Modelling Semantic Context of OOV Words in Large Vocabulary Continuous Speech Recognition
Imran Ahamad Sheikh, Dominique Fohr, Irina Illina, Georges Linares

To cite this version:
hal-01461617

HAL Id: hal-01461617
https://inria.hal.science/hal-01461617
Submitted on 8 Feb 2017

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.
Modelling Semantic Context of OOV Words in Large Vocabulary Continuous Speech Recognition

Imran Sheikh, Student Member, IEEE, Dominique Fohr, Irina Illina, and Georges Linarès

Abstract—The diachronic nature of broadcast news data leads to the problem of Out-Of-Vocabulary (OOV) words in Large Vocabulary Continuous Speech Recognition (LVCSR) systems. Analysis of OOV words reveals that a majority of them are Proper Names (PNs). However PNs are important for automatic indexing of audio-video content and for obtaining reliable automatic transcriptions. In this paper, we focus on the problem of OOV PNs in diachronic audio documents. To enable recovery of the PNs missed by the LVCSR system, relevant OOV PNs are retrieved by exploiting the semantic context of the LVCSR transcriptions. For retrieval of OOV PNs, we explore topic and semantic context derived from Latent Dirichlet Allocation (LDA) topic models, continuous word vector representations and the Neural Bag-of-Words (NBOW) model which is capable of learning task specific word and context representations. We propose a Neural Bag-of-Weighted Words (NBOW2) model which learns to assign higher weights to words that are important for retrieval of an OOV PN. With experiments on French broadcast news videos we show that the NBOW and NBOW2 models outperform the methods based on raw embeddings from LDA and Skip-gram models. Combining the NBOW and NBOW2 models gives a faster convergence during training. Second pass speech recognition experiments, in which the LVCSR vocabulary models gives a faster convergence during training. Second pass and Skip-gram models. Combining the NBOW and NBOW2 models which is also the motive for open vocabulary systems [7], [8]. However the OOV detection methods are trained with features obtained after speech recognition, for example posterior scores and word confusion, thus requiring speech data and their automatic transcriptions during training. Moreover, hybrid LM and open vocabulary systems may require careful selection of sub-word units and LM estimation which can sometimes lead to increased error rates [8]. Vocabulary selection based approaches propose a relevant vocabulary for speech recognition based on additional text data. Vocabulary selection has been proposed in order to minimise OOV rate for a domain specific corpus [9]–[12] and for daily update systems [13]. Document specific vocabulary selection methods [14]–[18] are more dynamic as they propose context specific vocabulary.

We adopt the document specific vocabulary selection approach for handling OOV PNs. Following a two pass approach, in the first pass speech recognition is performed on the audio document with the base vocabulary and LM. Then a list of relevant OOV PNs is inferred based on the semantic/topic context derived from the 1-best word hypothesis of the first pass recognition. Then a second pass speech recognition is performed with an updated vocabulary and LM which include the relevant OOV PNs. Unlike OOV detection and hybrid LM methods, which can only provide OOV positions and their sub-word level transcriptions, our approach is able to directly recognise the OOV PNs. Furthermore, our approach can be readily used to enhance transcriptions from open vocabulary and hybrid LM systems, although this is not in the scope of this paper.

This paper summarises and extends our continued research on learning semantic context of OOV words in LVCSR and provides a more comprehensive and conclusive analysis. In our previous works [1], [19] we have shown that methods based on Latent Dirichlet Allocation (LDA) [20] topic space can perform well for retrieval of relevant OOV PNs. Arguing
that these context representations which are learned in an unsupervised manner are not the most optimal for the task of retrieving OOV PNs, in [21] we presented neural network context models which were trained with the objective of maximising the OOV PN retrieval performance. Following our work in [21], in this paper we perform a detailed exploration for training the Neural Bag-of-Words (NBOW) model [22] and our Neural Bag-of-Weighted-Words (NBOW2) model [21], for retrieving the target OOV PNs. (For a given audio document several OOV PNs can be relevant. Those actually present in the audio are referred as target OOV PNs.) Complementary to our previous works [1], [19], [21], the main contributions of this paper are (a) a detailed discussion on training the NBOW group of models, focusing on techniques for speeding up their convergence and boosting their performance, (b) a complete analysis on the choice of hyper-parameters for all the models, (c) a comparison to the document specific context approach of [19] in terms of retrieval of less frequent PNs, and (d) comparison under different WER conditions (e) evaluation of the context models in terms of speech recognition of new PNs.

The rest of the paper is organised as follows. Section II presents a background discussing our approach for handling OOV PNs, previous methods based on LDA & Skip-gram model and related works. Section III describes the neural network based discriminative context representations that are explored in detail in this paper. The experiment protocol and the model training procedure are described in Section IV. The OOV PN retrieval results are discussed in Section V, followed by the speech recognition evaluation in Section VI and conclusion in Section VII.

II. BACKGROUND

A. Adopted Approach

Earlier proposed document specific vocabulary selection methods [14]–[18] query web search engines to retrieve relevant documents and then choose the new vocabulary words using term frequency, document frequency and co-occurrence based features. In contrast, we propose to model the context of OOV PNs in order to retrieve OOV PNs relevant to a test audio document. Fig. 1 is a block diagram illustration of our approach. Diachronic text news is collected from the Internet to build a diachronic text corpus which contains documents with new, i.e. OOV, PNs. The diachronic text corpus is used to learn a context model which captures relationships between the LVCSR In-Vocabulary (IV) words and the OOV PNs. This is the training phase. During the test phase, the audio document is processed by the LVCSR (with the base vocabulary and LM) to obtain the (first pass) 1-best hypothesis. Given this text output, the context model and the complete list of OOV PNs, the context of the spoken content is inferred and a context based ranking is performed to choose the relevant OOV PNs. This list of relevant OOV PNs is then used to update the vocabulary and LM of the LVCSR to perform a second pass speech recognition for recognising the missed PNs.

B. Semantic and Topic Context Models

Semantic context models have a long history in natural language processing [23]. Latent Semantic Analysis (LSA) [24] and Latent Dirichlet Allocation (LDA) [20] have been the most prominent methods for extracting underlying semantic and topic structures from documents. LSA derives a semantic vector space by decomposition of a word co-occurrence matrix, while PLSA and LDA are improved versions of LSA deriving topics using hierarchical Bayesian analysis. LDA is a complete generative model and previous work has shown that LDA outperforms PLSA and LSA for document classification [20] and word prediction [25] tasks. We had similar observations in our task and therefore we chose LDA to model topic context of OOV PNs. Given a corpus of $D$ documents with vocabulary of size $V$ and a number of topics $T$ to be modelled, the joint distribution corresponding to the generative process of the LDA model is:

$$p(w, z, \theta, \phi | \alpha, \beta) = p(\phi | \beta) p(\theta | \alpha) p(z | \theta) p(w | \phi_z)$$

where $z$ is the hidden topic assignment to word $w$ in a document $d$, $\theta = [\theta_{d|}\]_{D \times T}$ is the multinomial topic distribution for each document $d$ and $\phi = [\phi_{w|}\]_{V \times T}$ is the multinomial topic distribution for each word, both across $T$ topics. $\alpha$ and $\beta$ are Dirichlet priors for $\theta$ and $\phi$ respectively. The topic model parameters $\theta$, $\phi$ and the word topic assignments $z$ can be estimated using a Gibbs sampling algorithm [26].

