A Neural Model for Predicting Dementia from Language

Weirui Kong KONGW@ALUMNI.UBC.CA
Department of Computer Science
University of British Columbia
Vancouver, BC, Canada

Hyeju Jang HYEJUJ@CS.UBC.CA
Department of Computer Science
University of British Columbia
Vancouver, BC, Canada

Giuseppe Carenini CARENINI@CS.UBC.CA
Department of Computer Science
University of British Columbia
Vancouver, BC, Canada

Thalia Field THALIA.FIELD@UBC.CA
Department of Medicine
University of British Columbia
Vancouver, BC, Canada

Abstract

Early prediction of neurodegenerative disorders such as Alzheimer's disease (AD) and related dementias is important in developing early medical supports and social supports, and may identify ideal stages for testing novel therapeutics aimed at preventing disease progression. Currently, a diagnosis is based on clinical expertise and cognitive screening tests, which have limited accuracy in earlier stages of disease, or invasive and resource-intensive testing, such as lumbar puncture or specialized neuroimaging. Changes in speech and language patterns can occur in dementia in its earliest stages and may worsen as the disease progresses. This has led to recent attempts to create automatic methods that predict dementia through language analysis. In addition to features extracted from language samples, previous works have improved the prediction accuracy by introducing some task-specific features. But task-specific features prevent the model from generalizing to other tests. In this paper, we apply a neural model (Hierarchical Attention Networks) to the dementia prediction task. Remarkably, the model requires no task-specific feature and achieves state-of-the-art classification result on a widely used dementia dataset of spoken language. We also perform a detail analysis to interpret how a prediction is made. Interestingly, the same neural model does not work well on a corpus of written text, suggesting that dementia prediction from language may require different methods depending on the genre of the source language.

1. Introduction

Dementia is a progressive cognitive impairment caused by neurodegenerative disease, which affects more than 46.8 million people around the world (A.D. International, 2015). Among diverse types of dementia, Alzheimer's disease (AD), which accounts for 60% - 80% of all

dementia diagnosis, is among the most financially costly diseases in developed countries (Chambers et al., 2016). Although there is not yet a cure for AD, research suggests that novel therapeutics will be most effective if given early in the disease course (Posner et al., 2017).

However, predicting AD especially in its early stages is difficult. A diagnosis of dementia involves clinical opinion based on functional status, cognitive performance on standardized tests (Williams et al., 2013) and resource-intensive specialized tests, such as lumbar puncture or advanced neuroimaging (Nensa et al., 2014). In developing countries, access to some or all of these resources may not be available, and this is reflected in the higher than average rates of undiagnosed dementia in those regions. All around the world, only approximately 25% of the 46.8 million dementia population receive a formal diagnosis (A.D. International, 2015). Therefore, a non-invasive diagnostic tool that is inexpensive and easy to administer is of great importance to dementia patients, especially those in developing countries.

Clinical Relevance One promising direction is to design a tool that can assist in prediction of preclinical disease by using automated analysis of language. Language is one of the first faculties afflicted by the disease and subtle changes in language are observed a year or more before dementia is diagnosed, according to longitudinal studies on people with AD (Forbes-McKay and Venneri, 2005; Oulhaj et al., 2009). These changes include, for example, low grammatical complexity, limited vocabulary and frequent word finding problems.

Given that linguistic deficits are early signs of dementia, researchers have developed dementia prediction systems based on language by applying machine learning (ML) and natural language processing (NLP). Most prior work built computational models on the dataset DementiaBank (Becker et al., 1994), a publicly available dataset that contains audio recordings and transcripts of participants (people with dementia and healthy controls) describing the Cookie Theft picture (Figure 1). Prior work used not only acoustic features and various linguistic features but also task-specific features such as information units. Information units (Croisile et al., 1996) are objects and actions appearing in the picture (e.g., mother, stool, overflowing, and drying), which are usually pre-defined by human experts. Information unit features measure how well a participant captures key concepts in the picture. Based on these task-specific features as well as linguistic and acoustic features, prior models using traditional classification methods such as logistic regression have been shown to give reliable AD prediction (Fraser et al., 2016; Masrani, 2018). Although the task-specific features are effective for dementia prediction, one major disadvantage is that they are specific to a particular picture. If participants are asked to describe a different picture, information units in the picture need to be re-defined.

Technical Significance To improve performance of dementia prediction without using task-specific features, in this paper we propose to apply a neural network model, which does not require any feature engineering. The input is raw text (picture description transcripts) and the model output is the predicted label (dementia people or healthy controls). The attention mechanism of the neural model automatically learns to emphasize words and sentences that are helpful for prediction. This not only improves classification performance, but is also helpful for analyzing our experiment results. Our contributions can be summarized as follows:

- We apply a picture-agnostic neural method to the DementiaBank dataset, and obtain comparable results to traditional models that use task-specific features. By including a demographic feature (age), our model achieves state-of-the-art performance, improving the classification accuracy of the top-performer traditional method which also uses age, from 84.4% to 86.9%.
- We interpret the predictions made by our models from different angles, including model visualization and statistical tests. We find that the attention model attends more strongly to the information unit words defined by human experts. Yet, the attention focuses on only a specific subset of such words, for which we still do not have a satisfactory explanation.
- We also apply our approach to a dementia blog corpus, and discuss the results in comparison to prior work. In essence, on this corpus of written text, the neural method is not a competitive solution.

