AUTOMATIC DIAGNOSIS OF ALZHEIMER'S DISEASE USING NEURAL NETWORK LANGUAGE MODELS

Julian Fritsch [1,2], *Sebastian Wankerl* [3], *Elmar Nöth* [3]

[1] Idiap Research Institute, Martigny, Switzerland
[2] École polytechnique fédérale de Lausanne (EPFL), Lausanne, Switzerland
[3] Friedrich-Alexander-University Erlangen-Nuremberg, Germany
julian.fritsch@idiap.ch

ABSTRACT

In today's aging society, the number of neurodegenerative diseases such as Alzheimer's disease (AD) increases. Reliable tools for automatic early screening as well as monitoring of AD patients are necessary. For that, semantic deficits have been shown to be useful indicators. We present a way to significantly improve the method introduced by Wankerl et al. [1]. The purely statistical approach of n-gram language models (LMs) is enhanced by using the rwthlm toolkit to create neural network language models (NNLMs) with Long Short Term-Memory (LSTM) cells. The prediction is solely based on evaluating the perplexity of transliterations of descriptions of the Cookie Theft picture from DementiaBank's Pitt Corpus. Each transliteration is evaluated on LMs of both control and Alzheimer speakers in a leave-one-speaker-out cross-validation scheme. The resulting perplexity values reveal enough discrepancy to classify patients on just those two values with an accuracy of 85.6% at equal-error-rate.

Index Terms— Alzheimer's disease, automatic diagnosis, language models, neural network language models, Long-Short Term Memory

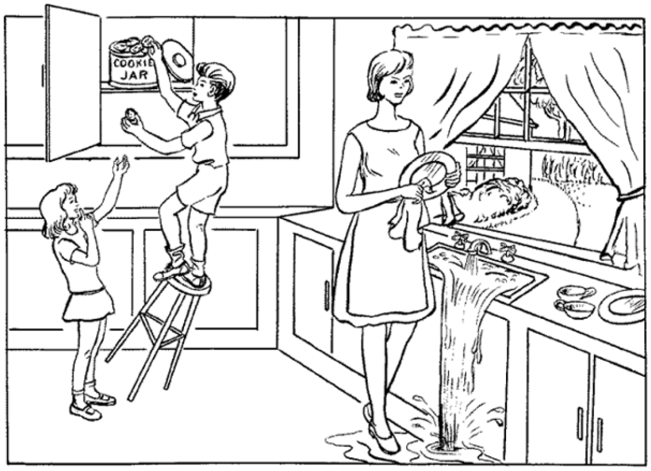

Fig. 1. Cookie Theft Picture [4]

1. INTRODUCTION

According to the World Alzheimer Report 2016, 47 million people live with dementia worldwide today, while this number is expected to increase to more than 131 million by 2050 [2]. This increase can be retraced to aging populations, but also to a huge majority of people that so far have not received a diagnosis. Since no cure exists, medication at best allows to alleviate or decelerate symptoms. To choose the right treatment, diagnosing the onset of the disease plays the pivotal role. AD accounts for around 60% of all cases of dementia. Its most typical cognitive deficit is memory loss, typically progressing to loss of cognition [3]. Yet, changes of cognitive capabilities, such as oral and written language that lie beyond what is expected due to age, can be measured. To investigate these changes, analyzing spontaneous speech has proven reasonable. That is why the Cookie Theft picture (Figure 1) description task is widely used in neurological tests as it is considered a very natural approximation to spontaneous discourse. The constrained context allows to use linguistic approaches, but also good comparability to other studies. The Cookie Theft picture description was first included in the popular Boston Diagnostic Aphasia Examination protocol [4]. Similar to intelligence tests it may be described on different levels of abstraction, ranging from very simple to more abstract depending on the person's cognitive capabilities, e.g. identifying the woman as the children's mother or the risk of the boy falling of the stool. We present a method that relies on transcribed Cookie Theft picture descriptions. We aim to distinguish Alzheimer and control group, even though estimating a person's cognitive capabilities is the goal. To assess cognitive capabilities several assessment scales exist, the most common being the mini-mental state examination (MMSE) protocol. It contains 30 questions regarding e.g. orientation, calculation or comprehension. The median for healthy people with at least nine years of schooling is 29. Mild dementia is associated with a score of 21-26, moderate dementia with a score of 10-20, severe dementia with a score less than 10 [5]. When conducting the examination, daily condition or discomfort might distort the patient's performance. Mitchell [6] reports a sensitivity of the MMSE in a clinical setting of 79.8%. Therefore, the goal of this analysis is not to reproduce the MMSE result, but to reach human parity in examining patients. We assume that a good correlation of our method with MMSE scores is useful, but not mandatory. Pre-selection of people that have a risk of developing a dementia is the main use-case of this work. Ideally, a test could be taken on a tablet by asking a person to describe the picture. Once the person provides a picture description, an automatic speech recognizer transcribes the picture before the presented method could