More recently alternative methods to learn word and context vector representations, based on predicting the context in which words appear, have become popular [27], [28]. These representations have been shown to perform effectively for a range of text processing tasks [29]. The models of Mikolov et al. [27], [30] have become popular due to their ability to handle large amounts of unstructured data with reduced computational costs. They proposed two models of which we use the Skip-gram model, because in our task the word vectors from this model perform slightly better than those from the CBOW model. The Skip-gram model is trained with an objective function to maximise the likelihood of predicting the surrounding words given the center word. Denoting $C(w)$ as the context of a word $w$ in the corpus, the objective function
is given as:

\[
\arg\max_\Theta \prod_{w \in \text{corpus}} \prod_{c \in C(w)} p(c|w; \Theta)
\]

where \( \Theta \) is the model parameter composed of word vectors corresponding to words when they are part of input and when they are part of the output context.

\[\text{(2)}\]

\( \text{C. Retrieving OOV Proper Names Using Unsupervised Context Representations} \)

To retrieve OOV PNs relevant to an audio document we aim to learn a semantic/topic context space which captures relationship between the In-Vocabulary (IV) words, PNs and the OOV PNs. Then the LVCSR word hypothesis by \( h \) and OOV PNs in diachronic text corpus (and topic model vocabulary) by \( \hat{\theta} \). The latent topic mixture of \( h \), i.e. \( p(t|h) \), is inferred by sampling the topic assignments for words in \( h \) using the word-topic distribution \( \phi \) learned during training. Given \( p(\hat{v}_i|t) \) from \( \phi \), the likelihood of an OOV PN \( (\hat{v}_i) \) in the diachronic text corpus is calculated as:

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

To retrieve OOV PNs we calculate \( p(\hat{v}_i|h) \) for each \( \hat{v}_i \) and then use it as a score to rank OOV PNs relevant to \( h \).

\[\text{(3)}\]

\( \text{1) Retrieval based on LDA:} \) In our OOV PN retrieval task, LDA topics are trained on the diachronic text corpus. Let us denote the LVCSR word hypothesis by \( h \) and OOV PNs in diachronic text corpus (and topic model vocabulary) by \( \hat{\theta} \). The latent topic mixture of \( h \), i.e. \( p(t|h) \), is inferred by sampling the topic assignments for words in \( h \) using the word-topic distribution \( \phi \) learned during training. Given \( p(\hat{v}_i|t) \) from \( \phi \), the likelihood of an OOV PN \( (\hat{v}_i) \) in the diachronic text corpus is calculated as:

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

\[\text{(3)}\]

\( \text{2) Retrieval with Predictive Word Vectors (AverageVec):} \) During training, Skip-gram word vectors are trained for the words in the diachronic text corpus. Given the word vectors and their linearity property, we obtain a representation for a test document by taking the average of all word vectors in the document. This representation is referred to as AverageVec. The \( K \) dimensional vector representation \( h \) of the LVCSR hypothesis is compared with the vector \( \hat{v}_i \) for each of the OOV PNs to calculate a score as:

\[
s_i = \frac{\sum_{k=1}^{K} h_k \hat{v}_{ik}}{\sqrt{\sum_{k=1}^{K} (h_k)^2} \sqrt{\sum_{k=1}^{K} (\hat{v}_{ik})^2}}
\]

\[\text{(4)}\]

The score \( s_i \) is used to rank and retrieve the OOV PNs \( \hat{v}_i \).

\[\text{1) However, to improve computational efficiency they use another function and training mechanism, albeit with the same objective [31].} \]

\[\text{D. Related Work} \]

There have been efforts to incorporate semantic contextual information into the LM for speech recognition [32]. Complementary to these advances in LM we explore the use of semantic/topic context to address the OOV problem in LVCSR. The task of retrieving OOV and PNs relevant to an audio document has been presented in previous works. PNs have been modelled with the LDA topic model [33], and a related approach [34] based on vector space representation similar to LSA has been tried. However, these approaches estimate one LDA/LSA context model for each PN which restricts them to only frequent PNs. This problem was partly addressed in our previous work [1], [35] by training a global topic model and including re-ranking techniques to improve retrieval of less frequent PNs. Later we showed that document similarity based methods perform even better, especially for retrieval of less frequent PNs [19]. Further, word embedding based methods to retrieve relevant PNs have been tried for audio documents with multiple news events appearing one after another [36]. Compared to these works, we explore neural network models trained to retrieve relevant OOV PNs for audio documents with a single news event.

Our methodology in this paper is related to the recent approaches for text classification with neural networks. In this context, fully connected feed forward networks [22], [37], Convolutional Neural Networks (CNN) [38]–[40] and also Recurrent Neural Networks (RNN) [41]–[45] have been applied. On the 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 [45], [46]. 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 word sequences due to word errors and secondly have no direct information about the position of OOVs. Moreover, in contrast to the tasks considered in most state-of-the-art text classification works, our task has a large number of output classes (OOV PNs) and the distribution of documents per OOV PN is very skewed [35].

The work of Ling [47] is related to our proposal of using different word weights in a neural network model. However, they use word position based weights to improve vectors learned by the Continuous Bag-Of-Words (CBOW) [30] model. Our NBOW2 model learns a context anchor vector to assign task specific word importance weights. Also related are works on learning to pay attention in a sequence of input, as applied in text [48] as well as speech [49], image [50] and protein sequence analysis [51].

\[\text{III. Discriminative Context Representations} \]

Models with discriminative context representations are trained to maximise the retrieval of relevant OOV PNs by using the target OOV PNs as the labels to be predicted by the model. These models can also be seen as the AverageVec setup of Section II-C2 with the word vectors trained to maximise the retrieval of relevant OOV PNs.
A. Neural Bag-of-Words Model

The Neural Bag-of-Words (NBOW) model [52] [22] is a fully connected neural network model which takes an input text \(X\) containing a set of words \(w\) and generates probability estimates for the \(L\) output labels. The NBOW model has two hidden layers, one corresponding to the input and another one to the output. The first hidden layer has a \([V \times K]\) matrix containing \(K\) dimensional vectors corresponding to each of the words in the chosen input vocabulary of size \(V\). With a sparse BOW input vector, with words present in the input set to 1 and others set to 0, the vector-matrix product at the first hidden layer translates into the sum of the vectors corresponding to input words. In practice the average of the word vectors is used instead:

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

(5)

The average vector \(z\) is fed into the output layer to estimate probabilities for the output labels as \(\hat{y} = \text{softmax}(zW^O + b)\), where \(W^O\) is a \([K \times L]\) matrix and \(b\) is a bias vector, and \(\text{softmax}(l) = \exp(l)/ \sum_{j=1}^{L} \exp(l_j)\).  

