DOCUMENT LEVEL SEMANTIC CONTEXT FOR RETRIEVING OOV PROPER NAMES

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

[*]MultiSpeech Group, LORIA-INRIA, 54500 Villers-lès-Nancy, France
[+]LIA, University of Avignon, 84911 Avignon, France
{imran.sheikh, irina.illina, dominique.fohr}@loria.fr, georges.linares@univ-avignon.fr

ABSTRACT

Recognition of Proper Names (PNs) in speech is important for content based indexing and browsing of audio-video data. However, many PNs are Out-Of-Vocabulary (OOV) words for LVCSR systems used in these applications due to the diachronic nature of data. By exploiting semantic context of the audio, relevant OOV PNs can be retrieved and then the target PNs can be recovered. To retrieve OOV PNs, we propose to represent their context with document level semantic vectors; and show that this approach is able to handle less frequent OOV PNs in the training data. We study different representations, including Random Projections, LSA, LDA, Skip-gram, CBOW and GloVe. A further evaluation of recovery of target OOV PNs using a phonetic search shows that document level semantic context is reliable for recovery of OOV PNs.

Index Terms— OOV, proper names, semantic, indexing

1. INTRODUCTION

Large Vocabulary Continuous Speech Recognition (LVCSR) based audio indexing approaches allow search, navigation, browsing and structuring of audio-video datasets [1]. However, such datasets are diachronic and LVCSR processing can be challenging due to the variations in linguistic content and vocabulary. Thus leading to *Out-Of-Vocabulary* (OOV) words for LVCSR. In previous works [2–6] it has been observed that a majority of OOV words (56-72%) are Proper Names (PNs). But PNs are important for indexing. In this paper, we focus on the problem of retrieval of relevant OOV PNs for recovery of the target OOV PNs. (We refer to all the PNs not in LVCSR vocabulary as *OOV PNs* and the OOV PNs present in the test audio as *target OOV PNs*.)

When processing diachronic audio with LVCSR systems, the number of OOV PNs can be very high [7]. Even if good amount of training data is available, appending the LVCSR vocabulary and updating the *Language Model* (LM) is not always a feasible solution [2]. To recognize the target OOV PNs in an audio document, we first find a list of OOV PNs which are relevant to this audio document, by using diachronic text resources. OOV PNs are inferred based on the semantic/topic context of the audio document. (We consider datasets which contain one news event per document. News documents with multiple news events are part of another study [7].) The *reduced* list of relevant OOV PNs can then be used to recover target OOV PNs using phone matching [8], or additional speech recognition pass [9]; or spotting PNs in speech [10]. In this paper, we evaluate recovery of the target OOV PNs using a phonetic search. (LVCSR decoding with updated LM can give better results [11] but this is not our current focus.)

Retrieval and recovery of OOV PNs in audio documents, based on the topic and semantic context, has been proposed previously. *Latent Dirichlet Allocation* (LDA) and *Latent Semantic Analysis* (LSA) was used to train context models for PNs in [12, 13]. However, these approaches work for PNs with significant amount of training documents. Retrieval methods adapted to handle less frequent PNs were proposed in [14, 15]. In [16] topic context was used to update LM for LVCSR based term detection. But as mentioned, updating LVCSR LM is not always feasible [2, 7]. The originality of our work is the use of document level semantic representations to model OOV PNs. We follow this approach, because it can model less frequent OOV PNs which have less training instances and secondly because this approach (unlike the rare OOV PN re-ranking technique proposed in [14, 15]) can be applied to different semantic representations, as discussed in this paper. As compared to [14] we study document level semantic context derived from different prominent representations including (a) LSA, LDA (b) *Continuous Bag-Of-Words* (CBOW) and *Skip-gram* distributed representations of words learned by neural networks [17] and (c) *GloVe* [18].

The paper is organized as follows: Section 2 describes our approach to represent and recover OOV PNs with document context. Section 3 presents the different document representations. Sections 4 & 5 present the experiment setup and results.

