Using Neural Word Embeddings in the Analysis of the Clinical Semantic Verbal Fluency Task

Nicklas Linz, Johannes Tröger and Jan Alexandersson
German Research Center for Artificial Intelligence (DFKI), Germany
<f>.<l>@dfki.de

Alexandra König
Memory Center, CoBTeK - IA CHU Université Côte d'Azur, France
akonig03@gmail.com

Abstract

The *Semantic Verbal Fluency Task* is a common neuropsychological assessment for cognitive disorders: patients are prompted to name as many words from a semantic category as possible in a time interval; the count of correctly named concepts is assessed. Patients often organise their retrieval around semantically related clusters. The definition of clusters is usually based on hand-made taxonomies and the patient's performance is manually evaluated. In order to overcome limitations of such an approach, we propose a statistical method using distributional semantics. Based on transcribed speech samples from 100 French elderly, 53 diagnosed with *Mild Cognitive Impairment* and 47 healthy, we used distributional semantic models to cluster words in each sample and compare performance with a taxonomic baseline approach in a realistic classification task. The distributional models outperform the baseline. Comparing different linguistic corpora as basis for the models, our results indicate that models trained on larger corpora perform better.

1 Introduction

Verbal fluency is amongst the most widely adapted neuropsychological standard tests and is routinely applied in the asessment of neurocognitive disorders. Its subform, category fluency or *semantic verbal fluency* (SVF), demands the assessed person to produce as many different items from a given category as possible within a given time interval, e.g., "as many animals as possible in 60 seconds". A substantial number of clinical studies confirm the discriminative power of SVF for brain pathologies including Alzheimer's disease (AD) (Pakhomov et al., 2016; Raoux et al., 2008; Auriacombe et al., 2006), AD's probable predecessor amnestic *mild cognitive impairment* (MCI), schizophrenia (Robert et al., 1998), as well as focal brain lesions (Troyer et al., 1998). In order to differentiate between multiple pathologies, semantic measures have been established which serve as additional markers next to the raw fluency word count (Gruenewald and Lockhead, 1980; Troyer et al., 1997). There is a broad agreement that these semantic measures serve as indicators for underlying cognitive processes. On a behavioural level, words are spoken in spurts, forming temporal clusters. A cluster is followed by a pause, implying (1) the lexical search between clusters, and (2) retrieval of words within a cluster. On a cognitive level, this is interpreted as follows: executive search processes happen between temporal clusters, (1) *switching*, and semantic memory retrieval processes happen within temporal clusters, (2) *clustering*.

(*cat* - *dog*) - (*cow* - *horse*)
$(Cluster_1)\ Switch_1\ (Cluster_2)$

Temporal cluster are closely related to semantic clusters, as "words that comprise these temporal clusters tend to be semantically related" (Troyer et al., 1997, p. 139). Traditionally, semantic clusters are defined by predefined semantic subcategories. After clustering the words, the clusters' mean size and the number of switches between clusters are computed.

However, multiple studies investigating the same subject group report a great variance of cluster sizes and switch counts. This can be explained through the subjective clustering criterion (Troyer et al., 1997) which leaves some room for interpretation regarding the clustering and thereby directly affecting both measures, switches and cluster size. Statistical semantic analysis automatically and reliably providing clusters is a powerful solution to this problem.

This paper explores the possibility of using distributional semantics in the analysis of SVF tasks with a focus on clustering and switching patterns. This is in contrast to *taxonomic models* which are based on predefined subcategories and might not be able to capture the full complexity of semantic connections made by humans. We investigate the application and performance of word2vec (Mikolov et al., 2013) by which words are embedded into a vector space and where the cosine distance in this space is used as a metric for semantic similarity. This allows for an automatic identification of semantic clusters as well as the computation of switches and cluster size. To indicate the feasibility of this approach within the particular scenario of automated SVF analysis for clinical MCI detection, we compare a set of statistical classification experiments building upon multiple variations of word2vec models to an implementation of the taxonomic approach provided by Troyer et al. (1997).

