Interspeech 2018
2-6 September 2018, Hyderabad

Detecting signs of dementia using word vector representations

Bahman Mirheidari[1], Daniel Blackburn[2], Traci Walker[3], Annalena Venneri[2], Markus Reuber[4], and Heidi Christensen[1,5]

[1]Department of Computer Science, University of Sheffield, [2]Sheffield Institute for Translational Neuroscience (SITraN), University of Sheffield, [3]Department of Human Communication Sciences, University of Sheffield, [4]Academic Neurology Unit, University of Sheffield, Royal Hallamshire Hospital, [5]Centre for Assistive Technology and Connected Healthcare (CATCH), University of Sheffield, Sheffield, UK

{bmirheidari2, heidi.christensen}@sheffield.ac.uk

Abstract

Recent approaches to word vector representations, e.g., 'w2vec' and 'GloVe', have been shown to be powerful methods for capturing the semantics and syntax of words in a text. The approaches model the co-occurrences of words and recent successful applications on written text have shown how the vector representations and their interrelations represent the meaning or sentiment in the text. Most applications have targeted written language, however, in this paper, we investigate how these models port to the spoken language domain where the text is the result of (erroneous) automatic speech transcription. In particular, we are interested in the task of detecting signs of dementia in a person's *spoken* language. This is motivated by the fact that early signs of dementia are known to affect a person's ability to express meaning articulately for example when they engage in a conversation – something which is known to be cognitively very demanding. We analyse conversations designed to probe people's short and long-term memory and propose three different methods for how word vectors may be used in a classification setup. We show that it is possible to identify dementia from the output of a speech recogniser despite a high occurrence of recognition errors.

Index Terms: clinical applications of speech technology, pathological speech, dementia detection

1. Introduction

The number of people with dementia is increasing rapidly all around the world. An estimated 50 million people were living with dementia in 2017, and this is predicted to almost double each 20 years soaring to 131.5 million by 2050 [1].

Early detection of dementia is a challenging task due to the lack of reliable bio-markers, overlapping symptoms with normal ageing and low accuracy of existing cognitive screening tools. The term dementia is used to describe a set of symptoms arising from a range of progressive diseases such as Alzheimer's Disease. One of the often noticed symptoms are problems with a person's memory and language. This is known to affect object naming, noun production and rates of verb usage. In general, loss of vocabulary, impoverished/simplified syntax/semantics, and overuse of semantically empty words are commonly found in the language of people with dementia ([2, 3, 4]). Despite this, relatively little research has focused on formal testing of the communication and conversational ability of patients notwithstanding this playing a significant role in the expert assessment carried out by neurologists when diagnosing.

To address this, we have been investigating the extent to which automatic analysis of structured conversations (answering predefined memory-probing questions) involving people with memory problems can reveal signs of dementia in a person's speech and language. The hope is to be able to incorporate such an automatic tool into the diagnostic pathway in current memory services and thereby provide a low-cost, easy-to-use and efficient route to the appropriate treatment.

In preliminary work using Conversation Analysis (CA) on the interactions between patients and neurologists, Elsey *et al.* found that identifying a small set of manual features enabled the distinction between patients with neurodegenerative disorder (ND) [1] and those with Functional Memory Disorder (FMD) [5, 6]. These are the two most common diagnostic categories seen at many memory clinics because doctors in primary care lack the expertise and tools to distinguish between them, and hence will refer too many to memory clinics, where up to 50% of patients seen do not have dementia. This leads to costly, stressful, time consuming and ultimately unnecessary examinations and scans. Distinguishing between FMD and ND is a far harder problem than distinguishing between healthy controls and e.g., people with Alzheimer's disease. Elsey *et al.* found that looking at the way FMD and ND patients are able to accurately answer memory-probing questions such as what they did in the weekend or when they last had a problem with their memory are very accurate indicators of which group they belong to and hence which treatment they should be offered. The CA approach, however, needed manual transcriptions of the conversations and a qualitative analysis performed by an expert; thus it is not feasible for large-scale use.

