Detecting Dementia through Retrospective Analysis of Routine Blog Posts by Bloggers with Dementia

Vaden Masrani
Dept. of Computer Science
University of British Columbia
vadmas@cs.ubc.ca

Gabriel Murray
Dept. of Computer Information Systems
University of the Fraser Valley
gabriel.murray@ufv.ca

Thalia Field
Dept. of Neurology
University of British Columbia
thalia.field@ubc.ca

Giuseppe Carenini
Dept. of Computer Science
University of British Columbia
carenini@cs.ubc.ca

Abstract

We investigate if writers with dementia can be automatically distinguished from those without by analyzing linguistic markers in written text, in the form of blog posts. We have built a corpus of several thousand blog posts, some by people with dementia and others by people with loved ones with dementia. We use this dataset to train and test several machine learning methods, and achieve prediction performance at a level far above the baseline.

1 Introduction

Dementia is estimated to become a trillion dollar disease worldwide by 2018, and prevalence is expected to double to 74.7 million by 2030 (Prince, 2015). Dementia is a clinical syndrome caused by neurodegenerative illnesses (e.g. Alzheimer's Disease, vascular dementia, Lewy Body dementia). Symptoms can include memory loss, decreased reasoning ability, behavioral changes, and – relevant to our work – speech and language impairment, including fluency, word choice and sentence structure (Klimova and Kuca, 2016).

Recently, there have been attempts to combine clinical information with language analysis using machine learning and NLP techniques to aid in diagnosis of dementia, and to distinguish between types of pathologies (Jarrold et al., 2014; Rentoumi et al., 2014; Orimaye et al., 2014; Fraser et al., 2015; Masrani et al., 2017). This would provide an inexpensive, non-invasive and efficient screening tool to assist in early detection, treatment and institution of supports. Yet, much of the work to date has focused on analyzing spoken language collected during formal assessment, usually with standardized exam tools.

There has been comparatively little work done on analyzing *written* language spontaneously generated by people with dementia. In coming years, there will be an increased number of tech-savvy seniors using the internet, and popular online commentators will continue to age. There will therefore be a growing dataset available in the form of tweets, blog posts, and comments on social media, on which to train a classifier. Provided our writers have a verified clinical diagnosis of dementia, such a dataset would be large, inexpensive to acquire, easy to process, and require no manual transcriptions.

There are downsides to using written language samples as well. Unlike spoken language, written text can be edited or revised by oneself or others. People with dementia may have "good days" and "bad days," and may write only on days when they are feeling lucid, and therefore written samples may be biased towards more intact language. Furthermore, we do not have an accompanying audio file and patients are not constrained to a single topic; people with dementia may have greater facility discussing familiar topics. A non-standardized dataset will also prevent the collection of common test-specific linguistic or acoustic features. However, working with a very large dataset may be able to mitigate the effects of these limitations.

In this work we gather a corpus of blog posts publicly available online, some by people with dementia and others by the loved ones of people with dementia. We extract a variety of linguistic features from the texts, and compare multiple machine learning methods for detecting posts written by people with dementia. All models perform well above the baseline, demonstrating the feasibility of this detection task.

Proceedings of the BioNLP 2017 workshop, pages 232–237,
Vancouver, Canada, August 4, 2017.

2 Related Work

Early signs of dementia can be detected through analysis of writing samples (Le et al., 2011; Riley et al., 2005; Kemper et al., 2001). In the "Nun Study" researchers analyzed autobiographies written in the US by members of the School Sisters of Notre Dame between 1931-1996. Those nuns who met criteria for dementia had lower grammatical complexity scores and lower "idea density" in their autobiographies.

Le et al. (2011) performed a longitudinal analysis of the writing styles of three novelists: Iris Murdoch who died with Alzheimer's disease (AD), Agatha Christie (suspected AD), and P.D. James (normal brain aging). Measurements of syntactic and lexical complexity were made from 51 novels spanning each of the author careers. Murdoch and Christie exhibited evidence of linguistic decline in later works, such as vocabulary loss, increased repetition, and a deficit of noun tokens (Le et al., 2011).

Despite evidence that linguistic markers predictive of dementia can be found in writing samples, there have been no attempts to train models to classify dementia based on writing alone. Previous work has been successful in training models using transcribed utterances from patients undergoing formal examinations, but this data is difficult to acquire and many models use audio and/or test-specific features which would not be available from online text (Rentoumi et al., 2014; Orimaye et al., 2014; Fraser et al., 2014; Roark et al., 2011). State-of-the-art classification accuracy of 81.92% was achieved by Fraser et al. (2015) with logistic regression using acoustic, textual, and test-specific features on 473 samples from DementiaBank dataset, an American cohort of 204 persons with dementia and 102 controls describing the "Cookie Theft Picture", a component of the Boston Diagnostic Aphasia Examination (Becker et al., 1994; Giles et al., 1996). More recently, these results have been extended via domain adaptation by Masrani et al. (Masrani et al., 2017).