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

(7)

where \(\cdot\) represents the dot product and \(f\) scales the importance weights to \([0, 1]\). The word importance weight \(\alpha_w\) is a function of the distance of that word \(w\) from \(a\) in the context space, ensuring that the calculation of \(\alpha_w\) takes into account the contextual word similarities and it is not biased by the frequency of occurrence of words in the training corpus. Regarding the function \(f\), common activation functions can be used such as sigmoid, softmax (as in [54]) and even hyperbolic tangent. In our experiments we found that the sigmoid function \(f(x) = (1+e^{-x})^{-1}\) is a better choice in terms of convergence speed and accuracy. We further discuss the choice of \(f\) in Section V.

B. Proposed Neural Bag-of-Weighted-Words (NBOW2) Model

The NBOW model learns word vectors specialised for the retrieval task, however we feel that it fails to explicitly use the information that certain words are more important than others for retrieval of an OOV PN. We thus propose the NBOW2 model, with the motivation of enabling the NBOW model to learn and use PN specific word importance weights. To learn the word importance weights, a weighted sum composition of the input word sequence \(X\) is introduced as follows.

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

(6)

where \(\alpha_w\) are scalar word importance weights for each word \(w \in X\). The weights \(\alpha_w\) are obtained by integrating a new \(K\) dimensional vector \(a\) into the model, and using the following operation:

For the retrieval of relevant OOV PNs, the words from the LVCSR hypothesis are given at the input and the softmax probabilities at the output are used as scores to rank the OOV PNs. During training of the model, the IV words in a document from the diachronic text corpus are given at the input and each co-occurring OOV PN in the document is set at the output in turns. The NBOW model is trained to minimise the categorical cross-entropy loss [53]. The categorical cross-entropy error function is commonly used for single label classification and some documents can have more than one OOV PN. In this case the training document is replicated for each OOV PN.
Fig. 3 shows a representation of the NBOW2 model as used in our task to retrieve relevant OOV PNs. The inputs, input embedding matrix, outputs, output matrix and also the training loss function are similar to that of NBOW. However, the procedure to obtain the document context vector has changed. After the lookup of the word vectors for input text, a dot product is performed between each input word vector and the vector \( a \), and the scalar values from the dot product are passed through the function \( f \). The scalar word importance weights are multiplied with the input word vectors and a weighted sum composition representing the document context vector is obtained.

C. Combination of the NBOW and NBOW2 Models

We further propose the NBOW2+ model in which the NBOW and NBOW2 document context vectors are concatenated together. NBOW2+ has two input matrices \( (W^I, W^2) \) and hence maintains two \( K \) dimensional word vectors \( v^I_w \) and \( v^2_w \) for each input word \( w \). It has one \( K \) dimensional anchor vector \( a \) similar to NBOW2 and one matrix \( W^O \) and one bias \( b \) in the output layer. The document context vector \( z \) is obtained as the concatenation of two context vectors \( z_1 \) and \( z_2 \) as follows:

\[
\begin{align*}
  z_1 &= \frac{1}{|X|} \sum_{w \in X} v^I_w, \\
  z_2 &= \frac{1}{|X|} \sum_{w \in X} a_w v^2_w \\
  z &= [z_1 \, z_2]
\end{align*}
\]

As the document context vector is a concatenation of the two \( K \) dimensional context vectors, the output layer parameters \((W^O, b)\) have a dimension of \( 2K \). The training procedure and loss function are the same as NBOW and NBOW2.

IV. EXPERIMENTAL PROTOCOL

A. Corpus

Table I presents three realistic broadcast news datasets which will be used as the training, validation and test sets in our study. These datasets also highlight the motivation for handling OOV PNs. The datasets are collected from two different sources: (a) the French newspaper L’Express\(^2\), and (b) the French website\(^3\) of the Euronews television channel. The L’Express dataset contains text news whereas the Euronews dataset contains text news as well as news videos and their text transcriptions. In our study the L’Express dataset will be used as the diachronic text corpus to train context/topic models in order to infer the OOV PNs relevant to Euronews videos (the test set). Euronews text documents, denoted as ‘validation’ in Table I, are used as a validation set in our experiments.

1) Pre-processing the diachronic text corpus to train context models: The TreeTagger part-of-speech tagging tool, which is reported to perform with an accuracy of 95.7% on French data [55], is used to automatically tag PNs in the text. The PNs and non PN words which occur in the lexicon of our LVCSR system are tagged as IV and the remaining PNs are tagged as OOV. For training the context models, words in the diachronic text corpus are lemmatised and filtered by removing PNs and non PN words occurring less than 3 times. A stop-list of common words and non-content words is applied and a part-of-speech based filter is employed to choose words tagged as PN, noun, adjective, verb and acronym. The context and topic models are trained with this filtered vocabulary.

2) OOV PN Statistics: As shown in the Table I, 72% (3.1K out of 4.3K) of OOV words in the Euronews video dataset are PNs and about 64% (1.9K out of 3K) of the videos contain OOV PNs. The total number of OOV PNs to be retrieved for the Euronews videos, 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 text corpus which has 9.3K new (OOV) PNs. This is because our diachronic text corpus and the test corpus have only a partial overlap in terms of coverage of topics and OOV PNs.

Selection of diachronic text corpus is a crucial step as it decides (a) the coverage of the target OOV PNs, and (b) the number of possible OOV PNs to rank. Too much variety in the diachronic text corpus may lead to a (unnecessarily) long list of possible OOV PNs. Our analysis in [56] discusses more in this direction. Prior knowledge about the task and the domain can help in selection of a better diachronic text corpus.

B. LVCSR systems

In our experiments we use two LVCSR systems with different WERs. One is a GMM-HMM based LVCSR system, which has a higher WER compared to the second LVCSR system based on DNN-HMM. These two LVCSR systems are used to demonstrate the effect of errors in the LVCSR hypothesis on OOV retrieval performance. Both these systems are trained to perform automatic segmentation and speech-to-text transcription of French broadcast news audio. A brief description of the two systems is as follows.