We subsequently demonstrated the feasibility of an *automatic* system to perform the same task [7, 8]. In more recent work, we investigated the possibility of creating a conversation-based diagnostic test where even the neurologist is "automated", i.e., they are replaced by an animated head on a computer screen (an *Intelligent Virtual Agent (IVA)*) [9]. Although we found some differences between the neurology-led and IVA-led conversations (for example the length of people's answers are different), we also found that the IVA-led conversations contain sufficient discriminative information to support distinguishing between the two diagnostic categories.

This paper expands on previous work by exploring the use of recently proposed word vector representations that are known to encode the meaning expressed in text. We investigate how best to make use of such models for our conversational domain, and to what degree the proposed methods generalise to other

[1]here caused by dementia

10.21437/Interspeech.2018-1764

domains/databases. The original manual features from [5] required some understanding/intelligence on behalf of the human extracting them. For example one feature was concerned with the patient's *"Inability to answer"* which required a judgement as to how and in which context phrases like "Don't know" were used. In [8], this feature was 'translated' into a set of automatically extracted but in effect *shallow* text-based features counting things like the number of empty words used or the use of "Don't know" type phrases. In this paper, we show how the word vector representations can be used to model differences between the FMD, ND and other groups of patients encountered at memory clinics such as people with Mild Cognitive Impairment (MCI) and Depressive Pseudo-Dementia (DPD) – two diagnostic categories with very similar symptoms. This may lead the way for deeper features/models of the conversational patterns observed.

2. Dementia detection

Recently, a number of studies have focused on the automatic detection of dementia using what has become the standard classification pipeline of starting with the audio signal and involving diarization, automatic speech recognition (ASR), feature extraction/selection followed by classification. Most research has focused on distinguishing between healthy controls and people with Alzheimer's disease, the most common cause of dementia with some studies also involving people diagnosed with MCI. Systems have used a combination of feature types extracted from the acoustic signal (such as duration, pauses and general voice quality parameters), as well as from the text (such as those based on part of speech tags, phonetic and word identity). Promising results have been reported on a variety of data with systems initially developed for manual transcripts [10, 11, 12, 13] and later on ASR transcripts with WERs often in the high 40s [14, 15, 16]. All of the text/transcript features are somewhat shallow in nature though, and effectively modelling semantic, lexical and linguistic characteristics without much notion of word co-occurrence except at a relatively simple level such as Wankerl *et al*'s use of n-grams [17].

This paper describes novel work exploring how word vector representation, based on word co-occurrence patterns, may be deployed to capture the deeper meaning of a conversation. To broaden up the task and generalisability of the proposed solutions, we include additional diagnostic categories and compare our two parallel corpora (neurology-led vs. IVA-led) with one of the *de facto* publicly available databases in this domain, the DementiaBank corpus [18].

3. Word vectors

Machine learning algorithms work on vectors of numbers, and word embedding is a technique which is widely used to convert text to numbers; instead of a word, a series of numbers are used. Traditionally, techniques such as bag-of-words (BOW) [19] and Frequency Inverse Document Frequency (TF-IDF) [20] were used for word embedding with some success. Recently, more successful approaches have used deep learning techniques to produce vectors representing words. Two recently introduced techniques are 'w2vec' [21, 22] and 'GloVe' [23] which are both based on the co-occurrences of words, taking into account the context (neighbouring words) in a text.

The 'w2vec' is trained using a simple three layer deep neural network (input, hidden and output layers). It can learn the word vectors using two techniques: skip-gram and continuous bag-of-words (CBOW). The skip-gram aims at predicting the context from a given word, while the CBOW attempts to predict a word given a context. Generally, the skip-gram can capture more information than that captured by the semantics of a single word, and using the negative sub-sampling technique the skip-gram technique can outperform the CBOW. Mikolov et al, demonstrated that the resulting 'w2vec' vectors exhibits some interesting properties, for instance, that vector("King") - vector("Man") + vector("Woman") is very close to the vector("Queen"). Despite the amazing advantages of the 'w2vec', it has some limitations including not taking into account the global co-occurrence of the words in the whole corpus. The 'GloVe' (Global Vectors) adds the benefits of the matrix factorisation approaches to the skip-gram to capture the global statistical information. Instead of focusing only on the probabilities of words in the context, the ratio of co-occurrence probabilities are taken into account. In fact, the 'GloVe' attempts to associate the logarithm of ratios of co-occurrence probabilities with the vector differences. The authors of the 'w2vec'[2] and 'GloVe'[3] have both shared their pre-trained models for public use.