This work was funded through European Union's Horizon H2020 Marie Skłodowska-Curie Actions Innovative Training Network European Training Network (MSCA-ITN-ETN) project TAPAS under grant agreement No. 766287.

be used to analyze the transliteration. As this procedure is quick, it could be repeated regularly and on a large scale and thereby reduce the costs for the health care system. Furthermore, by predicting the disease's progress in terms of MMSE score could help to determine the optimal use of pharmacological treatments and to deliver effective care.

2. RELATED WORK

Weiner et al. [7] presents an approach to dementia detection that compares feature extraction from manual and automatic transcriptions created by automatic speech recognition. The study analyzes data from 74 participants, who provided a total of 230 hours of audio recordings, over 3 sessions in a period of 15 years associated to 3 groups: 80 control, 13 aging-associated cognitive decline and 5 AD recordings. They contain biographic interviews and cognitive diagnoses without interviewer speech. Using an in-house speech recognition toolkit, the overall word-error rate is 58.2%. Nevertheless, it is shown that the majority of linguistic features created from automatic transcriptions outperform their manual version. The results are presented using the unweighted average recall (UAR) metric, a Gaussian classifier and by performing a leave-one-person-out cross-validation. The overall best result of $UAR = 0.623$ is achieved by the automatic version of the within-speaker perplexity, which are calculated using the SRILM toolkit [8]. This underlines, that linguistic features using automatic transcriptions are robust against transcription quality.

A more extensive approach to diagnose AD has been taken by Fraser et al. [9]. A large variety of features on semantic, acoustic, syntactic and information impairment were extracted from a selection of Alzheimer and control patients from DementiaBank's Pitt corpus (cf. sec. 4). The features were grouped and analyzed using a factor analysis. On a total of 370 features an accuracy of 58.51% was achieved, but using logistic regression to identify the 35 most discriminative features yielded an accuracy of 81.92%.

Asgari et al. [10] does a linguistic analysis on unstructured conversational speech of mild cognitive impairment (MCI) and controls. Data was collected during the course of a 6-week randomized clinical trial of daily online video chats on preselected topics. The author uses data from 14 MCI and 27 control participants. The data is automatically transcribed using an ASR system. The word stems of the resulting texts are grouped using the Berkeley aligner into 5 top-level categories to cluster topic-related words or of 68 subcategories per top-level category; about 40% of words were not classifiable. The occurrences of words per category are counted and normalized by the total number of words to build a 68-dimensional vector. This method is referred to as Linguistic Inquiry and Word Count. For classification, Support Vector Machine and Random Forests are used. Due to the imbalance between MCI and control participants a 5-fold cross-validation is repeated until the accuracy converged. In a variety of experiments, the overall best classification accuracy is 84%.

Klumpp et al. [11] reports results on the same selection of Cookie Theft picture description transliterations from DementiaBank's Pitt Corpus as presented in this work. The transliterations are stemmed to form one bag-of-words vectors of all occuring 546 words per transliteration. The bag-of words vectors are input to a three-layer neural network that outputs a probability that a transliteration belongs to the Alzheimer class. The experimental scheme is a leave-one-speaker cross-validation, identical to the protocol of this work. The overall accuracy is 84.4%, which shows that even though the word order is neglected, the global vocabulary usage is suitable to discriminate between Alzheimer and control patients.

