbs as adult controls.

3.10 Syntactic Complexity - Composite measures of MLU, syntactic errors and verbs

Ahmed et al, and Ahmed, Haigh et al found differences in syntactic complexity between adults with MCI and controls, and between MCI and moderate AD stages. The differences in syntactic complexity were not significant when individual measures were tested, but were apparent using a composite score consisting of MLU, words in sentences, syntactic errors, nouns with determiners, and verbs with inflections.

Lu's Syntactic Complexity Analyser Ygnve measure

3.11 Semantic features

. We compute semantic similarity using the average and minimum cosine distance between each pair of one-hot embeddings of utterances, and the cosine cutoff (i.e., the number of pairs of utterances whose the cosine distance is

below a certain threshold). We compute word specificity and ambiguity based on tree depth and the number of senses in WordNet [54]. We also extract multiple WordNet measures of similarity: Resnik [68], Jiang-Coranth [69], Lin [47], Leacock-Chodorow [70], and Wu-Palmer [71].

3.12 Syntactic features

Guinn, Singer and Habash [5] found that syntactic features such as Noun rate, verb rate, adjective rate and pronoun rate were not significantly different between interviewers and those with dementia in their dataset, although they did note that there was a slightly higher but non significant use of pronouns.

3.13 Pragmatic features.

We train a general 100-topic latent Dirichlet allocation (LDA) model [72] on the Wikipedia corpus for generalizability. LDA is a generative statistical model used to determine unlabeled topics in a document. For each transcript, we extract the probabilities of each LDA topic. Next, we extract features related to rhetorical structure theory (RST), which is a classic framework for discourse parsing in which partitions of text are arranged in a tree structure by pragmatic relations such as Elaboration or Contrast [73].

3.13.1 Go-ahead Utterances

Go-ahead utterances are defined as short one or two syllable responses that do not contribute to the conversation beyond a minimal response. The function of these go-ahead utterances can be to validate what the other person is saying, or to agree / disagree with what is being said. Another function can be that they wish speaker is indicating that they have nothing futher to add to the conversation and a signal to the other speaker to continue with what they are saying. In Curry, Singer and Habash's research, they found that in comparing interviewers and those with dementia, that interviewers used significantly fewer go-ahead utterances. In comparing controls and those with dementia, they found that there was a relative lack of go-ahead utterances in the controls which implies that controls had a lot more contributions to make in their conversations than those with dementia.

3.14 Formulaic Language

Fraser, Meltzer and Rudzicz (2015) [**?**] looked at connected speech using the DementiaBank corpus. They found that there were four factors which informed the classification of participants as either healthy or AD. These four factors were semantic impairment, acoustic abnormality, syntactic impairment and information impairment and were based on existing measures of semantic and syntactic complexity. Zimmerer (2016) [**?**] looked at whether language was more formulaic in those suffering from AD. He proposed that those who suffer from AD rely on formulatic sentences, for example 'Noun-Verb-Noun', and this is done to reduce language complexity. He noticed a significant difference in the use of formulaic sentences between AD and Healthy Controls.

3.15 Number of syllables and Characters

3.16 Number of fillers

3.17 Readability

Flesch reading score, Flesch-kincaid grade level

3.18 Polarity

3.19 Frequency

Mean values of frequency, age of acquisition, imageability, familiarity, arousal, dominance and valence based on lexical norms

3.20 Dysfluencies

Curry, Singer and Habash noted that in comparing controls and those with Dementia, that those with Dementia had a significantly higher number of pauses per word and a much higher incidence of words that were truncated in mid-speech.

In comparing interviewers with those with dementia, they also showed other signs of difficulties with fluency with higher rates of incomplete words, filler words and repeated words.

3.21 What are the goals of the studies that employ ML or MS techniques for prognosis of dementia and comorbidities?

4 Discussion and conclusions

4.1 Discussion of the current evidence

One of the criticisms of traditional learning models is it's reliance on features that are generated for them. As we can see from the brief look at some of the research in this field, researchers have taken a number of different approaches with success. However, there does not seem to be a consensus. As machine learning in general moves away from traditional machine learning models to deep learning techniques, it poses the question can deep learning assist with this problem?

One of the main benefits of deep learning is that it does not rely on pre defined features being fed to the model. Instead, deep learning models take raw data in with some amount of preprocessing and generates it's own features. This move away from relying on features solves one of the difficulties we have with the current psychological literature, namely that there are some disagreement on the which

One of the drawbacks of attempting to use natural language processing and machine learning in this context is the lack of data. This is the case with all machine learning techniques, but more so with deep learning. There have been ways to 'create' more data such as data augmentation.

Current state-of-the-art diagnostic measures of AD are invasive (CSF analysis), expensive (neuroimaging), and time-consuming (neuropsychological assessment). Furthermore, these measures are limited to speciality clinics and thus have limited accessibility as frontline screening and diagnostic tools for AD. More importantly, nonspecialists are often inaccurate at identifying early AD and MCI. Thus, there is an increasing need for additional noninvasive and/or cost-effective tools, allowing effective frontline identification of subjects in the preclinical or early clinical stages of AD who could be suitable for monitoring in speciality clinics and for early treatment. Implementation of effective screening instruments will allow diagnosis earlier in the course of dementia, even at the point when memory function is still essentially within the normal range. This strategy would enable an earlier, and potentially more effective, prevention and treatment of AD with a special focus to preserve cognitive functions.

