Similarity Measures for the Detection of Clinical Conditions with Verbal Fluency Tasks

Felipe S. F. Paula[1], **Rodrigo Wilkens**[2], **Marco A. P. Idiart**[3], **Aline Villavicencio**[1,4]
[1] Institute of Informatics, Federal University of Rio Grande do Sul (Brazil)
[2] CENTAL, Université catholique de Louvain (Belgium)
[3] Institute of Physics, Federal University of Rio Grande do Sul (Brazil)
[4] School of Computer Science and Electronic Engineering, University of Essex (UK)
felipesfpaula@gmail.com, rodrigo.wilkens@uclouvain.be,
marco.idiart@gmail.com, avillavicencio@inf.ufrgs.br

Abstract

Semantic Verbal Fluency tests have been used in the detection of certain clinical conditions, like Dementia. In particular, given a sequence of semantically related words, a large number of switches from one semantic class to another has been linked to clinical conditions. In this work, we investigate three similarity measures for automatically identify switches in semantic chains: semantic similarity from a manually constructed resource, and word association strength and semantic relatedness, both calculated from corpora. This information is used for building classifiers to distinguish healthy controls from clinical cases with early stages of Alzheimer's Disease and Mild Cognitive Deficits. The overall results indicate that for clinical conditions the classifiers that use these similarity measures outperform those that use a gold standard taxonomy.

1 Introduction

In the diagnosis of clinical conditions, language production along with socio-educational and cognitive factors have been regarded as providing important clues about the health of the semantic memory and of the mental lexicon (Troyer et al., 1998). Some neuropsychiatric protocols for the assessment of clinical conditions like Alzheimer's Disease (AD) and Mild Cognitive Deficits (MCD) often adopt Semantic Verbal Fluency (SVF) (Zhao et al., 2013), since linguistic impairments in such conditions are most likely located at the semantic level (Taler and Phillips, 2008). In these tests participants are asked to produce words related to a given theme (e.g. animals or supermarket items) in a short period of time (e.g. one minute) avoiding repetitions. The answers tend to contain subgroups (Bousfield and Sedgewick, 1944), referred to as clusters and their borders as switches. For instance, a sequence like *dog, mouse, cat, horse, pig*, and *cow* could be divided into two clusters with a switch: pets (dog, mouse, and cat) and farm animals (horse and pig). Clues like the size of semantic clusters and the number of switches (Troyer et al., 1998) have been correlated with clinical conditions (Murphy et al., 2006; Pekkala et al., 2008; Price et al., 2012; Bertola et al., 2014b), and, in some cases, data derived from SVF tests have indicated dementia five years before its onset (Raoux et al., 2008).

The analysis of clusters and switches requires manual annotation by specialists, based on preexisting manually constructed taxonomies, in a process that can be very time consuming and prone to coverage limitations. In this paper we investigate three similarity measures for detecting switches in word sequences: semantic similarity using a manually constructed resource, as well as word association strength and semantic relatedness both calculated from corpora. We then apply this information to distinguish different clinical groups using classifiers in a fully automated way. This paper is structured as follows: in §2, we review the detection of neuropsychiatric diseases with SVF tests. In §3 we discuss the data and the switch detection strategies. In §4 reports results. We finish with conclusions and future work.

2 Related Works

The cluster and switch dynamic is a classic source of information for separating clinical groups in SVF tests, due to their deep connections to executive functions and semantic memory (Troyer et al., 1998). Clinical detection approaches are widely based on SVF tests and analyze word productivity (Murphy et al., 2006), word repetitions (Raoux et al., 2008; Pekkala et al., 2008; Henry and Phillips, 2006), and number of clusters and switches (Gocer March and Pattison, 2006; Price et al., 2012).

Computational approaches for prediction of switches in SVFs have used information about semantic relatedness from distributional semantic models (Linz et al., 2017). Prediction of semantic clusters has been done with clustering algorithms using LSA similarity between pairs of words. These clusters were then used to detect bipolarity and schizophrenia (Rosenstein et al., 2015).