3. THEORY AND BACKGROUND

3.1. N-gram language models

N-gram LMs attempt to reflect the frequency with which each word or word sequence occurs in natural text. The probability for a sequence s is estimated by the product of the probabilities of the words that compose the sequence. This is simplified by only considering a history of $n-1$ preceding words:

$$P(s) = \prod_{i=1}^{l} P(w_i|w_1^{i-1}) \approx \prod_{i=1}^{l} P(w_i|w_{i-n+1}^{i-1}) \qquad (1)$$

where w_i denotes word i, w_i^j denotes words $w_i, ..., w_j$ and n is the number of preceding words taken into account. When setting n to one, two or three models are referred to as unigram, bigram, trigram models, respectively. One of the key issues in n-gram language modeling is the sparsity of training data. Considering for example, a text corpus with a vocabulary size of 100000 and trigram word sequences gives $100,000^3 - 1 = 10^{15} - 1$ potential combinations. The relative frequency as an estimate of the probability for a sequence of words s that does not occur in the training data results in assigning $P(s) = 0$, where the probability clearly should be larger than zero. For that, smoothing, meaning a redistribution of the probability mass alleviates that issue [12].

3.2. Neural network language models

As for n-grams, the goal of building an NNLM is to learn the joint probability of sequences of words. However, NNLMs are designed to overcome the sparse-data problem by using a different representation of words: Words are encoded by vectors, which will learn to encode words with similar meaning close to each other in a continuous vector space, so called word embeddings. This concept was first introduced by Mikolov et al. [13]. By relating words to each other and by that also sentences, a generalization forms that addresses the sparse data problem.

The general process boils down to the following steps: Each word in the vocabulary V is one-hot encoded by a $|V|$-dimensional vector. These vectors are input to a projection layer, which concatenates the words and maps them to their vector space representation. The projection layer is followed by variable number of hidden layers, which is followed by an output layer of vocabulary size $|V|$ target probabilities. The output probabilities $P(w_i|h_i)$ for each word w_i given history h_i are computed by applying the softmax function. In this case, the probability mass is distributed over only the target words, while unknown words are assigned zero unless one introduces a token for unknown words [14]. Finally, the network's parameters are adapted using gradient-based optimization algorithms.

When NNLMs are built, the following has to be considered: Using feed-forward layers will, according to the chain rule, result in predicting words of all predecessors of a sequence without taking the respective order into account [15]. On the other hand recurrent layers are very convenient for processing sequential data. LSTM layers are a special kind of recurrent layers, that expand the concept of recurrence by introducing so called gates, to optionally let information through and remember long-term dependencies. The LSTM cell's design, e.g. hidden state and forget gate, are suited to learn when to remember and when to remove information [16]. This allows variable context lengths, which was shown to outperform n-grams, feed-forward and recurrent neural networks [15].

3.3. Language model evaluation – perplexity

The performance of an LM in predicting a test sample, meaning a sequence of words $S = w_1, ..., w_N$, is often measured by calculating its perplexity:

$$PPL(S) = P(S)^{-\frac{1}{N}} \tag{2}$$

where $P(S)$ denotes the probability assigned to sequence S. The perplexity is inversely related to the average probability a model assigns to a sequence. Note, that the perplexity is normalized by the number of words N. When evaluating an LM, a well-fitting LM tends to assign high probabilities to the test sequence S. Therefore, lower perplexity indicates better performance through better predictability of the test sequence [12].

4. DATA

The DementiaBank's Pitt Corpus [17] contains audio recordings and the corresponding transcriptions of English Cookie Theft picture descriptions. From a total of 292 participants 194 suffered from some sort of dementia and 98 healthy speakers serve as control group. Preconditions were a minimum age of 44 years, an initial MMSE score of 10, at least 7 years of education and no history of disorders of the nervous system. The examinations were conducted on a yearly basis up to seven times, each time recording a variety of mental examinations such as MMSE. Unfortunately, there are recordings without a corresponding MMSE score. The participants diagnosed with dementia are mainly of Alzheimer's type or probable Alzheimer's. However, there are a few people with another type of dementia who were excluded from this study. Thus, 168 patients, who were diagnosed with AD or probable AD, contributing 255 transcriptions remain. This group consists of 55 males and 113 females. In the control group, 98 speakers provided 244 transcriptions, consisting of 31 males and 67 females. In the transcriptions, fillers, such as *uhm,* uhh, repetitions, paraphrases, grammatical mistakes and requests uttered by the participant were kept to obtain a more accurate copy of the actual recording. Annotations which are not directly linked to the utterances of the test subject (e.g. clears throat), are removed.

