Improved Neural Bag-of-Words Model to Retrieve Out-of-Vocabulary Words in Speech Recognition
Imran Sheikh, Irina Illina, Dominique Fohr, Georges Linares

To cite this version:

HAL Id: hal-01384488
https://hal.archives-ouvertes.fr/hal-01384488
Submitted on 20 Oct 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Improved Neural Bag-of-Words Model to Retrieve Out-of-Vocabulary Words in Speech Recognition

Imran Sheikh∗+, Irina Illina∗, Dominique Fohr+, Georges Linarèsla+

∗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. 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 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

---

1For a given audio documents several OOV PNs can be relevant. The ones actually present in the audio are referred as target OOV PNs.
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]². 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:

\[
z = \frac{1}{|X|} \sum_{w \in X} \alpha_w \cdot v_w.
\]

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

\[
\alpha_w = f(v_w \cdot a)
\]

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. The NBOW2 model will thus 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.

- **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 \((\alpha, \beta)\) are first chosen. Topic model parameters \( \alpha \) and \( \beta \) 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:

\[
p(\tilde{v}_i|h) = \sum_{t=1}^{T} p(\tilde{v}_i|t) \cdot p(t|h).
\]

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 \( K \) 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 \( s_i = \text{CosSim}(h, \tilde{v}_i) \), where \( \text{CosSim}(\cdot, \cdot) \) 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 using the Skip-gram word vectors; hence the term anchor. The target OOV PN coverage can be further increased by augmenting additional text datasets, as discussed in another work [27].

²Word vectors from Skip-gram model give a better performance in our task than the word vectors from the CBOW model [2].
The ANTS [28] LVCSR system is used to perform automatic segmentation and speech-to-text transcription of the test set. 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.

### 3.3. Experiment Setup

The ANTS [28] LVCSR system is used to perform automatic segmentation and speech-to-text transcription of the test set. 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 ($W$, $b$, 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 much 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)\(^5\). 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. (--- markers indicate end of first training phase and begin of second training phase)

### 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 audio set. 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

<table>
<thead>
<tr>
<th>Type</th>
<th>L'Express (train)</th>
<th>Euronews (valid)</th>
<th>Euronews (test)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time Period</td>
<td>Text</td>
<td>Text</td>
<td>Video</td>
</tr>
<tr>
<td>OOV PN unigrams</td>
<td>9.3K</td>
<td>3.4K</td>
<td>3.1K</td>
</tr>
<tr>
<td>OOV PN Documents</td>
<td>26.5K</td>
<td>1.9K</td>
<td>1.9K</td>
</tr>
<tr>
<td>Total OOV PN count</td>
<td>107K</td>
<td>6.9K</td>
<td>6.2K</td>
</tr>
</tbody>
</table>

\(^5\)A word dropout probability $p$ does not necessarily translate to leaving out $p\%$ of the input words.

\[\text{Total OOV PN count} \]
\[\text{OOV PN Documents} \]
\[\text{OOV PN unigrams} \]

\[\text{Time Period Jan - Jun 2014} \]

\[\text{Validation Set Error} \]
\[\text{Reference Transcriptions} \]
\[\text{LVCSR Transcriptions} \]

\[\text{Number of OOV PNs} \]

\[\text{Number of OOV PNs} \]

\[\text{Number of OOV PNs} \]

\[\text{Number of OOV PNs} \]
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 ($\rho$) 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 than those with $p = 0.99$. Further we also trained the (400) NBOW and NBOW+ models with fixed number of epochs (100) in the first and (50) second training phases. The NBOW+ model still achieved a maximum MAP (0.553) comparable to NBOW (0.556), the difference being 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 NBOW+ model as depicted in Figure 1.

Table 2: The maximum MAP obtained by NBOW models of 100 and 400 dimensional word vectors in given number of epochs (2 training phases). * denotes difference in MAP is statistically insignificant [32] compared to MAP with $p = 0.9$.

<table>
<thead>
<tr>
<th>word dropout probability ($p$)</th>
<th>100 epoch</th>
<th>400 epoch</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>0.500</td>
<td>0.525</td>
</tr>
<tr>
<td>0.25</td>
<td>0.497*</td>
<td>0.519*</td>
</tr>
<tr>
<td>0.5</td>
<td>0.514</td>
<td>0.533</td>
</tr>
<tr>
<td>0.75</td>
<td>0.536</td>
<td>0.561</td>
</tr>
<tr>
<td>0.9</td>
<td>0.540</td>
<td>0.568</td>
</tr>
</tbody>
</table>

4.3. Word Importance weights of NBOW2 model

We present Figure 3 to discuss (a) the scalar word importance weights $a_w$ (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 kelm (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 prés. 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 F1-score.

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

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).

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: F1 scores for recovery of target OOV PNs.
7. References


