INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Improved Neural Bag-of-Words Model to Retrieve Out-of-Vocabulary Words in Speech Recognition

Imran Sheikh[*+], *Irina Illina*[*], *Dominique Fohr*[*], *Georges Linarès*[+]

[*]Université de Lorraine, LORIA, UMR 7503, Vandoeuvre-lès-Nancy, F-54506, France
[*]Inria, Villers-lès-Nancy, F-54600, France
[*]CNRS, LORIA, UMR 7503, Vandoeuvre-lès-Nancy, F-54506, France
[+]Laboratoire Informatique d'Avignon, University of Avignon
{imran.sheikh, irina.illina, dominique.fohr}@loria.fr, georges.linares@univ-avignon.fr

Abstract

Many Proper Names (PNs) are Out-Of-Vocabulary (OOV) words for speech recognition systems used to process diachronic audio data. To enable recovery of the PNs missed by the system, relevant OOV PNs can be retrieved by exploiting the semantic context of the spoken content. In this paper, we explore the Neural Bag-of-Words (NBOW) model, proposed previously for text classification, to retrieve relevant OOV PNs. We propose a Neural Bag-of-Weighted-Words (NBOW2) model in which the input embedding layer is augmented with a context anchor layer. This layer learns to assign importance to input words and has the ability to capture (task specific) key-words in a NBOW model. With experiments on French broadcast news videos we show that the NBOW and NBOW2 models outperform earlier methods based on raw embeddings from LDA and Skip-gram. Combining NBOW with NBOW2 gives faster convergence during training.
Index Terms: lvcsr, oov, proper names

1. Introduction

The diachronic nature of news content causes frequent variations in the linguistic content and vocabulary, leading to *Out-Of-Vocabulary* (OOV) words problem for *Large Vocabulary Continuous Speech Recognition* (LVCSR). Simply appending the LVCSR vocabulary and updating the *Language Model* (LM) will (a) require a good amount of training data and/or (b) affect the LVCSR performance and complexity. An analysis of the OOV words reveals that majority of OOV words (56-72% [1]) are Proper Names (PNs). However PNs are important for obtaining accurate automatic transcriptions as well as automatic indexing of audio-video content. In this paper, we focus on the problem of retrieval of OOV PNs relevant to an audio document.

To retrieve OOV PNs relevant to an audio document we rely on their semantic context. During the training phase, diachronic text news with new (i.e., OOV) PNs are collected from the internet. This set of text documents, referred as a *diachronic text corpus*, is used to learn a context vector space which captures relationships between the *In-Vocabulary* (IV) words, PNs and the OOV PNs. During the test, the LVCSR hypothesis of the audio document is projected into the context space and then relevant OOV PNs are inferred. In our previous work [1] we have shown that methods based on *Latent Dirichlet Allocation* (LDA) topic space can perform well for retrieval of the *target OOV PNs*[1]. Alternative methods to learn word and context representations [2, 3], based on predicting the context in which words appear, have become popular. These representations have been shown to perform effectively in a range of applications and tasks [4].

In this paper, we present an improved *Neural Bag-of-Words* (NBOW) model [5]. The simple, yet impressive, NBOW model takes an average of the word vectors in the input sequence and performs classification with a fully connected layer. We propose to replace the average with a weighted sum, where the weights applied to each word (vector) are learned during the training of the model. We refer to this as the *Neural Bag-of-Weighted-Words* (NBOW2) model. With experiments on French broadcast news videos, we show that (a) the proposed NBOW2 model learns meaningful word importance weights, (b) the NBOW2 model, like the NBOW model, outperforms the baseline methods based on raw embeddings from LDA and Skip-gram, (c) NBOW2+ model, which combines the context vectors from the NBOW and NBOW2 models, gives faster convergence during training (d) the improved retrieval performance translates to an improvement in the recovery of the target OOV PNs.

