Vector-space topic models for detecting Alzheimer's disease

Maria Yancheva
Department of Computer Science,
University of Toronto
Toronto, Ontario, Canada
yancheva@cs.toronto.edu

Frank Rudzicz
Toronto Rehabilitation Institute; and
Department of Computer Science,
University of Toronto
Toronto, Ontario, Canada
frank@cs.toronto.edu

Abstract

Semantic deficit is a symptom of language impairment in Alzheimer's disease (AD). We present a generalizable method for automatic generation of information content units (ICUs) for a picture used in a standard clinical task, achieving high recall, 96.8%, of human-supplied ICUs. We use the automatically generated topic model to extract semantic features, and train a random forest classifier to achieve an F-score of 0.74 in binary classification of controls versus people with AD using a set of only 12 features. This is comparable to results (0.72 F-score) with a set of 85 *manual* features. Adding semantic information to a set of standard lexicosyntactic and acoustic features improves F-score to 0.80. While control and dementia subjects discuss the same topics in the same contexts, controls are more informative per second of speech.

1 Introduction

Alzheimer's disease (AD) is the most common cause of neurodegenerative dementia, and affects more than 24.3 million people worldwide (Ballard et al., 2011). Importantly, early detection enables some therapeutic intervention and disease-modifying treatment (Sperling et al., 2011). Longitudinal studies of people with autopsy-confirmed AD indicate that linguistic changes are detectable in the prodromal stages of the disease; these include a decline in grammatical complexity, word-finding difficulties, and semantic content deficiencies, such as low idea density (i.e., the ratio of semantic units to the total number of words in a speech sample), and low efficiency (i.e., the rate of semantic units over the duration of the speech sample) (Bayles and Kaszniak, 1987; Snowdon et al., 1996; Le et al., 2011; Ahmed et al., 2013b). In the present study, we investigate methods of automatically assessing the semantic content of speech, and use it to distinguish people with AD from healthy older adults.

A standard clinical task for eliciting spontaneous speech, with high sensitivity to language in early AD, is picture description. In it, a participant is asked to provide a free-form verbal description of a visual stimulus (Goodglass and Kaplan, 1983; Bayles and Kaszniak, 1987). The picture is associated with a set of human-supplied information content units (**hsICUs**) representing components of the image, such as subjects, objects, locations, and actions (Croisile et al., 1996). The semantic content of the elicited speech can then be scored by counting the hsICUs present in the description. Previous studies found that, even in the earliest stages, descriptions by those with AD are less informative compared to those of healthy older adults, producing fewer information units out of a pre-defined list of units, and having less relevant content and lower efficiency (Hier et al., 1985; Croisile et al., 1996; Giles et al., 1996; Ahmed et al., 2013a).

Using a pre-defined list of annotated hsICUs is subject to several limitations: (i) it is *subjective* — different authors use a different number of hsICUs for the same picture (e.g., from 7 to 25 for *Cookie Theft* in the Boston Diagnostic Aphasia Examination (BDAE)) (Hier et al., 1985; Croisile et al., 1996; Forbes-McKay and Venneri, 2005; Lai et al., 2009); (ii) it *may not be optimal* for detecting linguistic impairment — the manually-annotated hsICUs are neither exhaustive of all details present in the picture, nor necessarily reflective of the content units which differ most across groups; (iii) it is *not generalizable* — hsICUs are specific to a particular picture, and new visual stimuli (e.g.,

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 2337–2346,
Berlin, Germany, August 7-12, 2016.

required for longitudinal assessments) need to be annotated manually. In addition to requiring time and effort, this may result in inconsistencies, since the methodology for identifying hsICUs was never clearly defined in previous work.

Automatic scoring of semantic content in speech to detect cognitive impairment has so far required manual hsICUs. Hakkani-Tür et al. (2010) used unigram recall among hsICUs in the Western Aphasia Battery's *Picnic* picture (Kertesz, 1982) and obtained a correlation of 0.93 with manual hsICU counts. Pakhomov et al. (2010) counted N-grams ($N = 1, 2, 3, 4$) extracted from a list of hsICUs for the *Cookie Theft* picture to assess semantic content in the speech of patients with frontotemporal lobar degeneration. Fraser et al. (2016) counted instances of lexical tokens extracted from a list of hsICUs, using dependency parses of *Cookie Theft* picture descriptions, and combined them with other lexicosyntactic and acoustic features to obtain classification accuracy of 81.9% in identifying people with AD from controls. While those automated methods for scoring the information content in speech used manual hsICUs, we have found none that attempted to produce ICUs automatically.

In this paper, we present a generalizable method for automatically generating information content units for any given picture (or spontaneous speech task), using reference speech. Since clinical data can be sparse, we present a method for building word vector representations using a large general corpus, then augment it with local context windows from a smaller clinical corpus. We evaluate the generated ICUs by computing recall of hsICUs and use the constructed topic models to compare the speech of participants with and without dementia, and compute topic alignment. Second, we automatically score new picture descriptions by learning semantic features extracted from these generated ICU models, using a random forest classifier; we assess performance with recall, precision, and F-score. Third, we propose a set of clinically-relevant features for identifying AD based on differences in topic, topic context, idea density and idea efficiency.