5. METHODS

To create LMs the rwthlm toolkit is used [18]. The rwthlm toolkit is an open source C++ library that implements the concept of NNLM, using especially LSTMs. The toolkit offers all features necessary to train an NNLM, most importantly a customizable network architecture. Supported layer types are linear, feed-forward, standard recurrent and LSTMs. By definition the first layer must be a linear layer with identity activation function. The main focus of this work was on using LSTM layers, since recurrence allows for flexible context lengths. Results were assessed using a linear and an LSTM layer of the same size N. Network optimization highly depends on a good initial learning rate. Conceptually, faster convergence is achieved by a high initial learning rate, but too high values cause the perplexity to fluctuate. This behavior is handled by the toolkit by decreasing the learning rate as soon as a full epoch leads to a degradation in perplexity. When creating LMs, the toolkit expects a vocabulary list, a file containing the training text data, a file containing the development text data and an initial learning rate. The learning rate was tuned individually to the used architecture, so that the development perplexity after the first epoch is minimal. The development set was created by randomly selecting ten speakers of the respective group, who were then excluded from the training data. Note, that this may lead to different numbers of transliterations used as development set, as not every speaker provided the same number of picture descriptions. Additionally, instead of applying some early stopping criterion to either learning rate or development perplexity per epoch, the maximal number of epochs is set to 20.

The presented experimental setup was introduced in [1], for which the selection of 168 Alzheimer's and 98 control participants, i.e. 255 and 244 transliterations of the Pitt corpus is used. For both groups, all transliterations per group are used to create an Alzheimer's LM $\mathcal{M}_{Alzheimer}$ and a control LM $\mathcal{M}_{Control}$ by using the rwthlm toolkit. These are supposed to represent the typical Alzheimer's and healthy speech respectively. Then, a leave-one-speaker-out cross-validation is performed, which means that for all speakers, an LM $\mathcal{M}_{-s}$ is created, using all transliterations of that group but those of speaker s and the speakers excluded for the development set. The cross-validation is conducted per speaker, as speakers that provided multiple recordings are expected to use similar phrases that are not always representative for the whole group and would distort the results. The transliterations provided by speaker s are then used to evaluate a test set perplexity, also by using the rwthlm toolkit. A speaker s is always tested on $\mathcal{M}_{-s}$, giving a perplexity p_{own}. Additionally, the speaker is tested on the respective other LM, giving p_{other}, using $\mathcal{M}_{Alzheimer}$, if belonging to the control group, or $\mathcal{M}_{Control}$, if belonging to the Alzheimer's group. Naturally, p_{own} is expected to be lower than p_{other}, as the own group's LM is supposed to represent the test set better than the other group's LM. Finally, the difference between the perplexity evaluated on an Alzheimer's LM and the control's LM is calculated for each speaker, giving following distinction for the groups:

$$p_{diff} = \begin{cases} p_{own} - p_{other} & if\ s \in Alzheimer's\ group \\ p_{other} - p_{own} & if\ s \in Control's\ group \end{cases} \tag{3}$$

In order to actually classify the transliterations, the perplexity difference p_{diff} is considered. A binary decision is made by setting a threshold, such that both groups have equal error rate (EER). The results are assessed at EER, as this gives a robust estimate of a classifiers performance. Note, that to classify an unknown sample, p_{diff} would be obtained by subtracting the perplexity evaluated on $\mathcal{M}_{Control}$ from the perplexity provided by $\mathcal{M}_{Alzheimer}$.

6. RESULTS

The experimental setup, as described in sec. 5, can be executed for different neural network architectures. We are aware that comparing the accuracy of different architectures means implicitly optimizing on the test set. This is done for a lack of enough data. Different layer sizes N were evaluated. The overall best result was achieved using a linear and an LSTM layer of size $N = 150$, resulting in 72 wrongly classified transliterations at EER. This means 85.6% accuracy regarding a total of 499 recordings. The 36 wrongly classified transliterations per group in both cases come from 27 different speakers. In the same experimental setup, when creating trigram LMs using SRILM the overall accuracy is 77.1% with 114 wrongly classified transliterations [1]. This means that the result could be improved by 8.5% absolute corresponding to 42 wrongly classified transliterations less.