1.1. Related Work

The task of retrieval of OOV and PNs relevant to an audio document has been presented in previous works. These include methods based on probabilistic topic models applicable to common PNs [6, 7] and those addressing even the less frequent PNs [1, 8]. Word embedding based methods to retrieve relevant PNs have been tried for audio documents with multiple news events [9]. More recently [10] document similarity based methods have been shown to perform better, especially for retrieval of less frequent PNs. Compared to these works we explore neural networks trained to retrieve relevant OOV PNs, for audio documents with a single or coherent news event.

Our methodology in this paper is related to the recent approaches of text classification with neural networks. In this context, fully connected feed forward networks [5, 11], *Convolutional Neural Networks* (CNN) [12, 13, 14] and also *Recurrent Neural Networks* (RNN) [15, 16, 17, 18, 19] have been applied. On one hand, the approaches based on CNN and RNN capture rich compositional information, and have been outperforming the state-of-the-art results in text classification; on the other hand they are computationally intensive and require careful hyper-parameter selection and/or regularisation [20, 19]. For our task we rely on document level bag-of-words architectures mainly because they are suitable to process LVCSR transcriptions of audio documents, which are firstly prone to noise in

[1]For a given audio documents several OOV PNs can be relevant. The ones actually present in the audio are referred as *target OOV PNs*.

http://dx.doi.org/10.21437/Interspeech.2016-1219

word sequence due to word errors and secondly have no direct information about position of OOVs. Moreover, in contrast to the tasks in most state-of-the-art works in text classification, our task has a large number of output classes (OOV PNs) and the distribution of documents per OOV PN is very skewed [8].

We found that the work of Ling [21] is related to our proposal of using different weights for words. However, they use word position based weights to improve vectors learned by the *Continuous Bag-Of-Words* (CBOW) [22] model. Our NBOW2 model learns a context anchor vector to assign task specific word importance weights. The NBOW2 model is a variation of our D-CBOW2 model reported earlier [23] . We will highlight the differences between these models and show that the NBOW2 model gives a better retrieval performance. However, an explicit comparison of the two is not in scope of this paper.

2. Neural Bag-of-Words Model

The Neural Bag-of-Words (NBOW) model [24, 5] is a fully connected neural network which maps input text X, a sequence of words, to one of k output labels. The input to this model is in BOW form where the index of input words are set to 1 or number of occurrences of that word. The first hidden layer has d dimensional word vectors for each word in the chosen task vocabulary. Given the word vectors v_w for the words $w \in X$, the output of this layer is an average of the input word vectors: $z = 1/|X| \sum_{w \in X} v_w$. The average vector z is then fed to a fully connected layer to estimate probabilities for output labels as $\hat{y} = softmax(W_l\ z + b)$, where W_l is $k \times d$ matrix, b is a bias vector and $softmax(q) = exp(q)/\sum_{j=1}^{k} exp(q_j)$. The NBOW model is trained to minimise the categorical cross-entropy loss [25]. In our task to retrieve OOV PNs relevant to an audio document, the NBOW model is trained using a diachronic text corpus from the internet. During training IV words and PNs in a text document are given at the input and the co-occurring OOV PNs in the document are set at the output. During test, the LVCSR word hypothesis of the audio document is given at the input and the softmax probabilities at the output are used as scores to rank and retrieve the OOV PNs.

2.1. Proposed Neural Bag-of-Weighted-Words Model

While the NBOW model learns word vectors specialised for the task, we feel that it fails to *explicitly* use, as well as provide, the information that certain words are more important than the others for the given task. We thus propose the NBOW2 model, with the motivation to enable the NBOW model to learn and use task specific word importance weights. As compared to the NBOW model, our proposed NBOW2 model is a weighted sum composition of the input word sequence (X), calculated as:

$$z = \frac{1}{|X|} \sum_{w \in X} \alpha_w\ v_w \tag{1}$$

α_w are the scalar word importance weights for each word $w \in X$, obtained by introducing a global context anchor vector a as:

$$\alpha_w = f(v_w \cdot a) \tag{2}$$

where $(\cdot)$ represents a dot product and f scales the importance weights to $[0, 1]$. We believe that the vector a, which is itself learned and updated along with the word vectors, will act as a reference for the separation and the composition of the word vectors into a context vector; hence the term anchor. For function f, common activation functions sigmoid, softmax (as in [23]) and even hyperbolic tangent can be used. From our experiments we found that the sigmoid function $f(x) = (1+e^{-x})^{-1}$ is a better choice in terms of model convergence and accuracy. As compared to the NBOW model, the NBOW2 model will include an additional anchor vector, with the training and testing both using Equations 1 and 2.