2 Methodology

2.1 Data

DementiaBank is one of the largest public, longitudinal datasets of spontaneous speech from individuals with and without dementia. It was collected at the University of Pittsburgh (Becker et al., 1994) and contains verbal descriptions of the standard *Cookie Theft* picture (Goodglass and Kaplan, 1983), along with manual transcriptions.

In our study, we use 255 speech samples from participants diagnosed with probable or possible AD (collectively referred to as the 'AD' class), and 241 samples from healthy controls (collectively referred to as the 'CT' class), see Table 1. We remove all CHAT-format annotations (MacWhinney, 2015), filled pauses (e.g., '*ah*' and '*um*'), phonological fragments (e.g., '*b b boy*' becomes '*boy*'), repairs (e.g., '*in the in the kitchen*' becomes '*in the kitchen*'), non-standard forms (e.g., '*gonna*' becomes '*going to*'), and punctuation (e.g., commas are removed). These corrections are all provided in the database. We ignore transcripts of the investigator's speech, as irrelevant. Subject data were randomly partitioned into training, validation, and test sets using a 60-20-20 split.

Table 1: Distribution of dataset transcriptions.

Class	Subjects	Samples	Tokens
AD	168	255	24,753
CT	98	241	26,654
Total	266	496	51,407

2.2 Human-supplied ICUs (hsICUs)

We combine all hsICUs in previous work for the *Cookie Theft* picture (Hier et al., 1985; Croisile et al., 1996; Forbes-McKay and Venneri, 2005; Lai et al., 2009) with hsICUs obtained from a speech language pathologist (SLP) at the Toronto Rehabilitation Institute (TRI). The annotations of the SLP overlap completely with previously identified hsICUs, except for one (*apron*). The first three columns of Table 2 summarize these manually-produced hsICUs.

2.3 Automatic generation of ICUs

Our novel method of identifying ICUs is based on simple topic modelling using clusters of global word-vector representations from picture descriptions. First, we train a word-vector model on a large normative general-purpose corpus, allowing us to avoid sparsity in the clinical data's word-word co-occurrence matrix. Then, we extract the vector representations of words in the Dementia-

Table 2: Information units above the double line are human-supplied ICUs (hsICUs) found in previous work, except those marked with † which were annotated by an SLP for this study; those below are additionally analyzed. Over 1,000 clustering configurations based on word vectors extracted from *Control* and *Dementia* reference transcriptions, μ is the mean of the scaled distance (Eq. 1) of each hsICU to its closest cluster centroid, σ is the standard deviation, and $\delta = (\mu_{dementia} - \mu_{control})$. Statistical significance of δ was tested using an independent two-sample, two-tailed t-test; *** $= p < .001$, ** $= p < .01$, * $= p < .05$, ns = not significant.

			Control		**Dementia**			
Type	**ID**	**hsICU**	μ	σ	μ	σ	δ	p
Subject	S1	boy	-0.510	0.102	-0.860	0.204	**-0.350**	***
Subject	S2	girl	-0.357	0.203	-0.545	0.284	**-0.187**	***
Subject	S3	woman	0.171	0.468	0.140	0.433	-0.031	ns
Subject	S4	mother	-0.533	0.206	-0.187	0.300	**0.345**	***
Place	P1	kitchen	0.667	0.650	0.901	0.710	**0.234**	***
Place	P2	exterior	1.985	0.601	1.947	0.530	-0.039	ns
Object	O1	cookie	-1.057	0.221	-0.943	0.230	**0.114**	***
Object	O2	jar	0.243	0.486	0.146	0.453	**-0.097**	***
Object	O3	stool	-0.034	0.674	-0.162	0.623	**-0.128**	***
Object	O4	sink	-0.839	0.433	-0.600	0.631	**0.239**	***
Object	O5	plate	0.564	0.593	0.639	0.608	**0.076**	**
Object	O6	dishcloth	*4.509*	1.432	*3.989*	1.154	**-0.521**	***
Object	O7	water	-0.418	0.582	-0.567	0.530	**-0.149**	***
Object	O8	cupboard	0.368	0.613	0.453	0.637	**0.085**	**
Object	O9	window	-0.809	0.425	-0.298	0.452	**0.511**	***
Object	O10	cabinet	2.118	0.556	2.154	0.496	0.036	ns
Object	O11	dishes	0.037	0.503	-0.083	0.406	**-0.120**	***
Object	O12	curtains	-0.596	0.594	0.121	0.707	**0.717**	***
Object	O13	faucet	1.147	0.567	1.016	0.547	**-0.131**	***
Object	O14	floor	-0.466	0.384	-0.932	0.451	**-0.466**	***
Object	O15	counter	0.202	0.427	0.449	0.323	**0.247**	***
Object	O16	apron†	-0.140	0.433	0.181	0.688	**0.321**	***
Action	A1	boy *stealing* cookies	1.219	0.373	0.746	0.462	**-0.473**	***
Action	A2	boy/stool *falling* over	-0.064	0.465	-0.304	0.409	**-0.240**	***
Action	A3	woman *washing* dishes	-0.058	0.539	0.009	0.611	**0.068**	**
Action	A4	woman *drying* dishes	-0.453	0.469	-0.385	0.541	**0.068**	**
Action	A5	water *overflowing* in sink	0.147	0.804	0.282	0.791	**0.135**	***
Action	A6	girl's actions towards boy, girl *asking* for a cookie	0.800	0.555	0.620	0.861	**-0.179**	***
Action	A7	woman *daydreaming*, unaware or unconcerned about overflow	0.049	0.774	0.092	0.561	0.043	ns
Action	A8	dishes already washed *sitting* on worktop	-0.224	0.535	-0.597	0.426	**-0.373**	***
Action	A9	woman being *indifferent* to the children	0.781	0.795	0.881	0.585	**0.100**	**
Relation		brother	2.297	0.510	1.916	0.344	**-0.380**	***
Relation		sister	0.862	0.273	0.737	0.349	**-0.125**	***
Relation		son	2.140	0.443	1.818	0.312	**-0.322**	***
Relation		daughter	0.916	0.356	0.904	0.421	-0.012	ns