1) Automatic News Transcription System (ANTS): The ANTS [57] LVCSR system is based on context dependent

\(^1\)K denotes Thousand and M denotes Million
\(^2\) L’Express unigrams occurring less than 2 times are ignored
\(^3\) L’Express unigrams occurring in less than 3 documents are ignored; documents with more than 20 and less than 500 terms

Note: OOV and OOV PN statistics are computed after term-document filtering

<table>
<thead>
<tr>
<th>Type of Documents</th>
<th>L’Express (train)</th>
<th>Euronews (valid)</th>
<th>Euronews (test)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time Period</td>
<td>Jan - Jun 2014</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of Documents(^1)</td>
<td>45K</td>
<td>3.1K</td>
<td>3K</td>
</tr>
<tr>
<td>Vocabulary Size(^2)</td>
<td>150K</td>
<td>42K</td>
<td>45K</td>
</tr>
<tr>
<td>Corpus Size (word count)</td>
<td>24M</td>
<td>580K</td>
<td>700K</td>
</tr>
<tr>
<td>PN unigrams(^3)</td>
<td>57K</td>
<td>12K</td>
<td>11K</td>
</tr>
<tr>
<td>Total PN count</td>
<td>1.45M</td>
<td>54K</td>
<td>42K</td>
</tr>
<tr>
<td>OOV unigrams(^3)</td>
<td>12.4K</td>
<td>4.9K</td>
<td>4.3K</td>
</tr>
<tr>
<td>Documents with OOV(^3)</td>
<td>32.3K</td>
<td>2.25K</td>
<td>2.2K</td>
</tr>
<tr>
<td>Total OOV count(^3)</td>
<td>141K</td>
<td>9.1K</td>
<td>8K</td>
</tr>
<tr>
<td>OOV PN unigrams(^3)</td>
<td>9.3K</td>
<td>3.4K</td>
<td>3.1K</td>
</tr>
<tr>
<td>Documents with OOV PN(^3)</td>
<td>26.5K</td>
<td>1.9K</td>
<td>1.9K</td>
</tr>
<tr>
<td>Total OOV PN count(^3)</td>
<td>107K</td>
<td>6.9K</td>
<td>6.2K</td>
</tr>
</tbody>
</table>

TABLE I: Broadcast news diachronic datasets

2http://www.lexpress.fr/  
3http://fr.euronews.com/
GMM-HMM phone models trained on 200 hours of broadcast news audio files. It uses the Julius [58] speech recognition engine as the backend. The lexicon is based on French newspaper (*LeMonde*) news articles up to 2008 and contains 260k pronunciations for the 122k words. Using the SRILM toolkit [59], a 4-gram language model is estimated on text corpora of about 1800 million words. The automatic transcriptions of the test set obtained by ANTS have an average WER of 43.3% as compared to the reference transcriptions available from the source (http://fr.euronews.com). We found that these reference transcriptions are approximate for many test files. We obtained a WER of 33.8% on a set of 10 manually transcribed audio files which were chosen randomly from the test set.

2) Kaldi based Automatic Transcription System (KATS): The KATS LVCSR system is based on context dependent DNN-HMM phone models trained on the same dataset as ANTS. It uses the Kaldi [60] speech recognition engine at the backend. The lexicon is the same as that of ANTS. A bi-gram language model is estimated on the same text corpora as that used for training the ANTS language model. The automatic transcriptions of the test set audio news obtained by KATS have an average WER of 27.9% as compared to the reference transcriptions available from the source. We obtained a WER of 14.6% on the set of 10 manually transcribed audio files.

C. OOV PN Retrieval Performance Measures

To measure the performance of retrieval of OOV PNs relevant to an audio document we use measures based on Recall and Mean Average Precision (MAP) [61], which are commonly used to evaluate information retrieval systems. As mentioned earlier, for a given audio document several OOV PNs can be relevant. Those actually present in the audio are referred to as target OOV PNs. For our task we calculate recall (*R*) and precision (*P*) as:

\[ R = \frac{\# \text{ of target OOV PNs retrieved}}{\# \text{ total target OOV PNs}} \]

\[ P = \frac{\# \text{ of target OOV PNs retrieved}}{\# \text{ total OOV PNs retrieved}} \]

The MAP for a set of *Q* test (or validation) set documents is calculated as:

\[ MAP = \frac{\sum_{q=1}^{Q} \overline{P}(q)}{Q} \]  \hspace{1cm} (9)

where \( \overline{P}(q) \) is the average precision score for each document *q*. Given the ranked list of OOV PNs for a document \( \overline{P}(q) \) is calculated as:

\[ \overline{P}(q) = \frac{\sum_r P(r) \cdot \text{rel}(r)}{\# \text{ target OOV PNs in } q} \]  \hspace{1cm} (10)

where *P(r)* is the precision at rank *r*, \( \text{rel}(r) \) is an indicator function equaling 1 if the OOV PN at rank *r* is a target OOV PN or 0 otherwise.

Recall and MAP curves give different interpretation of results. After retrieval of the relevant OOV PNs, the top-*N* relevant OOV PNs are to be used for recovery/recognition of the target OOV PNs. To recover the target OOV PNs one can use an additional speech recognition pass [15], [36]; or spotting PNs in speech [62]. In each of these approaches, the retrieval ranks/scores may or may not be used. This is where the recall and MAP curves make a difference. The recall value at an operating point (*N* in the top-*N* choice) is not sensitive to the rank of the retrieved OOV PNs whereas the MAP value takes into account the retrieval ranks. For instance, in our experiments (see Fig. 4) if we take the top 5% (top 465) of the retrieved OOV PNs, all the methods will have same recall, but MAP will highlight the differences.

For detailed analysis, the retrieval results of the best model configurations will be shown as a graph of recall and MAP for the top-*N* retrieved OOV PNs (see Fig. 4). While calculating MAP the target OOV PNs not in the top-*N* OOV PN list get a precision score \( P(r) \) of zero. For direct comparison of two models, or model configurations, the maximum MAP achieved by the model will be used.

The statistical significance of the difference between the MAP values of two models is measured using Student’s paired t-test and randomisation test [63]. The null hypothesis is that there is no difference between the two models and they produce identical retrieval results. The null hypothesis is rejected if the *p*-value is less than 0.05 for both the tests [63]. For the randomisation test we generate 100,000 random permutations of the results of the two models under test.

D. Selection of Model Hyper-parameters

The LDA model has three hyper-parameters (a) \( \alpha \) the Dirichlet prior for document-topic distributions, (b) \( \beta \) the Dirichlet prior for topic-word distributions, and (c) \( T \) the number of topics which is also the size of the word and document topic vectors. There are works in literature [26], [64] which discuss the selection of LDA hyper-parameters and they are generally based on the log probability achieved by the model on a held out dataset. In our task we choose symmetric Dirichlet priors and select the hyper-parameters based on the maximum MAP achieved on our validation set. Table II shows the maximum MAP values obtained for a range of values for \( \alpha \), \( \beta \) and \( T \). With these hyper-parameters the maximum MAP varies between 0.229 and 0.370. Beyond these there is degradation or the improvement is insignificant.

The Skip-gram model has a crucial hyper-parameter, the context window size, apart from the word vector size. Table III shows the maximum MAP values obtained for a range of context window size and word vector size. We tried until window size of 20, limited by the length of the smallest documents in our datasets. For AverageVec the maximum MAP varies between 0.254 and 0.347 with different hyper-parameters.