3. Experiments and Results

3.1. Baseline Methods

The first baseline is the LDA based method, (*Method I*) from our previous work [1], briefly described below. We also examine an extension of this method to Skip-gram word vectors[2].

- **LDA Topic Space Representations:** LDA topic model is trained on the diachronic text corpus. A topic vocabulary, the number of topics (T) and Dirichlet priors (α, β) are first chosen. Topic model parameters θ and ϕ are then estimated using Gibbs sampling algorithm [26]. During test, the latent topic mixture $p(t|h)$ of the LVCSR word hypothesis (h) is inferred. Then the likelihood of an OOV PN ($\tilde{v}_i$) in the diachronic corpus is calculated, using $p(\tilde{v}_i|t)$ from ϕ, as: $p(\tilde{v}_i|h) = \sum_{t=1}^{T}\ p(\tilde{v}_i|t)\ p(t|h)$. To retrieve relevant OOV PNs, $p(\tilde{v}_i|h)$ is calculated for each OOV PN $\tilde{v}_i$ and used as a score to rank the OOV PNs.
- **Raw Skip-gram Vectors:** Skip-gram word vectors are trained for the words in the diachronic corpus. Given the word vectors and their linearity property, we obtain a representation for a test document by taking an average of vector of words in the document. This representation is referred to as *AverageVec*. The d dimensional vector representation of the LVCSR hypothesis (h) is compared with the vector ($\tilde{v}_i$) for each of the OOV PNs to calculate a score $s_i = CosSim(h, \tilde{v}_i)$, where $CosSim(\cdot, \cdot)$ is the cosine similarity. Score s_i is used to rank the OOV PNs $\tilde{v}_i$. The NBOW model can be seen as the AverageVec setup trained in a supervised manner but it is clear that this supervision is free of labelling costs.

3.2. Experiment Corpus

Table 1 presents the main characteristics of the French language diachronic broadcast news datasets which will be used as the train, validation and test sets in our study. Detailed description of the datasets is available in our previous works [27, 23]. The *L'Express* dataset will be used as a diachronic corpus to train context/topic models, in order to infer the OOV PNs relevant to *Euronews* videos. Non video text articles from *Euronews* are used as a validation/development set. The words and PNs which occur in the lexicon of our *Automatic News Transcription System* (ANTS) [28] are tagged as IV and remaining PNs are tagged as OOV. The total number of OOV PNs to be retrieved in the test set, obtained by counting unique OOV PNs per video, is 4694. Out of 4694, up to 2010 (42%) OOV PNs can be retrieved with the *L'Express* diachronic train set. The target OOV PN coverage can be further increased by augmenting additional text datasets, as discussed in another work [27].

3.3. Experiment Setup

The ANTS [28] LVCSR system is used to perform automatic segmentation and speech-to-text transcription of the test set.

[2]Word vectors from Skip-gram model give a better performance in our task than the word vectors from the CBOW model [2].

Table 1: Diachronic (French) news datasets used for experiments

	L'Express (train)	*Euronews* (valid)	*Euronews* (test)
Type	Text	Text	Video
Time Period		Jan - Jun 2014	
OOV PN unigrams	9.3K	3.4K	3.1K
OOV PN Documents	26.5K	1.9K	1.9K
Total OOV PN count	107K	6.9K	6.2K

The automatic transcriptions of the test audio news obtained by ANTS have an average *Word Error Rate* (WER) of 40% as compared to the reference transcriptions available from *Euronews*.

The train, dev and test texts are pre-processed as in our previous works [27, 23]. For comparison, the number of LDA topics and the dimensionality of the different neural context models are chosen to be equal and set to 400. A window size of 15 is chosen for training Skip-gram vectors. This selection of model hyper-parameters is based on performance on the validation set.