Bank corpus, and optionally augment them with local context windows from the clinical dataset.

We use GloVe v1.2 (Pennington et al., 2014) to obtain embedded word representations and train on a combined corpus of Wikipedia 2014[1] + Gigaword 5[2]. The trained model consists of 400,000 word vectors, in 50 dimensions.

Transcriptions in DementiaBank are lowercased and tokenized using NLTK v3.1, and each word token is converted to its vector space representation using the trained GloVe model. There are a total of 26,654 word vectors (1,087 unique vectors) in the control data, and 24,753 (1,131 unique) in the dementia data. Since we aim to construct a model of semantic content, only nouns and verbs are retained prior to clustering. The resulting dataset consists of 9,330 word vectors (801 unique vectors) in the control data, and 8,021 (843 unique) in the dementia data.

We use k-means clustering with whitening, initialization with the Forgy method, and a distortion threshold of 10^{-5} as the stopping condition, where distortion is defined as the sum of the distances between each vector and its corresponding centroid. We train a *control cluster model* on the control training set (see Fig. 1 for a 2D projection of cluster vectors using principal component analysis), and a *dementia cluster model* on the dementia training set. Clusters represent topics, or groups of semantically related word vectors, discussed by the respective group of subjects. While prior work is based on hsICUs that are *expected* to be discussed by *healthy* speakers, we construct a separate cluster model for the control and dementia groups since it is unclear whether the topics discussed by both groups overlap. We vary k $(= 1, 5, 10, 15, 20, 30, 40, 50)$, completing 1,000 runs for each value, and use the Elbow method to select the optimal number of clusters on the respective validation set. The optimal setting, $k = 10$, optimizes the tradeoff between the percentage of variance explained by the clusters, and their total number. The resulting clusters represent topics that can be compared against hsICUs.

3 Experiments

3.1 Recall of hsICUs

In order to assess (i) how well the automatically generated clusters match clinical hsICUs for this

[1]http://dumps.wikimedia.org/enwiki/20140102/

[2]https://catalog.ldc.upenn.edu/LDC2011T07

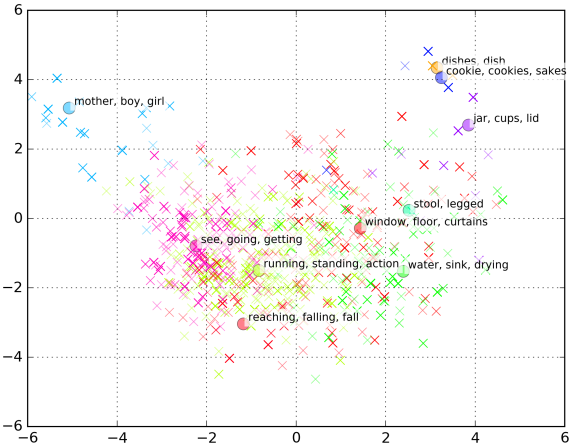

Figure 1: *Control* cluster model. The word vectors belonging to a given cluster are shown in the same colour. The most frequent words in each cluster are displayed.