Figure 2 shows the histogram of MMSE scores for our selection of speakers from the Pitt corpus, for all speakers from whom an MMSE score is available. These are 234 recordings from 166 AD

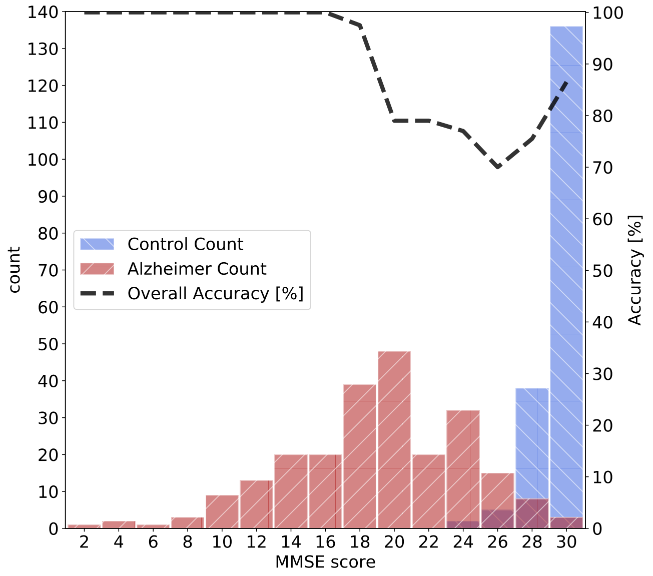

Fig. 2. Histogram of all MMSE scores in the Pitt corpus and overall accuracies per MMSE score.

Table 1. Correlations between MMSE scores and perplexity difference p_{diff}.

	r	ρ
Alzheimer	0.433	0.547
Control	0.112	0.109
All	0.656	0.771

and 181 recordings from 94 healthy speakers. The dashed black line indicates the overall accuracy per MMSE score.

Table 1 shows Pearson's and Spearman's correlation between MMSE scores and the perplexity difference values p_{diff}. The correlation values for the Alzheimer group indicate that p_{diff} can be used to predict a person's MMSE score. The control group's correlation is low. Considering that the MMSE scores only range between 24 and 30, but the perplexity difference values still vary to a larger extend, a low correlation value is to be expected. Moreover the correlation of all MMSE values and all respective perplexity differences indicates the potential usage to predict a person's MMSE score without knowing a diagnosis.

The performance of a classifier can also be evaluated by analyzing its ROC curve. In order to simulate a screening scenario we compare our result's ROC curve obtained from all speakers with the ROC curve when only evaluating transliterations with an MMSE score from 21 to 30. This means speakers are either healthy or diagnosed with a mild dementia. 78 transliterations from 59 speakers remain in the AD group. Figure 3 shows the two ROC curves. At an overall accuracy of 85.6% of all transliteration, the false positive rate (FPR) is 15%, the true positive rate (TPR) is at 86%, meaning that while only classifying 15% of control speakers wrongly, 86% of Alzheimer speakers are classified correctly. In the screening scenario an overall accuracy of 79.5% at an equal-error-percentage per group and 66 wrongly classified transliterations, 20% FPR correspond to 21% TPR. At 90% TPR, the FPR on all transliterations is still only 25%, whilst on the subset from 21 to 30 MMSE scores the FPR is 35%. The AUC is 0.92 compared to 0.87 in the screening scenario.

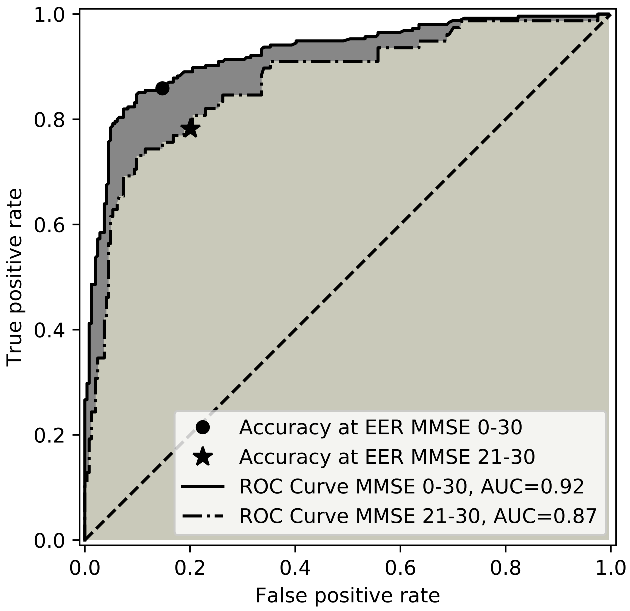

Fig. 3. Comparison of accuracies and ROC curves of the results from all speakers and speakers with an MMSE score from 21 to 30.