Our baseline methods (Section 3.1) are denoted as LDA and AverageVec. Proposed models (Section 2) are denoted as NBOW and NBOW2. Additionally we propose a combination of the NBOW and NBOW2 models, denoted as NBOW2+, in which the NBOW and NBOW2 averaged context vectors are concatenated together, both during training and test.

3.4. Training the NBOW group of models

For the NBOW group of models, we first train Skip-gram word vectors on the diachronic text corpus and use them to initialise the input layer vectors. This gives better performance than random initialisation. Then the NBOW, NBOW2 and NBOW2+ models are trained in 2 phases. In the first training phase, only the output parameters (W_l, b), and the anchor vector (a) for the NBOW2 and NBOW2+ models, are trained, keeping the input word vectors fixed. After this first training phase, all the model parameters are trained and updated in the second training phase. We found that a model trained in two phases in such a manner gives a better retrieval performance compared to the same model in which all the parameters are trained in one go.

To control the training of all the NBOW models an early stopping criterion [29] based on the validation set error is used. Early stopping is applied in both the first and the second training phases. Further we applied a dropout at the input layer (*word dropout*) [19, 5]. With experiments on validation set we chose a word dropout probability of 0.9 (from among 0, 0.25, 0.5, 0.75 and 0.9)[3]. All NBOW models were trained with gradient descent algorithm with ADADELTA [30].

Figure 1 shows a graph of validation set errors, of the NBOW, NBOW2 and NBOW2+ models, as the training progresses. It must be noted that this error is like a classification error stating whether an OOV PN in the validation set document is given the highest output probability or not. It will help to analyse the learning of the models. For instance it can be observed that the NBOW2+ model gives a faster convergence without compromise in error rate. More detailed discussion on this and the retrieval performance will follow in next sections.

[3] A word dropout probability p does not necessarily translate to leaving out p% of the input words.

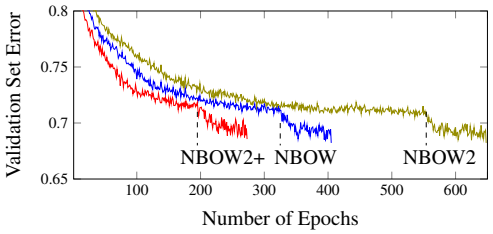

Figure 1: Validation set errors during training of NBOW, NBOW2 and NBOW2+ models. (- - - markers indicate end of first training phase and begin of second training phase)

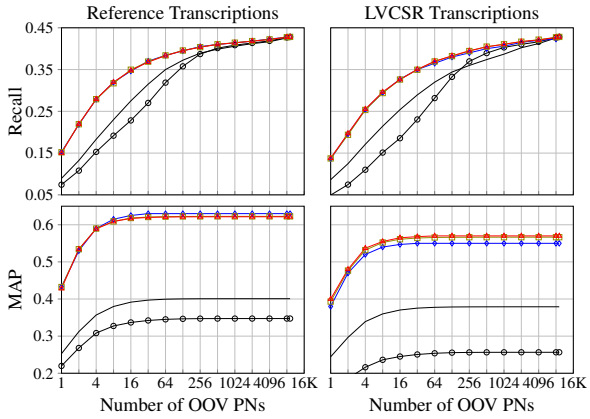

Figure 2: OOV PN retrieval performance for *Euronews* audio test set. —— is LDA, —○— AverageVec, —◇— NBOW, —□— NBOW2, —△— NBOW2+. (NBOW, NBOW2 and NBOW2+ give similar performance and their graphs are overlapping.)