2 Related Work

Recently, first computational approaches to analyse SVF have been proposed (Woods et al., 2016). The classical measure for SVF performance is word count per minute; sometimes the one minute is split into four 15s time frames. In qualitative analysis of SVF performance this count can be modelled as a combination of two components: the mean cluster size and the number of switches between clusters. The two measures relate to the word count as depicted below; The semantic clustering criterion is the main determiner for both measures. The following section will briefly discuss the two concurring approaches for semantic clustering: taxonomy/ subcategory-based semantic clustering and statistical clustering/ chaining.

$$\text{Word Count} = \text{Mean Cluster Size} \times (\text{Number of Switches} + 1)$$

2.1 Subcategory-based clustering

Troyer et al. (1997) first described a taxonomy-based semantic clustering approach, which despite obvious shortcomings is still extremely popular within clinical research (Troyer et al., 1998; Gomez and White, 2006; Bonner et al., 2010). In this approach words, i.e., animals, can belong to one or more predefined subcategories. The categories are based on living environment, zoological categories, and human use. A cluster is then defined as successively generated words belonging to the same subcategory. If a word could be assigned to two clusters, meaning it is part of the subcategory of the previous and the next cluster, it is counted as belonging to both. In case one cluster is contained by another one, only the bigger one is scored. Adaptations have been suggested, which extend the inclusion rules (Ledoux et al., 2014), the minimal cluster size (Robert et al., 1998), or the handling of repetitions and intrusions (Mueller et al., 2015). However, the fundamental mechanisms remain the same and some prominent limitations are: (1) recognising non-category based associations is not catered for: phonemically similar words (e.g. *donkey & monkey*) or animals that occur together in popular culture (e.g. *panther*, *crane & aardvark*, as in the cartoon series *The Pink Panther*); (2) human-made taxonomies are error prone and likely to be incomplete. In the Troyer et al. (1997) system, there is only one category for *water animals* and therefore, *frog* and *dolphin* appear in the same semantic cluster which may not capture the differences between both animals well; (3) there is a high effort to build a model for a new category which leads to usage within a single category. However, availability of different semantic categories (e.g., tools & supermarket) is of high clinical value for re-testing patients as it prevents confounding training effects.

For a detailed discussion, see Woods et al. (2016).

2.2 Statistical clustering and chaining

To avoid the above-mentioned shortcomings, statistical methods have been applied in order to obtain semantic clusters. However, careful revision of these approaches reveals that many do not actually implement semantic clustering, but rather what we would call *semantic chaining*. In semantic chains, the semantic chain adherence decision is solely based on the previous word.

chain: (*cat* - *dog* - *wolf*) - (*cow*) vs. cluster: (*cat* - *dog*) - (*dog* - *wolf*) - (*cow*)

To our knowledge, Hills et al. (2012) are the only authors who explicitly differentiates between a *static* and *fluid* switch model—a clustering and a chaining model. In this study, the model of Troyer et al. (1997) is used to evaluate clustering and chaining models. A chaining model is built on the basis of the BEAGLE (Jones and Mewhort, 2007) model, a holographic word embedding trained on Wikipedia. To the best of our knowledge, there has been no research into building a clustering model instead of a chaining one based on distributional semantics.

Ledoux et al. (2014) use Latent Semantic Analysis (LSA), based on the LSA website[1] of the Colorado University to compute similarity within clusters and between clusters, to verify their adaptation of Troyer's method. Woods et al. (2016) use Explicit Semantic Analysis (ESA) (Gabrilovich and Markovitch, 2009)—a vector embedding trained on co-occurrence of words in Wikipedia articles—to identify chaining behaviour for different demographics based on pairwise cosine similarity.

In summary, though very powerful for automation of SVF tasks, statistical approaches are only as good as the linguistic material they are trained on. Most approaches discussed above were trained on Wikipedia articles. However, this might not be the most suitable training material for a model that should capture semantic associations made by humans.

Therefore, we compare the discriminative performance of qualitative SVF parameters derived from statistical models based on word2vec to the approach by Troyer et al. (1997) as prominent baseline and subcategory-based approach. Additionally, we investigate the performance of two different text corpora as basis—the common Wikipedia-based approach vs. a less organised and less academic corpus. We also explore the performance of semantic clustering and semantic chaining implementations.