The rest of the paper is organized as follows. In Section 2, we review prior studies that focus on traditional ML and NLP approaches for dementia prediction. Then, in Section 3 we provide an overview of two datasets used for our experiments. After that, in Section 4 we introduce our model as well as prior top-performer traditional models for comparison. Experiments and result analysis are described in Section 5. Lastly, in Section 6 we conclude and suggest some future directions.

2. Related Work

With advances in ML algorithms and NLP techniques, building computational tools to automatically predict dementia from a sample of narrative speech has received growing attentions. Ahmed et al. (2013) proposed features that were helpful for identifying dementia from speech, using data collected in the Oxford Project to Investigate Memory and Aging (OPTIMA) study. They found that language was progressively impaired as the disease progressed and suggested using semantic, lexical content and syntactic complexity features for classification.

Orimaye et al. (2014) used diverse machine learning methods with lexical and syntactic features to distinguish between dementia patients and healthy adults on the DementiaBank dataset (Becker et al., 1994). They compared five different classifiers including support vector machines (SVMs), naive Bayes, decision trees, neural networks and Bayesian networks, and reported that SVMs showed the best performance with a F-score of 74%.

In another study, Al-Hameed et al. (2017) extracted acoustic features from the audio files of the DementiaBank dataset, building a regression model to predict clinical examination scores used for dementia prediction (Mini Mental State Examination scores, ranging from 0 to 30). This work used only acoustic features, and their regression model predicted MMSE scores with a mean absolute error less than 4.

Fraser et al. (2016) explored a broad spectrum of both linguistic and acoustic features, demonstrating the necessity of feature selection. They found that optimal performance was obtained when 35-50 features were used, and the performance dropped off dramatically with a feature set size larger than 50. They achieved an accuracy of 81.96% in distinguishing individuals with AD from those without.

As briefly mentioned in the Introduction section, the DementiaBank is associated with a set of human-defined information units representing key components of the Cookie Theft picture, such as subjects, objects, locations and actions (Croisile et al., 1996). Upon the information units, Masrani (2018) proposed a novel feature group called spatial neglect features. They vertically split the picture into two halves and computed features that measure spatial neglect, e.g., count of mentions of any information unit for each region. Combing their new feature group with linguistic, acoustic, information unit features and the demographic feature (age), followed by a feature selection step, they achieved the accuracy of 84.4%.

Our study differs from previous approaches in that, to avoid task-specific features and alleviate feature engineering, we apply a neural model, Hierarchical Attention Networks (HAN) (Yang et al., 2016), to the dementia prediction. HAN has been very successful in several text categorization tasks like sentiment estimation (Zhang et al., 2018) and topic classification (Tsaptsinos, 2017). By incorporating the demographic feature *age* within HAN, the model outperforms previous methods by a substantial margin on the DementiaBank dataset.

3. Data Sets

3.1. DementiaBank

The DementiaBank corpus was collected for the study of communication in dementia, between 1983 and 1988 at the University of Pittsburgh (Becker et al., 1994). It contains interview recordings and manually-transcribed transcripts of English-speaking participants describing the Cookie Theft picture (Figure 1). The participants are categorized into dementia patient and healthy control groups. Of the 309 dementia samples, 257 samples are classified as possible/probable AD, and the remaining samples as other types of dementia. Our study uses only the 257 AD samples and 242 healthy elderly control samples. Statistics about the DementiaBank samples used in this study are listed in Table 1.

Table 1: Demographics of DementiaBank dataset.

Diagnosis	**Samples**	**Mean Words**	**Mean Age**
AD	257	104.98 (s=59.8)	71.72 (s=8.47)
Control	242	113.56 (s=58.5)	63.95 (s=9.16)

3.2. Dementia Blog Corpus

While the DementiaBank corpus consists of transcribed spoken language, the Dementia Blog Corpus is derived from written language. This corpus was created by Masrani et al. (2017) by collecting blog posts written by authors with and without dementia. In particular, they scraped the text of 2805 posts from 6 public blogs up to April 4th, 2017. Three blogs were written by dementia patients, and three written by family members of dementia patients were used as control. There are a total of 1654 samples written by persons with dementia and 1151 from healthy controls. Table 2 summarizes statistics of the Dementia Blog dataset.

