CORTEX XXX (2013) 1–8

Available online at www.sciencedirect.com

SciVerse ScienceDirect

Journal homepage: www.elsevier.com/locate/cortex

Special issue: Research report

Machine learning approaches to diagnosis and laterality effects in semantic dementia discourse

Peter Garrard [a,*], *Vassiliki Rentoumi* [a], *Benno Gesierich* [b], *Bruce Miller* [b] *and Maria Luisa Gorno-Tempini* [b]

[a] *Stroke and Dementia Research Centre, St George's, University of London, Cranmer Terrace, London SW17 ORE, UK*
[b] *UCSF Memory and Aging Center, Sandler Neurosciences Center, 675 Nelson Rising Lane, Suite 190, San Francisco, CA, USA*

ARTICLE INFO

Article history:
Received 15 January 2013
Reviewed 22 February 2013
Revised 12 March 2013
Accepted 15 May 2013
Published online xxx

Keywords:
Semantic dementia
Discourse
Laterality
Machine learning
Information gain

ABSTRACT

Advances in automatic text classification have been necessitated by the rapid increase in the availability of digital documents. Machine learning (ML) algorithms can 'learn' from data: for instance a ML system can be trained on a set of features derived from written texts belonging to known categories, and learn to distinguish between them. Such a trained system can then be used to classify unseen texts. In this paper, we explore the potential of the technique to classify transcribed speech samples along clinical dimensions, using vocabulary data alone. We report the accuracy with which two related ML algorithms [naive Bayes Gaussian (NBG) and naive Bayes multinomial (NBM)] categorized picture descriptions produced by: 32 semantic dementia (SD) patients versus 10 healthy, age-matched controls; and SD patients with left- (n = 21) versus right-predominant (n = 11) patterns of temporal lobe atrophy. We used information gain (IG) to identify the vocabulary features that were most informative to each of these two distinctions.

In the SD versus control classification task, both algorithms achieved accuracies of greater than 90%. In the right- versus left-temporal lobe predominant classification, NBM achieved a high level of accuracy (88%), but this was achieved by both NBM and NBG when the features used in the training set were restricted to those with high values of IG. The most informative features for the patient versus control task were low frequency content words, generic terms and components of metanarrative statements. For the right versus left task the number of informative lexical features was too small to support any specific inferences. An enriched feature set, including values derived from Quantitative Production Analysis (QPA) may shed further light on this little understood distinction.

© 2013 Elsevier Ltd. All rights reserved.

1. Introduction

Rapid growth in the availability of digital documents, such as web pages, blogs and emails, has highlighted the importance of methods for automatic text classification. Imposing order on unstructured collections of texts facilitates storage, search, browsing and future re-use. Current approaches to text classification rely on machine learning (ML) techniques, which build automatic text classifiers by learning the characteristics of the categories of interest (such as topic or genre) from a set

* *Corresponding author.*
E-mail addresses: pgarrard@sgul.ac.uk, peter.garrard@gmail.com (P. Garrard).
0010-9452/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.cortex.2013.05.008

Please cite this article in press as: Garrard P, et al., Machine learning approaches to diagnosis and laterality effects in semantic dementia discourse, Cortex (2013), http://dx.doi.org/10.1016/j.cortex.2013.05.008

of pre-classified documents. The many practical applications of ML-based text classification include authorship attribution, in which a text of unknown origin is attributed to one of a set of candidate authors. In some instances the approach has achieved accuracy comparable to that of human experts (Sebastiani, 2002). The potential of ML in the classification of language samples generated by different clinical populations has been less extensively investigated, though the problem of distinguishing among samples obtained from patients with different clinical syndromes is clearly analogous to that of authorship attribution.

Semantic dementia (SD) is a progressive neurodegenerative syndrome characterized by the relatively isolated degradation of the semantic component of long term declarative memory (Hodges et al., 1992; Snowden et al., 1992; Tulving, 1972; Warrington, 1975). Disruption of a system so central to language inevitably impacts on the production of discourse, giving rise to a pattern of spontaneous speech that is fluent, phonologically and grammatically correct, and makes use of a high frequency, often generic vocabulary (e.g., 'thing', 'bit' or 'stuff') (Bird and Lambon Ralph, 2000; Hodges et al., 1992; Meteyard and Patterson, 2009).

