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

Imran Sheikh, Dominique Fohr, Irina Illina, Georges Linares

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 and language model are updated with the retrieved OOV PNs, demonstrate the effectiveness of the proposed context models.**

***Index Terms*—large vocabulary continuous speech recognition, out-of-vocabulary, proper names, semantic context**

I. Introduction

Broadcast news data are diachronic in nature, characterised by continuous changes in information and content. The frequent variations in linguistic content and vocabulary pose a challenge to *Large Vocabulary Continuous Speech Recognition* (LVCSR). All possible known words cannot be included in the vocabulary and *Language Model* (LM) of an LVCSR system because (a) there are many infrequent and new words, particularly *Proper Names* (PNs), which are not well represented in training data, and (b) it would increase the LVCSR search space and complexity without guaranteeing a decrease in the *Word Error Rate* (WER). Therefore a practical choice is to leave out a part of the vocabulary, which then leads to *Out-Of-Vocabulary* (OOV) words in LVCSR. An analysis of the OOV words reveals that a majority of OOV words (56-72% [1]) are PNs. However PNs are important for automatic indexing of audio-video content as well as for obtaining accurate and reliable automatic transcriptions. In this paper, we focus on the problem of retrieval and recovery of OOV PNs in diachronic audio documents.

Methods addressing OOV words in LVCSR systems can be categorised into (a) OOV detection based approaches and (b) vocabulary selection based approaches. OOV detection based approaches [2]–[6] aim to detect the presence of OOV words and/or locate OOV regions in the LVCSR hypothesis, followed by a search for the matching OOV word. These approaches mainly use hybrid language models which can hypothesise both in-vocabulary word and sub-word units, 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

Manuscript received ; revised . The work was supported by the French National Research Agency (ANR) ContNomina project under contract ANR-12-BS02-0009.

Imran Sheikh, Dominique Fohr and Irina Illina are with the Multispeech (Inria/CNRS/Université de Lorraine) project-team at LORIA, UMR 7503, Vandoeuvre-lès-Nancy, F-54506, France (e-mail: {imran.sheikh, irina.illina, dominique.fohr}@loria.fr).

Georges Linarès is with the Laboratoire d'Informatique d'Avignon (LIA), University of Avignon, 84000 Avignon, France (e-mail: georges.linares@univ-avignon.fr).

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.

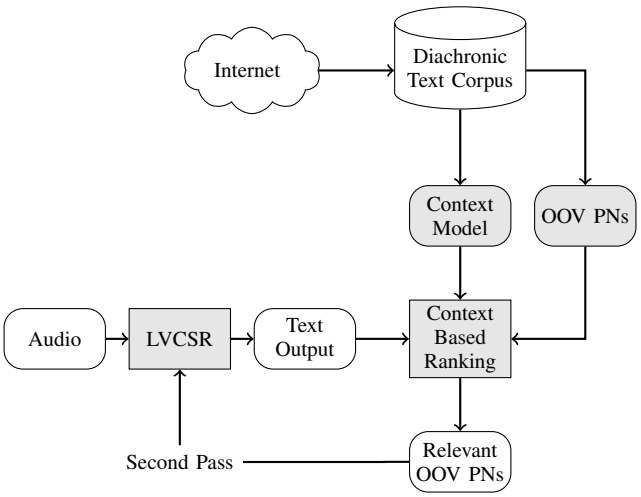

Fig. 1: Block diagram illustration of our approach for recognition of Out-Of-Vocabulary (OOV) Proper Names (PNs).

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:

$$p(w, z, \theta, \phi|\alpha, \beta) = p(\phi|\beta)\ p(\theta|\alpha)\ p(z|\theta)\ p(w|\phi_z) \quad (1)$$

where z is the hidden topic assignment to word w in a document d, $\theta = [\theta_{dt}]_{D\times T}$ is the multinomial topic distribution for each document d and $\phi = [\phi_{vt}]_{V\times T}$ is the multinomial topic distribution for each word, both across T topics. α and β are Dirichlet priors for θ and ϕ respectively. The topic model parameters θ, ϕ 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:[1]:

$$\arg\max_{\Theta} \prod_{w \in corpus} \left[\prod_{c \in C(w)} p(c|w;\Theta) \right] \quad (2)$$

where Θ 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.