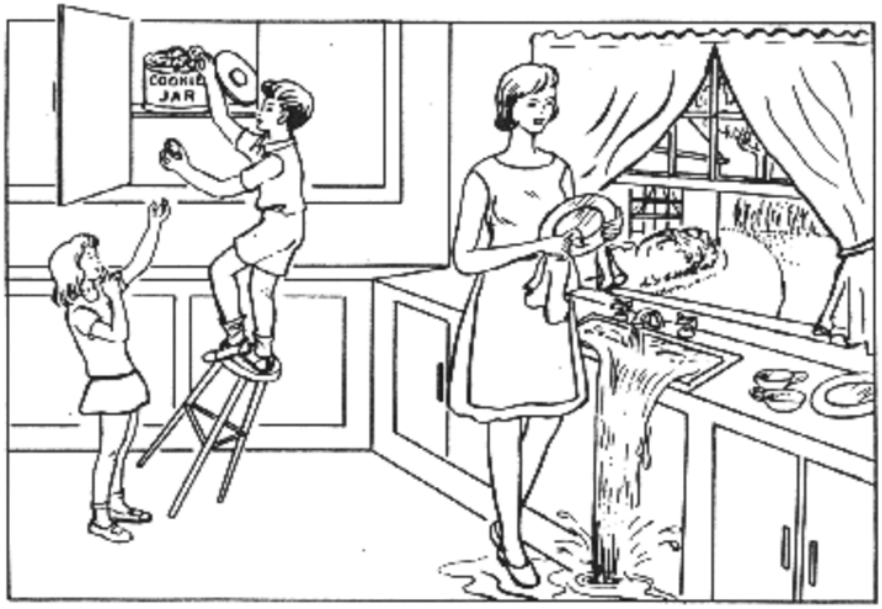


Figure 1: The Cookie Theft picture.

Table 2: Blog information as of April 4th, 2017.

URL (http://*.blogspot.ca)	**Posts**	**Mean Words**	**Diagnosis**	**Gender/Age**
living-with-alzhiemers	344	263.03 (s=140.28)	AD	M, 72 (approx)
creatingmemories	618	242.22 (s=169.42)	AD	F, 61
parkblog-silverfox	692	393.21 (s=181.54)	Lewy Body	M, 65
journeywithdementia	201	803.91 (s=548.34)	Control	F, unknown
earlyonset	452	615.11 (s=206.72)	Control	F, unknown
helpparentsagewell	498	227.12 (s=209.17)	Control	F, unknown

4. Methods for Dementia Prediction

In Section 4.1 we review traditional methods proposed by previous studies. Then in Section 4.2 we briefly describe the original HAN model and our modified version.

4.1. Traditional Models for AD Prediction

We compare our neural network method against four models including a simple demographic-based baseline and three prior models using different features. Among them, Masrani (2018)-S reported the state-of-the-art performance.

- **Age only:** The age is an important predictor of dementia according to Gao et al. (1998). Our demographic-based baseline uses only the ages of participants as features to demonstrate the predictiveness of the age feature.

- **Fraser et al. (2016):** This model uses a total of 370 features including linguistics, acoustic, and task-specific features. More specifically, these features include *Context-Free-Grammar Rules, Syntactic Complexity, Information Unit, Acoustic*, etc.
- **Masrani (2018)-B:** This model uses acoustic and linguistic features, which are mostly from Fraser et al. (2016), together with the demographic feature (age). There are 315 features in total. This model is used as a baseline in (Masrani, 2018).
- **Masrani (2018)-S:** This model uses spatial neglect features, which are task-specific features, together with features from Masrani (2018)-B. This model is the current best performer for dementia prediction on the DementiaBank corpus.

Table 3 summarizes different feature groups leveraged in the traditional methods.

Table 3: Features used by traditional methods. Info: information unit features. Spatial: spatial neglect features.

Dataset	Methods	Linguistic	Acoustic	Info	Spatial	Age
DementiaBank	Age only	×	×	×	×	✓
	Fraser et al. (2016)	✓	✓	✓	×	×
	Masrani (2018)-B	✓	✓	✓	×	✓
	Masrani (2018)-S	✓	✓	✓	✓	✓
Dementia Blog	Masrani et al. (2017)	✓	×	×	×	×

All models above except the demographic-based baseline performed a feature selection step. We use logistic regression to evaluate these models because it outperformed other classification algorithms in both (Fraser et al., 2016) and (Masrani, 2018).

4.2. Hierarchical Attention Networks

Our method is based on the Hierarchical Attention Networks (HAN). Figure 2 illustrates the overall architecture of HAN for dementia prediction. The model input are words from one interview sample (i.e., a description of the Cookie Theft picture). The model output is the probability distribution over two categories, AD and healthy. The model consists of a word sequence encoder, a word-level attention layer, a sentence encoder and a sentence-level attention layer.

We briefly introduce the functionality of each layer. For more details, refer to (Yang et al., 2016). The word encoder uses the bidirectional GRU (Bahdanau et al., 2014), an efficient implementation of recurrent neural network (RNN). It encodes each word in one sentence into a hidden vector, given the context of other words in the sentence. Then the word-level attention layer puts different weights on each word vector, producing a weighted hidden vector of the sentence. Once we get all the sentence vectors of the input sample, we feed them into another bidirectional GRU, i.e., a sentence encoder. This sentence encoder along with the sentence-level attention layer builds a weighted vector (denoted by v in Figure 2) for the whole document, which is the latent representation of an input sample

by applying attention mechanism to both word level and sentence level. Finally a linear layer projects v to a 2-dimensional vector, on which a softmax operation is performed. The output is the probability distribution for AD and healthy. Negative log likelihood of the correct labels is used as the training loss.

We evaluate the performance of two models, one is the original HAN model and the other incorporates demographic information by concatenating v with the age of the interviewee. Since the scale of age ([50, 90] in our dataset) is much larger than the values of elements of v (typically in [-1, 1]), we standardize the age, making it zero mean and unit variance before concatenating it with v.

