How can machine learning and natural language processing techniques aid the process of diagnosing Mild Cognitive Impairment and Alzheimer's Disease?

Jomar Alcantara

2018-07-13

Contents

1 Introduction

Alzheimer's Disease (AD) is a neurodegenerative disease in which the brain develops neurofibrillary tangles and neuritic plaques along with the loss of cortical neurons and synapses. The hallmark clinical symptoms are cognitive deficits such as problems with episodic and semantic memory, organising and planning, difficulties with language and visuospatial deficits[1]. In addition, these symptoms are often accompanied by emotional difficulties such as depression and irritability and behavioural difficulties. Whilst there are several variants of Dementia, AD remains the most common type of Dementia and will be the focus of this review.

According to a recent report commissioned by the Alzheimer's Society in 2015, they estimate the prevalence of AD in the UK at approximately 815,000 people. This represents 1 in 14 of the population aged 65 or over and 1 in 79 in the general population. They estimate an annual healthcare spend on 4.3 billion of which approximately 85 million is spend solely on diagnosis and that the total impact of AD (excluding the costs associated with early onset dementia) is 26.3 billion annually. Globally, this picture is a lot bleaker. A recent report suggests that in 2015 there were 46 million people with a diagnosis of AD and that number is expected to hit 131.5 million by 2050 [2]. The report also states that the worldwide cost of AD in 2018 is estimated to be in the region of one trillon US dollars.

Current thinking suggests that the cognitive deficits associated with AD often begin before the clinical symptoms of the disease become apparent. Researchers propose that neurofibrillary tangles and other associated physiological effects of AD develop over time until a threshold is reached and clinical symptoms become apparent [3]. If this is the case, it should be possible to detect subtle cognitive changes in language and memory before a clinical diagnosis can be formed. Given that language is less intrusive to test and requires a lot of the cognitive processes that may be impacted by AD, a lot of research has focused on the decline in the use of language in those with AD.

One of the most famous pieces of research was by Berisha and Liss (2015) [4] who examined speeches and public interviews of former US president Ronald Reagan. They found that Reagan's speeches towards the end of his presidency suffered from difficulties in word-finding, inappropriate phrases and uncorrected sentences which are hallmarks of language deterioration associated with Alzheimer's Disease. Another classical study by Snowdon et al (1996) [5] looked at whether linguistic ability in early life was asoociated with cognitive function and AD in later life. They found that idea density (defined as the number of expressed propositions divided by the number of words) was a key predictor in predicting whether nuns would go on to develop AD in later life. They found that those who would go on to develop AD all had low idea density in early life and they found no AD present in those with high idea density.

The range of language deficits in those who suffer with AD are wide and varied and differ as the disease develops. According to the DSM 5 [1], those with Mild dementia suffer from noticeable word finding difficulty. They may

substitute general terms for more specific terms and may avoid the use of specific names of acquaintances. There may be grammatical errors involving subtle omission or incorrect use of articles, prepositions, auxiliary verbs, etc. Those who have progressed from Mild to Major depression also have difficulties with expressive or receptive language. They will often use general-use phrases such as "that thing" and "you know what I mean" and prefers general pronouns rather than names. With severe impairment, sufferers may not even recall names of closer friends and family. Idiosyncratic word usage, grammatical errors, and spontaneity of output and economy of utterances occur. Echolalia (meaningless repetition of another person's spoken words) and automatic speech typically precede mutism. With the wide range of deficits someone with AD can suffer, it makes sense to try to categorise these deficits in some way.

Emery [6] completed a literature review in this area and she looked at four levels of language: Phonology, Morphology, Syntax and Semantics. Her review looked at the idea that language and the processes involved in language are hierarchical. She proposes that language goes from simple units of construction, and build layers of complexity and sophistication. She found that people with AD generally had intact Phonology and Morphology but more impaired Syntax and Semantics. Emery concludes that language decline is hierarchical and is related to the complexity of the language task given to a participant. She and that language decline is hierarchical in that the language forms we learn last (the most complex language forms) are the first to deteriorate.

It is clear from both the clinical diagnostic criteria and supporting research that language is impacted in those with AD. However, one of the costs of analysing language is that there is a huge burden on trained practitioners, be it clinical psychologists, audio transcribers and text encoders to facilitate the process of collecting data and analysis. The field of machine learning and natural language processing has been suggested as a way toimprove the accuracy and lessen the human cost of this research as well as provide new insights into the difficulties that AD suffer in terms of language decline [7].