4. Discussion

4.1. OOV PN Retrieval Performance

Figure 2 shows the *Recall* and *Mean Average Precision* (MAP) [31] performance of retrieval of OOV PNs for all the methods. The graphs shown are for the reference transcriptions (left) and the LVCSR transcriptions (right) of the *Euronews* test set audio. The X-axis represents the number of OOV PNs selected from the diachronic corpus i.e. the 'N' in the top-N retrieved results. The Y-axis represents recall (top) and MAP (bottom) of the target OOV PNs. We can observe that our previous method based on LDA [1] performs better than AverageVec, which follows a similar methodology but with Skip-gram word vectors. The recall and MAP retrieval performance for NBOW, NBOW2 and NBOW2+ models is very similar and their graphs are overlapping. But as mentioned earlier, combining the NBOW and NBOW2 representations in the NBOW2+ model gives a faster convergence in training. Overall the three models clearly outperform the baseline methods in terms of recall and MAP, both for reference and LVCSR transcriptions. (Similarly they outperformed the document similarity method [10] which improves performance for less frequent PNs, at the cost of searching the diachronic corpus. The best of these methods achieved a maximum MAP of 0.519 for reference and 0.462 for LVCSR.)

4.2. Scrutinising the training of NBOW model

We try to examine how much the choice of training hyper-parameters affects the NBOW model training. Table 2 depicts the effect of applying word dropout. The comparison is in terms

of maximum MAP achieved for the LVCSR hypothesis. For space constraints we show only for 100 and 400 word vector dimension NBOW models. Firstly, it is clear that word dropout improves the MAP performance significantly and secondly we cannot attribute the increase in number of training epochs to increase in word dropout. Next we increased the ADADELTA decay constant (ρ) from 0.99, as in our experiments, to 0.95. The training takes fewer epochs but the MAP performance also reduces. For instance with word dropout probability p = 0.9, the 100 and 400 dimensional models take 185 and 351 epochs respectively and achieve a maximum MAP of 0.45 and 0.5 respectively. These MAP values are significantly lower that those with ρ = 0.99. Further we also trained the (400) NBOW and NBOW2+ models with fixed number of epochs (100) in the first and (50) second training phases. The NBOW2+ model still achieved a maximum MAP (0.553) comparable to NBOW (0.556), the difference being statistically insignificant. And these MAP values are significantly lower than that (0.568) obtained from training with early stopping. From these experiments, we can conclude that to obtain better retrieval performance with the NBOW model we need a longer training, which can be reduced by the NBOW2+ model as depicted in Figure 1.

Table 2: The maximum MAP obtained by NBOW models of 100 and 400 dimension word vectors in given number of epochs (2 training phases). * denotes difference in MAP is statistically insignificant [32] compared to MAP with p = 0.0.

		word dropout probability (p)				
		0.0	0.25	0.5	0.75	0.9
100	MAP	0.500	0.497*	0.514	0.536	0.540
	epoch	459	403	415	458	568
400	MAP	0.525	0.519*	0.533	0.561	**0.568**
	epoch	481	482	398	417	410

4.3. Word Importance weights of NBOW2 model

We present Figure 3 to discuss (a) the scalar word importance weights α_w (b) the choice of the function f, for the NBOW2 model (see Equation 2). Considering a sample test document, the left graph of Figure 3 shows the weights assigned by the NBOW2 model with f as sigmoid activation and the right graph shows the weights assigned by f as softmax activation. Firstly, it is clear from these graphs that the NBOW2 models have learned that different words have different degree of importance. For example, for this document with the missing OOV PN *kehm* (https://de.wikipedia.org/wiki/Sabine_Kehm), as per the left graph the top four important words are *michael*, *formule*, *critique* and *hospitaliser* and the four least important words are *rester*, *tenir*, *monde* and *présent*. Secondly, the NBOW2 model with f as softmax tends to assign higher weights to fewer words and weights close to zero to the remaining words. This could be one reason for its relatively bad retrieval performance [23].

4.4. Recovery of OOV PNs with Keyword Search (KWS)

We perform an evaluation of the retrieved list of OOV PNs using an automatic Keyword Search (KWS) method [33] which enables searching of OOV words in an LVCSR lattice. First, a list of relevant OOV PNs is retrieved with the models presented in this paper. In the second step, KWS is performed on the entire LVCSR lattice of the audio file, for each OOV PN in the list of relevant OOV PNs. Evaluation is done in terms of F_1-*score*.