3 Methodology and Results

As authors before us, we are left with a lack of hard metrics to reliably compare the performance of semantic similarity models. Mikolov et al. (2013) propose a benchmark task for evaluating word2vec models, but it is not suitable to judge the applicability to our task. To get around this conundrum, we adhere to the following line of reasoning: Whatever approach performs best at our task at hand, that is discriminating between MCI and healthy subjects, is the approach we should use in analysis. This method is obviously limited by the amount of data that is available for evaluation, and results have to be interpreted with this in mind. We are going to compare two different distributional semantic models with different hyper parameters on a French data set.

3.1 Data

The corpus used consists of 100 samples from older persons: 53 patients diagnosed with MCI (M_{Age}=76.8 ±7.2; 28F/ 25M; $M_{FluencyCount}$=14.63 ± 5.76) and 47 healthy control subjects (HC) with a subjective memory complaint (M_{Age}=72.4 ±7.9; 40F/ 7M; $M_{FluencyCount}$=18.86 ± 5.57). Patients are given 60s to name as many animals as they can. All performances have been recorded and transcribed. The data have been collected in the context of the Dem@Care project (Karakostas et al., 2014).

[1] http://lsa.colorado.edu/

Table 1: Hyper parameters of trained word2vec models (CBoW=Continous Bag of Words; Skip=Skip-Gram Model), classification results for chaining and clustering implementations (Pre=Precision; Rec=Recall; F1=F_1 Score; highest values are marked in bold) and Pearson correlation coefficient between clustering and chaining-based features (switch counts=r_{Switch}; mean cluster size=r_{Size}).

		Hyper parameters			**Chain**			**Cluster**			**Correlation**	
Model	Size	Algorithm	Cutoff	Dimensionality	Pre	Rec	F1	Pre	Rec	F1	r_{Switch}	r_{Size}
FraWac	1.6 B	CBoW	100	200	0.75	0.79	**0.77**	**0.73**	0.80	0.76	0.90	0.87
		Skip	100	200	0.66	0.75	0.70	0.70	0.83	0.76	0.90	0.85
		Skip	100	500	0.72	0.72	0.72	0.68	0.68	0.68	0.91	0.88
		Skip	200	500	0.71	0.72	0.69	0.71	**0.84**	**0.77**	0.90	0.75
Wiki	600 M	CBoW	100	1000	0.67	0.71	0.69	0.67	0.75	0.71	0.99	0.95
		CBoW	200	1000	**0.77**	0.74	0.75	0.71	0.69	0.70	0.96	0.87
		Skip	100	1000	0.68	**0.80**	0.74	0.71	0.72	0.72	0.91	0.84
		Skip	200	1000	0.70	0.74	0.72	0.67	0.76	0.71	0.84	0.77
Troyer	-	-	-	-	-	-	-	0.71	0.74	0.72	-	-

3.2 Models

We compare a set of models, all of them learned using word2vec (Mikolov et al., 2013). Word2vec is a word-embedding based on a shallow, two-layer neural network trained to embed words in a vector space, where the cosine distance is a measure for semantic similarity. We compare models trained on two different linguistic corpora: (1) models based on the FraWac corpus (Baroni et al., 2009), a large corpus collected by a web crawler and (2) models based on a dump of the French Wikipedia. Pre-trained models are taken from here[2]. All varying word2vec hyper parameters are reported in Table 1. For all models, the context window was set to 5 tokens and negative sampling was used.

3.3 Clustering and Chaining

On the basis of these models and the cosine distance in the resulting vector space we compute semantic clusters/chains in the following way:

Let a_1, a_2, …, a_n be the sequence of animals produced by patient p. Let $\vec{a_1}$, $\vec{a_2}$, …, $\vec{a_n}$ be their representations in the vector space and let $a_1, \ldots, a_{n-1}$ form a semantic cluster/chain. a_n is part of this cluster/chain if

Cluster

$$\left|\frac{\langle \vec{\mu}, \vec{a_n} \rangle}{||\vec{\mu}|| \cdot ||\vec{a_n}||}\right| > \delta_p$$

Chain

$$\left|\frac{\langle \vec{a_{n-1}}, \vec{a_n} \rangle}{||\vec{a_{n-1}}|| \cdot ||\vec{a_n}||}\right| > \delta_p$$

with

$$\vec{\mu} = \frac{1}{n-1} \cdot \sum_{\vec{x} \in \{\vec{a_1}, \ldots, \vec{a_{n-1}}\}} \vec{x}$$

$$\delta_p = \frac{n!}{(n-2)!} \cdot \sum_{\vec{x}, \vec{y} \in \{\vec{a_1}, \ldots, \vec{a_n}\}} \left|\frac{\langle \vec{x}, \vec{y} \rangle}{||\vec{x}|| \cdot ||\vec{y}||}\right|$$

One of the main problems of using distributional semantic models to determine clusters/chains is finding a sensible cut-off value δ. We decided to use the mean distance between any animal produced by a subject. An ad-hoc global cut-off value would be hard to determine, since similarity scores tend to vary a lot.