The purpose of this review is to seek to understand what techniques have been used when considering the application of Machine Learning and Natural Language processing to aid the diagnosis of Mild Cognitive Impairment (MCI) and AD. A search of the literature was conducted using ProQuest (PsychArticles), SCOPUS, Web of Science. The following results were found (Table 1). All papers were then reviewed for relevance by reading the abstract and full text where appropriate and a shortlist was compiled. An additonal search through references of shortlisted papers was also conducted. Papers were included where researchers used machine learning to classify participants as MCI, AD or Healthy using language. We excluded any papers that focused on other forms of dementia or cognitive impairment, as well as any papers in which the language being analysed was not English. This resulted in 17 journal articles and/or conference papers which form this review.

Database	Number of Results	Search Terms
ProQuest(PsychArticles)	1484 Results	Language AND Decline AND Dementia
ProQuest(PsychArticles)	486 Results	Language AND Decline AND Dementia AND Speech
ProQuest(PsychArticles)	159 Results	Machine Learning AND Dementia AND Language
Web of Science	1207 Results	Language AND Decline AND Dementia
Web of Science	151 Results	Language AND Decline AND Dementia AND Speech
Web of Science	34 Results	Machine Learning AND Dementia AND Language
Scopus	791 Results	Language AND Decline AND Dementia
Scopus	91 Results	Language AND Decline AND Dementia AND Speech
Scopus	29 Results	Machine Learning AND Dementia AND Langauge

Table 1: Search Terms and Number of Results.

2 Types of Language Assessment

One of the key debates when looking at how to analyse language is the type of task provided to elicit language production in participants. In the literature researchers have primarily focused on Picture Description tasks but have also suggested other ways in which we might collect data.

2.1 Picture Description Tasks

One of the most commonly used tasks to measure language is the Picture Description task. An example of this is part of the Boston Diagnostic Aphasia Examination (BDAE), called the Boston Cookie Theft picture description task [8]. In this task participants are asked to describe a picture presented to them in as much detail as possible. The picture itself depicts a familiar domestic scene and would not require participants to use any vocabulary beyond that learned in childhood. It was originally designed to assess Aphasia, but has shown itself to be useful in the assessment of language for the purposes of diagnosis of AD as well [9]

The picture description task does a fine job of eliciting descriptive language, however because of the limited content has limited use. The task in itself just a descriptive task, and therefore elicits a certain type of language. There is some disagreement as to the benefits of this using this methodology. This task is reported as being useful to lexico-semantic disorders [7, 10] as the language being generated is primarly nouns and deixis (words to identify items and words to put those items into context). However, Ash [11]felt that there was no difference in using this task vs Story Narration (described below). In explaining the differences, it is worth noting that these researchers were using differing variables and this could explain their different perspectives.

In terms of Machine Learning research in this area, a number of researchers have used transcripts based on picture description tasks [12, 13, 14, 15] and have successfully extracted linguistic features that could differentiate between AD and controls.

2.2 Narrative description task

The story narration task is designed to study a participant's ability to describe and elaborate on a story which is depicted using a series of pictures. The stories depicted are usually based on children's books or famous stories such as Cinderella. [16]

This task requires ordering the story in a structured and coherent framework. It also requires comprehension and understanding of the stories characters and the events depicted, as well as an awareness of a character's goals and internal responses to given events. This task is particularly useful, as the procedure reduces the demands on memory and is therefore able rule out memory as a confounding variable for any results observed. As noted above, Ash [11] felt that this task was interchangable with the Picture Description task and because this task requires elaboration rather than simple description, is a sturdier test of lexical and semantic abilities as well as syntactic complexity. [17]

Given the relative strengths of the Narrative description task vs Picture description task, there are few pieces of research research that have used Machine Learning to analyse features from Narrative picture tasks [16]. This could be down to the availability of data and the absence of any meaningful sets of transcripts of participants performing this task. However, this could be an interesting direction to take research in the future to see if features generated from this task could be used to predict MCI or AD.

2.3 Interviews

Interviews can also be used to elicit language, the idea of employing questions to guide a conversation between speakers. There are three types of interviews: unstructured, structured and semi-structured. Structured interviews tend to produce very limited speech and therefore has never been used in this area [7]. Unstructured interviews are open ended and generally do not conform to any particular pattern. They use generic themes such as family or hobbies to guide the conversation. Whilst this is the most ecologically valid form of conversation and therefore language generation, it's unstructured nature means that the protocol cannot be consistent and therefore reproduced. Semi-structured interviews are therefore preferred other forms of interview as a middle ground. The semi structured nature of these interviews means that there is some replicability but does not constrain the participant in answering questions.

The analysis of interviews can be difficult to analyse as both the content can vary even between participants. It is also difficult to measure as there are no pre-defined task goals in comparison to the other two methods. Nevertheless, this is the most naturalistic setting for looking at language production and can be used to look at the syntactic and semantic parts of language generation [10]. There have been some attempts to use interviews to to assess language production in AD with promising results [18, 19].