Based on the maximum MAP obtained, and for comparison of the models, the number of LDA topics and the size of Skipgram word vectors are chosen to be 400. The Skip-gram model trained with a context window size of 20 is chosen and the LDA model with \( \alpha = 0.01 \) and \( \beta = 0.01 \) is chosen.

We found that the NBOV, NBOV2 and NBOV2+ models with word vectors of size 400 also gave the best validation performance. Apart from word vector size there are other crucial hyper-parameters to be chosen for these models. These will be discussed in detail in Section IV-E and Section V-B.
TABLE II: Selection of LDA hyper-parameters (c.f. Section II-C1) based on maximum MAP on the validation set. (Chosen model is in bold.)

<table>
<thead>
<tr>
<th>α</th>
<th>β</th>
<th>number of topics (T)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>0.01</td>
<td>0.244 0.319 0.352</td>
</tr>
<tr>
<td></td>
<td>0.25</td>
<td>0.298 0.338 0.357</td>
</tr>
<tr>
<td>0.25</td>
<td>0.248</td>
<td>0.260 0.288 0.280</td>
</tr>
<tr>
<td>0.1</td>
<td>0.244</td>
<td>0.318 0.334 0.359</td>
</tr>
<tr>
<td></td>
<td>0.1</td>
<td>0.313 0.343 0.357</td>
</tr>
<tr>
<td>0.25</td>
<td>0.252</td>
<td>0.276 0.271 0.267</td>
</tr>
<tr>
<td>0.25</td>
<td>0.239</td>
<td>0.284 0.333 0.336</td>
</tr>
<tr>
<td></td>
<td>0.251</td>
<td>0.297 0.336</td>
</tr>
<tr>
<td>0.25</td>
<td>0.229</td>
<td>0.270 0.354</td>
</tr>
</tbody>
</table>

TABLE III: Selection of Skip-gram hyper-parameters for AverageVec (Section II-C2) based on maximum MAP on the validation set. (Chosen model in bold.)

<table>
<thead>
<tr>
<th>context window</th>
<th>word vector dimension</th>
<th>100</th>
<th>200</th>
<th>300</th>
<th>400</th>
<th>500</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0.254 0.290 0.294</td>
<td>0.301 0.302</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>0.276 0.313 0.322</td>
<td>0.318 0.324</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>0.295 0.323 0.333</td>
<td><strong>0.347</strong> 0.345</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

E. Training the NBOW group of models

In this section we discuss in general the choices made for training the NBOW, NBOW2 and NBOW2+ models. It includes some crucial hyper-parameters which can affect the retrieval performance significantly. A more model specific discussion and comparison is made in Section V-B.

1) Initialisation: It is well known that good initialisation and pre-training of hidden layer weights are crucial for training deep neural networks [53], [65]. While the NBOW model is not deep, we examined if initialisation is crucial and if it affects the performance of the NBOW model in our task. We will present the results for the NBOW model with input word vectors initialised (a) randomly and (b) with Skip-gram word vectors pre-trained on the diachronic text corpus. The vectors corresponding to output OOV PNs are randomly initialised. (Initialising these with Skip-gram word vectors did not give any significant performance improvements.)

2) Full Training v/s Two Phase Training: We explore two methods of training the NBOW, NBOW2 and NBOW2+ models: (a) full training and (b) two phase training. In full training all the network parameters including the input matrix, output matrix, output bias vector of NBOW model (c.f. Section III-A), and additionally the anchor vector for NBOW2 and NBOW2+ models (c.f. Section III-B and III-C), are trained and updated using back-propagation.

The two phase training method has a first training phase in which only the output parameters ($W^O$, $b$), and the vector $a$ for the NBOW2 and NBOW2+ models, are updated by keeping the input word vectors fixed to pre-trained Skip-gram word vectors. In the second training phase all the model parameters including the word vectors are updated. The motivation behind the two phase training is again a better initialisation and convergence. The first training phase is supposed to take the randomly initialised output parameters to a better state for simultaneously training all the network parameters.

3) Learning Rate and Stopping Criteria: All the NBOW models are trained with gradient descent algorithm with ADADELTA [66]. ADADELTA provides an adaptive per-dimension learning rate for gradient descent and is robust to noisy gradient information. We tested two values of the ADADELTA decay constant ($\rho$), 0.99 and 0.95, and used $\rho = 0.99$ in all our experiments as it gives a lower validation error rate and a better retrieval performance.

To control the training of all the NBOW models an early stopping criterion [67] based on the validation set error is used. Early stopping is used in full training as well as both the first and the second training phases of two phase training.

4) Dropout at Input: The dropout technique [68] has been shown to significantly reduce overfitting and give major improvements over other regularisation methods in deep neural networks. While the NBOW model and the proposed NBOW2 and NBOW+ models are not deep architectures, we are interested to analyse if the dropout mechanism helps us to avoid overfitting and add robustness to the document level BOW input. We applied dropout at the input layer to (a) synthesise variations of training set documents, and (b) simulate deletion errors in LVCSR hypothesis. With experiments on the validation set we chose a word dropout probability of 0.9 (among 0, 0.25, 0.5, 0.75 and 0.9)\(^4\). We found that word dropout has been recently tried and gave improvements in text classification tasks [22], [45].

V. OOV PN RETRIEVAL RESULTS AND DISCUSSION

A. OOV PN Retrieval Performance

Fig. 4 shows the recall and MAP performance of retrieval of OOV PNs for different methods discussed in this paper. The performance on reference transcriptions (left), the ANTS LVCSR transcriptions (middle) and the KATS LVCSR transcriptions (right) of the Euronews test set audio are shown, in order to demonstrate the effect of errors in the LVCSR hypothesis on the OOV retrieval performance. In the case of reference transcriptions, the OOV PNs are removed and only the IV words are retained. The X-axis represents the number of OOV PNs selected from the diachronic text 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. For each of the methods, the models giving best performance on the validation set are chosen (see Section IV-D and V-B). The number of dimensions of the context/topic space is 400 (see Section IV-D).

We observe that our previous method based on LDA [1] performs better than AverageVec. The recall and MAP retrieval performance for NBOW, NBOW2 and NBOW2+ models is very similar and their graphs are overlapping. We will discuss in detail in Section V-B the difference in performance of the NBOW2 and NBOW2+ models as compared to the NBOW

\(^4\)Using a fixed number of epochs in the first phase of the two phase training did not give a better performance.