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 of the audio document will be projected into the context space to infer relevant OOV PNs. In this section we present OOV PN retrieval methods based on representations from LDA and Skip-gram models. It must be noted that the representations from the LDA model are actually multinomial distributions (or topic distribution vectors), learned in an unsupervised manner following a Bayesian parameter estimation setup. Similarly, the Skip-gram model learns word vectors with an objective to maximise the average log probability of predicting the surrounding context word given the center word.

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 $\tilde{v}_i$. 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 ϕ learned during training. Given $p(\tilde{v}_i|t)$ from ϕ, the likelihood of an OOV PN ($\tilde{v}_i$) in the diachronic text corpus is calculated as:

$$p(\tilde{v}_i|h) = \sum_{t=1}^{T} p(\tilde{v}_i|t)\, p(t|h) \quad (3)$$

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

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 $\tilde{v}_i$ for each of the OOV PNs to calculate a score as:

$$s_i = \frac{\sum_{k=1}^{K} h_k\, \tilde{v}_{ik}}{\sqrt{\sum_{k=1}^{K} (h_k)^2}\, \sqrt{\sum_{k=1}^{K} (\tilde{v}_{ik})^2}} \quad (4)$$

The score s_i is used to rank and retrieve the OOV PNs $\tilde{v}_i$.

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

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].

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:

$$z = \frac{1}{|X|} \sum_{w \in X} v_w \tag{5}$$

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 $\text{softmax}(l) = \exp(l)/\sum_{j=1}^{L} \exp(l_j)$.

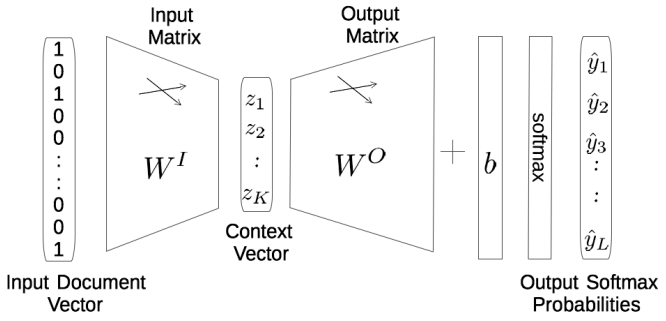

Fig. 2: The Neural Bag-of-Words (NBOW) Model.

Fig. 2 shows a representation of the NBOW model as used in our task to retrieve relevant OOV PNs. The input (word embedding) matrix W^I has vectors corresponding to the IV words & PNs ($W^I \equiv \{v_1, v_2, v_3 ... v_V\}$) and the output matrix W^O has vectors corresponding to OOV PNs. The input is a sparse BOW vector with 1's representing the IV words and PNs present in a training/test document. The average vector $z \equiv \{z_1, z_2 ... z_K\}$ represents the context vector for the document. A vector-matrix product between the average/context vector and the output/OOV PN matrix (W^O) is equivalent to comparison of the input document and the OOV PNs in the context space.

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. As discussed in previous works [37] the cross-entropy loss function leads to better classification performance and faster convergence as compared to the pairwise error function which tries to minimise the ranking loss in multi label classification.

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.

$$z = \frac{1}{|X|} \sum_{w \in X} \alpha_w \; v_w \tag{6}$$

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

$$\alpha_w = f(v_w \cdot a) \tag{7}$$

where $(\cdot)$ represents the dot product and f scales the importance weights to $[0, 1]$. The word importance weight α_w is a function of the distance of that word w from a in the context space, ensuring that the calculation of α_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.