Our methods are similar to Fraser et al. (2015), with the main difference being the dataset used and their inclusion of audio and test-specific features, which are not available in our case. To the best of our knowledge, ours is the first comparison of models trained exclusively on unstructured written samples from persons with dementia.

3 Experimental Design

In this section, we describe the novel blog corpus and experimental setup.

3.1 Corpus

We scraped the text of 2805 posts from 6 public blogs as described in Table 1. Three blogs were written by persons with dementia (First blogger: male, AD, age unknown. Second blogger: female, AD, age 61. Third blogger: Male, Dementia with Lewy Bodies, age 66) and three written by family members of persons with dementia to be used as control (all female, ages unknown). Other demographic information, such as education level, was unavailable. From each of the three dementia blogs, we manually filtered all texts not written by the owner of the blog (e.g. fan letters) or posts containing more images than text. We were left with 1654 samples written by persons with dementia and 1151 from healthy controls. The script to download the corpus is available at `https://github.com/vadmas/blog_corpus/`.

3.2 Classification Features

Following Fraser et al. (2015), we extracted 101 features across six categories from each blog post. These features are described below.

Parts Of Speech (14) We use the Stanford Tagger (Toutanova et al., 2003) to capture the frequency of various parts of speech tags (nouns, verbs, adjectives, adverbs, pronouns, determiners, etc). Frequency counts are normalized by the number of words in the sentence, and we report the sentence average for a given post. We also count not-in-dictionary words and word-type ratios (noun to verb, pronoun to noun, etc).

Context Free Grammar (45) Features which count how often a phrase structure rule occurs in a sentence, including NP→VP PP, NP→DT NP, etc. Parse trees come from the Stanford parser (Klein and Manning, 2003).

Syntactic Complexity (28) Features which measure the complexity of an utterance through metrics such as the depth of the parse tree, mean length of word, sentences, T-Units and clauses and clauses per sentence. We used the L2 Syntactic Complexity Analyzer (Lu, 2010).

Psycholingustic (5) Psycholinguistic features are linguistic properties of words that effect word

URL	Posts	Mean words	Start Date	Diagnosis
https://creatingmemories.blogspot.ca/	618	242.22 (s=169.42)	Dec 2003	AD
http://living-with-alzhiemers.blogspot.ca/	344	263.03 (s=140.28)	Sept 2006	AD
http://parkblog-silverfox.blogspot.ca/	692	393.21 (s=181.54)	May 2009	Lewy Body
http://journeywithdementia.blogspot.ca/	201	803.91 (s=548.34)	Mar 2012	Control
http://earlyonset.blogspot.ca/	452	615.11 (s=206.72)	Jan 2008	Control
http://helpparentsagewell.blogspot.ca/	498	227.12 (s=209.17)	Sept 2009	Control

Table 1: Blog Information.

processing and learnability (Salsbury et al., 2011). We used five psycholinguisic features: *Familiarity*, *Concreteness*, *Imageability*, *Age of acquisition*, and the *SUBTL* , which is a measure of the frequency with which a word is used in daily life (Kuperman et al., 2012; Brysbaert and New, 2009a; Salsbury et al., 2011). Psycholinguistic word scores are derived from human ratings[1] while the SUBTL frequency norm[2] is based on 50 million words from television and film subtitles (Brysbaert and New, 2009b).

Vocabulary Richness (4) We calculated four metrics which capture the range of vocabulary in a text: type-token ratio, Brunet's index, a length insensitive version of the type-token ratio, Honore's statistic, and the moving-average type-token ratio (MATTR) (Asp and De Villiers, 2010; Covington and McFall, 2010). These metrics have been shown to be effective in previous AD research (Bucks et al., 2000; Fraser et al., 2015)

Repetitiveness (5) We represent sentences as TF-IDF vectors and compute the cosine similarity between sentences. We then report the proportion of sentence pairs below three similarity thresholds (0, 0.3, 0.5) as well as the min and average cosine distance across all pairs of sentences.