image, and (ii) how much the two generated topic models differ, we analyze the vector space distance between each hsICU and its closest cluster centroid ($d_{Euclidean}$) in each of the *control* and *dementia* models. Since some clusters are more dispersed than others, we need to scale the distance appropriately. To do so, for each cluster in each model, we compute the mean distortion, μ_{cl}, of the vectors in the cluster, and the associated standard deviation σ_{cl}. For each hsICU vector, we compute the scaled distance between the vector and its closest cluster centroid in each generated model as follows:

$$d_{scaled} = \frac{(d_{Euclidean} - \mu_{cl})}{\sigma_{cl}} \quad (1)$$

The scaled distance is equivalent to the number of standard deviations above the mean — a value below zero indicates hsICUs which are very close to an automatically generated cluster centroid, while a large positive value indicates hsICUs that are far from a cluster centroid. To account for the fact that k-means is a stochastic algorithm, we perform clustering multiple times and average the results. Table 2 shows the mean, μ, and standard deviation, σ, of d_{scaled}, for each hsICU, over 1,000 cluster configurations for each model.

To quantify the recall of hsICUs using each generated cluster model, we consider hsICUs with $\mu \leq 3.0$ to be *recalled* (i.e., the distance to the assigned cluster centroid is not greater than those of 99.7% of the datapoints in the cluster, given a Gaussian distribution of distortion). The recall of

hsICUs, for both the control and dementia models, is 96.8%. Since the optimal number of generated clusters is $k = 10$, while the number of hsICUs is 31, multiple hsICUs can be grouped in related *themes* (e.g., one automatically generated cluster corresponds to the description of animate subjects in the picture, capturing four hsICUs: S1–S4). Both the control and dementia models do not recall hsICU O6, *dishcloth*, which suggests that it is a topic that neither study group discusses. All remaining hsICUs are recalled by both the control and dementia models, indicating that the hsICU topics are discussed by both groups.

However, to assess whether they are discussed to the same *extent*, i.e. to evaluate whether the two topic models differ, we conducted an independent two-sample two-tailed t-test to compare the mean scaled distance, μ, of each hsICU to its closest cluster centroid, in each cluster model (see δ in Table 2). As anticipated, since they involve inference of attention, the control model is better at accounting for the topics of the overflowing sink and the mother's indifference: *overflowing* ($t(1998) = -3.78$, $p < .001$); *sink* ($t(1998) = -9.85$, $p < .001$); *indifferent* ($t(1998) = -3.20$, $p < .01$). While there is no significant difference in the term *woman* between the two groups, the control model predicts the term *mother* better than the dementia model ($t(1998) = -30.05$, $p < .001$). To investigate whether healthy participants are more likely to identify relations between the subjects than participants with cognitive impairment, we repeated the recall experiment with the following new hsICUs: *brother*, *sister*, *son*, *daughter*. Interestingly, the dementia cluster model contains a cluster which aligns significantly more closely, than any in the control model, with all four of these relation words: *brother* ($t(1998) = 19.53$, $p < .001$); *sister* ($t(1998) = 8.93$, $p < .001$); *son* ($t(1998) = 18.78$, $p < .001$). While the control participants mention relation words as often as the participants with dementia[3], the generated cluster models show that the ratio of relation words to non-relation words is higher for the dementia group[4].

The new hsICU, *apron*, which was not identified in previous literature but was labelled by an SLP for this study, is significantly more likely to be discussed by the control population ($t(1998) = -12.46$, $p < .001$), suggesting at the importance of details for distinguishing cognitively impaired individuals. In a similar vein, control participants are significantly more likely to identify objects in the background of the scene, such as the *window* ($t(1998) = -26.04$, $p < .001$), *curtains* ($t(1998) = -24.54$, $p < .001$), *cupboard* ($t(1998) = -3.03$, $p < .01$), or *counter* ($t(1998) = -14.59$, $p < .001$).

3.2 Cluster model alignment

While prior work counted the *frequency* with which fixed topics are mentioned, our data-driven cluster models allow greater exploration of differences between the set of topics discussed by each subject group, and the alignment between them. Since prior work has found that subjects with cognitive impairment produce more irrelevant content, we quantify the amount of dispersion within each cluster through the standard deviation of its distortion and its type-to-token ratio (TTR), as shown in Table 3. Further, we compute directional alignment between pairs of clusters in each model. For each cluster in one model, alignment is determined by computing the closest cluster in the other model for each vector, and taking the majority assignment label (see a in Table 3). To quantify the alignment, the Euclidean distance of each vector to the assigned cluster in the other model is computed, scaled by the mean and standard deviation of the cluster distortion; the mean of the scaled distance, μ_a, is reported in Table 3.

To quantify the alignment of clusters in each model, we consider clusters to be *recalled* if their distance to the closest cluster in the other model is $\mu_a \leq 3$. Notably, all control clusters (C0-C9) are recalled by the dementia model, while one dementia cluster, D7, is not recalled by the control model. This exemplifies the fact that while the dementia group mentions all topics discussed by controls, they also mention a sufficient number of extraneous terms which constitute a new heterogeneous topic cluster, having the highest TTR.