SVF tests have also been computationally modeled in terms of graphs with nodes corresponding to words and edges to the temporal connections between them. Topological measures, such as, the number of nodes and edges, shortest path, diameter, and density were

Proceedings of NAACL-HLT 2018, pages 231–235
New Orleans, Louisiana, June 1 - 6, 2018. ©2018 Association for Computational Linguistics

used to distinguish the control from clinical groups diagnosed with schizophrenia and manic depression disorder (Mota et al., 2012), AD and MCD (Bertola et al., 2014b).

In this work we use similarity measures based on the **association strength** between two words, their **semantic similarity** and their **semantic relatedness** for detecting switches in SVFs involving AD and MCD groups.[1]

3 Methods

3.1 SVF Dataset

The SVF dataset (Bertola et al., 2014a) contains the responses of 100 participants (mean age of 75.78, $sd = 7.13$) of both genders and of similar levels of education. The participants are classified into four groups of 25 individuals. One is a control group with normal cognitive performance, and three are groups with clinical conditions according to assessment guidelines (de Paula et al., 2013; McKhann et al., 1984; Winblad et al., 2004): Amnestic Mild Cognitive Deficit (aMCD), Multi-domain Mild Cognitive Deficit (mMCD) and Alzheimer's Disease (AD). Since the groups are homogeneous, there is no significant differences between members of the same group. Additionally, we also considered a fifth group, the Cognitively Impaired (CI) group, that includes randomly selected participants from the three clinical groups. The responses of each participant are annotated following the guidelines adopted by Troyer et al. (1998); Bertola et al. (2014b).

3.2 Switch identification

In this paper we explore different types of similarity for detecting switches in SVF. An SVF can be divided in semantic chains, which we define as sequences of consecutive words whose similarity falls above a certain threshold (Morris and Hirst, 1991; Pakhomov and Hemmy, 2014). Different semantic chains are separated by switches[2]. Switches form the basis for training classifiers to distinguish control from clinical cases in the SVF dataset (Bertola et al., 2014a). We use Random Forest classifiers (Breiman, 2001) trained with the following features: the number of switches, n; the largest chain size, $c_{max} = \max(c_a)$; the average chain length, $\bar{c} = \frac{1}{n+1}\sum_{a=1}^{n+1} c_a$; the fraction of occurrence of the smallest chain, $f_{min} = \#(c_{min})/(n+1)$, where $\#(c)$ indicates the number of chains of size c in the SVF test of a participant.

Results are reported in terms of average area under the receiver operator characteristic curve (AUC) from 10 times 10-fold-cross validation.[3]

To determine the effectiveness of different types of similarity measures for switch identification we examine semantic similarity from a manually constructed resource, as well as two measures derived from corpora: word association strength, and semantic relatedness. Semantic similarity is determined from the shortest path that connects two words according to the WordNet (Fellbaum, 1998; Perkins, 2010) hypernym taxonomy. The association strength is calculated using the positive value of the Pointwise Mutual Information (PMI) (Church and Hanks, 1990), and the semantic relatedness using the cosine similarity between two GloVe word embeddings (Pennington et al., 2014).‘

WordNet provides a high quality manual resource but is not available for all languages. In this work we translated the SVF responses from Brazilian Portuguese to English.[4] Similarity using association strength and semantic relatedness can be constructed from raw corpora, which makes them an attractive alternative for low-resourced languages like Portuguese. In this work we used a corpus built from the Portuguese Wikipedia[5], which was lemmatized and had high frequency function words removed. After preprocessing, the corpus contained more than 118 million tokens, and 44,000 types. PMI for word pairs was calculated using a sliding window of size 7 over the corpus. GloVe[6] word embeddings were constructed using default parameters, with the exception of the window size and vector dimension which were set to 7 and 300, respectively.