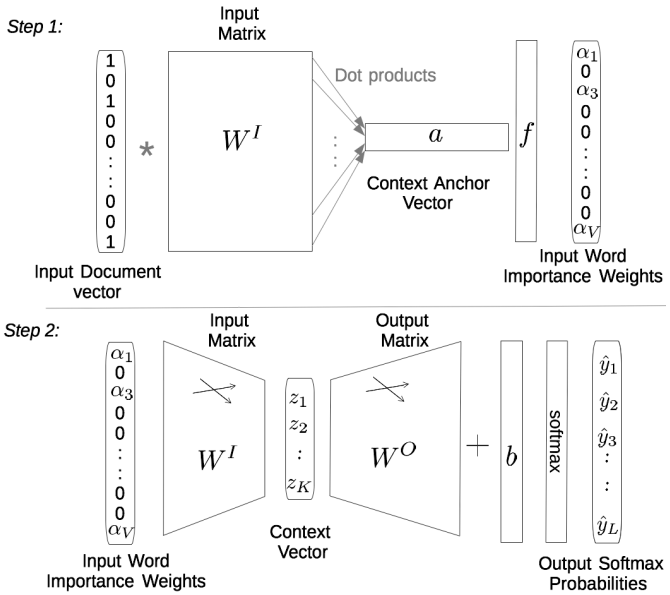

Fig. 3: Neural Bag-of-Weighted-Words (NBOW2) Model.

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_1^I, W_2^I) and hence maintains two K dimensional word vectors v_w^1 and v_w^2 for each input word w. It has one K dimensional anchor vector a similar to NBOW2 and one matrix W^O and one bias vector 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:

$$z_1 = \frac{1}{|X|}\sum_{w\in X} v_w^1 \quad , \quad z_2 = \frac{1}{|X|}\sum_{w\in X} \alpha_w \, v_w^2 \qquad z = [z_1 z_2] \tag{8}$$

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]http://www.lexpress.fr/

[3]http://fr.euronews.com/

TABLE I: Broadcast news diachronic datasets

	L'Express (train)	*Euronews* (valid)	*Euronews* (test)
Type of Documents	Text	Text	Video
Time Period		Jan - Jun 2014	
Number of Documents[1]	45K	3.1K	3K
Vocabulary Size[2]	150K	42K	45K
Corpus Size (word count)	24M	550K	700K
PN unigrams[2]	57K	12K	11K
Total PN count	1.45M	54K	42K
OOV unigrams[3]	12.4K	4.9K	4.3K
Documents with OOV[3]	32.3K	2.25K	2.2K
Total OOV count[3]	141K	9.1K	8K
OOV PN unigrams[3]	9.3K	3.4K	3.1K
Documents with OOV PN[3]	26.5K	1.9K	1.9K
Total OOV PN count[3]	107K	6.9K	6.2K

[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

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

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:

$$MAP = \frac{\sum_{q=1}^{Q} \overline{P}(q)}{Q} \tag{9}$$

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:

$$\overline{P}(q) = \frac{\sum_r P(r)\ rel(r)}{\#\text{ target OOV PNs in q}} \tag{10}$$

where $P(r)$ is the precision at rank r, $rel(r)$ 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) α the Dirichlet prior for document-topic distributions, (b) β 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 α, β 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 values for context window 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 Skip-gram 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 NBOW, NBOW2 and NBOW2+ 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.)

α	β	number of topics (T) 100	200	300	400	500
0.01	0.01	0.244	0.319	0.352	**0.368**	0.370
	0.1	0.254	0.298	0.338	0.357	0.365
	0.25	0.248	0.260	0.288	0.280	0.268
0.1	0.01	0.244	0.318	0.334	0.359	0.365
	0.1	0.245	0.313	0.343	0.357	0.348
	0.25	0.252	0.276	0.271	0.267	0.286
0.25	0.01	0.239	0.284	0.333	0.336	0.354
	0.1	0.251	0.297	0.336	0.329	0.330
	0.25	0.229	0.270	0.354	0.262	0.250

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.)

context window	word vector dimension 100	200	300	400	500
10	0.254	0.290	0.294	0.301	0.302
15	0.276	0.313	0.322	0.318	0.324
20	0.295	0.323	0.333	**0.347**	0.345

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 (ρ), 0.99 and 0.95, and used ρ =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].

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 NBOW2+ 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)[5]. 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.