3 How do we analyse language?

Part of the reason we need to pay attention to how we ask participants to generate data is understanding how we wish to analyse the data afterwards. As discussed above, the different methodologies to collect data generate different types of data. There are two main approaches which we have looked at to analyse language, using frequencies of words and combinations of words and measures of syntax and semantics. There are other less common methods of analysing language but these are beyond the scope of this review.

3.1 On using frequencies of words and non-words

The basic premise behind using frequencies to measure changes in language is that certain types of words and non words are more prominent in those suffering with AD than healthy controls.

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 [20] 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) [13] investigated the use of machine learning algorithms to detect differences in syntactic, lexical and n-gram linguistic biomarkers to distinguish between those with a diagnosis of AD and healthy controls, they found significant differences in the uses of all three types of biomarkers but their main finding supported n-grams as the most significant predictor. Because Orimaye used the picture description task as their method of data collection, the n-grams generated are only specific to the task given and cannot be generalised.

Asgari, Kaye and Dodge (2017) [18] 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 [21]. They were able to successfully used machine learning algorithms to distinguish between these two groups with an accuracy of 84%.

3.2 Measures of Syntactic and Semantic Complexity

Another approach to linguistic analysis in this field is the idea of measuing syntactic and semantic complexity. According to Emery (2000) [6] in which she states that Semantic and Syntactic skills deteriorate first in people with MCI and AD. If this is true, then psychological measures of semantic and syntactic skills should be able to pick up signs of deterioration and act as markers for possible MCI and AD. An example of these measures is the concept of idea density. Formally, idea density is defined as the average number of propositions per sentence [22] and this was used to successfully differentiate between people who would later go on to develop AD [5]. Type to Token Ratio, Brunet's Index and Honore's Statistic are two more examples of validated measures which are used to measure the lexical diversity of a given piece of text. This has also shown to be effective in differentiating between MCI, AD and Controls, with those with language impairments [23] and this has carried through in research involving machine learning [24, 20].

Fraser, Meltzer and Rudzicz (2015) [15] 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. Zimmerer (2016) [12] looking 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' is a formulaic sentence structure and by relying on this, AD suffers reduce the cognitive load of trying to come up with a sentence structure.

It appears that both frequency based approaches and approaches that use measures of linguistic complexity have both been shown to be effective in various experimental settings. The fact that there are two vastly different perspectives to tackling this question, and even within these perspectives there are a range of various methodologies leaves the field with a sense of confusion as to the most effective way to solving the problem.

4 Conclusions and Future Work

Identification of language impairment is important in Dementia because it aids diagnosis of specific types of dementia, which in turn can alter the prognosis and change the management of the degenerative disorder. As these differences in language are quite subtle, the varying subtypes of dementia are frequently misdiagnosed and can sometimes be missed altogether. Given the burden on the diagnosis of dementia on clinicians, it appears to be useful to find some non-invasive protocols for the early diagnosis of dementia. It has already been shown that analysis of speech and language has shown markers that pre-date the official diagnosis of dementia [5, 4]. A significant 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. Firstly, the vast

majority of the research described above looks at pre-existing datasets like the DementiaBank corpus. Whilst this is useful for 'backtesting the data', it would be useful for this process to be tested live to see whether these results can be replicated now.

One of the critisicms of the literature to date is the features that are being used to measure language. For example, Zimmerer (2016) [12] describes connectivity in such as a way that correlates directly with what Mueller (2018) [14] 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 conceptualise as part of the larger problem. 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.

Future research should be directed towards developing non-intrusive ways of detecting subtle changes in language based in the home, such that any deterioration that could indicate the presence of MCI or AD and be flagged up early in the progression of any potential cogntive impairment for further review via referral. Given the amount of processing power and the sophistication of machine learning, in particular deep learning algorithms, machine learning approaches seem to be the most logical approach for achieving this aim. Further, despite the quality of datasets 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 we could claim are limited in scope. 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.

References

[1] American Psychiatric Association. Diagnostic and statistical manual of mental disorders: DSM-5, 2013.

[2] Martin Prince, Anders Wimo, M Guerchet, Gemma-Claire Ali, Yu-Tzu Wu, and Matthew Prina. World Alzheimer Report 2015 The Global Impact of Dementia. Technical report, 2015.

[3] Peter J Nestor, Philip Scheltens, and John R Hodges. Advances in the early detection of Alzheimer's disease. *Nature Medicine*, 12(8):961–966, 2006.