\(^5\)Word dropout probability $p$ does not necessarily translate to leaving out $p\%$ of the input words in our implementation.
Fig. 4: Recall and MAP performance of OOV PN retrieval for Euronews audio test set. — is LDA, — AverageVec, — NBO, — NBO2, — NBO2+. NBO, NBO2 and NBO2+ models initialised with Skip-gram word vectors and trained in two phases (c.f. Section IV-E2 and Table IV). and — overlapped by — NBO2+.

model. Overall the three models clearly outperform the other methods in terms of recall and MAP, both for reference and LVCSR transcriptions. It is interesting to note that the LDA based method is very robust to LVCSR word errors while the NBO group of models are slightly affected. In the case of LDA the difference between the maximum MAP of the reference transcription and that of the LVCSR transcription is statistically insignificant. The difference is significant for the NBO models. The AverageVec method is highly affected by LVCSR word errors.

B. Scrutinising the training of NBO and NBO2+ models

In this section, we first analyse how the choice of training conditions, namely (a) word dropout and (b) two phase training affect the performance of the NBO model. We present Table IV for this discussion. Then with the help of Fig. 5 and Table V we compare the training convergence and retrieval performance of the NBO, NBO2 and NBO2+ models.

1) Robustness with Word Dropout: The effect of applying word dropout can be observed from Table IV. It is clear that word dropout improves the MAP; a higher dropout rate giving a higher MAP. We can observe that the NBO model initialised with Skip-gram word vectors (Sg-1p) takes a smaller number of training epochs and gives better MAP performance than the NBO model with random initialisation (Rand1-p). However, applying word dropout gives larger relative improvements in Rand1-p as compared to Sg-1p. For instance the MAP value for reference transcriptions, i.e. MAP-TR, improves by 15% (0.076 points) for Rand1-p and by 6.75% (0.038 points) for Sg-1p for word dropout of 0.9 as compared to no word dropout. Secondly, we can observe that the improvement in MAP with word dropout is relatively larger for LVCSR transcriptions. For instance if we compare the MAP value for reference and ANTS LVCSR transcriptions, i.e. MAP-TR and MAP-TA, the improvements are 15% (0.076) v/s 25% (0.107) for Rand1-p, 6.75% (0.038) v/s 11.8% (0.058) for Sg-1p and 3.3% (0.020) v/s 8.2% (0.043) for Sg-2p.

<table>
<thead>
<tr>
<th>word dropout probability (p)</th>
<th>0.0</th>
<th>0.25</th>
<th>0.5</th>
<th>0.75</th>
<th>0.9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rand1-p</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>epochs</td>
<td>175</td>
<td>217</td>
<td>249</td>
<td>320</td>
<td>276</td>
</tr>
<tr>
<td>MAP-V</td>
<td>0.458</td>
<td>0.482</td>
<td>0.502</td>
<td>0.537</td>
<td>0.530</td>
</tr>
<tr>
<td>MAP-TR</td>
<td>0.500</td>
<td>0.522</td>
<td>0.549</td>
<td>0.578</td>
<td>0.576</td>
</tr>
<tr>
<td>MAP-TA</td>
<td>0.419</td>
<td>0.435</td>
<td>0.464</td>
<td>0.505</td>
<td>0.526</td>
</tr>
<tr>
<td>MAP-TK</td>
<td>0.457</td>
<td>0.473</td>
<td>0.500</td>
<td>0.533</td>
<td>0.542</td>
</tr>
<tr>
<td>Sg-1p</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>epochs</td>
<td>112</td>
<td>147</td>
<td>152</td>
<td>149</td>
<td>155</td>
</tr>
<tr>
<td>MAP-V</td>
<td>0.511</td>
<td>0.522</td>
<td>0.535</td>
<td>0.541</td>
<td>0.543</td>
</tr>
<tr>
<td>MAP-TR</td>
<td>0.563</td>
<td>0.569</td>
<td>0.576</td>
<td>0.587</td>
<td>0.601</td>
</tr>
<tr>
<td>MAP-TA</td>
<td>0.491</td>
<td>0.483</td>
<td>0.502</td>
<td>0.531</td>
<td>0.549</td>
</tr>
<tr>
<td>MAP-TK</td>
<td>0.523</td>
<td>0.522</td>
<td>0.532</td>
<td>0.551</td>
<td>0.566</td>
</tr>
<tr>
<td>Sg-2p</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>epochs</td>
<td>481</td>
<td>482</td>
<td>398</td>
<td>417</td>
<td>410</td>
</tr>
<tr>
<td>MAP-V</td>
<td>0.551</td>
<td>0.553</td>
<td>0.562</td>
<td>0.574</td>
<td>0.585</td>
</tr>
<tr>
<td>MAP-TR</td>
<td>0.602</td>
<td>0.598</td>
<td>0.605</td>
<td>0.615*</td>
<td>0.622</td>
</tr>
<tr>
<td>MAP-TA</td>
<td>0.525</td>
<td>0.519</td>
<td>0.533</td>
<td>0.561*</td>
<td>0.568</td>
</tr>
<tr>
<td>MAP-TK</td>
<td>0.555</td>
<td>0.552</td>
<td>0.561</td>
<td>0.578*</td>
<td>0.586</td>
</tr>
</tbody>
</table>

TABLE IV: Maximum MAP obtained by the NBO model (400 dimension word vectors) trained with an early stopping criterion. ‘epochs’ denote the total number of training epochs taken by the model to converge. Suffixes V, TR, TA and TK (to MAP) denote the performance on the validation set, the reference transcriptions of test set, ANTS LVCSR and KATS LVCSR transcriptions of test set respectively. Rand and Sg denote random and Skip-gram word vector initialisation. 1p and 2p denote one and two phase training. The best configuration is highlighted in bold. * denotes statistically insignificant difference compared to the best configuration.
2) Two phase training and the Improvement with NBOW2+ model: In Section IV-E2 we proposed to train the NBOW models in two phases. The MAP results in Table IV show that the best retrieval performance is obtained with this two phase training method. However, it takes a larger number of training epochs compared to training the NBOW model in one phase (Sg-1p). With the help of Fig. 5, we show that this problem is addressed by the NBOW2+ model. Fig. 5 shows a graph of validation set errors of the NBOW, NBOW2 and NBOW2+ models, as training progresses. It can be observed from Fig. 5 that all three models (NBOW, NBOW2 and NBOW2+) converge to the same point but at different convergence rates. While both NBOW and NBOW2 models take a larger number of training epochs, the NBOW2+ model gives a faster convergence without compromise in error rate. To support this argument we present Table V which compares the MAP achieved by the NBOW, NBOW2 and NBOW2+ models with 400 dimension word vectors and trained with a word dropout probability of 0.9.

As a counter experiment we examined if the ADADELTA decay constant ($\rho$) can speed up the two phase training convergence. We observed from our experiments that the ADADELTA decay constant ($\rho$) of 0.95 takes fewer training epochs as compared to decay constant ($\rho$) of 0.99, but at the cost of reduced MAP performance. For instance with word dropout of 0.9, the 400 dimensional NBOW model takes 351 epochs and achieves a maximum MAP of 0.5 as compared to 410 epochs and 0.568 MAP obtained with $\rho = 0.99$.