2. METHODOLOGY

Our main objective is to identify a list of OOV PNs relevant to the given audio document. Relevant OOV PNs are retrieved based on semantic context. Then (*to assess the retrieved list of OOV PNs*) we try to recover the target OOV PNs using the retrieved list. For this a phonetic search is performed on the LVCSR hypothesis with each of the retrieved OOV PNs.

- **Retrieval of relevant OOV PNs**: To retrieve OOV PNs relevant to an audio document we rely on its semantic context. Several methods have been proposed in literature to obtain word embeddings corresponding to semantic context of words. These embeddings are derived from word co-occurrence statistics [18] and hence are dependent on the frequency of occurrence of words in training data [15]. As a result, these methods do not learn good representations for less frequent words (and OOV PNs) or mostly discard them in training. In our task, we therefore use *the semantic representation of a document as the semantic context of the OOV PN* occurring in this document. During training, diachronic text news are collected from the internet and indexed with new (i.e., OOV) PNs. These set of text documents indexed with OOV PNs is referred as *diachronic corpus*. A semantic vector representation is learned for each of the documents in the diachronic corpus and stored as a semantic context vector for the OOV PNs in that document. OOV PNs occurring in more than one diachronic document will have multiple context vectors or in other words document specific context vectors. Multiple OOV PNs in a diachronic document will share a common context vector. During test the semantic vector representation of the LVCSR hypothesis of the audio document (H) is compared with the context vectors (C_i) for each of the OOV PNs. For retrieval, relevance score $s \approx \max_i\{CosineSimilarity(H, C_i)\}$ is used. This approach applies to different document representations.

- **Phonetic Search for Target OOV PNs**: To recover the target OOV PNs a phonetic search is performed on the LVCSR hypothesis, for each of the retrieved OOV PNs. Phonetic form of the LVCSR word hypothesis is obtained and the retrieved OOV PNs are converted into their phonetic forms using a *Grapheme to Phoneme* (G2P) converter. We employ a search based on the classical *k-differences approximate string matching* algorithm [19]. In our case the algorithm makes a decision based on the phonetic string match score. It should be noted that it is not required to search the entire LVCSR hypothesis. The error and OOV regions in the LVCSR hypothesis can be hypothesised (as in [20]) and only these regions can be searched. Further, the phonetic search can be improved using techniques like searching N-best results or phone lattice, using phone confusion matrix, etc. (which is not the main focus of this paper).

3. DOCUMENT LEVEL CONTEXT VECTORS

In this section we briefly present the different document level representations used to model context of OOV PNs. Random Projection is our non-semantic baseline representation, whereas LSA, LDA, CBOW, Skip-gram and GloVe are the semantic counterparts used to represent context of OOV PNs.

- **Random Projections of TF-IDF vector**: It is classical to represent text documents as vector of Term Frequency-Inverse Document Frequency (TF-IDF) values of the words in the vocabulary. It has been shown that Random Projections can efficiently reduce the dimensionality of TF-IDF vectors while still preserving their original similarities and distances [21]. With random projection, the N-dimensional TF-IDF vectors of D documents in a corpus ($X_{D\times N}$) are projected to a K-dimensional ($K << N$) subspace as: $X^{RP}_{K\times D} = R_{K\times D}X_{D\times N}$, where $R_{K\times D}$ is a random matrix with random unit vectors.

- **LSA and LDA**: LSA [22] and LDA [23] have been prominent unsupervised methods to obtain semantic/topic space representations from documents. In LSA TF-IDF matrix of documents are projected into a (K dimensional) semantic space by performing *Singular Value Decomposition* (SVD). LDA is a generative probabilistic model which learns a (K) topic space in which documents are expressed as a mixture of topics and each topic is a distribution over words.

- **CBOW and Skip-gram**: Distributed approaches based on neural networks are being widely adopted for learning word embeddings in semantic space. The work of Mikolov et al. [17] has become popular due to its ability to handle large amounts of unstructured data with reduced computational costs and perform with high accuracy. It proposed two models: (a) the Skip-gram model, with an objective function to maximize the likelihood of prediction of contextual words given the centre word; (b) the CBOW model which predicts the centre word given the surrounding words. In our work we use these word embeddings to represent documents. During training the CBOW/Skip-gram word embeddings are learned for all the words in the diachronic corpus. Given these word embeddings and their linearity property, we obtain a representation for a document by taking an average over all the vocabulary words in the document. This document representation is referred to as AverageVec. (Paragraph Vector, a distributed model to represent sentences, paragraphs and documents has been proposed [24]. But it gives a poor performance in our experiments.)