One of the main applications of the word vector encoding techniques is sentinent analysis - the problem of identifying opinions or moods in a piece of text. A popular benchmarks for sentiment analysis is the ICL Internet Movie Database (IMDB) containing 50000 movie reviews associated with positive or negative sentiments (half for training and half for testing). Inspired by the language modelling and probabilistic latent topic models, [24] introduced a model for the vector representation and achieved an accuracy of 88.89% for the binary classification task. [25] attempted to extend the 'w2vec' model to make vectors representing paragraphs or documents ('doc2vec'). The main idea was to add an extra token (ID) for each document to the content while training the BOW or skip-gram model. They reported 92.58% accuracy for the sentiment analysis task of IMDB, however, other researchers have struggled to reproduce the same outcomes [26]. Combining CNNs (Convolutional Neural Networks) with BLSTMs (Bidirectional Long Short Term Memory) networks [27] resulted in a classification rate of 89.7% for the IMDB sentiment analysis task. Random embedding substitution obtained 88.98% accuracy by using a normal LSTMs and 89.71% using BLSTM. [28] also reported 89% accuracy using CNN and LSTM. In addition to the sentiment analysis, the word vectors have been used in various NLP tasks such as: Semantic Queries [29], Semantic Textual Similarity [26], document analysis [30], and text understanding [31].

Recently, word vectors have been used in a number of different tasks involving spoken language. [32] applied the 'Doc2Vec' to the ASR outputs of non-native English speaker taking the TOEFL internet-based test (iBT) to score (measure) the responses, and they observed a considerable amount of improvements comparing to using TF-IDF features. [33] used the 'GloVe' embedding to initialise the final dense layer of their deep neural network to directly convert acoustic features to words and reported reasonably low WER on the Switchboard Call Home dataset.

The use of word vectors for detection of pathologies and para-linguistic information in speech is very novel. [34] used the 'GloVe' word vectors to detect depression from the transcripts produced by the ASR on the de-identified speech (modifying voice characteristics for privacy reasons). They split each

[2] http://mccormickml.com/2016/04/12/googles-pretrained-word2vec-model-in-python/

[3] https://nlp.stanford.edu/projects/glove/

turn of a speaker into a series of words. Each turn was then represented by summing up the normalised 'GloVe' word vectors of the turn. They also considered applying a weight coefficient to the vectors allowing them to assign more importance to the rarer words. Then, they reduced the dimension of the turn vectors by using the PCA algorithm and used an SVM classifier to classify between depression and non-depression speech. They gained 80% classification accuracy using de-identified speech recognised by ASR (with 37.3% WER).

4. Detecting dementia with word vectors

For classification tasks we need to use the word vector representations of the individual words in a transcript in a way that enables us to distinguish the different classes. This section describes the four different approaches we have investigated. The first two are based on composing a vector from the individual word vectors and using these vectors to train a classifier as per usual; the third method uses a cosine similarity as a measure of how different a word vector is to typical word vectors found in the labelled/known classes; the fourth approach models the vectors from the first two approaches in a sequential model.

Assume a corpus C consists of n documents, $D_i; 1 \leq i \leq n$. Each document consists of a number of words, $D_i = w_{i1}, ..., w_{ij}$ and each word can be converted to a vector V with d dimensions as $V(w_{ij})$ using one of the pre-trained word vector algorithms like 'w2vec' or 'GloVe'. Ignoring the non important words in a text (stop list) as well as replicated words, we can make a new vector by calculating the average of the word vectors appended to the variance of the word vectors as:

$$AV(D_i) = [\mu(D_i), \sigma(D_i)] \tag{1}$$

where μ and σ are the average and variance and $AV(D_i)$ has dimensions $2*d$. The first proposed approach uses the $AV(D_i)$ vectors for training a classifier.