[2] http://fauconnier.github.io/

3.4 Classification

We train different classifiers, one for each model using Support Vector Machines (SVMs) with a radial basis kernel. This is mainly because we only have two features (Hsu et al., 2010). Moreover, since our data set is small, we perform a stratified 10-fold cross validation. As features we use the mean size of clusters identified and the number of switches between clusters. For results, see Table 1.

4 Discussion

This paper set out to compare the discriminative performance of qualitative SVF parameters derived from statistical models based on neural word embeddings with the traditional subcategory-based approach by Troyer et al. (1997). Therefore, we implemented Troyer's approach as a baseline deriving the semantic clustering criterion from predefined subcategories. We compared this to a group of statistical approaches based on a patient-dependent clustering criterion derived from word2vec models. Through both approaches, we automatically calculated mean cluster size and number of switches based on transcripts of two groups' SVF recordings: MCI and healthy controls. In order to examine both approaches' feasibility within the given scenario, we trained classifiers, showing results clearly in favour of the statistically derived feature set. This is in line with reported feasibility benefits of this approach (Woods et al., 2016; Hills et al., 2012). However, to the best of our knowledge, no study so far compared both approaches based on the discriminative performance they achieve, given a clinical classification scenario; so far, either one of both approaches has been used to validate the features derived by the other approach and vice versa. Nonetheless, maybe the most straight forward way of comparing both approaches is by applying them to a relevant clinical scenario—which SVF has actually been designed and used for—and deciding based on their performance in the classification task at hand.

Additionally, we investigate the performance of two different text corpora as basis for the word2vec models. Our results show that the classifiers using features based on the FraWac corpus models (Baroni et al., 2009) achieve higher F1 scores than the ones based on the Wikipedia models. Although it is difficult to derive a conclusion from this rather exploratory result, possible explanations might be that the FraWac corpus is simply larger, or that it represents a less (artificially) academic and therefore more natural linguistic resource.

Finally, considering different effects through semantic chaining vs. semantic clustering, we yield no interpretable results favouring either one of the implementations. Our experiments yield throughout high correlation indices between both implementations across both SVF dependent variables/ features: switch count & mean cluster size. This is in line with Hills et al. (2012), who also find no clear pattern.

5 Conclusion

To conclude, this paper presents a clinical application of neural word embeddings rendering a statistical approach to the traditionally manual analysis of semantic verbal fluency tasks. Our results demonstrate the feasibility and therefore economic validity of such an approach, having especially relevant implications for remote automatic screening applications like in Tröger et al. (2017). The strong dependency between both qualitative SVF measures, switch count & mean cluster size, and simple word count performance, still remains a challenge for understanding their respective diagnostic values. Future research should therefore explore measures which are based on the here-presented encouraging approach and which go beyond the triangular relation of SVF switches, cluster size and word count.

Acknowledgements

This research was partially funded by the EIT Digital Wellbeing Activity 17074, *ELEMENT*. The data was collected during the EU FP7 *Dem@Care* project, grant agreement 288199.

References

Auriacombe, S., N. Lechevallier, H. Amieva, S. Harston, N. Raoux, and J.-F. Dartigues (2006). A Longitudinal Study of Quantitative and Qualitative Features of Category Verbal Fluency in Incident Alzheimer's Disease Subjects: Results from the PAQUID Study. *Dementia and geriatric cognitive disorders 21*(4), 260–266.

Baroni, M., S. Bernardini, A. Ferraresi, and E. Zanchetta (2009). The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. *Language Resources and Evaluation 43*(3), 209–226.

Bonner, M. F., S. Ash, and M. Grossman (2010). The New Classification of Primary Progressive Aphasia into Semantic, Logopenic, or Nonfluent/Agrammatic Variants. *Current Neurology and Neuroscience Reports 10*(6), 484–490.