- **GloVe**: LSA/LDA derive semantic spaces by performing factorization of document level co-occurrence statistics. Whereas CBOW and Skip-gram scan context windows across the corpus and produce linear directions of meaning. GloVe (Global Vectors) model was proposed [18] to produce linear dimensions of meaning and capture global corpus statistics. The improved performance of GloVe reported in [18] motivates us to use GloVe. As in case of CBOW/Skip-gram we obtain a document representation from GloVe by averaging the words in the document.

4. DIACHRONIC BROADCAST NEWS DATASETS

We present two realistic diachronic news datasets which are our training and test sets. These datasets, described in

Table 1, also highlight the motivation for handling OOV PNs. The L'Express dataset is collected from the website (lexpress.fr) of the French newspaper *L'Express* whereas the Euronews dataset is collected from the French website (fr.euronews.com) of the *Euronews* TV channel. L'Express dataset contains text news whereas Euronews has news videos and their text transcriptions. TreeTagger [25] is used to automatically tag PNs in the text. Words and PNs which occur in the lexicon of our *Automatic News Transcription System* (ANTS) [26] are tagged as In-Vocabulary (IV) and remaining PNs are tagged as OOV. ANTS lexicon is based on French newspaper (*LeMonde*) news articles until 2008 and contains 122K unique words. As shown, 64% of OOV words in Euronews videos are PNs and 47% videos contain OOV PNs.

Table 1. Broadcast news diachronic datasets.

	L'Express	Euronews
Type of Documents	Text	Video
Time Period	Jan 2014 - Jun 2014	
Number of Documents*	45K	3K
Vocabulary Size (unigrams)	150K	18K
Corpus Size (total word count)	24M	600K
Number of PN unigrams+	40K	2.2K
Total PN count	1.3M	19K
Documents with OOV	43K	2172
Number of OOV unigrams+	55K	1588
Total OOV count	450K	7710
Documents with OOV PN	36K	1415
Number of OOV PN unigrams+	17K	1024
Total OOV PN count	200K	3128

*K =10^3 and M =10^6; +unigrams occurring once excluded

5. EXPERIMENTS AND RESULTS

In our experiments the L'Express dataset is used as diachronic corpus. Audio from the Euronews video dataset is used as the test set. The ANTS [26] LVCSR system is used to perform automatic segmentation and speech-to-text transcription of the test audio news. The automatic transcriptions of the test audio news obtained by ANTS have an average *Word Error Rate* (WER) of 40% as compared to the reference transcriptions.

Diachronic corpus vocabulary is lemmatized and filtered by removing PNs occurring only once, non PN words less than 4 times, and using a stop-list of common and non-content French words. Moreover, only words tagged as PN, noun, adjective, verb and acronym are retained. The filtered vocabulary has 40K PNs and 28K words. Out of the 40K PNs 17K are not present in the ANTS LVCSR lexicon and are tagged as OOV PNs. Four OOV PNs, present in test set and not present in diachronic corpus, are excluded from experiments. Context vectors discussed in Section 3 are trained with this filtered vocabulary. We tried different numbers of dimensions/topics (in range 20-1K), the best performance is obtained for 300. LDA hyper-parameters and window size for CBOW, Skip-gram and GloVe are chosen for best performance.

5.1. Performance of OOV PN Retrieval

OOV PN retrieval using context vectors, as discussed in Section 2, is evaluated on the 1415 test audio news (in Euronews dataset) which contains OOV PNs. As shown in Table 1, these 1415 documents consist of 1024 unique OOV PN unigrams occurring a total of 3128 times. However, the total number of OOV PNs to be retrieved, obtained by counting unique OOV PNs per document, is 2300. Out of the 2300, 476 (20%) occur 5 times or fewer in the diachronic corpus.