The second approach is similar but based on a feature vector derived as the difference between $AV(D_i)$ and a vector combined over all training documents in each class. That is, for a supervised classification task with m known classes, $c_{1...m}$, we can make m combined AV vectors: $AV(c_l); 1 <= l <= m$. The feature vector in this second approach is found by summing the differences between $AV(D_i)$ and each $AV(c_l)$. We refer to this vector as $DiffAV$:

$$DiffAV(D_i) = \sum_{l=1}^{m} (AV(c_l) - AV(D_i) \tag{2}$$

As a third approach for representing documents, we calculate the cosine similarity between the word vectors of a document and the word vectors of each class. The value of the cosine similarity will be normalised (sum up to one). We refer to this as $CosWV$ (cosine word vectors) and define it as:

$$CosV(c_l, D_i) = \sum_{j=1}^{k} \sum_{t=1}^{r} cos(V(w_{ij}), V(w_{lt})) \tag{3}$$

$$CosWV(D_i) = \left[\frac{1}{M} CosV(c_1, D_i), .., \frac{1}{M} CosV(c_m, D_i) \right] \tag{4}$$

where $M = Max(CosV)$. For the fourth and final approach, we extract fixed length frames of the whole document using a sliding window over the text (we have used 80 words and a 25% overlap) and computing the AV and $DiffAV$ vector of each frame gives us $SeqAV(D_i) = [AV(D_{i1}), ..., AV(D_{if})]$ and $SeqDiffAV(D_i) = [DiffAV(D_{i1}), ..., DiffAV(D_{if})]$.

5. Experimental setup

In addition to the **IMDB** dataset, which contains 50000 text entries with feedback (either positive or negative reviews) about movies, we have used four other datasets that are summarised in Table 1: i) **DementiaBank** [18]: 473 text/audio files of people with Alzheimer's Disease and healthy controls describing the 'Cookie Theft'' picture; ii) **Hallam** [8]: 45 neurologist-patient conversations recorded at the memory clinic at the Royal Hallamshire Hospital (Sheffield, UK) with the following diagnostic categories: FMD, ND and Depressive Pseudo-Dementia (DPD)[4]; iii) **IVA** [9]: 18 IVA-patient conversations also recorded at the Royal Hallamshire Hospital with the following diagnostic categories: FMD, ND and MCI; and iv) **Seizure** [35]: 241 neurologist-patient conversations with different types of seizure diagnosis (note that we only used this for boosting ASRs' acoustic and language models). The IMDB comes with separate training and test sets (25000 each), however the other, smaller datasets are split into training and test sets using the standard k-fold cross validation (k=10).

Table 1: *Datasets info including Len.:the total length in hours/mins, Utts.:number of utterances, Spks.:number of speakers, and Avg .Utts.:Average utterance length in seconds.*

Dataset(No)	Len.	Utts.	Spks.	Avg. Utts.
DemBank(473)	8h	473	255	61.1s
Hallam(45)	12h	8970	117	4.8s
IVA(18)	3h 15m	785	40	14.9s
Seizure(241)	50h 16m	28k	597	6.3s

The spoken language based datasets are recognised by ASRs trained with the Kaldi toolkit [36]. We followed the TDDN-LSTM recipe after boosting our data sets with around 50 hours from the Seizure dataset. We used k-fold with k=5 as a cross validation approach to train the ASRs. Table 2 shows the average WER for the three data sets using HMM/GMM and DNN models respectively. For all dataset, the DNNs outperformed significantly the HMM based ASRs and we obtained 25.6% WER for the IVA dataset.

Table 2: *WER for the three datasets.*

Dataset	HMM/GMM	DNN
DemBank	60.6%	45.3%
Hallam	57.9%	43.8%
IVA	43.3%	25.6%

6. Results

We initially used the 'w2vec' model pre-trained on the Google News dataset (3 million vocabulary size, 100 billions words and 300 vector size) and the 'GloVe' model pre-trained on the Common Crawl (2.2 million vocabulary size, 840 billion words, and 300 vector size). However, in almost all cases the 'GloVe' word vectors outperformed 'w2vec' vectors. Therefore, we have only presented the 'GloVe' results in the following.