From these experiments we can conclude that (a) two phase training leads to better retrieval performance with the NBOW and NBOW2 models, although it requires a longer training, and (b) the NBOW2+ model, which combines the average and weighted average contexts of NBOW and NBOW2 models, can significantly reduce this training time without compromise in the MAP performance.

C. Word Importance weights of the NBOW2 model

We present Fig. 6 to discuss (a) the scalar word importance weights $\alpha_w$ learned by the NBOW2 model and (b) the choice of the function $f$ for the NBOW2 model (see Equation (7)). It shows a graph of the importance weights of words in a document from the test set. The left graph of Fig. 6 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 model learns and assigns different degrees of importance for different words. For example this test document is about the accident of Formula one driver Michael Schumacher and it has a missing OOV PN, ‘Kehm’ (Sabine Kehm is the spokesperson for Michael Schumacher). If we analyse the list of words as per the left graph, the top four important words are *medical, formula, critique and hospitaliser*. Moreover the second clear observation is that the NBOW2 model assigns higher weights to words which are important for retrieval of the OOV PN. The same holds true for the NBOW2 model with softmax $f$. Moreover the second clear observation is that the NBOW2 model with $f$ as softmax tends to assign higher weights to fewer words and weight close to zero for most other words. While this feature seems interesting, it leads to a relatively bad OOV PN retrieval performance [54]. We hypothesise that this happens because the NBOW2 model with softmax $f$ ignores (gives low importance value to) too many words from the input which affects its discriminative ability, especially when (a) the LVCSR hypothesis has many word errors and (b) the document contains OOV PNs from different semantic/topic contexts.

\begin{table}[h]
\centering
\begin{tabular}{|c|c|c|c|c|}
\hline
 & NBOW & NBOW2 & NBOW2+ \\
\hline
epochs & 276 & 123 & 210 \\
\hline
MAP-V & 0.530 & 0.474 & 0.519 \\
\hline
MAP-TR & 0.576 & 0.507 & 0.574 \\
MAP-TA & 0.526 & 0.402 & 0.526 \\
MAP-TK & 0.542 & 0.440 & 0.546 \\
\hline
epochs & 155 & 166 & 161 \\
\hline
MAP-V & 0.543 & 0.541 & 0.547 \\
MAP-TR & 0.601 & 0.599 & 0.601 \\
MAP-TA & 0.549 & 0.549 & 0.545 \\
MAP-TK & 0.566 & 0.566 & 0.566 \\
\hline
epochs & 410 & 648 & 273 \\
\hline
MAP-V & 0.585* & 0.587* & 0.593 \\
MAP-TR & 0.622* & 0.622* & 0.621 \\
MAP-TA & 0.568* & 0.566* & 0.569 \\
MAP-TK & 0.586* & 0.586* & 0.588 \\
\hline
\end{tabular}
\caption{Maximum MAP obtained by the NBOW, NBOW2 and NBOW2+ models with 400 dimension word vectors trained with word dropout probability $p = 0.9$ and an early stopping criterion. ‘epochs’ denote the total number of training epochs taken by the model to converge. Suffixes V, TR, TA and TK (to MAP) denote the performance on the validation set, reference transcription in test set, ANTS LVCSR and KATS LVCSR transcriptions of test set respectively. Rand and Sg denote random and Skip-gram word vector initialisation. 1p and 2p denote one and two phase training. The best configuration is highlighted in bold. * denotes statistically insignificant difference compared to the best configuration.}
\end{table}
D. Retrieval of Less Frequent OOV PNs

As discussed in our previous work [1], [35], the LDA topic based representation is biased against retrieval of less frequent OOV PNs. To address this problem we have previously proposed a re-ranking method [1] for less frequent OOV PNs and later showed that document specific context representation of OOV PNs [19] can give higher improvements in retrieval of less frequent OOV PNs. The document context representation method requires computing the document similarity with all the documents in the diachronic text corpus during test time but it is nonetheless effective.

We compared the overall retrieval performance as well as the retrieval performance of less frequent OOV PNs of the document context method with that of the NBOW model. The best of the document context method, based on Skip-gram word vectors with a vector size of 400, achieved a maximum MAP of 0.519 for reference and 0.462 for ANTS LVCSR as compared to 0.622 and 0.568 respectively for the NBOW model. Fig. 7 shows the distribution of ranks of OOV PNs based on their frequency of occurrence (document frequency) in the diachronic text corpus used for training the context models. As evident from Fig. 7, the NBOW model performs better in retrieval of less frequent OOV PNs. Quantitatively, there are about 24% (479 out of 2010) of OOV PNs which appear in 10 or fewer documents in the L’Express diachronic corpus and maximum MAP of 0.285, 0.088, 0.325 and 0.382 are achieved by the retrieval methods of graphs (a), (b), (c) and (d) respectively in Fig. 7.

VI. RECOGNITION OF OOV PNS

The list of relevant OOV PNs retrieved by the context model is to be used for recognition or recovery of the missed PNs. In our previous works we evaluated the effectiveness of the list of relevant OOV PNs obtained from context models by performing a keyword search based recovery. In [19] we performed a phonetic search for the top-$N$ relevant OOV PNs in the 1-best LVCSR hypothesis and in [21] we performed a Finite State Transducer (FST) based keyword search in the LVCSR lattice. While keyword search based recovery enables a faster evaluation, it results in many false alarms. In this paper we perform a second pass speech recognition to recognize the OOV PNs by updating the LVCSR system with the list of relevant OOV PNs retrieved by the context model.

A. Updating LVCSR for Recognition of OOV PNs

Updating the LVCSR system for new words requires updating the pronunciation lexicon and the LM n-gram probabilities. To update the pronunciation lexicon, automatic Grapheme-to-Phoneme G2P converters can be used. We trained the Sequitur G2P converter [69] on our original pronunciation lexicon and used it to generate up to 3 pronunciations of each new OOV PN. Estimating LM probabilities for new words is a non-trivial and open problem. Most of the proposed methods rely on similarity between IV and OOV words [70]–[72] or use word classes in the LM [73]–[75]. In our second pass speech recognition experiments we added OOV PNs as new unigrams without changing the existing unigram probabilities and leaving out the higher order n-grams for OOV PNs. The unigram probabilities are adjusted by taking a part of the $\langle unk \rangle$ probability and assigning it to an OOV PN as follows:

$$p_{oov \text{- unigram}} = p_{\langle unk \rangle} \times \frac{\delta}{\# OOV \text{ PNs}}$$  \hspace{1cm} (11)

where $\delta$ is the fraction of $\langle unk \rangle$ probability assigned to all the OOV PNs to be added. This approach to add OOV PNs is similar to a class LM with a class in unigrams. (A detailed comparison to other methods is not in the scope of this paper.)