However the literature has identified a number of challenges when approaching this problem. Firstly, the clinical features that combine to meet the diagnostic criteria for Dementia or it's variants are continuous in nature and heterogenous between patients and are also impacted by other variables. For example as shown above, cognitive performance is affected in part by a patients educational attainment and a patients ability to live independently is impacted by a their physical health as much as their cognitive health. The challenge therefore is to find features that are minimally impacted by other factors, or that can be controlled for by a experimental design such as a matched pairs design to control for educational attainment.

Another challenge lies with the recruitment of suitable individuals who may notice a decline in cognition to the point where we might classify them as having MCI, but these individuals deduce that there is little to no value to admitting there is a problem and seeking help whilst their symptoms are 'manageable'.

Finally there is a large amount of variability in the presentations of those with MCI and early dementia, and this is compounded by an similar amount of variability in the criteria researchers have used for experimental groups and the approaches researchers have used to tackle this problem. This had led to a confused literature. Recently a call for research that has consistent inclusion / exclusion criteria has been made along with some proposed definitions of MCI and it's subgroups [**?**]. Researchers have identified the analysis of language impairment as an area of promise to explore in the diagnosis of MCI and early AD and recent developments in natural language processing and machine learning techniques have the potential to assist in this research. Indeed, given the increased burden on the diagnosis of MCI and AD on professionals there has been a call to use technology to potentially ease this burden [3]. A small but growing amount of research has gone into the use of machine learning techniques to potentially look at the automated classification of participants with MCI and/or AD, however this is a new area of research and there are some gaps in our knowledge.

4.2 Methodological Issues

4.3 Limitations

4.4 The future of the field

One of the areas for research to study is a careful examination of the features that are being used to measure language deterioration. For example, Zimmerer (2016) [**?**] describes connectivity in such as a way that correlates directly with what Mueller (2018) [**?**] calls Fluency. Whilst these are very nuanced measures which differ slightly in the form they take, the sheer range of measures and features being produced make it difficult to organise and explore what is truly going on in those with MCI and early AD.. Some work needs to be done in producing a consistent list of measures that are validated using existing datasets and can be used for future research moving forward.

Teng et al suggests that work should focus on the MCI population and concentrate on developing a consensus neuropsychological battery that could yield predictable rates of progression to AD [**?**]. This, in conjunction with the development of a model of language, sensitive enough to detect subtle deterioration in language use to act as an additional cognitive marker to aid diagnosis could potentially move some way to providing this.

Future research should also be directed towards developing non-intrusive ways of detecting subtle changes in natural language such that any perceived deterioration that could indicate the presence of MCI or AD could be flagged up early. Machine learning approaches seem to be the most logical approach for achieving this aim as language could be collected in non-intrusive ways and passed to a machine learning algorithm for preliminary classification. Despite the excellent quality of datasets, for example the DementiaBank dataset, being used to 'backtest' these algorithms, further research should look at generating additional datasets to increase the validity of the results found so far as well as using other methods to generate data other than Picture Description tasks which some researchers could claim are limited in scope. Finally, the recent resurgence in the use of neural networks and deep learning could provide the answer to the confused literature in terms of features. A key benefit of deep learning is it's ability to automate the process of feature engineering. So there is an opportunity to explore the use of deep learning, to not only develop new features but also validate existing features independently.

This area of research is extremely promising in its early results and the impact of successful research would be life changing for both individuals and the health of the worlds aging population in general.

5 Conclusions

We can see that both Natural Language Processing and Machine Learning techniques have a lot to offer, Indeed there has been a lot of research which have used pre-existing datasets to explore this area with promising results. One of the difficulties with the current research is the approach of trying to discriminate between those AD and healthy controls. This is not necessarily a problem in the real world as it is trivially easy to do for trained clinicians. A more interesting, but potentially harder to problem to solve is to discriminate between those with MCI and healthy controls, and more importantly to track their decline over time. Further, to date the authors are unaware of any research in which these techniques are applied to newly created samples of language.

References

[1] Martin Prince, Anders Wimo, Guerchet M, Ali GC, Wu YT, and Prina M. World Alzheimer Report 2015 The Global Impact of Dementia An analysis of prevalence, incidence, cost and trends. *Alzheimer's Disease International*, 2015.

[2] Vanessa Taler and Natalie A. Phillips. Language performance in Alzheimer's disease and mild cognitive impairment: A comparative review. *Journal of Clinical and Experimental Neuropsychology*, 30(5):501–556, 2008.

[3] Veronica Boschi, Eleonora Catricalà, Monica Consonni, Cristiano Chesi, Andrea Moro, and Stefano F. Cappa. Connected speech in neurodegenerative language disorders: A review. *Frontiers in Psychology*, 8(MAR), 2017.