Studies of magnetic resonance (MR) imaging of the brains of patients with SD, at individual and group level, have consistently identified bilateral, asymmetric temporal lobe atrophy (Galton et al., 2001; Garrard and Hodges, 2000; Gorno-Tempini et al., 2004; Mummery et al., 2000) in which the left hemisphere is often (though not always see, e.g., Evans et al., 1995 and Thompson et al., 2003) the more profoundly affected. Comparisons between the subgroup with predominantly left- and predominantly right-sided atrophy (referred to henceforth as 'L > R' and 'R > L' respectively) have identified a number of features that appear to characterize the latter. These differences have largely emphasized the social and emotional impairment that frequently accompanies SD, the absence of insight (Chan et al., 2009; Perry et al., 2001; Thompson et al., 2003), and differences in the type of semantic knowledge disrupted (person specific rather than general) (Evans et al., 1995; Gentileschi et al., 2001; Josephs et al., 2008; Joubert et al., 2006).

In a recent study, Wilson et al. (2010) used Quantitative Production Analysis (QPA) (Saffran et al., 1989) to analyse transcripts of discourse samples obtained from patients with primary progressive aphasia (PPA) (Gorno-Tempini et al., 2011), including 25 with a diagnosis of SD. The SD transcripts differed most markedly from those of controls and other PPA syndromes on a lexical dimension, with striking differences identified in the increased used of pronouns, verbs, and high frequency nouns. Deviations from normal performance on selected aspects of syntactic structure and complexity were least marked in the SD subtype.

Studies examining the differential effects of L > R and R > L temporal lobe atrophy on language production are fewer in number: using measures derived from voxel based morphometry (VBM) (Good et al., 2001), Mummery et al. (2000) found that measures of semantic performance selectively correlated with atrophy in left-temporal lobe structures; Lambon Ralph et al. (2001) found that, in R > L cases, anomia and single word comprehension tended to deteriorate *pari passu*, while a L > R group showed disproportionately severe anomia. The same study also identified differences in the types of error made on a naming test by the two patient groups: the responses of L > R patients were more likely to be circumlocutory or omitted altogether, while patients in the R > L group committed more coordinate errors (such as production of 'goat' in response to a picture of a horse). To date, however, there have been no systematic comparisons of the effects of R > L versus L > R temporal lobe atrophy on the production of connected discourse in SD.

Garrard and Forsyth (2010) hypothesized that the language of SD patients and controls would be distinguishable on the basis of the lexical frequency data from transcripts of connected speech. They used principal components analysis (PCA) to identify two latent variables in a high dimensional 'discourse space', the values of which distinguished between connected speech samples obtained from patients with SD and those produced by controls. The study found that the vocabulary used by SD patients differed from those of controls along at least two major dimensions. The region of this two-dimensional 'discourse space' occupied by the control transcripts was characterized by the use of specific content-bearing terms and the deployment of a variety of grammatical function words. In contrast, the patient transcripts were associated with values correlating with use of the pronouns 'HE' and 'SHE', generic terms such as 'SOMETHING', the deictic words 'HERE' and 'THERE', the light verb 'DOING', and components of the phrase 'I DO NOT KNOW'. The two dimensions therefore captured not only the lexical-semantic deficit, but also the syntactically simplified character of SD discourse that has emerged from a number of manual analyses (Benedet and Patterson, 2006; Patterson and Macdonald, 2006).

In the current paper, we broaden this methodological approach by applying and comparing the performance of two (Bayesian) ML methods, in the classification of spoken discourse samples produced by SD patients and controls. In addition, we apply the same methods to the problem of distinguishing between speech samples produced by SD patients with L > R and R > L patterns of temporal lobe atrophy.

2. Materials and methods

The data for the study consisted of transcribed samples of connected speech, and structural MR imaging obtained from 32 patients meeting diagnostic and imaging criteria for SD (Gorno-Tempini et al., 2011; Hodges et al., 1992) and from 10 age matched, cognitively normal controls (NC). All participants were fluent English speakers, recruited through the Memory and Aging Center at UCSF after giving written informed consent. The study was approved by the institutional review board.

Within the SD group were 21 individuals classified on clinical and imaging grounds (see Section 2.2) as showing a L > R pattern of temporal lobe asymmetry, and 11 with the R > L pattern. The demographic characteristics of the three participant groups, as well as the performance of the SD groups on standardized tests of semantic memory, are summarized in Table 1. SD patients with a L > R pattern showed significantly greater impairment on naming and single word comprehension than the R > L subgroup, but the two SD

Please cite this article in press as: Garrard P, et al., Machine learning approaches to diagnosis and laterality effects in semantic dementia discourse, Cortex (2013), http://dx.doi.org/10.1016/j.cortex.2013.05.008

Download English Version:

https://daneshyari.com/en/article/7315652

Download Persian Version:

https://daneshyari.com/article/7315652

Daneshyari.com