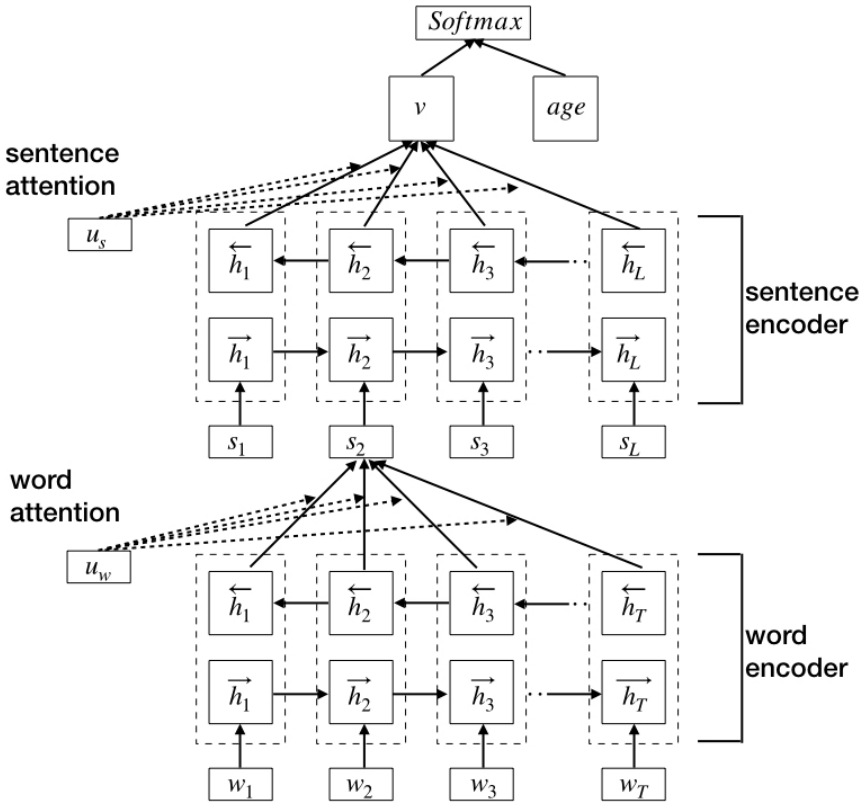


Figure 2: Hierarchical attention network for dementia prediction.

5. Experiments

5.1. Experiment Settings

Due to the limited size of the DementiaBank dataset, we performed 10-fold cross validation, in which a 10% of the data (test set) were used for evaluation, and the remaining 90% (training set) were used for feature selection and constructing the model. The reported per-

formance is an average across the 10 folds. For evaluation metrics, we computed prediction accuracy, precision, recall and F-score.

In addition to the simple demographic-based baseline that uses age as the only feature, six models were tested for comparison: Fraser et al. (2016)'s model which used linguistic, acoustic and information units; Masrani (2018)-B which used linguistic, acoustic, information units, and age features; Masrani (2018)-S which obtained the best results among previous studies by adding spatial neglect features; a bidirectional GRU model, and our two HAN based models.

The bidirectional GRU model does not use a hierarchical structure like HAN does. Similarly to HAN, it has a word encoder, including the same word level attention mechanism. However, instead of using a sentence encoder, it builds a document representation via a max-pooling operation across sentence embeddings. The document representation is then fed to a linear layer and a softmax function to produce the prediction. We test this bi-GRU model as a baseline to investigate the effect of the hierarchical architecture of HAN.

To get the best results, all six approaches involved a model selection on the training data, within each step of the 10-fold cross validation procedure. For the first three traditional models, they selected N features with the highest Pearson's correlation coefficients between each feature and the binary class in the training set. This subset of features were used for building the classifier. For the HAN based models and the bi-GRU baseline, within the training set we further reserved 10% of the samples for validation. We then trained a model on the remaining training samples for many iterations, storing the model parameters after each iteration. The validation data was used for selecting a model that achieved the lowest validation loss.

For the hyper parameters of the HAN models, we set the word embedding dimension to be 300 and the GRU dimension to be 100. The word embeddings were initialized randomly. For training, we used SGD (stochastic gradient descent) with momentum of 0.9 and learning rate of 0.1. The GRU baseline had the same setting as the HAN models. Those hyper parameters were not fine-tuned.

5.2. Experiment Results on DementiaBank

Table 4 summarizes the results. The HAN model achieved performance of 0.815 in both accuracy and F-score. When combined with the age feature, the HAN-AGE model resulted in a remarkable boost in performance, 2.5% improvement in accuracy and 3% improvement in F-score over Masrani (2018)-S. In addition, the HAN model showed a significant increase in performance compared to the bi-GRU baseline, demonstrating the effect of the hierarchical structure in HAN.

5.2.1. Analysis of Effects of Data Set Size

In general, training deep neural network models require large data. To investigate if the HAN models are robust to the size of the training data, we evaluated the two HAN-based models and Masrani (2018)-S from the last experiment with different proportions of the data set. We also included a logistic regression classifier with age being the only feature. Figure 3 reports test accuracy when we repeated the previous experiment with 5%, 15%, 25%, 50% and 75% of the original DementiaBank dataset. For each proportion setting,

Table 4: Binary classification with 10-fold cross-validation. Note that results of Fraser's model and Masrani's model are from the original papers.