[3] An independent two-sample two-tailed t-test of the effect of group on the *number of occurrences* of each relation word shows no statistical significance: *son* ($t(494) = 0.65$, $p > .05$), *daughter* ($t(494) = 0.63$, $p > .05$), *brother* ($t(494) = 0.97$, $p > .05$), *sister* ($t(494) = 1.65$, $p > .05$).

[4] An independent two-sample two-tailed t-test of the effect of group on this ratio shows a significant difference in the ratio of *sister* to *mother*, with the control group having a lower ratio ($t(494) = -4.10$, $p < .001$).

Table 3: Cluster statistics for control (C*) and dementia (D*) models, with computed cluster alignment. *Cluster words* are the 5 most frequently occurring words. f_{vec} is the fraction of all vectors which belong to the given cluster. μ_{cl} and σ_{cl} are the mean and standard deviation of the cluster distortion. f_n is the fraction of nouns among cluster vectors; $(1 - f_n)$ is the fraction of verbs. *TTR* is the type-to-token ratio. a is the ID of the aligned cluster, and μ_a is the mean scaled distance to the aligned cluster centroid.

	ID	Cluster words	f_{vec}	μ_{cl}	σ_{cl}	f_n	TTR	a	μ_a
Control	C0	window, floor, curtains, plate, kitchen	0.14	5.42	1.18	0.94	0.14	D4	0.69
	C1	dishes, dish	0.04	1.62	1.11	1.00	0.01	D1	0.01
	C2	running, standing, action, hand, counter	0.18	4.97	1.25	0.57	0.22	D8	0.16
	C3	water, sink, drying, overflowing, washing	0.17	5.18	1.13	0.66	0.09	D6	0.04
	C4	stool, legged	0.03	0.53	1.26	0.96	0.01	D4	-0.28
	C5	mother, boy, girl, sister, children	0.11	3.49	1.08	1.00	0.04	D2	-0.08
	C6	cookie, cookies, sakes, cream	0.06	2.00	1.15	1.00	0.01	D0	-0.08
	C7	jar, cups, lid, dried, bowl	0.04	3.88	2.30	0.97	0.04	D5	0.63
	C8	see, going, getting, looks, know	0.18	3.84	1.16	0.38	0.13	D3	0.18
	C9	reaching, falling, fall, summer, growing	0.05	4.18	1.41	0.38	0.16	D8	0.21
Dementia	D0	cookie, cookies, cake, baking, apples	0.07	2.18	0.74	1.00	0.02	C6	0.09
	D1	dishes, dish, eating, bowls, dinner	0.05	1.42	1.72	0.98	0.03	C1	0.05
	D2	boy, girl, mother, sister, lady	0.11	3.63	1.25	0.99	0.05	C5	0.20
	D3	going, see, getting, get, know	0.24	3.67	1.06	0.38	0.11	C8	-0.11
	D4	stool, floor, window, chair, curtains	0.10	5.10	1.00	0.97	0.13	C0	0.08
	D5	jar, cups, jars, dried, honey	0.04	2.00	2.26	0.98	0.03	C7	-0.44
	D6	sink, drying, washing, spilling, overflowing	0.14	5.36	1.20	0.52	0.19	C3	0.36
	D7	mama, huh, alright, johnny, ai	0.01	6.24	1.34	0.95	0.55	C8	**4.13**
	D8	running, fall, falling, reaching, hand	0.18	4.97	1.29	0.47	0.25	C2	0.15
	D9	water, dry, food	0.05	0.39	1.13	1.00	0.01	C3	-0.59

3.3 Local context weighted vectors

Since there is significant overlap in the topics discussed between the control and dementia groups, we proceed by investigating whether the overlapping topics are discussed in the same contexts. To this end, we augment the word vector representations with local context windows from DementiaBank. Each word vector is constructed using a linear combination of its global vector from the trained GloVe model, and the vectors of the $\pm N$ surrounding context words, where each context word is weighted inversely to its distance from the central word:

$$\phi_w = v_w + \sum_{i=-N}^{-1} \alpha_i \times v_i + \sum_{i=1}^{N} \alpha_i \times v_i \qquad (2)$$

Here, ϕ_w is the local-context-weighted vector for word w, v_w is the GloVe vector for word w, v_i is the GloVe vector for word i within the context of w, and α_i is the weighting of word i, inversely and linearly proportional to the distance between context and central word. Following previous work (Fraser and Hirst, 2016), we use a context window of size $N = 3$. We extract local-context-weighted vectors for all control and dementia transcripts, and construct two topic models as before.

To quantify whether the dementia contexts differ significantly from the control contexts for the *same word*, we extract all word usages as local-context-weighted vectors, and find the centroid of the control usages, along with the mean and standard deviation of the control vectors from their centroids. Then, we compute the average scaled Euclidean distance, d_{scaled}, of the dementia vectors from the control centroid, as in Eq. 1. Words with $d_{scaled} > 3$ (i.e., where the dementia context vectors are further from the control centroid than the majority of control context vectors) are considered to have different context usage across the control and dementia groups.