Formally the switch is a binary function $\psi(x_i)$ that operates on the sequence of N words $(w_1, w_2, \cdots, w_N)$ produced by a subject in the SVF test. There is a switch between consecutive words w_i and w_{i+1} when their similarity $x_i = s(w_i, w_{i+1})$ falls below a threshold, in which case $\psi(x_i) = 1$, otherwise $\psi(x_i) = 0$. In this paper we explore three heuristics for the switch function:

Detection based on the global mean. The threshold is given by the average similarity of the list.

$$\psi_{global}(x_i) = H\left(\frac{1}{N-1}\sum_{j=1}^{N-1} x_j - x_i\right)$$

where $H(x) = 1$ if $x \geq 0$ and $H(x) = 0$ otherwise.

Detection based in the local mean. The threshold is given by the average similarity of the last k pairs of words.

$$\psi_k(x_i) = H\left(\frac{1}{k}\sum_{j=1}^{k} x_{i-j} - x_i\right)$$

[1] Although lexical and distributional characteristics of SVFs, like the total number of words and their frequencies, may be effective indicators of clinical conditions, in this paper we focus on switch information and how it can be approximated.

[2] For simplicity sake we consider that a chain may have a single word in which case it has length one.

[3] The models were trained with the Caret Package: topepo.github.io/caret

[4] Given the limitations in WordNet coverage, animals that were not found were replaced by similar animals found in WordNet and with the same frequency profile.

[5] Wikipedia dump corpus from June of 2015

[6] nlp.stanford.edu/projects/glove/

Hibrid detection. We combine the local and global approach in a voting system where a switch is considered if it receives at least v votes from previously switch criteria. Here we consider a combination of global with locals $k = 2$ and 3:

$$\psi_{vot_v}(x_i) = H(\psi_{global}(x_i) + \psi_2(x_i) + \psi_3(x_i) - v)$$

where v can be 1, 2 (majority voting), and 3 (total agreement).

4 Results

Evaluation is carried out at two levels of granularity: a rough-grained classification for the detection of a clinical condition in general (control vs. CI group), and a fine-grained classification for one of the three conditions (aMCD, mMCD and AD groups). Table 3.2 displays the average AUC per heuristic for the different sources, with the highest scores shown in bold along with other scores that are not statistically different, considering p-values adjusted with the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995). The last line of each subtable shows the scores obtained by training the classifiers with the gold standard manual annotation with the taxonomy used by Troyer et al. (1998) (GS in the tables).

Overall, in terms of the type of similarity both the semantic similarity (WordNet) and word association strength (PMI) were significantly better than the gold standard manual annotation for the rough-grained classification and for two of the three clinical cases (mMCD was the exception). This indicates the complementary nature of these additional types of similarity beyond what the smaller and possibly stricter GS taxonomy can offer. Examining the specific groups, the lower scores for aMCD and mMCD also seem to reflect the potential progression of these condition from the control to the more severe impairments of the AD group (aMCD < mMCD < AD).

Among the different measures, the strict total agreement voting (ψ_{vot_3}) provides the best results with association strength for the rough-grained classification (Table3.2(a)), and for the fine-grained classifications of the mMCD (Table3.2(c)) and AD groups (Table3.2(d)). These results suggest that a more conservative identification of switches leading to larger chains provides a better approximation for these three groups.

For the two intermediate clinical groups, aMCD and mMCD, the use of local average information from a small window including only the previous word (ψ_1) also produces good results. However, there is no consensus regarding the source of switch identification, as for aMCD both semantic similarity and association strength were effective, and for mMCD it was semantic relatedness that provided a better characterization of the groups.

Finally, for the AD group various combinations of measures and sources of semantic information lead to effective distinction from the control group, with the best results using the strict total agreement voting. These results are indicative of AD as the clinical group with strongest cognitive impairment in relation to the control.

For a qualitative assessment of the results, we also examine the vocabulary overlap among the groups, using the Jaccard index as shown in Table 4, which presents the average Jaccard index between subjects across all groups. It shows a higher agreement among the control than among the other groups. This is compatible with the discussion by Brandt and Manning (2009) who identified a more systematic strategy for vocabulary exploration in the control than in 'the clinical groups.