Model	Accuracy	Precision	Recall	F-score
Baseline (age only)	0.595	0.591	0.729	0.653
Fraser et al. (2016)	0.820	-	-	-
Masrani (2018)-B	0.822	-	-	0.824
Masrani (2018)-S	0.844	-	-	0.846
bi-GRU baseline	0.748	0.750	0.811	0.768
HAN	0.815	0.839	0.818	0.815
HAN-AGE	0.869	0.859	0.904	0.876

we ran 5 independent experiments (randomly selecting the target subset of the data) and computed the mean and standard deviation. We can see that age is very informative, since a majority class classifier would have an accuracy around 0.5. Note that the performance of HAN drops dramatically when limited training data is used, whereas the HAN-AGE model is much less sensitive to the size of training data. The HAN-AGE model maintains a relatively high performance even with only 5% of the data samples.

5.2.2. Analysis of Attention

During the training process, the attention mechanism makes the HAN model learn which words are important in predicting a given label. To explore this information captured by the attention mechanism, we first performed a qualitative analysis by visualizing the hierarchical attention layers on a small subset of our data (see Figure 4).

In the visualization, each line represents a sentence. Blue denotes the sentence attention weight and red denotes the word attention weight. Figure 4 shows that the model tends to select words like `overflowing, stool, mother`, and `drying`, and their corresponding sentences. Interestingly, these words belong to the set of information units defined by human experts for the Cookie Theft picture. To analyze how much information captured by the attention mechanism overlaps with the human defined information units, we performed further quantitative analysis.

In particular, we performed a statistical test to investigate if HAN pays more attention to information unit words, compared to other words. In order to do this, we considered two categories to which every word token[1] in our dataset belongs to: (i) the word is either in the set of information unit words or not (ii) the word is the most attended in its sentence or not. We then went through all the word tokens in the dataset and counted the frequencies of these two categories. Table 5 shows the resulting contingency table. Now the χ^2 test can tell us whether the two categories are dependent on each other. More technically, it can tell us whether there is a statistical significant difference between the expected frequencies (in parenthesis) and the observed frequencies in the two categories.

1. A word token is a specific occurrence of a word type in a text, for instance the text "the boy is telling the girl but the girl is not listening" contains 8 word types and 12 word tokens.