5.1.1. Retrieval using Word v/s Document Semantic Context

We compare the performance of retrieval of OOV PNs when using document level context vectors, as discussed in Section 2, to that when using OOV PN (word) context vectors, as proposed in earlier works [14, 15]. To study the difference in the performance we plot the distribution of the ranks obtained by the target OOV PNs versus the frequency of occurrence of the target OOV PNs in the diachronic corpus used for training.

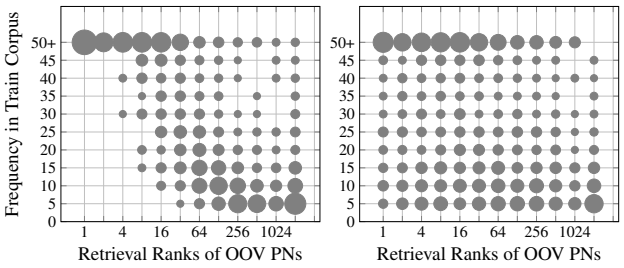

Fig. 1. Rank-Frequency Distribution for Retrieval with LDA based Word (left) and Document (right) Context Vectors

Figure 1 shows the rank-frequency distribution when using word (left) and document (right) context vectors from LDA. (This comparison is shown only for LDA but holds true for other representations discussed in this paper.) Word level semantic context vectors perform better for frequent OOV PNs whereas document level semantic context vectors are more uniform across different OOV PNs and hence better for retrieval of less frequent OOV PNs. The improvement in retrieval performance can be measured in terms of Recall [27] at *5% operating point*, an operating point chosen for analysis and to restrict phonetic search. Using document context vectors over word context vectors gives a 5% absolute improvement in recall with LDA based representation. This improvement is 12% with Skip-gram. This justifies the use of document level semantic context vectors, along with the fact that this approach can be used for different representations.

5.1.2. Performance of Different Semantic Representations

Figure 2 shows the OOV PN retrieval performance in terms of *Recall* and *Mean Average Precision* (MAP) [27] on the Reference and LVCSR transcriptions of the test set. In the graphs

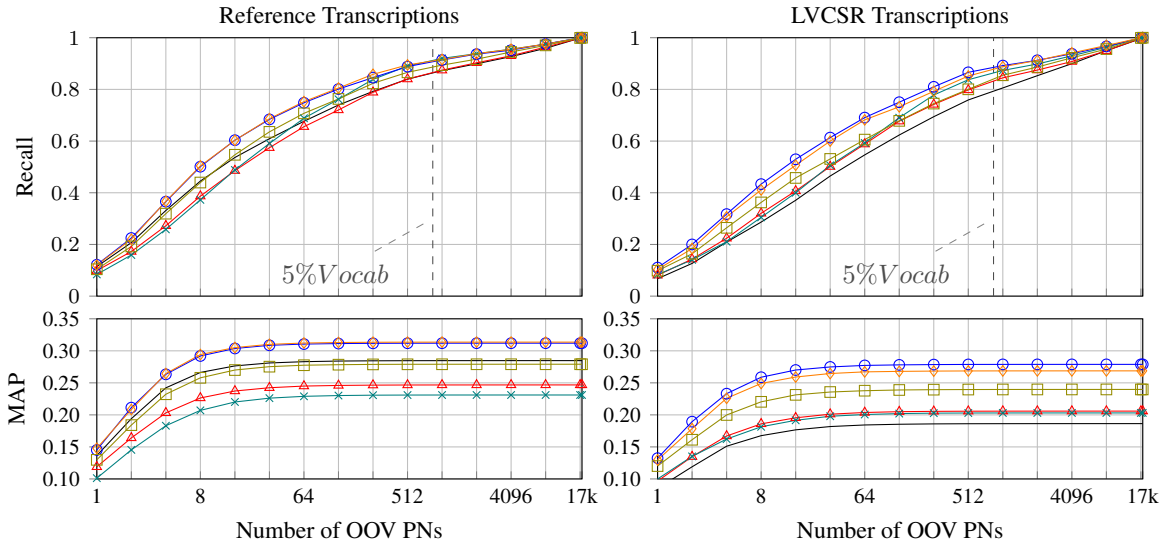

Fig. 2. OOV PN retrieval with different representations (—— RP, —△— LSA, —×— LDA, —○— CBOW, —◇— Sg, —□— GloVe).

in Figure 2, the X-axis represents the number of top-N OOV PNs selected from the diachronic corpus. The Y-axis represents recall and MAP of the target OOV PNs. The different context representations are denoted by their abbreviations.