Given that the switches derived by our best models were more effective for the detection of the clinical conditions than the gold standard, we explored the idea that maybe the human annotation could be further improved. To test that, we asked subjects to re-annotate 594 pairs of words for which there was disagreement between the gold standard and the predicted switches. Each pair was annotated by an average of 8.1 annotators ($sd = 2.28$) using four context words. When compared with the gold standard, the new annotation resulted in a change of judgment for 12.7% of the word pairs, with higher agreement with the switches predicted by our heuristics. For instance, for $\psi_{vot3}(x_i)$ it increased agreement in 11% for WordNet similarity, 15% for GloVe relatedness, and 16% for PMI word association strength.

These results confirm the effectiveness of semantic similarity and association strength as indicators of clinical conditions. Moreover, the results suggest that these measures also capture the progression of these conditions and changes in strategies adopted for vocabulary production (Brandt and Manning, 2009), since aMCD can progress to mMCD, which may evolves to others, such as AD and Parkinson disease.

5 Conclusions and Future Work

In this paper we examined the use of three similarity measures (association strength, semantic similarity, and semantic relatedness) for detection of switches in SVF tests, and their effectiveness in detecting clinical conditions. Random forest classifiers trained using the predicted switches were able to successfully identify clinical conditions, and in a fine-grained evaluation were particularly effective for distinguishing the control from clinical group. Our results also outperformed the graph-based approach used by Bertola et al. (2014b) over the same dataset.

Future work includes investigation of the accuracy of these methods for different clinical conditions, and languages. However, the results obtained here show the potential of the method as a tool to help health professionals in diagnosing clinical groups.

	(a) CI			(b) aMCD		
	WordNet	Glove	PMI	WordNet	Glove	PMI
ψ_{global}	0.64 (0.22)	0.66 (0.19)	0.66 (0.19)	0.44 (0.26)	0.56 (0.28)	**0.66 (0.28)**
ψ_1	0.65 (0.21)	0.71 (0.17)	0.68 (0.20)	**0.68 (0.25)**	0.50 (0.29)	**0.65 (0.27)**
ψ_2	0.66 (0.22)	0.66 (0.19)	0.70 (0.18)	0.50 (0.30)	0.60 (0.27)	**0.65 (0.27)**
ψ_3	**0.75 (0.19)**	0.68 (0.18)	0.66 (0.20)	0.59 (0.27)	0.58 (0.30)	0.57 (0.29)
ψ_{vot1}	**0.74 (0.17)**	0.71 (0.18)	0.62 (0.20)	**0.63 (0.27)**	**0.62 (0.27)**	0.46 (0.28)
ψ_{vot2}	**0.72 (0.19)**	0.55 (0.21)	0.69 (0.20)	**0.64 (0.28)**	0.45 (0.28)	**0.63 (0.26)**
ψ_{vot3}	**0.72 (0.18)**	0.62 (0.18)	**0.76 (0.14)**	0.61 (0.28)	0.40 (0.28)	0.54 (0.29)
GS	0.68 (0.17)			0.58 (0.27)		

	(c) mMCD			(d) AD		
	WordNet	Glove	PMI	WordNet	Glove	PMI
ψ_{global}	0.60 (0.27)	0.55 (0.27)	0.54 (0.30)	0.87 (0.17)	0.78 (0.24)	0.80 (0.23)
ψ_1	0.56 (0.30)	**0.75 (0.26)**	0.66 (0.28)	0.71 (0.25)	0.81 (0.21)	0.76 (0.22)
ψ_2	0.65 (0.28)	**0.70 (0.25)**	0.65 (0.27)	0.81 (0.21)	0.83 (0.19)	0.77 (0.25)
ψ_3	**0.71 (0.25)**	0.51 (0.27)	**0.68 (0.28)**	**0.91 (0.15)**	0.85 (0.24)	0.82 (0.20)
ψ_{vot1}	**0.70 (0.26)**	0.60 (0.30)	0.56 (0.28)	0.87 (0.22)	0.86 (0.20)	0.78 (0.23)
ψ_{vot2}	**0.70 (0.26)**	0.46 (0.26)	0.64 (0.24)	0.89 (0.16)	0.77 (0.22)	0.77 (0.21)
ψ_{vot3}	0.67 (0.24)	0.59 (0.25)	**0.73 (0.21)**	0.87 (0.18)	0.84 (0.21)	**0.93 (0.13)**
GS	0.67 (0.24)			0.82 (0.22)		