3.3 Training and Testing

We perform a 9-fold cross validation by training each model on all the posts of four blogs and testing on the remaining two, where we assure that each test set contains the posts of one control blog and one dementia blog. Within each fold we perform a feature selection step before training where we select for inclusion into the model the first k features which have the highest absolute correlation with the labels in the training fold.

[1] http://websites.psychology.uwa.edu.au/school/MRCDatabase/uwa_mrc.htm

[2] http://subtlexus.lexique.org/

4 Results

For each machine learning model, we calculate the ROC curve and the area under the curve (AUC), comparing with a random performance baseline AUC of 0.5. The AUC results are shown in Figure 1, with all models well above the baseline of 0.5. The best performing models are logistic regression and neural networks, with average AUC scores of 0.815 and 0.848, respectively.

The SUBTL measure of vocabulary richness was the feature most correlated with the outcome variable in eight out of nine folds. Figure 2 shows the SUBTL scores for each blog post in the corpus, arranged by blog and with the bloggers with dementia shown in the top row. A lower score indicates a richer vocabulary. We can see that the bloggers with dementia have a less rich vocabulary. Interestingly, however, the longitudinal trend does not show their vocabularies worsening during the time-period captured in this corpus. The analysis of other features highly informative for the target prediction is ongoing, and additional findings will be discussed at the workshop.

5 Conclusion

We have shown that it is possible to distinguish bloggers with dementia from those without, on a novel corpus of blog data. We extracted linguistic features from the texts and compared a large number of machine learning methods, all of which performed well above the baseline. While feature analysis is ongoing, we have made some interesting observations about the effect of the SUBTL measure of vocabulary richness. Future work will include liaising with patient and caregiver support groups to expand this new dementia corpus, inclusion of a topic clustering preprocessing step to control for variation across content, and further longitudinal analysis.

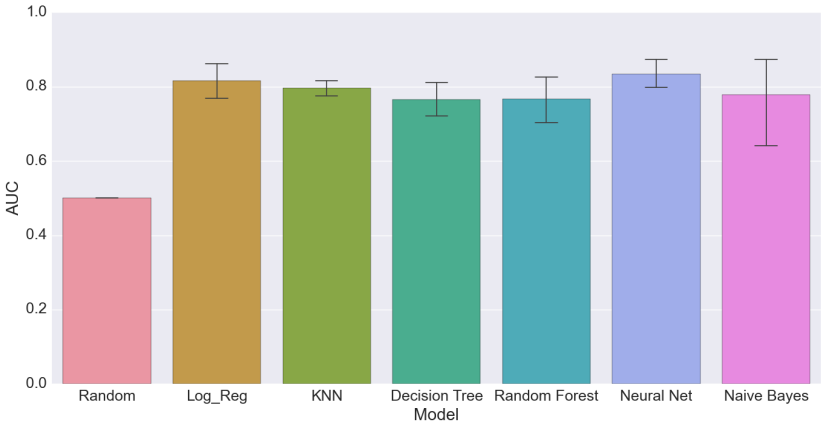

Figure 1: Comparison of models. We show the mean AUC and 90% confidence intervals across a 9-fold CV. All the posts of a blog appear in either the training or test set, but not both.

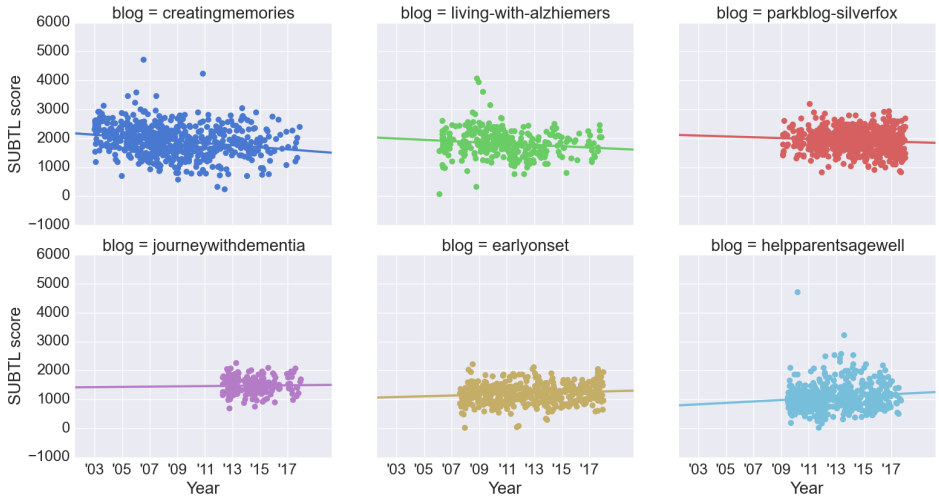

Figure 2: SUBTL word scores for each post in a given blog. Bloggers with dementia (AD or Dementia w/ Lewy Bodies) appear in the top row.