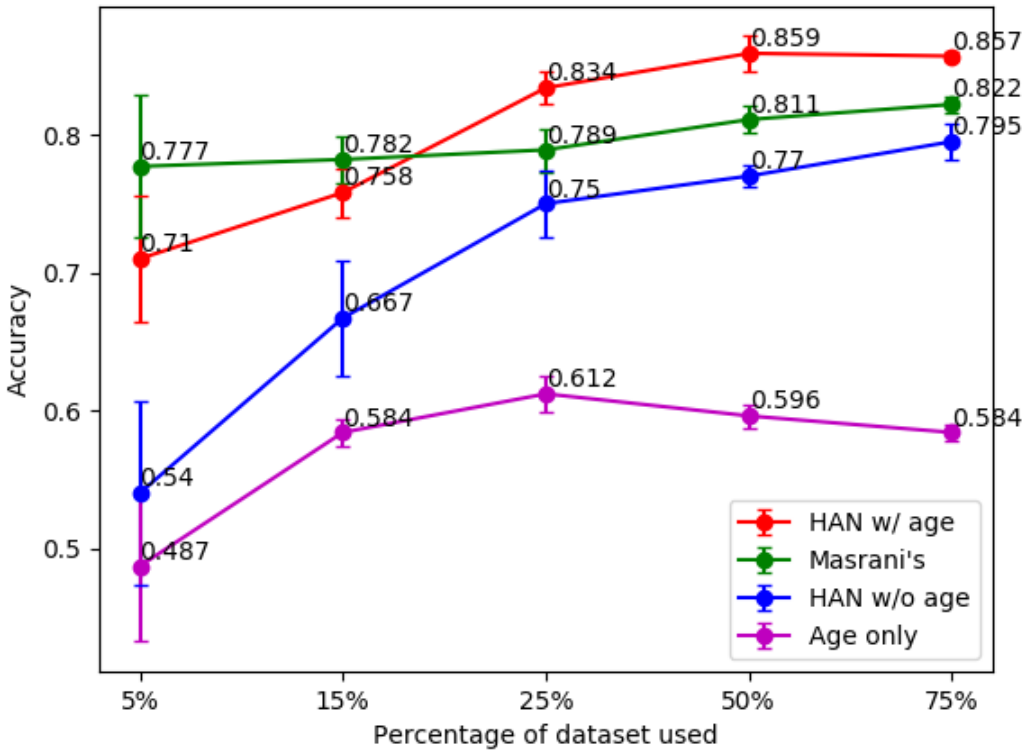


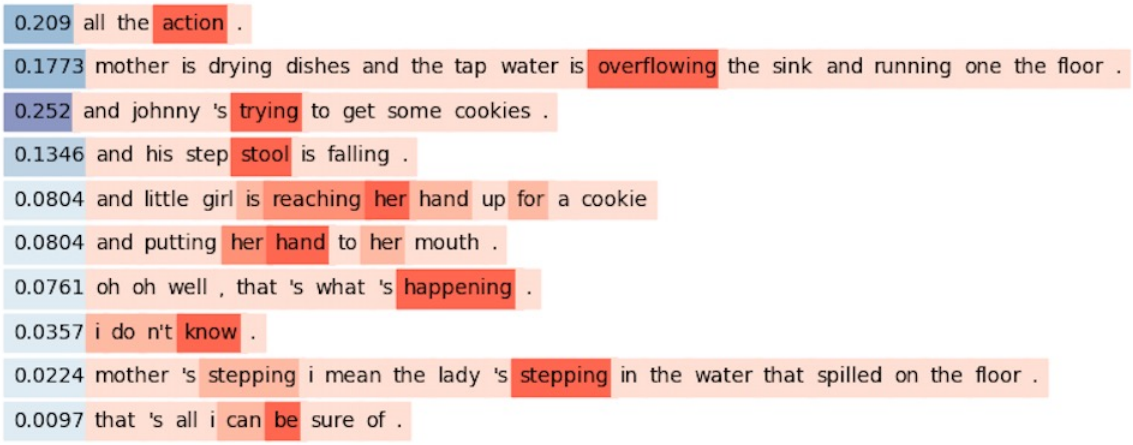
Figure 3: Test accuracy by varying training data proportions.

The result $\chi^2 = 663, p < 0.00001$ shows that the two categories are dependent on each other, i.e., information unit does affect the attention level, with the number of information unit words being the most emphasized (1481), being much bigger than its expectation value (823). So HAN appears to be able to capture similar information to the one specified by human experts.

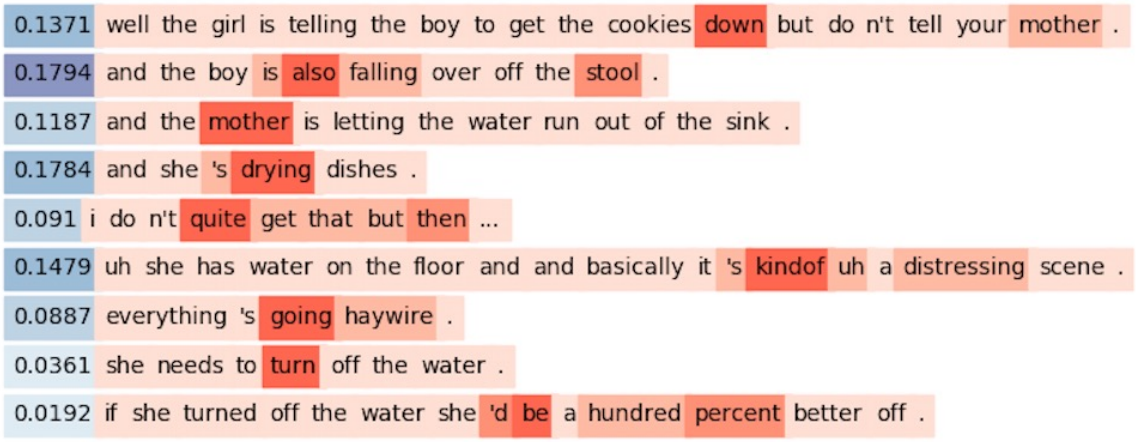
Table 5: Contingency table (numbers in parenthesis are expectation values).

	Most emphasized	**Not most emphasized**	**Total**
Information unit	1481 (823)	7599 (8257)	9080
Non information unit	4889 (5547)	56270 (55612)	61159
Total	6370	63869	70239

Now an interesting question that is still open is whether the attention model is uniformly paying more attention to all the information unit words or it is focusing on a specific subset of the information unit words. To answer this question, we define and compute the *attention frequency* and the *random frequency* for each of the 20 human-defined information units.



(a) Sample id: 059-2 Diagnosis: control Prediction: control



(b) Sample id: 007-3 Diagnosis: AD Prediction: AD

Figure 4: Visualization of attention.

More specifically, for an information unit word, the *attention frequency* was computed as the number of times it was the word with the highest word attention weight in a sentence. Let S_w denote the set of all sentences containing word w and $weight(c, s)$ be the attention weight of word token c in sentence s, we can formalize the computation of *attention frequency* for word type w as

$$Attention\text{-}Frequency(w) = \sum_{s \in S_w} \mathcal{I}[w = \arg\max_c weight(c, s)],$$

where $\mathcal{I}$ is an indicator function.

In contrast, the *random frequency* was computed as the expected number of times the word would have the highest word attention weight, if weights were assigned randomly

within each sentence. Therefore it is defined as follows:

$$Random\text{-}Frequency(w) = \sum_{s \in S_w} \frac{1}{|s|},$$

where S_w denotes the set of all sentences containing word w and $|s|$ is the length of the sentence. The rationale is that if attention weights are assigned at random, a word in a sentence will have the highest attention with probability $1/|s|$.

In Figure 5, the x-axis are 20 human defined information units and y-axis shows their respective frequencies. The results indicate that the model does not attend to all information unit words uniformly. It strongly attends to words like `woman, window, stool, sink, water, wash, cookie, exterior` and `plate`, but pays less attention to words like `dishes, boy, girl`, etc. than what would be expected by their random appearance. Currently, we do not have a satisfactory explanation for why the word attention model is attending more to that specific subset of information unit words. Further investigation is left as future work.