[4] Visar Berisha, Shuai Wang, Amy LaCross, and Julie Liss. Tracking Discourse Complexity Preceding Alzheimer's Disease Diagnosis: A Case Study Comparing the Press Conferences of Presidents Ronald Reagan and George Herbert Walker Bush. *Journal of Alzheimer's Disease*, 45(3):959–963, 2015.

[5] Curry Guinn, Ben Singer, and Anthony Habash. A comparison of syntax, semantics, and pragmatics in spoken language among residents with Alzheimer's disease in managed-care facilities. *IEEE SSCI 2014 - 2014 IEEE*

Symposium Series on Computational Intelligence - CICARE 2014: 2014 IEEE Symposium on Computational Intelligence in Healthcare and e-Health, Proceedings, (January):98–103, 2015.

A Article Table

Paper Number	Paper Title	Authors
1	A Comparison of Syntax, Semantics, and Pragmatics in Spoken Language among Residents with Alzheimer's Disease in Managed-Care Facilities	Authors
2	Abnormalities of Connected Speech in Semantic Dementia vs Alzheimer's diseases	Authors
3	Aided Diagnosis of Dementia Type through Computer-Based Analysis of Spontaneous Speech	Authors
4	Alignment of spoken narratives for automated neuropsychological assessment	Authors
5	Automatic detection and rating of dementia of alzheimer type through lexical analysis of spontaneous speech	Authors
6	Changes in Style in Authors with Alzheimer's Disease	Authors
7	Computer-based evaluation of Alzheimer's disease and mild cognitive impairment patients during a picture description task	Authors
8	Computerized assessment of syntactic complexity in Alzheimer's disease: A case study of Iris Murdoch's writing	Authors
9	Computerized neuropsychological assessment in mild cognitive impairment based on natural language processing-oriented feature extraction	Authors
10	Contrastive conversational analysis of language production by Alzheimer's and control people	Authors
11	Declines in Connected Language Are Associated with Very Early Mild Cognitive Impairment: Results from the Wisconsin Registry for Alzheimer's Prevention	Authors
12	Deep-Deep Neural Network Language Models for Predicting Mild Cognitive Impairment	Authors
13	Domain Adaptation for Detecting Mild Cognitive Impairment	Authors
14	Diagnosing people with dementia using automatic Conversation Analysis	Authors
15	Deep language space neural network for classifying mild cognitive impairment and Alzheimer-type dementia	Authors
16	Enriching Complex Networks withWord Embeddings for Detecting Mild Cognitive Impairment from Speech Transcripts	Authors

17	Features and Machine Learning Classification of Connected Speech Samples from Patients with Autopsy Proven Alzheimer's Disease with and without Additional Vascular Pathology	Authors
18	Formulaic Language in People with Probable Alzheimer's Disease: A Frequency-Based Approach	Authors
19	Language Analysis of Speakers with Dementia of the Alzheimer's Type	Authors
20	Language analytics for assessing brain health: Cognitive impairment, depression and pre-symptomatic Alzheimer's disease	Authors
21	Learning Linguistic Biomarkers for Predicting Mild Cognitive Impairment using Compound Skip-grams	Authors
22	Learning Predictive Linguistic Features for Alzheimer's Disease and related Dementias using Verbal Utterances	Authors
23	Linguistic Features Identify Alzheimer's Disease in Narrative Speech	Authors
24	Longitudinal detection of dementia through lexical and syntactic changes in writing: a case study of three British novelists	Authors
25	Microlinguistic aspects of the oral narrative in patients with Alzheimer's disease	Authors
26	Multilingual word embeddings for the assessment of narrative speech in mild cognitive impairment	Authors
27	Natural Language Features for Detection of Alzheimer's Disease in Conversational Speech	Authors
28	NLP-Oriented Contrastive Study of Linguistic Productions of Alzheimer's and Control People	Authors
29	Predicting mild cognitive impairment from spontaneous spoken utterances	Authors
30	Predicting probable Alzheimer's disease using linguistic deficits and biomarkers	Authors
31	Propositional Idea Density in women's written language over the lifespan: Computerized analysis	Authors
32	Sentence segmentation in narrative transcripts from neuropsychological tests using recurrent convolutional neural networks	Authors
33	Speech processing approach for diagnosing dementia in an early stage	Authors
34	Spoken language biomarkers for detecting cognitive impairment	Authors
35	Talk2Me: Automated linguistic data collection for personal assessment	Authors
36	The Effect of Heterogeneous Data for Alzheimer's Disease Detection from Speech	Authors
37	Toward the Automation of Diagnostic Conversation Analysis in Patients with Memory Complaintss	Authors

39	Tracking Discourse Complexity Preceding Alzheimer's Disease Diagnosis: A Case Study Comparing the Press Conferences of Presidents Ronald Reagan and George Herbert Walker Bush	Authors
40	Vector-space topic models for detecting Alzheimer's disease	Authors
41	Vocabulary Size in Speech May Be an Early Indicator of Cognitive Impairment.	Authors