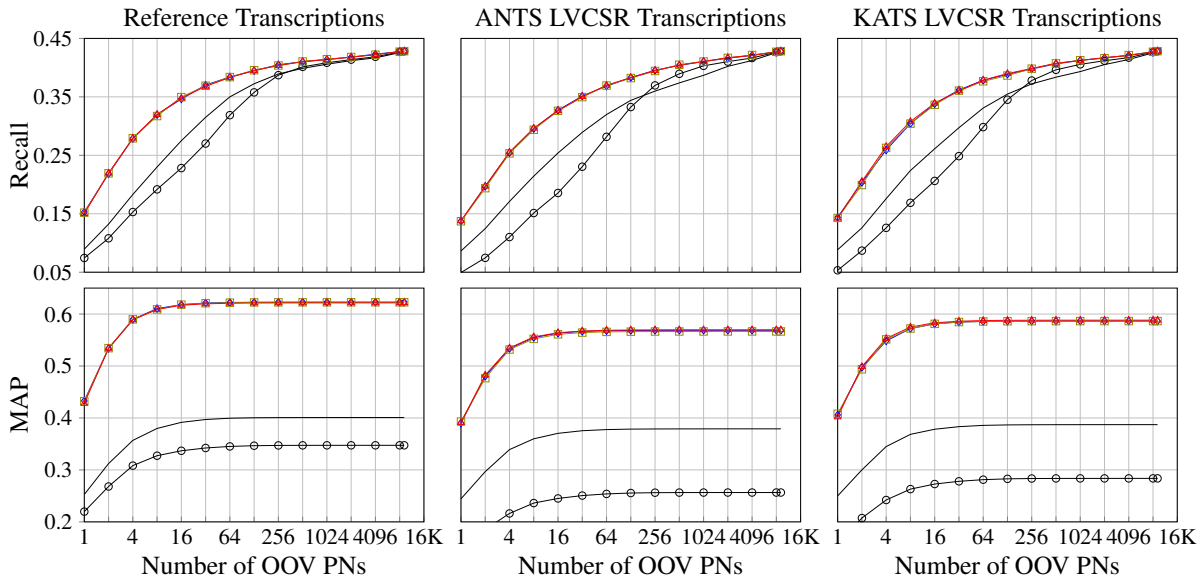

Fig. 4: Recall and MAP performance of OOV PN retrieval for *Euronews* audio test set. —— is LDA, —○— AverageVec, —◇— NBOW, —□— NBOW2, —△— NBOW2+. NBOW, NBOW2 and NBOW2+ models initialised with Skip-gram word vectors and trained in two phases (c.f. Section IV-E2 and Table IV). —◇— and —□— overlapped by —△— NBOW2+.

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 NBOW 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 NBOW models. The AverageVec method is highly affected by LVCSR word errors.

B. Scrutinising the training of NBOW and NBOW2+ 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 NBOW 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 NBOW, NBOW2 and NBOW2+ 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 NBOW model initialised with Skip-gram word vectors (Sg-1p) takes a smaller number of training epochs and gives better MAP performance than the NBOW model with random initialisation (Rand1-p). However, applying word dropout gives larger relative improvements in Rand-1p as compared to Sg-1p. For instance the MAP value for reference transcriptions, i.e. MAP-TR, improves by 15% (0.076 points) for Rand-1p 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 Rand-1p, 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 IV: Maximum MAP obtained by the NBOW 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.

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

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.

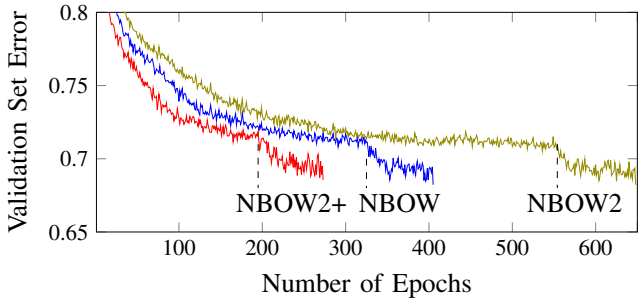

Fig. 5: Validation set errors during the two phase training of NBOW, NBOW2 and NBOW2+ models. (- - - markers indicate end of first and begin of second training phase)

As a counter experiment we examined if the ADADELTA decay constant (ρ) can speed up the two phase training convergence. We observed from our experiments that the ADADELTA decay constant (ρ) of 0.95 takes fewer training epochs as compared to decay constant (ρ) 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 ρ = 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 α_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

TABLE V: 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.

		NBOW	NBOW2	NBOW2+
R-1p	epochs	276	123	210
	MAP-V	0.530	0.474	0.519
	MAP-TR	0.576	0.507	0.574
	MAP-TA	0.526	0.402	0.526
	MAP-TK	0.542	0.440	0.546
Sg-1p	epochs	155	166	161
	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
Sg-2p	epochs	410	648	273
	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**

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 *michael*, *formule*, *critique* and *hospitaliser* and the four least important words are *rester*, *tenir*, *monde* and *présent*[6]. From this example, it is evident 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.