Figure 4 shows the F1-scores obtained with the top-N lists

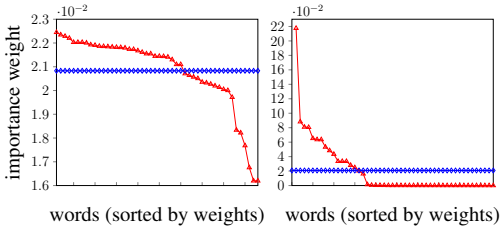

Figure 3: Word importance weights assigned by the NBOW2 model (—△—) in a sample document with 48 words. Two variations of the NBOW2 model are shown: (left) f as sigmoid and (right) f as softmax. —◇— denotes the all equal weights (1/48 = 0.0208) in the simple average by the NBOW model.

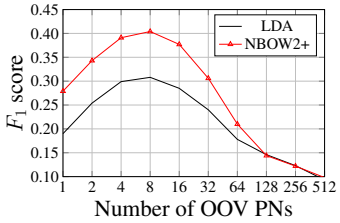

Figure 4: F_1 scores for recovery of target OOV PNs.

from NBOW2+ and LDA. We show only these two since the NBOW models have similar performance and LDA is shown to outperform AverageVec. The KWS algorithm has a matching score threshold which controls the operating characteristics, and hence recall/precision and F1-score, of the search. We show the best F1-scores corresponding to top-N OOV PN lists of different sizes (N). Beyond top-512 there is no significant difference in the F1-scores. Overall we can observe that the better the Recall and MAP of OOV PN retrieval, the better is the F-1 score[4].

5. Conclusion and Future Work

We examined the NBOW model for our task of retrieval of OOV PNs relevant to an audio document. We proposed a novel extension to the NBOW model, which enables it to learn the *words important for the given task*. With experiments on French broadcast news videos we showed that (a) the NBOW and NBOW2 models give improvements in retrieval performance as compared to the previous method (b) combining the NBOW and NBOW2 into a new model leads to a faster convergence in training. The improvements in retrieval were validated by performing recovery of the target OOV PNs with an automatic keyword search. These results motivate us to extend NBOW2 to Deep Averaging Networks (DAN) [5], which cascade additional fully connected layers to the NBOW model. Further, instead of using a single anchor vector to obtain word importance weights, we would like to explore models with multiple anchors.

6. Acknowledgements

This work is funded by the ContNomina project supported by the French National Research Agency (ANR) under contract ANR-12-BS02-0009. KWS experiments presented in this paper were carried out using the Grid'5000 testbed, supported by a scientific interest group hosted by Inria, CNRS, RENATER and other Universities and organisations (https://www.grid5000.fr).

[4] document similarity method mentioned at end of Section 4.1 achieved a maximum F-1 score of 0.342.

7. References

[1] I. Sheikh, I. Illina, D. Fohr, and G. Linarès, "OOV proper name retrieval using topic and lexical context models," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5291–5295.

[2] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems 26*, 2013, pp. 3111–3119.

[3] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543.

[4] M. Baroni, G. Dinu, and G. Kruszewski, "Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2014, pp. 238–247.

[5] M. Iyyer, V. Manjunatha, J. Boyd-Graber, and H. Daumé III, "Deep unordered composition rivals syntactic methods for text classification," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2015, pp. 1681–1691.

[6] B. Bigot, G. Senay, G. Linarès, C. Fredouille, and R. Dufour, "Person name recognition in ASR outputs using continuous context models," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 8470–8474.

[7] G. Senay, B. Bigot, R. Dufour, G. Linarès, and C. Fredouille, "Person name spotting by combining acoustic matching and LDA topic models," in *14th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2013, pp. 1584–1588.

[8] I. Sheikh, I. Illina, and D. Fohr, "Study of entity-topic models for OOV proper name retrieval," in *16th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2015, pp. 3506–3510.

[9] D. Fohr and I. Illina, "Continuous word representation using neural networks for proper name retrieval from diachronic documents," in *16th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2015, pp. 1344–1348.