Gabrilovich, E. and S. Markovitch (2009, March). Wikipedia-based Semantic Interpretation for Natural Language Processing. *J. Artif. Int. Res. 34*(1), 443–498.

Gomez, R. G. and D. A. White (2006). Using verbal fluency to detect very mild dementia of the Alzheimer type. *Archives of Clinical Neuropsychology 21*(8), 771 – 775.

Gruenewald, P. J. and G. R. Lockhead (1980). The Free Recall of Category Examples. *Journal of Experimental Psychology: Human Learning and Memory 6*, 225–240.

Hills, T. T., M. N. Jones, and P. M. Todd (2012, Apr). Optimal Foraging in Semantic Memory. *Psychol Rev 119*(2), 431–440.

Hsu, C.-W., C.-C. Chang, and C. jen Lin (2010). A Practical Guide to Support Vector Classification.

Jones, M. N. and D. J. Mewhort (2007, Jan). Representing Word Meaning and Order Information in a Composite Holographic Lexicon. *Psychol Rev 114*(1), 1–37.

Karakostas, A., A. Briassouli, K. Avgerinakis, I. Kompatsiaris, and M. Tsolaki (2014). The Dem@Care Experiments and Datasets: a Technical Report. Technical report, Centre for Research and Technology Hellas (CERTH).

Ledoux, K., T. D. Vannorsdall, E. J. Pickett, L. V. Bosley, B. Gordon, and D. J. Schretlen (2014). Capturing additional information about the organization of entries in the lexicon from verbal fluency productions. *Journal of Clinical and Experimental Neuropsychology 36*(2), 205–220.

Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado, and J. Dean (2013). Distributed Representations of Words and Phrases and their Compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems 26*, pp. 3111–3119. Curran Associates, Inc.

Mueller, K. D., R. L. Koscik, A. LaRue, L. R. Clark, B. Hermann, S. C. Johnson, and M. A. Sager (2015). Verbal Fluency and Early Memory Decline: Results from the Wisconsin Registry for Alzheimer's Prevention. *Archives of Clinical Neuropsychology 30*(5), 448.

Pakhomov, S. V., L. Eberly, and D. Knopman (2016). Characterizing cognitive performance in a large longitudinal study of aging with computerized semantic indices of verbal fluency. *Neuropsychologia 89*, 42 – 56.

Raoux, N., H. Amieva, M. L. Goff, S. Auriacombe, L. Carcaillon, L. Letenneur, and J.-F. 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.

Robert, P. H., V. Lafont, I. Medecin, L. Berthet, S. Thauby, C. Baudu, and G. Darcourt (1998). Clustering and switching strategies in verbal fluency tasks: Comparison between schizophrenics and healthy adults. *Journal of the International Neuropsychological Society 4*(6), 539–546.

Tröger, J., N. Linz, J. Alexandersson, A. König, and P. Robert (2017). Automated Speech-based Screening for Alzheimer's Disease in a Care Service Scenario. In *Proceedings of the 11th EAI International Conference on Pervasive Computing Technologies for Healthcare*, PervasiveHealth '17. ICST. in press.

Troyer, A. K., M. Moscovitch, and G. Winocur (1997). Clustering and Switching as Two Components of Verbal Fluency: Evidence From Younger and Older Healthy A dults. *neuropsychology 11*(1), 138.

Troyer, A. K., M. Moscovitch, G. Winocur, M. P. Alexander, and D. Stuss (1998). Clustering and switching on verbal fluency: the effects of focal frontal- and temporal-lobe lesions. *Neuropsychologia 36*(6), 499 – 504.

Troyer, A. K., M. Moscovitch, G. Winocur, L. Leach, and M. Freedman (1998). Clustering and switching on verbal fluency tests in Alzheimer's and Parkinson's disease. *Journal of the International Neuropsychological Society 4*(2), 137–143.

Woods, D. L., J. M. Wyma, T. J. Herron, and E. W. Yund (2016, 12). Computerized Analysis of Verbal Fluency: Normative Data and the Effects of Repeated Testing, Simulated Malingering, and Traumatic Brain Injury. *PLOS ONE 11*(12), 1–37.