[6]In English: *formula, critical, hospitalise, remain, stay, world, present*

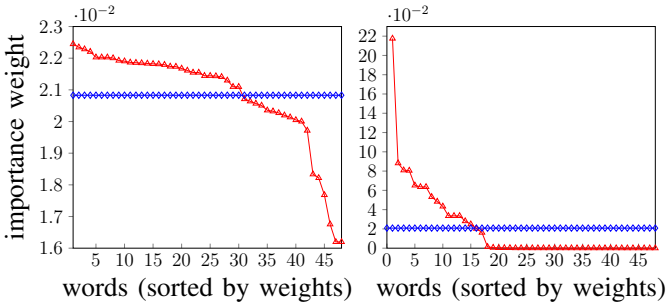

Fig. 6: Word importance weights assigned by the NBOW2 model (—△—) in a sample document with 48 words. Two variations of the NBOW2 model are shown: (left) f as sigmoid and (right) f as softmax. —◇— denotes the all equal weights (1/48 = 0.0208) in the simple average by the NBOW model.

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

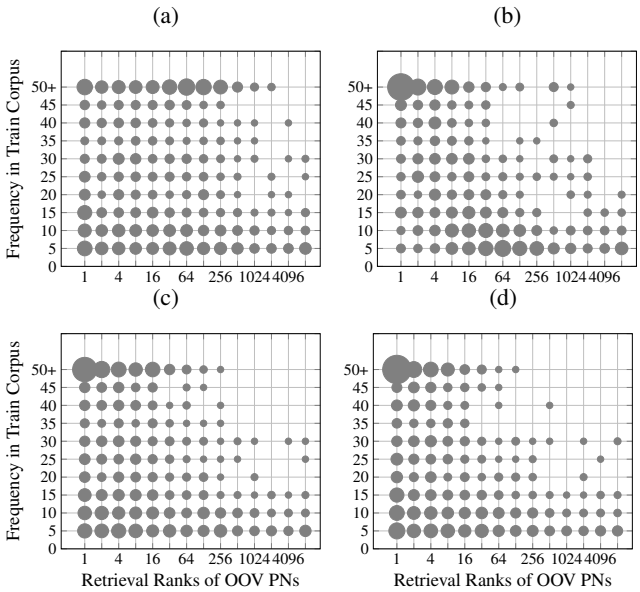

Fig. 7: Rank-Frequency distribution for retrieval of OOV PNs with (a) AverageVec (b) LDA (c) Document Specific OOV PN representation using Skip-gram word vectors [19] (d) NBOW.

we perform a second pass speech recognition to recognise 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 $<unk>$ probability and assigning it to an OOV PN as follows:

$$p_{oov-pn-unigram} = p_{<unk>} \times \frac{\delta}{\#\text{ OOV PNs}} \qquad (11)$$

where δ is the fraction of $<unk>$ 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 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.)

	LDA-128	NBOW2+-128
Recall@128	0.37	0.41
MAP@128	0.41	0.62

To tune the δ 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 δ, which gave an optimal performance for each of the methods. A higher value of δ 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 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.)

	No-OOV	LX-All	LDA -128	NBOW2+ -128	Oracle
OOV PNs added	0	9.3K	128	128	oracle
% OOVPNER	100.0	117.8	63.9	63.6	63.1
% PNER	61.6	67.8	57.0	56.8	56.7
% WER	52.7	52.8	52.0	51.9	51.8

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

[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] A. Rastrow, A. Sethy, B. Ramabhadran, and F. Jelinek, "Towards using hybrid word and fragment units for vocabulary independent LVCSR systems," in *ISCA INTERSPEECH*, 2009, pp. 1931–1934.

[3] L. Qin and A. Rudnicky, "OOV word detection using hybrid models with mixed types of fragments," in *ISCA INTERSPEECH*, 2012, pp. 2450–2453.

[4] C. Parada, M. Dredze, D. Filimonov, and F. Jelinek, "Contextual information improves OOV detection in speech," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2010, pp. 216–224.