[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] David A. Snowdon, Susan J. Kemper, James A. Mortimer, Lydia H. Greiner, David R. Wekstein, and William R. Markesbery. Linguistic ability in early life and cognitive function and Alzheimer's disease in late life: Findings from the Nun Study. *Journal of the American Medical Association*, 275(7):528–532, 1996.

[6] V O Emery. Language impairment in dementia of the Alzheimer type: a hierarchical decline? *International Journal of Psychiatry in Medicine*, 30(2):145–164, 2000.

[7] Veronica Boschi, Eleonora Catricalà, Monica Consonni, Cristiano Chesi, Andrea Moro, and Stefano F. Cappa. Connected speech in neurodegenerative language disorders: A review, 2017.

[8] E. Kaplan, H. Goodglass, and S. Weintraub. Boston Naming Test. In *The Corsini encyclopedia of psychology*, page 2009. 2010.

[9] Elaine Giles, Karalyn Patterson, and John R. Hodges. Performance on the Boston Cookie Theft picture description task in patients with early dementia of the Alzheimer's type: Missing information. *Aphasiology*, 10(4):395–408, 1996.

[10] Seyed Ahmad Sajjadi, Karalyn Patterson, Michal Tomek, and Peter J. Nestor. Abnormalities of connected speech in semantic dementia vs Alzheimer's disease. *Aphasiology*, 26(6):847–866, 2012.

[11] S Ash, E Evans, J O'Shea, K Chandrasekaran, A Boller, L Burkholder, E Camp, D Weinberg, J Haley, K Kitain, and C McMillan. Differentiating primary progressive aphasias in connected speech production. *Dementia and Geriatric Cognitive Disorders*, 34:246, 2012.

[12] Vitor C. Zimmerer, Mark Wibrow, and Rosemary A. Varley. Formulaic Language in People with Probable Alzheimer's Disease: A Frequency-Based Approach. *Journal of Alzheimer's Disease*, 53(3):1145–1160, 2016.

[13] Sylvester O. Orimaye, Jojo S.M. Wong, Karen J. Golden, Chee P. Wong, and Ireneous N. Soyiri. Predicting probable Alzheimer's disease using linguistic deficits and biomarkers. *BMC Bioinformatics*, 18(1):1–13, 2017.

[14] Kimberly D. Mueller, Rebecca L. Koscik, Bruce P. Hermann, Sterling C. Johnson, and Lyn S. Turkstra. Declines in connected language are associated with very early mild cognitive impairment: Results from the Wisconsin Registry for Alzheimer's Prevention. *Frontiers in Aging Neuroscience*, 9(JAN):1–14, 2018.

[15] Kathleen C. Fraser, Jed A. Meltzer, and Frank Rudzicz. Linguistic features identify Alzheimer's disease in narrative speech. *Journal of Alzheimer's Disease*, 49(2):407–422, 2015.

[16] Kathleen C. Fraser, Jed A. Meltzer, Naida L. Graham, Carol Leonard, Graeme Hirst, Sandra E. Black, and Elizabeth Rochon. Automated classification of primary progressive aphasia subtypes from narrative speech transcripts. *Cortex*, 55(1):43–60, 2014.

[17] Juliana Onofre De Lira, Karin Zazo Ortiz, Aline Carvalho Campanha, Paulo Henrique Ferreira Bertolucci, and Thaís Soares Cianciarullo Minett. Microlinguistic aspects of the oral narrative in patients with Alzheimer's disease. *International Psychogeriatrics*, 23(3):404–412, 2011.

[18] Meysam Asgari, Jeffrey Kaye, and Hiroko Dodge. Predicting mild cognitive impairment from spontaneous spoken utterances. *Alzheimer's and Dementia: Translational Research and Clinical Interventions*, 3(2):219–228, 2017.

[19] 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. In *IEEE SSCI 2014 - 2014 IEEE Symposium Series on Computational Intelligence - CICARE 2014: 2014 IEEE Symposium on Computational Intelligence in Healthcare and e-Health, Proceedings*, pages 98–103, 2015.

[20] C. Thomas, V. Keselj, N. Cercone, K. Rockwood, and E. Asp. Automatic detection and rating of dementia of Alzheimer type through lexical analysis of spontaneous speech. *IEEE International Conference Mechatronics and Automation, 2005*, 3(February 2005):1569–1574, 2005.

[21] James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. The Development and Psychometric Properties of LIWC2015. *UT Faculty/Researcher Works*, 2015.

[22] Walter Kintsch and Janice Keenan. Reading rate and retention as a function of the number of propositions in the base structure of sentences. *Cognitive Psychology*, 5(3):257–274, 1973.

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

[24] Shuai Wang. *Automatic Tracking of Linguistic Changes for Monitoring Cognitive-Linguistic Health*. PhD thesis, 2016.