References

Elissa D Asp and Jessica De Villiers. 2010. *When language breaks down: Analysing discourse in clinical contexts*. Cambridge University Press.

James T Becker, François Boiler, Oscar L Lopez, Judith Saxton, and Karen L McGonigle. 1994. The natural history of alzheimer's disease: description of study cohort and accuracy of diagnosis. *Archives of Neurology* 51(6):585–594.

Marc Brysbaert and Boris New. 2009a. Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for american english. *Behavior Research Methods* 41:977–990.

Marc Brysbaert and Boris New. 2009b. Moving beyond kučera and francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for american english. *Behavior research methods* 41(4):977–990.

Romola S Bucks, Sameer Singh, Joanne M Cuerden, and Gordon 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.

Michael A Covington and Joe D McFall. 2010. Cutting the gordian knot: The moving-average type–token ratio (mattr). *Journal of Quantitative Linguistics* 17(2):94–100.

Kathleen C Fraser, Graeme Hirst, Naida L Graham, Jed A Meltzer, Sandra E Black, and Elizabeth Rochon. 2014. Comparison of different feature sets for identification of variants in progressive aphasia. *ACL 2014* page 17.

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

Elaine Giles, Karalyn Patterson, and John R Hodges. 1996. 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.

William Jarrold, Bart Peintner, David Wilkins, Dimitra Vergryi, Colleen Richey, Maria Luisa Gorno-Tempini, and Jennifer Ogar. 2014. Aided diagnosis of dementia type through computer-based analysis of spontaneous speech. In *Proceedings of the ACL Workshop on Computational Linguistics and Clinical Psychology*. pages 27–36.

Susan Kemper, Lydia H Greiner, Janet G Marquis, Katherine Prenovost, and Tracy L Mitzner. 2001. Language decline across the life span: findings from the nun study. *Psychology and aging* 16(2):227.

Dan Klein and Christopher D Manning. 2003. Accurate unlexicalized parsing. In *Proc. of ACL, Sapporo, Japan*. Association for Computational Linguistics, pages 423–430.

Blanka Klimova and Kamil Kuca. 2016. Speech and language impairments in dementia. *Journal of Applied Biomedicine* 14(2):97–103.

Victor Kuperman, Hans Stadthagen-Gonzalez, and Marc Brysbaert. 2012. Age-of-acquisition ratings for 30,000 english words. *Behavior Research Methods* 44(4):978–990.

Xuan Le, 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.

Xiaofei Lu. 2010. Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics* 15(4):474–496.

Vaden Masrani, Gabriel Murray, Thalia Field, and Giuseppe Carenini. 2017. Domain adaptation for detecting mild cognitive impairment. In *Proc. of Canadian AI, Edmonton, Canada*.

Sylvester Olubolu Orimaye, Jojo Sze-Meng Wong, and Karen Jennifer Golden. 2014. Learning predictive linguistic features for alzheimer's disease and related dementias using verbal utterances. In *Proc. 1st Workshop. Computational Linguistics and Clinical Psychology (CLPsych)*.

Martin James Prince. 2015. *World Alzheimer Report 2015: the global impact of dementia: an analysis of prevalence, incidence, cost and trends*. London.

Vassiliki Rentoumi, Ladan Raoufian, Samrah Ahmed, Celeste A de Jager, and Peter Garrard. 2014. Features and machine learning classification of connected speech samples from patients with autopsy proven alzheimer's disease with and without additional vascular pathology. *Journal of Alzheimer's Disease* 42(s3).

Kathryn P Riley, David A Snowdon, Mark F Desrosiers, and William R Markesbery. 2005. Early life linguistic ability, late life cognitive function, and neuropathology: findings from the nun study. *Neurobiology of aging* 26(3):341–347.

Brian Roark, Margaret Mitchell, John-Paul Hosom, Kristy Hollingshead, and Jeffrey Kaye. 2011. Spoken language derived measures for detecting mild cognitive impairment. *IEEE Transactions on Audio, Speech, and Language Processing* 19(7):2081–2090.

Tom Salsbury, Scott A Crossley, and Danielle S McNamara. 2011. Psycholinguistic word information in second language oral discourse. *Second Language Research* 27(3):343–360.

Kristina Toutanova, Dan Klein, Christopher D Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proc. of NAACL, Edmonton, Canada*. Association for Computational Linguistics, pages 173–180.