[5] S. Kombrink, M. Hannemann, and Lukáš Burget, *Detection and Identification of Rare Audiovisual Cues*. Springer Berlin Heidelberg, 2012, ch. Out-of-Vocabulary Word Detection and Beyond, pp. 57–65.

[6] W. Chen, S. Ananthakrishnan, R. Prasad, and P. Natarajan, "Variable-span out-of-vocabulary named entity detection," in *ISCA INTERSPEECH*, 2013, pp. 3761–3765.

[7] M. Bisani and H. Ney, "Open vocabulary speech recognition with flat hybrid models," in *ISCA INTERSPEECH*, 2005, pp. 725–728.

[8] M. A. B. Shaik, A. E. D. Mousa, S. Hahn, R. Schlüter, and H. Ney, "Improved strategies for a zero OOV rate LVCSR system," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2015, pp. 5048–5052.

[9] A. Allauzen and J.-L. Gauvain, "Diachronic vocabulary adaptation for broadcast news transcription," in *9th European Conference on Speech Communication and Technology (INTERSPEECH'2005 - Eurospeech)*, 2005, pp. 1305–1308.

[10] C.-E. Liu, K. Thambiratnam, and F. Seide, "Online vocabulary adaptation using limited adaptation data," in *ISCA INTERSPEECH*, 2007, pp. 1821–1824.

[11] D. Jouvet and D. Langlois, "A machine learning based approach for vocabulary selection for speech transcription," in *16th International Conference on Text, Speech, and Dialogue (TSD)*, 2013, pp. 60–67.

[12] A. I. R. Ming Sun, Yun-Nung Chen, "Learning OOV through semantic relatedness in spoken dialog systems," in *ISCA INTERSPEECH*, 2015, pp. 1453–1457.

[13] C. Martins, A. Texeira, and J. Neto, "Dynamic language modeling for a daily broadcast news transcription system," in *IEEE Workshop on Automatic Speech Recognition Understanding*, 2007, pp. 165–170.

[14] O. S. S. Seneff, "A two-pass strategy for handling OOVs in a large vocabulary recognition task," in *ISCA INTERSPEECH*, 2005, pp. 1669–1672.

[15] S. Oger, G. Linarès, F. Béchet, and P. Nocera, "On-demand new word learning using world wide web," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2008, pp. 4305–4308.

[16] S. Meng, L.-F. Wang, Y.-M. Lin, G. Li, K. Thambiratnam, and F. Seide, "Vocabulary and language model adaptation using just one file," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2010, pp. 5410–5413.

[17] P. Maergner, A. Waibel, and I. Lane, "Unsupervised vocabulary selection for real-time speech recognition of lectures," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2012, pp. 4417–4420.

[18] I. Nkairi, I. Illina, G. Linarès, and D. Fohr, "Exploring temporal context in diachronic text documents for automatic OOV proper name retrieval," in *Language & Technology Conference*, 2013, pp. 540–544.

[19] 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, pp. 6050–6054.

[20] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, Mar. 2003.

[21] I. Sheikh, I. Illina, D. Fohr, and G. Linarès, "Improved neural bag-of-words model to retrieve out-of-vocabulary words in speech recognition," in *ISCA INTERSPEECH*, 2016, pp. 675–679.

[22] 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.

[23] P. D. Turney and P. Pantel, "From frequency to meaning: Vector space models of semantics," *J. Artif. Int. Res.*, vol. 37, no. 1, pp. 141–188, Jan. 2010.

[24] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal Of The American Society For Information Science*, vol. 41, no. 6, pp. 391–407, 1990.

[25] T. L. Griffiths, J. B. Tenenbaum, and M. Steyvers, "Topics in semantic representation," *Psychological Review*, vol. 114(2), pp. 211–244, 2007.

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

[27] 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*, 2013, pp. 3111–3119.

[28] 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.

[29] 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.

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

[31] Y. Goldberg and O. Levy, "word2vec explained: deriving Mikolov et al.'s negative-sampling word-embedding method," *CoRR*, vol. abs/1402.3722, 2014.

[32] A. O. Bayer and G. Riccardi, "Semantic language models for automatic speech recognition," in *Spoken Language Technology Workshop (SLT), 2014 IEEE*, Dec 2014, pp. 7–12.

[33] 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.

[34] 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.