Interestingly, all of the control cluster words are used in the *same contexts* by both healthy participants and those with dementia. However, the *average number of times* these words are used per transcript is significantly higher in the control group (1.07, $s.d. = 0.12$) than in the dementia group (0.77, $s.d. = 0.14$; $t(18) = 1.87$, $p < .05$).

While the two groups discuss the same topics generally and use the same words in the same contexts, not all participants in the dementia group identify all of the control topics or discuss them with the same frequency. A contextual analysis reveals that certain words are discussed in a distinct number of limited contexts, while others are discussed in more varied contexts. For in-

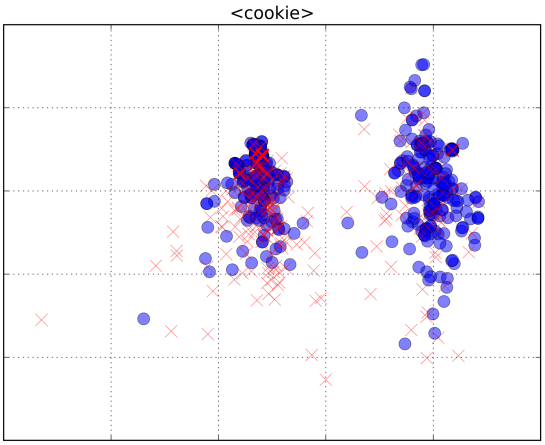

Figure 2: All usages of the word *cookie* in DementiaBank. Control usages are represented with blue circles; dementia with red crosses.

stance, while we identified a control cluster associated with the topic of the *cookie* in Section 3.2, there are two clearly distinct contexts in which this word is used, by both groups, as illustrated in Fig. 2. The two clusters in context space correspond to: (i) the usage of *cookie* in the compound noun phrase *cookie jar*, and (ii) referring to a single *cookie*, e.g. *reaching for a cookie*, *hand her a cookie*, *getting a cookie*.

3.4 Classification

To classify speakers as having AD or not, we extract the following types of features from our automatically-generated cluster models: (i) distance-based metrics for each of the control model clusters, C0–C9, (ii) distance-based metrics for each of the dementia model clusters, D0–D9, (iii) idea density, and (iv) idea efficiency. Given the vectors associated with a transcript's nouns and verbs, feature C_i (and equivalently, D_i) is computed by finding the average scaled distance, d_{scaled} (Eq. 1), of all vectors assigned to cluster C_i. A feature value below zero indicates that the transcript words assigned to the cluster are very well predicted by it (i.e., their distance from the cluster centroid is less than the average cluster distortion). Conversely, clusters which represent topics not discussed in the transcript have large positive feature values. We chose these distance-based metrics to evaluate topic recall in the transcript since a continuous measure is more appropriate for modelling the non-discrete nature of language and semantic similarity. We compute *idea density* as the number of expected topics mentioned[5] divided by the total number of words in the transcript, and *idea efficiency* as the number of expected topics mentioned divided by the total duration of the recording (in seconds). The expected topics used for computation of idea density and idea efficiency are the ICUs from the automatically-produced cluster models.

We perform classification using a random forest, whose parameters are optimized on the validation set, and performance reported on the test set. We vary the following experimental settings: *cluster model* (control; dementia; combined), *feature set* (distance-based; distance-based + idea density + idea efficiency), and *context* (no context; context with $N = 3$). A three-way ANOVA is conducted to examine the effects of these settings on average test F-score. There is a significant interaction between *feature set* and *context*, $F(1, 110) = 9.07$, $p < 0.01$. Simple main effect analysis shows that when using the extended feature set, vectors constructed without local context windows from the clinical dataset yield significantly better results than those with context ($p < 0.001$), but there is no effect when using only distance-based features ($p = 0.87$). There is no main effect of *cluster model* on test performance, $F(2, 117) = 2.30$, $p = 0.11$, which is expected since cluster alignment revealed significant overlap between the topics discussed by the control and dementia groups (Section 3.2). Notably, there is a significant effect of *feature set* on test performance, whereby adding the idea density and idea efficiency features results in significantly higher F-scores, both when using local context for vector construction ($p < 0.05$), and otherwise ($p < 0.001$).

As a baseline, we use a list of hsICUs extracted by Fraser et al. (2016) in a state-of-the-art automated method for separating AD and control speakers in DementiaBank. These features consist of (i) counts of lexical tokens representing hsICUs (e.g., *boy*, *son*, and *brother* are used to identify whether hsICU S1 (Table 2) was discussed, and (ii) Boolean values which indicate whether each hsICU was mentioned or not. Overall, this constitutes 85 features. Additionally, Fraser et al. (2016) identified a list of lexicosyntactic and acoustic (LS&A) features which are indicative of cognitive impairment. We compute the performance of each set of features independently, and then com-

[5] I.e., the number of word vectors in the transcript whose scaled distance is within 3 s.d.'s from the mean cluster distortion of at least one cluster.