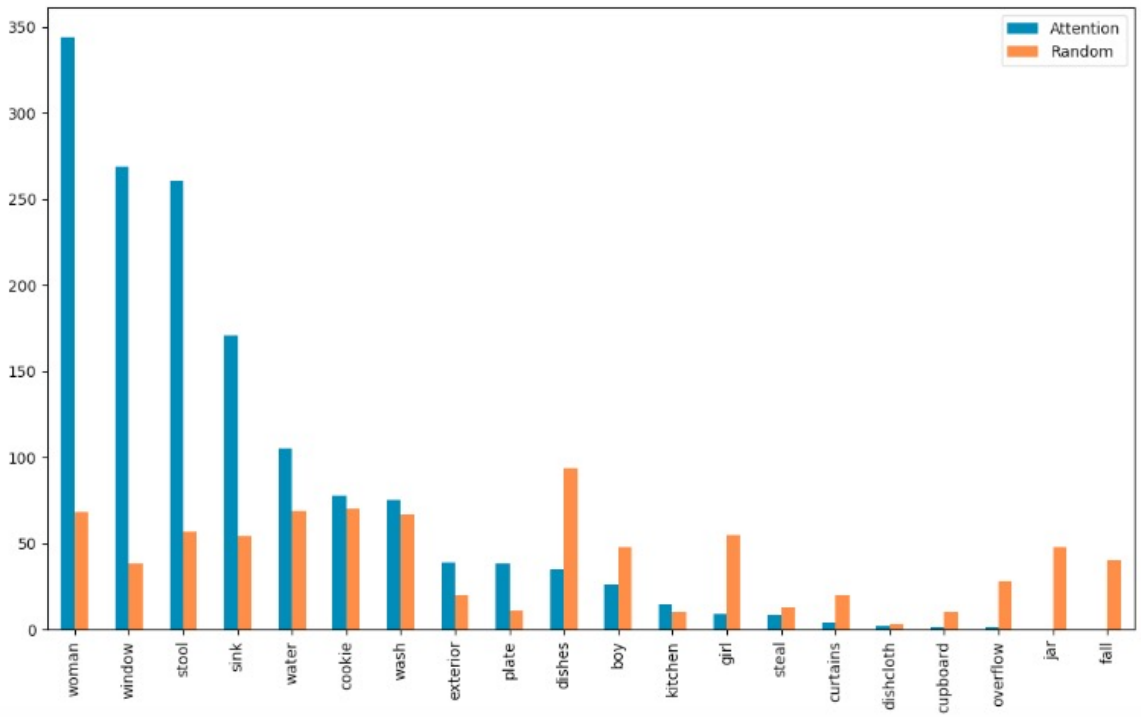


Figure 5: Attention frequency vs. random frequency.

5.3. Evaluation on the Blog Corpus

We evaluate HAN on the Dementia Blog Corpus to see how it performs on written language. We compare our model to Masrani et al. (2017)'s, which built and tested traditional models for predicting dementia on the blog dataset using only the linguistic features, as shown in Table 3. We used 9-fold cross validation as in (Masrani et al., 2017), where each test

fold contains all posts from one dementia blog and one control blog, and the posts from the remaining four blogs were used in the training fold. The model selection process was carried out as described in section 5.1.

Table 6: Binary classification with 9-fold cross-validation on blog corpus.

Model	**Accuracy**	**F-score**
Majority class	0.590	0.742
Masrani et al. (2017)	0.724	0.785
HAN	0.579	0.582

The experiment results are summarized in Table 6. The traditional model demonstrates that dementia can also be automatically predicted from written text in the form of blog posts. However, the HAN model fails in this task. A key difference that may explain this result, is that the samples in DementiaBank are descriptions of one single picture and so are all about the same topic (i.e., same objects and events, resulting in a corpus vocabulary of 1828 word types). In contrast, samples from the blog data cover a large variety of topics, ranging from regular medical appointments to re-connecting an old friend on Facebook (with a much larger vocabulary size of 27413). The HAN model succeeded in focusing on informative concepts shown in the Cookie Theft picture, with the help of the attention layers. However, for blog data there are no such concepts shared across all blog posts. Thus the data are likely not sufficient to cover such a much larger vocabulary, resulting in the extremely poor performance of HAN. On the contrary, the traditional machine learning method is quite effective on blog posts, likely because its large human engineered set of features also include features that are not lexically based (i.e., based on words), but instead capture task-independent aspects of language like syntactic constituents and syntactic complexity.

To further explore the large difference in performance between neural and traditional methods on blog data, we conducted an additional experiment. Unlike the original split setting where all posts from the same blog are contained either in the training fold or the test fold, here we shuffle all the posts regardless which blogs they belong to, and divide them into 10 folds for cross validation. In this scenario, posts from the same blog will very likely appear in both the training and testing data, creating a form of data contamination. Not surprisingly, the HAN model is very accurate on this artificial task, with an average accuracy and F-score as high as 0.934 and 0.944, respectively. This could be because HAN captures the writing style and topics of each blogger rather than informative patterns for dementia prediction.