[4]patients showing cognitive decline in line with that of dementia but caused by depression

Table 3: *Classification accuracy using Logistic Regression classifier and pre-defined GloVe word vectors.*

Dataset	AV	DiffAV
IMDB	86.3%	**86.4**%
DemBank(Man)	**69.8**%	68.7%
DemBank(ASR)	**58.4**%	57.7%
Hallam(Man)	**63.5**%	**63.5**%
Hallam(ASR)	**54.2**%	45.8%
IVA(Man)	**50.0**%	45.0%
IVA(ASR)	**55.0**%	50.0%

6.1. Average and variance vector approaches

Table 3 shows classification accuracy across datasets for approaches 1 and 2 (based on vectors AV and $DiffAV$) when applied to the manual and ASR transcripts and using the Logistic Regression classifier. For almost all of the datasets, the AV-based approach provides a classification accuracy that is better than or almost equal to that of $DiffAV$.

6.2. Word Cosine similarity

The third approach is based on the cosine similarity. The $CosWV$ accumulation calculates the cosine similarity between the words of each classes and the words of the text. This makes it applicable when there are fewer numbers of training samples, however, as the numbers increases the calculation takes a long time. We therefore only calculated $CosWV$ for the two smaller datasets, Hallam and IVA. Table 4 shows the accuracy rate using a Logistic Regression classifier. Comparing to Table 3 classification results gained by $CosWV$ are remarkably better than AV and $DiffAV$, especially for both IVA(Man) and IVA(ASR) which has improved with up to around 20%. Comparing ASR results with their manual transcript counterparts shows that results are the same or only slightly worse. That is, this approach appears to be robust to recognition errors (that are relatively high as seen in Table 2)

Table 4: *3-way classification accuracy using Logistic Regression classifier and $CosWV$ for Hallam and IVA datasets.*

Dataset	Accuracy
Hallam(Man)	66.5%
Hallam(ASR)	65.8%
IVA(Man)	70%
IVA(ASR)	70%

6.3. Sequence classification

Deep neural network (DNN) approaches are very successful for classification, but requires sufficient amounts of training data. We do not yet have enough of the Hallam and IVA conversations collected for this approach, but both the IMDB and DementiaBank datasets are big enough. Fixed length text were taken from the input text using the sliding window technique. We used a combination of CNNs and LSTM (Keras python library [37]) to build the sequence classification model (256 LSTM cells and 64 filters for the CNNs, 5 kernels and 2 max polling with 0.2% dropout rate). For IMDB, we achieved classification accuracy rates of 90.7% and 92.5% using $SeqAV$ and $SeqDiffAV$ respectively (Table 5). We trained a similar model for the DementiaBank with over 400 samples for training in a 10-fold cross validation (64 LSTM cells and the same CNN settings as the IMDB). Similarly to IMDB, both DemBank(Man) and DemBank(ASR) sequence classification accuracy rates are considerably higher than those achieved with the first two approaces (Table 3) and $SeqDiffAV$ always performs better than $SeqAV$.

Table 5: *Sequence Classification using CNN_LSTM and predefined GloVe word vectors.*

Dataset	SeqAV	SeqDiffAV
IMDB	90.7%	**92.5**%
DemBank(Man)	74.0%	**75.6**%
DemBank(ASR)	61.7%	**62.3**%

6.4. Comparing to previous work

In our previous work, we trained binary classifiers (ND/FMD) using acoustic, lexical and CA-inspired features. In order to compare our results with those results, we broke down the 3-way classifications into binary classifications for Hallam and IVA. Table 6 shows the results for Hallam. For Hallam(Man), the binary classifications all are over 70%. For the FMD vs. ND, the achieved 70.8% (Hallam(Man)) is around 9% lower than the 76.7% we gained using the 44 features (mix of 12 acoustic, 12 lexical and 20 CA-inspired features) [9]. Table 7 shows the results for IVA. For the FMD vs. ND, the achieved 77.8% (IVA(Man)) is slightly lower than the 81.8% from [9]. However, FMD vs. ND for IVA(ASR) results in 100% accuracy (compared to 90.9% using 44 features [9]). Note that in the previous work we firstly tuned the classifier to get the best results for the manual Hallamdata and then performed the classification. However, in this study we only used the default parameters for the classifier without tuning.