Table 4: Binary classification (AD:CT) using a random forest classifier, with 10-fold cross-validation. All cluster models are trained on vectors with no local context. LS&A are lexicosyntactic and acoustic features as described by Fraser et al. (2016). The reported precision, recall, and F-score are a weighted average over the two classes.

Model	Features	Accuracy	Precision	Recall	F-score
Baseline	hsICUs	0.73	0.74	0.73	0.72
Baseline	LS&A	0.76	0.77	0.76	0.76
Baseline	hsICUs + LS&A	0.80	0.80	0.80	**0.80**
control	distance-based	0.68	0.69	0.68	0.68
dementia	distance-based	0.66	0.67	0.66	0.66
combined	distance-based	0.68	0.69	0.68	0.68
control	distance-based + idea density + idea efficiency	0.74	0.76	0.74	0.74
dementia	distance-based + idea density + idea efficiency	0.74	0.75	0.74	0.74
combined	distance-based + idea density + idea efficiency	0.74	0.75	0.74	0.74
control	distance-based + idea density + idea efficiency + LS&A	0.79	0.79	0.79	0.79
dementia	distance-based + idea density + idea efficiency + LS&A	0.77	0.78	0.77	0.77
combined	distance-based + idea density + idea efficiency + LS&A	0.80	0.80	0.80	**0.80**

bine them. Table 4 summarizes the results; the first column indicates the cluster model (e.g., *control* indicates a cluster model trained on the control transcriptions), and the second column specifies the feature set. Our 12 automatically generated features (i.e., the combined set of distance-based measures, idea density, and idea efficiency) result in higher F-scores (0.74) than using 85 manually generated hsICUs (0.72); a two-sample paired t-test shows no difference (using control cluster model: $t(9) = 1.10$, $p = 0.30$; using dementia cluster model: $t(9) = 0.74$, $p = 0.48$) indicating the similarity of our method to the manual gold standard. Furthermore, we match state-of-the-art results (F-score of 0.80) when we augment the set of LS&A features with our automatically generated semantic features.

4 Discussion

We demonstrated a method for generating topic models automatically within the context of clinical assessment, and confirmed that low idea density and low idea efficiency are salient indicators of cognitive impairment. In our data, we also found that speakers with and without Alzheimer's disease generally discuss the *same topics* and in the *same contexts*, although those with AD give more spurious descriptions, as exemplified by the irrelevant topic cluster D7 (Table 3).

Using a fully automated topic generation and feature extraction pipeline, we found a small set of features which perform as well as a large set of manually constructed hsICUs in binary classification experiments, achieving an F-score of 0.80 in 10-fold cross-validation on DementiaBank. The features which correlate most highly with class include: idea efficiency (Pearson's $r = -0.41$), which means that healthy individuals discuss more topics per unit time; distance from cluster C4 ($r = 0.34$), which indicates that speakers with AD focus less on the topic of the *three-legged stool*; and idea density ($r = -0.26$), which shows that healthy speakers need fewer words to express the same number of topics.

While we anticipated that combining a large normative corpus with local context windows from a clinical corpus would produce optimal vectors, using the former exclusively actually performs better. This phenomenon is being investigated. This implies that word-vector representations do not *need* to be adapted with context windows in specific clinical data in order to be effective.

A limitation of the current work is its requirement of high-quality transcriptions of speech, since high word-error rates (WERs) could compromise semantic information. We are therefore generating automatic transcriptions of the DementiaBank audio using the Kaldi speech recognition toolkit[6]. So far, a triphone model with the standard insertion penalty (0) and language model scale (20) on DementiaBank gives the best average WER of $36.7 \pm 3.6\%$ with 10-fold cross-validation. Continued optimization is the subject of ongoing research but preliminary experiments with these transcriptions indicate significantly lower perfor-

[6] http://kaldi.sourceforge.net/

mance of the baseline model (0.68 F-score; $t(9) = 3.52$, $p < 0.01$). While the eventual aim is a completely automatic system, our methodology overcomes several major challenges in the manual semantic annotation of clinical images for cognitive assessment, even with manual transcriptions. Specifically, our methodology is fully objective, sensitive to differences between groups, and generalizable to new stimuli which is especially important if longitudinal analysis is to avoid the so-called 'practice effect' by using multiple stimuli.

Across many domains, to extract useful semantic features (such as idea density and idea efficiency), one needs to first identify information content units in speech or text. Our method can be applied to any picture or contentful stimuli, given a sufficient amount of normative data, with no modification. Although we apply this generalizable method to a single (albeit important) image used in clinical practice in this work, we note that we obtain better accuracies with this completely automated method than a completely manual alternative.