Table 1: Average scores and standard deviation for random forest classifiers trained to distinguish control from clinical groups. Switch detection with different sources of similarity (WordNet, GloVe and PMI) as well as gold standard taxonomy (GS). Control vs. Cognitive Impairment (CI), Control vs. Amnestic Mild Cognitive Deficit (aMCD), Control vs. Multi-domain Mild Cognitive Deficit (mMCD) and Control vs. Alzheimer's Disease (AD)

	CTRL	aMCD	mMCD	AD
CTRL	**0.27**	0.21	0.20	0.20
aMCD		0.22	0.19	0.19
mMCD			0.23	0.20
AD				0.24

Table 2: Jaccard index for vocabulary agreement between groups

Acknowledgments

We would like to thank Laiss Bertola for sharing the dataset. This research was partly supported by CNPq (projects 312114/2015-0, 423843/2016-8, and 133629/2018-0) and BEWARE (project 1610378).

References

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)* pages 289–300.

Laiss Bertola, Maria Luiza Cunha Lima, Marco A. Romano-Silva, Edgar N. de Moraes, Breno Satler Diniz, and Leandro F. Malloy-Diniz. 2014a. Impaired generation of new subcategories and switching in a semantic verbal fluency test in older adults with mild cognitive impairment. *Frontiers in Aging Neuroscience* 6. https://doi.org/10.3389/fnagi.2014.00141.

Laiss Bertola, Natalia B. Mota, Mauro Copelli, Thiago Rivero, Breno Satler Diniz, Marco A. Romano-Silva, Sidarta Ribeiro, and Leandro F. Malloy-Diniz. 2014b. Graph analysis of verbal fluency test discriminate between patients with alzheimer's disease, mild cognitive impairment and normal elderly controls. *Frontiers in Aging Neuroscience* 6. https://doi.org/10.3389/fnagi.2014.00185.

W. A. Bousfield and C. H. W. Sedgewick. 1944. An analysis of sequences of restricted associative responses. *The Journal of General Psychology* 30(2):149–165. https://doi.org/10.1080/00221309.1944.10544467.

Jason Brandt and Kevin J Manning. 2009. Patterns of word-list generation in mild cognitive impairment and alzheimer's disease. *The Clinical Neuropsychologist* 23(5):870–879.

Leo Breiman. 2001. Random forests. *Machine learning* 45(1):5–32.

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Comput. Linguist.* 16(1):22–29. http://dl.acm.org/citation.cfm?id=89086.89095.

Jonas Jardim de Paula, Laiss Bertola, Rafaela Teixeira Ávila, Lafaiete Moreira, Gabriel Coutinho, Edgar Nunes de Moraes, Maria Aparecida Camargos Bicalho, Rodrigo Nicolato, Breno Satler

Diniz, and Leandro Fernandes Malloy-Diniz. 2013. Clinical applicability and cutoff values for an unstructured neuropsychological assessment protocol for older adults with low formal education. *PLoS ONE* 8(9):e73167. https://doi.org/10.1371/journal.pone.0073167.

Christiane Fellbaum. 1998. *WordNet*. Wiley Online Library.

Evrim Gocer March and Philippa Pattison. 2006. Semantic verbal fluency in alzheimer's disease: approaches beyond the traditional scoring system. *Journal of Clinical and Experimental Neuropsychology* 28(4):549–566.

Julie D Henry and Louise H Phillips. 2006. Covariates of production and perseveration on tests of phonemic, semantic and alternating fluency in normal aging. *Aging, Neuropsychology, and Cognition* 13(3-4):529–551.