Table 6: *Binary classifications for Hallam using $CosWV$.*

Dataset	FMD/ND	FMD/DPD	ND/DPD
Hallam(Man)	**70.8**%	75.8%	71.6%
Hallam(ASR)	62.0%	**93.7**%	**75.9**%

Table 7: *Binary classifications for IVA using $CosWV$.*

Dataset	FMD/ND	FMD/MCI	ND/MCI
IVA(Man)	77.8%	57.1%	**81.25**%
IVA(ASR)	**100**%	**75.0**%	62.5%

7. Conclusions

In this exploratory and preliminary work, we have showed the potential of using word vectors to help in detecting signs of dementia in spoken language. Amongst the four approaches we introduced in this study, the word cosine similarity vector was the best, achieving over 65% accuracy for a challenging three-way classification task for the Hallam dataset and 70% for the IVA dataset. In addition, we showed that the method was robust to errors introduced from automatic speech recognisers across a number of datasets. In future work, we will add the word vector as a single feature to a set of other features (acoustic, lexical and CA-inspired) to improve the overall accuracy for both three-way and binary classifications.

8. References

[1] Alzheimer's Disease International, "Dementia statistics," 2018, accessed on March 10, 2018. [Online]. Available: https://www.alz.co.uk/research/statistics

[2] J. Appell, A. Kertesz, and M. Fisman, "A study of language functioning in Alzheimer patients," *Brain and Language*, vol. 17, no. 1, pp. 73–91, 1982.

[3] K. A. Bayles and A. W. Kaszniak, *Communication and cognition in normal aging and dementia*. Taylor & Francis Ltd London, 1987.

[4] H. E. Hamilton, *Conversations with an Alzheimers patient: An interactional sociolinguistic study*. Cambridge, England: Cambridge University Press, 1994.

[5] C. Elsey, P. Drew, D. Jones, D. Blackburn, S. Wakefield, K. Harkness, A. Venneri, and M. Reuber, "Towards diagnostic conversational profiles of patients presenting with dementia or functional memory disorders to memory clinics," *Patient Education and Counseling*, vol. 98, pp. 1071–1077, 2015.

[6] D. Jones, P. Drew, C. Elsey, D. Blackburn, S. Wakefield, K. Harkness, and M. Reuber, "Conversational assessment in memory clinic encounters: interactional profiling for differentiating dementia from functional memory disorders," *Aging & Mental Health*, vol. 7863, pp. 1–10, 2015.

[7] B. Mirheidari, D. Blackburn, M. Reuber, T. Walker, and H. Christensen, "Diagnosing people with dementia using automatic conversation analysis," in *Proc. Interspeech*, 2016, pp. 1220–1224.

[8] B. Mirheidari, D. Blackburn, K. Harkness, T. Walker, A. Venneri, M. Reuber, and H. Christensen, "Towards the automation of diagnostic conversation analysis in patients with memory complaints," *Journal of Alzheimer's Disease*, 2017.

[9] ——, "An avatar-based system for identifying individuals likely to develop dementia," *Proc. Interspeech*, pp. 3147–3151, 2017.

[10] L. Toth, G. Gosztolya, V. Vincze, I. Hoffmann, G. Szatloczki, E. Biro, F. Zsura, M. Pakaski, and J. Kalman, "Automatic detection of mild cognitive impairment from spontaneous speech using ASR," in *Proc. Interspeech*, 2015.

[11] J. Weiner, C. Herff, and T. Schultz, "Speech-Based Detection of Alzheimerś Disease in Conversational German," in *Proc. Interspeech*, 2016, pp. 1938–1942.

[12] K. C. Fraser, J. A. Meltzer, and F. Rudzicz, "Linguistic Features Identify Alzheimer s Disease in Narrative Speech," *Journal of Alzheimer's Disease*, vol. 49, pp. 407–22, 2015.

[13] T. Alhanai, R. Au, and J. Glass, "Spoken language biomarkers for detecting cognitive impairment," *arXiv preprint arXiv:1710.07551*, 2017.

[14] L. Toth, I. Hoffmann, G. Gosztolya, V. Vincze, G. Szatloczki, Z. Banreti, M. Pakaski, and J. Kalman, "A speech recognition-based solution for the automatic detection of mild cognitive impairment from spontaneous speech," *Current Alzheimer Research*, vol. 15, no. 2, pp. 130–138, 2018.