7. DISCUSSION

Just by comparing perplexity values, the database can be separated with an accuracy of 85.6% on a total of 499 transliterations. This approach exploits the assumption that Alzheimer patients describe the picture in an unforeseen way, which leads to more unpredictable language structures resulting in higher perplexity values. Figure 2 illustrates that clear cases, e.g. those with low and high MMSE scores, have a high recognition rate. The borderline cases at the transition from dementia to control group having an MMSE score between 24 and 26 are less obvious and therefore have a lower recognition rate. Yet, evaluating this approach on only healthy and mildly demented subjects shows still high enough accuracy to be used as a screening tool. This represents an improvement of the method presented by Wankerl et al. [1] on a scale, that it may be applicable in a realistic screening scenario. In addition, correlating all MMSE scores with the respective perplexity difference yields a Pearson's correlation of $r = 0.656$ and Spearman's correlation of $\rho = 0.771$. This shows, that this method can also be used to predict a patient's MMSE score.

Generally, this method is not supposed to replace a physician's examination, but to pre-select people that should undergo an examination. If this approach had to be fully automated, a speech recognizer would be prepended. So far, we are implicitly assuming 100% recognition rate, as we work on hand-transcribed data. As shown in [7], linguistic features may even be more robust on speech recognition output. Nevertheless it has to be investigated, how speech recognition errors influence the presented method's performance. Ultimately, to consecutively assess people's cognitive capabilities, this method could be extended by more pictures to avoid learning effects. A clear advantage of this purely statistical approach of using LMs for automatic dementia diagnosis is that it is not relying on any further linguistic or medical annotations. Therefore, it is likely, that this approach is not restricted to the English language, thus may work in another language as well.

8. REFERENCES

[1] S. Wankerl, E. Nöth, and S. Evert, “An n-gram based approach to the automatic diagnosis of alzheimer’s disease from spoken language,” in *Proc. Interspeech*, 2017.

[2] M. Prince, A. Comas-Herrera, M. Knapp, M. Guerchet, and M. Karagiannidou, *World Alzheimer Report 2016*, Alzheimer’s Disease International (ADI), London, 2016.

[3] A. Husband and W. Alan, “Different types of dementia,” *The Pharmaceutical Journal*, , no. 277, pp. 579–582, November 2006.

[4] H. Goodglass and E. Kaplan, “Boston diagnostic aphasia examination,” *Lea & Febiger*, 1983.

[5] M. F. Folstein, S. E. Folstein, and P. R. McHugh, “Mini mental state a practical method for grading the cognitive state of patients for the clinician,” *Journal of Psychiatric Research*, vol. 12, no. 3, pp. 189–198, 1975.

[6] A. J. Mitchell, “A meta-analysis of the accuracy of the mini-mental state examination in the detection of dementia and mild cognitive impairment,” *Journal of psychiatric research*, vol. 43, no. 4, pp. 411–431, 2009.

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

[8] A. Stolcke, “Srilm-an extensible language modeling toolkit,” in *Seventh international conference on spoken language processing*, 2002.

[9] K. C. Fraser and F. Meltzer, J. A. Rudzicz, “Linguistic features identify alzheimer’s disease in narrative speech,” *Journal of Alzheimer’s Disease*, vol. 49, no. 2, pp. 407–422, 2016.

[10] M. Asgari, J. Kaye, and H. Dodge, “Predicting mild cognitive impairment from spontaneous spoken utterances,” in *Alzheimer’s & dementia*. 2017, vol. 3, pp. 219–228, Elsevier.

[11] P. Klumpp, J. Fritsch, and E. Nöth, “Ann-based alzheimers disease classification from bag of words,” *Proc. ITG*, pp. 341–344, 2018.

[12] S. F. Chen and J. Goodman, “An empirical study of smoothing techniques for language modeling,” *Computer Speech & Language*, vol. 13, no. 4, pp. 359–394, 1999.

[13] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” *CoRR*, vol. 1301.3781, 2013.

[14] E. Arisoy, T. N. Sainath, B. Kingsbury, and B. Ramabhadran, “Deep neural network language models,” in *Proceedings of the NAACL-HLT 2012 Workshop: Will We Ever Really Replace the N-gram Model? On the Future of Language Modeling for HLT*. Association for Computational Linguistics, 2012, pp. 20–28.

[15] M. Sundermeyer, R. Schlüter, and H. Ney, “Lstm neural networks for language modeling,” in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[16] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[17] 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.

[18] M. Sundermeyer, R. Schlüter, and H. Ney, “rwthlm – the rwth aachen university neural network language modeling toolkit,” in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.