[35] I. Sheikh, I. Illina, and D. Fohr, "Study of entity-topic models for OOV proper name retrieval," in *ISCA INTERSPEECH*, 2015, pp. 3506–3510.

[36] D. Fohr and I. Illina, "Continuous word representation using neural networks for proper name retrieval from diachronic documents," in *ISCA INTERSPEECH*, 2015, pp. 1344–1348.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

[41] 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 (EMNLP)*, 2013, pp. 1631–1642.

[42] 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.

[43] 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.

[44] 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.

[45] A. M. Dai and Q. V. Le, "Semi-supervised sequence learning," in *Advances in Neural Information Processing Systems*, 2015, pp. 3079–3087.

[46] 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.

[47] 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 (EMNLP)*, 2015, pp. 1367–1372.

[48] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv:1409.0473*, 2014.

[49] W. Chan, N. Jaitly, Q. V. Le, and O. Vinyals, "Listen, attend and spell," *CoRR*, vol. abs/1508.01211, 2015.

[50] K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proceedings of International Conference on Machine Learning (ICML)*, 2015, pp. 2048–2057.

[51] S. K. Sønderby, C. K. Sønderby, H. Nielsen, and O. Winther, "Convolutional lstm networks for subcellular localization of proteins," in *Proceedings of the 2nd International Conference on Algorithms for Computational Biology*, 2015, pp. 68–80.

[52] 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.

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

[54] 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.

[55] A. Allauzen and H. Bonneau-Maynard, "Training and evaluation of pos taggers on the french multitag corpus," in *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC)*, May 2008.

[56] I. Sheikh, I. Illina, and D. Fohr, "How diachronic text corpora affect context based retrieval of oov proper names for audio news," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*, may 2016, pp. 3851–3855.

[57] 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.

[58] A. Lee and T. Kawahara, "Recent development of open-source speech recognition engine julius," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2009, pp. 131–137.

[59] A. Stolcke, "SRILM - an extensible language modeling toolkit," in *Proceedings International Conference on Spoken Language Processing*, November 2002, pp. 257–286.

[60] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2011.

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

[62] C. Parada, A. Sethy, M. Dredze, and F. Jelinek, "A spoken term detection framework for recovering out-of-vocabulary words using the web," in *11th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2010, pp. 1269–1272.

[63] 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.

[64] H. M. Wallach, D. M. Mimno, and A. McCallum, "Rethinking lda: Why priors matter," in *Advances in Neural Information Processing Systems*, 2009, pp. 1973–1981.

[65] H. Larochelle, Y. Bengio, J. Louradour, and P. Lamblin, "Exploring strategies for training deep neural networks," *J. Mach. Learn. Res.*, vol. 10, pp. 1–40, Jun. 2009.

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

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

[68] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, 2014.

[69] M. Bisani and H. Ney, "Joint-sequence models for grapheme-to-phoneme conversion," *Speech Communication*, vol. 50, no. 5, pp. 434 – 451, 2008.

[70] L. Orosanu and D. Jouvet, "Adding new words into a language model using parameters of known words with similar behavior," in *International Conference on Natural Language and Speech Processing*, 2015.

[71] L. Qin, "Learning out-of-vocabulary words in automatic speech recognition," Ph.D. dissertation, Language Technologies Institute, School of Computer Science, Carnegie Mellon University, 2013.

[72] G. Lecorv, G. Gravier, and P. Sbillot, "Automatically finding semantically consistent n-grams to add new words in lvcsr systems," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2011, pp. 4676–4679.

[73] A. Allauzen and J.-L. Gauvain, "Open vocabulary ASR for audiovisual document indexation," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2005, pp. 1013–1016.

[74] A. Pražák, P. Ircing, and L. Müller, "Language model adaptation using different class-based models," in *SPECOM 2007 Proceedings*, Moscow, 2007, pp. 449–454.

[75] W. Naptali, M. Tsuchiya, and S. Nakagawa, "Class-based n-gram language model for new words using out-of-vocabulary to in-vocabulary similarity." *IEICE Transactions*, vol. 95-D, no. 9, pp. 2308–2317, 2012.
[76] C. Troncoso and T. Kawahara, "Trigger-based language model adaptation for automatic transcription of panel discussions," *IEICE - Trans. Inf. Syst.*, vol. E89-D, no. 3, pp. 1024–1031, Mar. 2006.