Acknowledgments

The authors would like to thank Selvana Morcos, a speech language pathologist at the Toronto Rehabilitation Institute, for her generous help with providing professional annotations of information content units for the BDAE *Cookie Theft* picture.

References

S. Ahmed, C. A. de Jager, A. F. Haigh, and P. Garrard. 2013a. Semantic processing in connected speech at a uniformly early stage of autopsy-confirmed Alzheimer's disease. *Neuropsychology*, 27(1):79–85.

S. Ahmed, A. F. Haigh, C. A. de Jager, and P. Garrard. 2013b. Connected speech as a marker of disease progression in autopsy-proven Alzheimer's disease. *Brain*, 136(12):3727–3737.

C. Ballard, S. Gauthier, A. Corbett, C. Brayne, D. Aarsland, and E. Jones. 2011. Alzheimer's disease. *The Lancet*, 377(9770):1019–1031.

K. A. Bayles and A. W. Kaszniak. 1987. *Communication and cognition in normal aging and dementia*. Little, Brown, Boston.

J. T. Becker, F. Boller, O. L. Lopez, J. Saxton, and K. L. McGonigle. 1994. The natural history of Alzheimer's disease. *Archives of Neurology*, 51:585–594.

B. Croisile, B. Ska, M. J. Brabant, A. Duchene, Y. Lepage, G. Aimard, and M. Trillet. 1996. Comparative study of oral and written picture description in patients with Alzheimer's disease. *Brain and Language*, 53(1):1–19.

K. E. Forbes-McKay and A. Venneri. 2005. Detecting subtle spontaneous language decline in early Alzheimer's disease with a picture description task. *Neurological Sciences*, 26(4):243–254.

K. C. Fraser and G. Hirst. 2016. Detecting semantic changes in Alzheimer's disease with vector space models. In Dimitrios Kokkinakis, editor, *Proceedings of LREC 2016 Workshop: Resources and Processing of Linguistic and Extra-Linguistic Data from People with Various Forms of Cognitive/Psychiatric Impairments (RaPID-2016)*, pages 1–8, Portorož, Slovenia. Linköping University Electronic Press.

K. C. Fraser, J. A. Meltzer, and F. Rudzicz. 2016. Linguistic features identify Alzheimer's disease in narrative speech. *Journal of Alzheimer's Disease*, 49(2):407–422.

E. Giles, K. Patterson, and J. 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.

H. Goodglass and E. Kaplan. 1983. *The assessment of aphasia and related disorders*. Lea and Febiger, Philadelphia.

D. Hakkani-Tür, D. Vergyri, and G. Tur. 2010. Speech-based automated cognitive status assessment. In *11th Annual Conference of the International Speech Communication Association*, pages 258–261.

D. B. Hier, K. Hagenlocker, and A. G. Shindler. 1985. Language disintegration in dementia: effects of etiology and severity. *Brain and Language*, 25(1):117–133.

A. Kertesz. 1982. *The Western aphasia battery*. Grune and Stratton, New York.

Y. H. Lai, H. H. Pai, and Y. T. Lin. 2009. To be semantically-impaired or to be syntactically-impaired: linguistic patterns in Chinese-speaking persons with or without dementia. *Journal of Neurolinguistics*, 22(5):465–475.

X. Le, I. Lancashire, G. Hirst, and R. 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, may.

B. MacWhinney. 2015. *The CHILDES Project: Tools for analyzing talk*. Lawrence Erlbaum Associates, Mahwah, NJ, 3rd edition.

S. V. S. Pakhomov, G. E. Smith, D. Chacon, Y. Feliciano, N. Graff-Radford, R. Caselli, and D. S. Knopman. 2010. Computerized analysis of speech and language to identify psycholinguistic correlates of frontotemporal lobar degeneration. *Cognitive and Behavioral Neurology*, 23(3):165–177.

J. Pennington, R. Socher, and C. D. Manning. 2014. GloVe: Global vectors for word representation. In *Conference on Empirical Methods on Natural Language Processing (EMNLP)*, pages 1532–1543.

D. A. Snowdon, S. J. Kemper, J. A. Mortimer, L. H. Greiner, D. R. Wekstein, and W. R. Markesbery. 1996. Linguistic ability in early life and cognitive function and Alzheimer's disease in late life. Findings from the Nun Study. *JAMA: the Journal of the American Medical Association*, 275(7):528–532.

R. A. Sperling, P. S. Aisen, L. A. Beckett, D. A. Bennett, S. Craft, A. M. Fagan, T. Iwatsubo, C. R. Jack, J. Kaye, T. J. Montine, D. C. Park, E. M. Reiman, C. C. Rowe, E. Siemers, Y. Stern, K. Yaffe, M. C. Carrillo, B. Thies, M. Morrison-Bogorad, M. V. Wagster, and C. H. Phelps. 2011. Toward defining the preclinical stages of Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & Dementia: the Journal of the Alzheimer's Association*, 7(3):280–292.