The best recall and MAP performance on LVCSR transcriptions is obtained with the CBOW representation and the worst with Random Projections (RP) which is a non-semantic representation. It must be noted that the performance with RP is not the worst for reference transcriptions of the videos. The speech recognition errors introduced with the LVCSR degrade the document representation. But with semantic representation the degradation is lesser. The performance trend for LVCSR is: CBOW/Sg $>$ GloVe $>$ LDA/LSA $>$ RP. As opposed to the results on the analogy task in [18], GloVe does not outperform CBOW/Sg representations in our task. Furthermore, the Recall graph shows that document context based retrieval of OOV PNs can reduce the search list for target OOV PN drastically. The context vectors of the LVCSR hypothesis can recover up to 79-87% of the target OOV PNs within top 5% of retrieval results. Thus reducing the phonetic search to only 5% i.e. 850 OOV PNs from diachronic corpus.

5.2. Target OOV PN Recovery Performance

Target OOV PNs are searched in the LVCSR hypothesis as discussed in Section 2. The phonetic string corresponding to the LVCSR hypothesis is obtained using forced alignment and the OOV PNs are converted to phone strings with our G2P converter [28]. As mentioned, k-differences approximate matching algorithm [19] is used for performing phonetic search. The search was performed only in the error regions of the LVCSR hypothesis (obtained by alignment with manual transcriptions). Figure 3 shows the F_1-*scores* for recovery of the target OOV PNs. The best F_1-score is shown for different

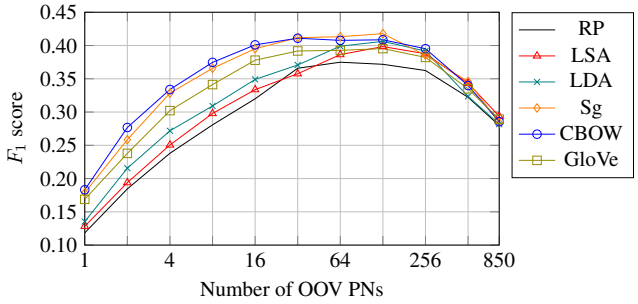

Fig. 3. OOV PN recovery with different representations.

number of OOV PNs retrieved by the context model.

The trends of Recall and MAP of Figure 2 reflect in the F_1-scores of Figure 3. CBOW and Sg context vector based retrieval results give better F_1-scores than GloVe, which is better than LDA/LSA. RP gives least F_1-scores. However, the F_1-score stops improving beyond top 64 retrieved OOV PNs due to the increase in false positives. The best F_1-score 0.41 is obtained with Skip-gram based document context vector.

6. CONCLUSION

We proposed an new approach to represent OOV PNs and their context with document level semantic vector representations for retrieval of OOV PNs relevant to an audio document. This approach can handle less frequent OOV PNs in the training data and gives better retrieval performance compared to word level semantic vector representations. Among semantic representations, AverageVec obtained from CBOW and Skip-gram neural network word embeddings perform better than the classical LDA/LSA topic/semantic spaces. Recovery of the target OOV PNs using a phonetic search gives an F_1-score upto 0.41, thus confirming that document level semantic context is reliable for recovery of OOV PNs in diachronic audio.

7. REFERENCES

[1] C. Alberti *et al.*, "An audio indexing system for election video material," in *IEEE ICASSP*, 2009, pp. 4873–4876.

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

[3] D. Palmer and M. Ostendorf, "Improving out-of-vocabulary name resolution," *Computer Speech & Language*, vol. 19, pp. 107 – 128, 2005.