Nicklas Linz, Johannes Tröger, Jan Alexandersson, and Alexandra König. 2017. Using neural word embeddings in the analysis of the clinical semantic verbal fluency task. In *IWCS 2017–12th International Conference on Computational Semantics–Short papers*.

Guy McKhann, David Drachman, Marshall Folstein, Robert Katzman, Donald Price, and Emanuel M. Stadlan. 1984. Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. *Neurology* 34(7):939–944.

Jane Morris and Graeme Hirst. 1991. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational linguistics* 17(1):21–48.

Natalia B Mota, Nivaldo AP Vasconcelos, Nathalia Lemos, Ana C Pieretti, Osame Kinouchi, Guillermo A Cecchi, Mauro Copelli, and Sidarta Ribeiro. 2012. Speech graphs provide a quantitative measure of thought disorder in psychosis. *PloS one* 7(4):e34928.

Kelly J Murphy, Jill B Rich, and Angela K Troyer. 2006. Verbal fluency patterns in amnestic mild cognitive impairment are characteristic of alzheimer's type dementia. *Journal of the International Neuropsychological Society* 12(4):570–574.

Serguei VS Pakhomov and Laura S Hemmy. 2014. A computational linguistic measure of clustering behavior on semantic verbal fluency task predicts risk of future dementia in the nun study. *Cortex* 55:97–106.

Seija Pekkala, Martin L. Albert, Avron Spiro III, and TIMO Erkinjuntti. 2008. Perseveration in alzheimer's disease. *Dementia and geriatric cognitive disorders* 25(2):109–114.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics (ACL). https://doi.org/10.3115/v1/d14-1162.

Jacob Perkins. 2010. *Python text processing with NLTK 2.0 cookbook*. Packt Publishing Ltd.

Sarah E Price, Glynda J Kinsella, Ben Ong, Elsdon Storey, Elizabeth Mullaly, Margaret Phillips, Lanki Pangnadasa-Fox, and Diana Perre. 2012. Semantic verbal fluency strategies in amnestic mild cognitive impairment. *Neuropsychology* 26(4):490.

Nadine Raoux, Hélène Amieva, Mélanie Le Goff, Sophie Auriacombe, Laure Carcaillon, Luc Letenneur, and Jean-François Dartigues. 2008. Clustering and switching processes in semantic verbal fluency in the course of alzheimer's disease subjects: Results from the paquid longitudinal study. *Cortex* 44(9):1188–1196.

Mark Rosenstein, Peter Foltz, Anja Vaskinn, and Brita Elvevåg. 2015. Practical issues in developing semantic frameworks for the analysis of verbal fluency data: A norwegian data case study. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. Association for Computational Linguistics, Denver, Colorado, pages 124–133. http://www.aclweb.org/anthology/W15-1215.

Vanessa Taler and Natalie A Phillips. 2008. Language performance in alzheimer's disease and mild cognitive impairment: a comparative review. *Journal of clinical and experimental neuropsychology* 30(5):501–556.

Angela K Troyer, Morris Moscovitch, Gordon Winocur, Michael P Alexander, and Don Stuss. 1998. Clustering and switching on verbal fluency: the effects of focal frontal- and temporal-lobe lesions. *Neuropsychologia* 36(6):499–504. https://doi.org/10.1016/s0028-3932(97)00152-8.

Berndt Winblad, Katie Palmer, Miia Kivipelto, V Jelic, Laura Fratiglioni, L-O Wahlund, A Nordberg, L Bäckman, Michael Albert, O Almkvist, et al. 2004. Mild cognitive impairment–beyond controversies, towards a consensus: report of the international working group on mild cognitive impairment. *Journal of internal medicine* 256(3):240–246.

Qianhua Zhao, Qihao Guo, and Zhen Hong. 2013. Clustering and switching during a semantic verbal fluency test contribute to differential diagnosis of cognitive impairment. *Neuroscience Bulletin* 29(1):75–82. https://doi.org/10.1007/s12264-013-1301-7.