[10] I. Sheikh, I. Illina, D. Fohr, and G. Linarès, "Document level semantic context for retrieving OOV proper names," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016.

[11] J. Nam, J. Kim, E. Loza Mencía, I. Gurevych, and J. Fürnkranz, "Large-scale multi-label text classification - revisiting neural networks," in *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD-14), Part 2*, vol. 8725, Sep. 2014, pp. 437–452.

[12] Y. Kim, "Convolutional neural networks for sentence classification," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1746–1751.

[13] R. Johnson and T. Zhang, "Effective use of word order for text categorization with convolutional neural networks," in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2015, pp. 103–112.

[14] P. Wang, J. Xu, B. Xu, C. Liu, H. Zhang, F. Wang, and H. Hao, "Semantic clustering and convolutional neural network for short text categorization," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2015, pp. 352–357.

[15] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts, "Recursive deep models for semantic compositionality over a sentiment treebank," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2013, pp. 1631–1642.

[16] K. M. Hermann and P. Blunsom, "The Role of Syntax in Vector Space Models of Compositional Semantics," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2013, pp. 894–904.

[17] L. Dong, F. Wei, C. Tan, D. Tang, M. Zhou, and K. Xu, "Adaptive recursive neural network for target-dependent twitter sentiment classification," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL, Baltimore, MD, USA, Volume 2: Short Papers*, 2014, pp. 49–54.

[18] K. S. Tai, R. Socher, and C. D. Manning, "Improved semantic representations from tree-structured long short-term memory networks," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2015, pp. 1556–1566.

[19] A. M. Dai and Q. V. Le, "Semi-supervised sequence learning," in *Advances in Neural Information Processing Systems 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds., 2015, pp. 3061–3069.

[20] Y. Zhang and B. Wallace, "A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification," *CoRR*, vol. abs/1510.03820, 2015.

[21] W. Ling, Y. Tsvetkov, S. Amir, R. Fermandez, C. Dyer, A. W. Black, I. Trancoso, and C.-C. Lin, "Not all contexts are created equal: Better word representations with variable attention," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 1367–1372.

[22] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *CoRR*, vol. abs/1301.3781, 2013.

[23] I. A. Sheikh, I. Illina, D. Fohr, and G. Linarès, "Learning to retrieve out-of-vocabulary words in speech recognition," *CoRR*, vol. abs/1511.05389, 2015.

[24] N. Kalchbrenner, E. Grefenstette, and P. Blunsom, "A convolutional neural network for modelling sentences," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2014, pp. 655–665.

[25] Y. Goldberg, "A primer on neural network models for natural language processing," *CoRR*, vol. abs/1510.00726, 2015.

[26] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proceedings of the National Academy of Sciences*, vol. 101, pp. 5228–5235, 2004.

[27] I. Sheikh, I. Illina, and D. Fohr, "How diachronic text corpora affect context based retrieval of oov proper names for audio news," in *Language Resources and Evaluation Conference (LREC)*, 2016.

[28] I. Illina, D. Fohr, O. Mella, and C. Cerisara, "The Automatic News Transcription System: ANTS some Real Time experiments," in *8th International Conference on Spoken Language Processing (INTERSPEECH'2004 - ICSLP)*, 2004, pp. 377–380.

[29] Y. Bengio, "Practical recommendations for gradient-based training of deep architectures," *CoRR*, vol. abs/1206.5533, 2012.

[30] M. D. Zeiler, "ADADELTA: an adaptive learning rate method," *CoRR*, vol. abs/1212.5701, 2012.

[31] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge, UK: Cambridge University Press, 2008.

[32] M. D. Smucker, J. Allan, and B. Carterette, "A comparison of statistical significance tests for information retrieval evaluation," in *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, 2007, pp. 623–632.

[33] G. Chen, S. Khudanpur, D. Povey, J. Trmal, D. Yarowsky, and O. Yilmaz, "Quantifying the value of pronunciation lexicons for keyword search in low resource languages," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 8560–8564.