[4] C. Parada, M. Dredze, and F. Jelinek, "OOV sensitive named-entity recognition in speech," in *INTERSPEECH*, 2011, pp. 2085–2088.

[5] A. Allauzen and J.-L. Gauvain, "Open vocabulary ASR for audiovisual document indexation," in *IEEE ICASSP*, 2005, pp. 1013–1016.

[6] F. Béchet, A. Nasr, and F. Genet, "Tagging unknown proper names using decision trees," in *38th Annual Meeting of ACL*, PA, USA, 2000, pp. 77–84.

[7] D. Fohr and I. Illina, "Continuous word representation using neural networks for proper name retrieval from diachronic documents," in *INTERSPEECH*, 2015.

[8] Y.-C. Pan, Y.-Y. Liu, and L.-S. Lee, "Named entity recognition from spoken documents using global evidences and external knowledge sources with applications on mandarin chinese," in *IEEE Workshop ASRU*, 2005, pp. 296–301.

[9] S. Oger, G. Linarès, F. Béchet, and P. Nocera, "On-demand new word learning using world wide web," in *IEEE ICASSP*, 2008, pp. 4305–4308.

[10] C. Parada, A. Sethy, M. Dredze, and F. Jelinek, "A spoken term detection framework for recovering out-of-vocabulary words using the web," in *INTERSPEECH*, 2010, pp. 1269–1272.

[11] G. Chen, S. Khudanpur, D. Povey, J. Trmal, D. Yarowsky, and O. Yilmaz, "Quantifying the value of pronunciation lexicons for keyword search in lowresource languages," in *IEEE ICASSP*, May 2013, pp. 8560–8564.

[12] G. Senay, B. Bigot, R. Dufour, G. Linarès, and C. Fredouille, "Person name spotting by combining acoustic matching and LDA topic models," in *INTERSPEECH*, 2013, pp. 1584–1588.

[13] B. Bigot, G. Senay, G. Linarès, C. Fredouille, and R. Dufour, "Person name recognition in ASR outputs using continuous context models," in *IEEE ICASSP*, 2013, pp. 8470–8474.

[14] I. Sheikh, I. Illina, and D. Fohr, "OOV proper name retrieval using topic and lexical context models," in *IEEE ICASSP*, 2015.

[15] ——, "Study of entity-topic models for OOV proper name retrieval," in *INTERSPEECH*, 2015.

[16] J. Wintrode and S. Khudanpur, "Combining local and broad topic context to improve term detection," in *IEEE SLT Workshop*, Dec 2014, pp. 442–447.

[17] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *Proceedings of Workshop ICLR*, 2013.

[18] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proceedings of 2014 EMNLP Conference*, 2014, pp. 1532–1543.

[19] G. Navarro, "A guided tour to approximate string matching," *ACM Comput. Surv.*, vol. 33, pp. 31–88, Mar. 2001.

[20] B. Lecouteux, G. Linars, and B. Favre, "Combined low level and high level features for out-of-vocabulary word detection." in *INTERSPEECH*, 2009, pp. 1187–1190.

[21] E. Bingham and H. Mannila, "Random projection in dimensionality reduction: Applications to image and text data," in *Proceedings of the Seventh ACM SIGKDD*, New York, NY, USA, 2001, pp. 245–250.

[22] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *J. Assoc. Inf. Sci. Technol.*, vol. 41, no. 6, pp. 391–407, 1990.

[23] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.

[24] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *Proceedings of the 31st ICML*, 2014, pp. 1188–1196.

[25] H. Schmid, "TreeTagger: A part-of-speech tagger and lemmatizer for several languages." 2014. [Online]. Available: http://hdl.handle.net/11372/LRT-323

[26] I. Illina, D. Fohr, O. Mella, and C. Cerisara, "Automatic News Transcription System: ANTS some Real Time experiments," in *INTERSPEECH*, 2004, pp. 377–380.

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

[28] I. Illina, D. Fohr, and D. Jouvet, "Multiple Pronunciation Generation using Grapheme-to-Phoneme Conversion based on Conditional Random Fields," in *SPECOM'2011*, Kazan, Russia, Sep. 2011.