[15] J. Weiner, M. Engelbart, and T. Schultz, "Manual and automatic transcriptions in dementia detection from speech," in *Proc. Interspeech*, 2017, pp. 3117–3121.

[16] L. Zhou, K. C. Fraser, and F. Rudzicz, "Speech recognition in Alzheimerś disease and in its assessment," in *Proc. Interspeech*, 2016, pp. 1948–1952.

[17] S. Wankerl, E. Nöth, and S. Evert, "An n-gram based approach to the automatic diagnosis of alzheimers disease from spoken language," in *Proc. Interspeech*, 2017.

[18] J. T. Becker, F. Boiler, O. L. Lopez, J. Saxton, and K. L. McGonigle, "The natural history of alzheimer's disease: description of study cohort and accuracy of diagnosis," *Archives of Neurology*, vol. 51, no. 6, pp. 585–594, 1994.

[19] Z. S. Harris, "Distributional structure," *Word*, vol. 10, no. 2-3, pp. 146–162, 1954.

[20] K. Sparck Jones, "A statistical interpretation of term specificity and its application in retrieval," *Journal of documentation*, vol. 28, no. 1, pp. 11–21, 1972.

[21] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[22] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[23] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. EMNLP*, 2014, pp. 1532–1543.

[24] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning word vectors for sentiment analysis," in *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1*. Association for Computational Linguistics, 2011, pp. 142–150.

[25] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *International Conference on Machine Learning*, 2014, pp. 1188–1196.

[26] J. H. Lau and T. Baldwin, "An empirical evaluation of doc2vec with practical insights into document embedding generation," *arXiv preprint arXiv:1607.05368*, 2016.

[27] Q. Shen, Z. Wang, and Y. Sun, "Sentiment analysis of movie reviews based on cnn-blstm," in *Int. Conf. on Intelligence Science*. Springer, 2017, pp. 164–171.

[28] A. Yenter and A. Verma, "Deep cnn-lstm with combined kernels from multiple branches for imdb review sentiment analysis," in *Ubiquitous Computing, Electronics and Mobile Communication Conference (UEMCON), 2017 IEEE 8th Annual*. IEEE, 2017, pp. 540–546.

[29] R. Bordawekar and O. Shmueli, "Using word embedding to enable semantic queries in relational databases," in *Proc. of the 1st Workshop on Data Management for End-to-End Machine Learning*. ACM, 2017, p. 5.

[30] D. Park, S. Kim, J. Lee, J. Choo, N. Diakopoulos, and N. Elmqvist, "Conceptvector: text visual analytics via interactive lexicon building using word embedding," *IEEE trans. on visualization and computer graphics*, vol. 24, no. 1, pp. 361–370, 2018.

[31] S. Gao, H. Zhang, and K. Gao, "Text understanding with a hybrid neural network based learning," in *Int. Conf. of Pioneering Computer Scientists, Engineers and Educators*. Springer, 2017, pp. 115–125.

[32] J. Tao, L. Chen, and C. M. Lee, "Dnn online with ivectors acoustic modeling and doc2vec distributed representations for improving automated speech scoring." in *Proc. Interspeech*, 2016, pp. 3117–3121.

[33] K. Audhkhasi, B. Ramabhadran, G. Saon, M. Picheny, and D. Nahamoo, "Direct acoustics-to-word models for english conversational speech recognition," *arXiv preprint arXiv:1703.07754*, 2017.

[34] P. Lopez-Otero, L. Docio-Fernandez, A. Abad, and C. Garcia-Mateo, "Depression detection using automatic transcriptions of de-identified speech," *Proc. Interspeech*, pp. 3157–3161, 2017.

[35] K. Ekberg and M. Reuber, "Can conversation analytic findings help with differential diagnosis in routine seizure clinic interactions?" *Communication & medicine*, vol. 12, no. 1, p. 13, 2015.

[36] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, Dec. 2011, iEEE Catalog No.: CFP11SRW-USB.

[37] F. Chollet *et al.*, "Keras: Deep learning library for theano and tensorflow," *URL: https://keras. io/k*, vol. 7, p. 8, 2015.