B. Recognition Experiment Setup

Since the second pass speech recognition experiments are to be performed for different OOV PN lists, we formed a smaller test set for these experiments. From the total 3000 Euronews videos (see Table I), we formed a subset of videos appearing...
in 4 randomly selected weeks. This test subset comprises a total of 467 videos of which 318 (videos) have one or more PN missed in the first pass speech recognition as they were OOV. It must be noted that there are 149 videos in this test set with no OOV PN; as would be the case in a real setup where it is not known beforehand if the video has OOV(s) or not. The 318 videos contain a total of 1023 OOV PN (non-unique) terms, out of which up to 483 can be recovered with the L’Express diachronic corpus. The total number of words and PNs to be recognised are 97935 and 5838, respectively.

We perform the second pass speech recognition with our ANTS system since it can easily perform a document specific LM update at runtime\(^7\). Our baselines will be ANTS one pass speech recognition without knowledge of OOV PNs, denoted as No-OOV, and ANTS one pass speech recognition which includes all 9.3K new OOV PNs from L’Express, denoted as LX-All. We compare these to second pass speech recognition with the ANTS system updated with the top-128 document specific relevant OOV PNs retrieved by the LDA and the NBOW2+ models. These will be denoted as LDA-128 and NBOW2+-128. We chose the point 128 for our analysis because after this point both the recall as well as the MAP curves are flat, and before this point there are big differences in the recall of the different retrieval methods. Moreover, Skip-gram-128 performance is not shown but we found that it is similar to that of LDA. Similarly, we expect that NBOW-128 and NBOW2-128 would perform similar to NBOW2+-128.

The Recall@128 and MAP@128, i.e. the recall and MAP with the top-128 retrieved OOV PNs, for the LDA-128 and the NBOW2+-128 setup are shown in Table VI. Since we are using only a subset of the original test set, the Recall@128 and MAP@128 values are different compared to those in Fig. 4, but NBOW2+ gives a better performance than LDA as observed in Fig. 4.

<table>
<thead>
<tr>
<th>TABLE VI: OOV PN retrieval performance on the test subset after the first pass using ANTS LVCSR. (These retrieval results will be used in the second pass recognition.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LDA-128</td>
</tr>
<tr>
<td>Recall@128</td>
</tr>
<tr>
<td>MAP@128</td>
</tr>
</tbody>
</table>

To tune the \(\delta\) parameter in (11), we used another subset of Euronews videos (not part of the test subset). After different trials we chose a value of 0.001 for \(\delta\), which gave an optimal performance for each of the methods. A higher value of \(\delta\) will improve the OOV PN recognition but also lead to increased false alarms.

We also present the PNER and WER results from an oracle setup. In the oracle setup we perform only one pass of ANTS speech recognition using an updated pronunciation lexicon and LM (using (11)). They are specific to each video from the test sub-set and include the OOV PNs which actually appear in the video. For comparison we add only those OOV PNs which can be obtained using the L’Express diachronic text corpus.

\(^7\)The KATS system is based on Kaldi which requires a lengthy (~6hours) compilation of the LM (HCLG) FST.

C. Recognition Results

Table VII shows the Proper Name Error Rate (PNER) after second pass speech recognition. PNER is obtained by first aligning the reference and hypothesised word level transcriptions and then calculating substitution, deletion, insertion errors, and thus the error rate only on the proper name terms. Similarly, OOVPNER is the error rate calculated only for OOV PNs. WER is the word error rate on this test set. It can be observed from Table VII that adding all OOV PNs from the diachronic corpus (LX-All) leads to an increased PNER and OOVPNER. The increased error rate is mainly due to insertion and substitution errors, and it can possibly be reduced with better LM update techniques. The LDA and NBOW2+ context models enable selection of relevant OOV PNs and hence the recognition of new PNs and reduction of PNER. While LDA and NBOW2+ models show similar PNER performance, we can see that NBOW2+ gives a lower OOVPNER. The NBOW2+ model leads to more correctly recognised OOV PNs. The performance of NBOW2+ is close to our Oracle setup, and after analysing errors in the Oracle setup we hypothesise that automatic G2P pronunciations of OOV PNs is another source of recognition errors. Furthermore, adding the new PNs into the vocabulary and language model did not have a negative impact on the WER. Instead the WER showed minor improvements, 0.7% and 0.8% absolute for LDA-128 and NBOW2+-128, with respect to the No-OOV case. The improvement in WER were due to recognition of OOV PNs and due to reduction in insertion and deletion errors.

<table>
<thead>
<tr>
<th>TABLE VII: Second pass PN recognition results. PNER denotes Proper Name Error Rate. OOVPNER denotes OOV Proper Name Error Rate. (For LDA-128 and NBOW2+-128, top-128 document specific OOV PNs, retrieved by the LDA and NBOW2+ models, are added to the lexicon and LM.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>No-OOV</td>
</tr>
<tr>
<td>--------------------------------------------------</td>
</tr>
<tr>
<td>OOV PNs added</td>
</tr>
<tr>
<td>% OOVPNER</td>
</tr>
<tr>
<td>% PNER</td>
</tr>
<tr>
<td>% WER</td>
</tr>
</tbody>
</table>

VII. Conclusion

Semantic context models can improve the recovery of OOV words by significantly reducing the list of possible OOVs for an audio document. We discussed methods based on LDA topic models, neural word vector representations and examined the NBOW model for the 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 retrieval of an OOV PN.

With experiments on French broadcast news videos we showed that our methods based on the LDA topic model and NBOW models can retrieve about 37% to 39% of the target OOV PNs within the top 1% of the 9321 possible OOV PNs obtained from a six month period of news from L’Express news website. (This translates to a recall rate of 85% to 90% considering only the target OOV PNs from the L’Express
diachronic text corpus.) The proposed methods based on LDA and NBOW group of models are robust to LVCSR word errors. The NBOW and NBOW2 models give improvements in retrieval performance as compared to the raw embedding methods. Two phase training and input word dropout techniques enable the NBOW models to achieve improved performance. Combining the NBOW and NBOW2 into a new model leads to a faster convergence in training. The relevant OOV PNs retrieved by the context models were further evaluated by performing a second pass speech recognition. These second pass speech recognition experiments demonstrated a 4.8% absolute reduction in proper name error rate, which would otherwise increase by 6.2% absolute or more if all OOV PNs were simply added to the LVCSR vocabulary. Further improvements are possible by using well designed language model adaptation schemes and by using diachronic text data from more sources.

These results motivate us to extend the NBOW2 model to deal with other scenarios including broadcast news audio with multiple news events. Keywords identified by NBOW2 and NBOW2+ models could be used in a setup of trigger based language models [76]. Moreover, we foresee more sophisticated techniques for learning and assigning word importance instead of using a single vector to obtain the importance weights.

ACKNOWLEDGMENT

Experiments presented in this paper were partly 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).

The authors thank Emmanuel Vincent and the anonymous reviewers for their comments that greatly improved the manuscript.

REFERENCES