6. Conclusions and Future Work

Early prediction of dementia is extremely important, as researchers believe that early diagnosis will be key to slowing and stopping the disease. Medical methods for early prediction, such as positron emission tomography (PET) and magnetic resonance imaging (MRI), are expensive and invasive

In this study, we tackled the problem of predicting dementia from language, which could result in much less costly and non-invasive solutions. We extended previous work based on traditional machine learning methods and engineered features, by applying a neural model on language samples of elderly people to classify dementia patients from healthy controls. By incorporating age as extra information, the model not only achieved the state-of-the-art performance on the DementiaBank dataset, but showed a decent prediction accuracy even when trained with a small portion of the available data. Visualization and statistical analysis revealed that the attention mechanism of the model manages to capture similar key concepts as the information unit features specified by human experts. Although our task-agnostic method was only tested on an English dataset describing the Cookie Theft picture, it could be generalized to other cultures and languages. This would be particularly useful for applying the method in developing countries, which have an even more pressing need for inexpensive solutions.

Meanwhile, the blog experiment results indicated that HAN is not a universal classifier for predicting dementia from language. In the task where samples are not all about a single topic, a traditional model that exploits linguistic features (e.g., *syntactic complexity, context-free grammar rules*) appears to be a better choice than HAN.

A key limitation of predicting dementia from language is the scarcity of related data sets. The DementiaBank dataset seems to contain sufficient data (257 AD samples and 242 controls) to train neural text categorization models like HAN. However, there are only 5 vascular dementia samples and no sample at all for other types of dementia (e.g., dementia with Lewy body). Automatic diagnosis of different sub-types of dementia will not be possible until more data is collected. This is something we are currently doing in the clinic.

References

A.D. International. Dementia statistics. https://www.alz.co.uk/research/statistics, 2015. Accessed: 2019-2-13.

Samrah Ahmed, Celeste A de Jager, Anne-Marie Haigh, and Peter Garrard. Semantic processing in connected speech at a uniformly early stage of autopsy-confirmed alzheimer's disease. *Neuropsychology*, 27(1):79, 2013.

Sabah Al-Hameed, Mohammed Benaissa, and Heidi Christensen. Detecting and predicting alzheimer's disease severity in longitudinal acoustic data. In *Proceedings of the International Conference on Bioinformatics Research and Applications 2017*, pages 57–61. ACM, 2017.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

James T Becker, François Boiler, Oscar L Lopez, Judith Saxton, and Karen L McGonigle. The natural history of alzheimer's disease: description of study cohort and accuracy of diagnosis. *Archives of Neurology*, 51(6):585–594, 1994.

Larry W Chambers, Christina Bancej, and Ian McDowell. *Prevalence and monetary costs of dementia in Canada: population health expert panel.* Alzheimer Society of Canada in collaboration with the Public Health Agency , 2016.

Bernard Croisile, Bernadette Ska, Marie-Josee Brabant, Annick Duchene, Yves Lepage, Gilbert Aimard, and Marc Trillet. Comparative study of oral and written picture description in patients with alzheimer's disease. *Brain and language*, 53(1):1–19, 1996.

Katrina E Forbes-McKay and Annalena Venneri. Detecting subtle spontaneous language decline in early alzheimers disease with a picture description task. *Neurological sciences*, 26(4):243–254, 2005.

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

Sujuan Gao, Hugh C Hendrie, Kathleen S Hall, and Siu Hui. The relationships between age, sex, and the incidence of dementia and alzheimer disease: a meta-analysis. *Archives of general psychiatry*, 55(9):809–815, 1998.

Vaden Masrani. Detecting dementia from written and spoken language. Master's thesis, University of British Columbia, 2018.

Vaden Masrani, Gabriel Murray, Thalia Field, and Giuseppe Carenini. Detecting dementia through retrospective analysis of routine blog posts by bloggers with dementia. *BioNLP 2017*, pages 232–237, 2017.

Felix Nensa, Karsten Beiderwellen, Philipp Heusch, and Axel Wetter. Clinical applications of pet/mri: current status and future perspectives. *Diagnostic and Interventional Radiology*, 20(5):438, 2014.

Sylvester Olubolu Orimaye, Jojo Sze-Meng Wong, and Karen Jennifer Golden. Learning predictive linguistic features for alzheimers disease and related dementias using verbal utterances. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 78–87, 2014.

Abderrahim Oulhaj, Gordon K Wilcock, A David Smith, and Celeste A de Jager. Predicting the time of conversion to mci in the elderly: role of verbal expression and learning. *Neurology*, 73(18):1436–1442, 2009.

Holly Posner, Rosie Curiel, Chris Edgar, Suzanne Hendrix, Enchi Liu, David A Loewenstein, Glenn Morrison, Leslie Shinobu, Keith Wesnes, and Philip D Harvey. Outcomes assessment in clinical trials of alzheimers disease and its precursors: readying for short-term and long-term clinical trial needs. *Innovations in clinical neuroscience*, 14(1-2):22, 2017.

Alexandros Tsaptsinos. Lyrics-based music genre classification using a hierarchical attention network. *arXiv preprint arXiv:1707.04678*, 2017.

Jennifer A Williams, Alyssa Weakley, Diane J Cook, and Maureen Schmitter-Edgecombe. Machine learning techniques for diagnostic differentiation of mild cognitive impairment and dementia. In *Workshops at the twenty-seventh AAAI conference on artificial intelligence*, 2013.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489, 2016.

Lei Zhang, Shuai Wang, and Bing Liu. Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):e1253, 2018.