OOV PROPER NAME RETRIEVAL USING TOPIC AND LEXICAL CONTEXT MODELS

Imran Sheikh[1], *Irina Illina*[1], *Dominique Fohr*[1], *Georges Linarès*[2]

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

ABSTRACT

Retrieving *Proper Names* (PNs) specific to an audio document can be useful for vocabulary selection and OOV recovery in speech recognition, as well as in keyword spotting and audio indexing tasks. We propose methods to infer and retrieve OOV PNs relevant to an audio news document by using probabilistic topic models trained over diachronic text news. LVCSR hypothesis on the audio news document is analysed for latent topics, which is then used to retrieve relevant OOV PNs. Using an LDA topic model we obtain *Recall* up to 0.87 and *Mean Average Precision* (MAP) of 0.26 with only top 10% of the retrieved OOV PNs. We further propose methods to re-score and retrieve rare OOV PNs, and a lexical context model to improve the target OOV PN rankings assigned by the topic model, which may be biased due to prominence of certain news events. Re-scoring rare OOV PNs improves Recall whereas the lexical context model improves MAP.

Index Terms— OOV, proper names, speech recognition

1. INTRODUCTION

A general observation has been that the majority of *Out-Of-Vocabulary* (OOV) words in audio news are *Proper Names* (PNs); 56-72% as studied independently in [1–5]. On the other hand, PNs in audio news are of prime importance to improve speech recognition and content based indexing. In this paper we focus on retrieving the most relevant OOV PNs for an audio news document by using probabilistic topic models trained over diachronic text news. *In-Vocabulary* (IV) words hypothesised by *Large Vocabulary Continuous Speech Recognition* (LVCSR) are analysed for latent topic and lexical context, which then helps to retrieve relevant OOV PNs. The list of retrieved OOV PNs can now be used to recover target OOV PNs using phone matching [6], or additional speech recognition pass [7]; or spotting PNs in speech [8].

We propose to use generative probabilistic models to learn relations between IV words, latent topics and PNs. The main contributions being (a) different methods for retrieval of OOV PNs using a topic model; (b) techniques to handle OOV PNs that have appeared only a few times in diachronic text news; (c) a lexical context model, which improves ranks of target OOV PNs retrieved based on topic. The rest of the paper is organised as follows. In Section 2 we discuss related works and the novelty of our approach. Section 3 and 4 present proposed techniques for retrieval of OOV PNs using a topic model and our lexical context model. Section 5 reports experiments, followed by a conclusion in Section 6.

2. RELATED WORK

OOV word recovery and vocabulary selection have been the interest of researchers for some time. OOV word recovery techniques have used LVCSR hypothesis to query search engines on the *World Wide Web* (WWW) [6–8]. From the retrieved documents, target OOV candidates are chosen using phone sequences observed in the pre-identified OOV region [6,8] or using words adjacent to the OOV region [7]. Vocabulary selection techniques use TF-IDF measures [9], frequency & recency of new words [10] or selection of all new PNs [11]. Adaptation of vocabulary to a new corpus, using linear combination of existing corpora have also been proposed [12–14].

We propose retrieval of OOV PNs for an audio news document with probabilistic models. We use *Latent Dirichlet Allocation* (LDA) [15] to model PN-topic relations and propose a new model to capture lexical context of OOV PNs. Previously, PNs have been modelled with LDA [16], and a similar approach [17] based on vector space representation similar to *Latent Semantic Analysis* (LSA) [18] has been tried. However these approaches estimate one LDA/LSA context model for each PN which restricts them to only frequent PNs. In our approach, we estimate a global topic model and present techniques to handle infrequent OOV PNs. Lexical context based on word co-occurrences [19] and mutual information [20] has been proposed previously to improve LVCSR. Compared to these works, our lexical context model relaxes the fixed window length constraint and follows a generative process like LDA.

3. OOV PN RETRIEVAL USING TOPIC MODELS

Our goal is to retrieve the most relevant OOV PNs for an audio news document (referred to as *test document*). To achieve this we rely on collection of diachronic text news; referred to as *diachronic corpus*. To learn relations between words, latent topics and OOV PNs, topic models are trained using the diachronic corpus as *training corpus*. Given any test document, IV words are hypothesised by LVCSR and the topic models are used to infer and retrieve the most relevant OOV PNs.

3.1. Topic Models

Latent Semantic Analysis (LSA) [18], Probabilistic LSA [21] and Latent Dirichlet Allocation (LDA) [15] have been the most prominent methods for extracting topics and underlying semantic structure from a collection of documents. While LSA derives semantic spaces from word co-occurrence matrices and operates using a spatial representation, PLSA and LDA derive topics using hierarchical Bayesian analysis. We have chosen LDA to capture PN-Topic relations as it is a well defined generative model and shown to outperform PLSA and LSA for document classification [15] and word prediction [22] tasks.

LDA is used to model topics in the diachronic corpus. For the diachronic corpus of D text documents, a topic vocabulary of size N_v, the number of topics T and Dirichlet priors are chosen. Topic model parameters θ and ϕ are estimated by Gibbs sampling [23]. $\theta = [\theta_{dt}]_{D \times T}$ is the topic distribution for each document d, and $\phi = [\phi_{vt}]_{N_v \times T}$ is the topic distribution to word v from the diachronic corpus, both across T topics.

3.2. Methods to Retrieve OOV PNs

In this section, we present different methodologies to retrieve OOV PNs for a test document by using topic models. It should be noted that these techniques can be applied to any probabilistic topic model, although we have chosen LDA. Let us denote LVCSR hypothesis of test document by h and OOV PNs in the diachronic corpus by $\tilde{v}_x$ (throughout this paper, $\tilde{}$ is used for variables associated to PNs and suffix x for OOV). In order to retrieve OOV PNs, we calculate $p(\tilde{v}_x|h)$, for each $\tilde{v}_x$ and then treat it as a score to rank OOV PNs relevant to h.

3.2.1. Method I: Using Test Document Topic Distribution

This method relies on the topic mixture inferred for the test document. Given the words observed in h, the latent topic mixture $[p(t|h)]_T$ can be inferred by re-sampling from the word-topic distribution ϕ learned during training. Likelihood of an OOV PN ($\tilde{v}_x$) can be calculated as:

$$p(\tilde{v}_x|h) = \sum_{t=1}^{T} p(\tilde{v}_x|t)\; p(t|h) \tag{1}$$

3.2.2. Method II: Using Document Similarity

This new method relies on topic similarity between h and each text document $d^{'}$ in diachronic corpus which consists the OOV PN $\tilde{v}_x$. Topic mixture $[p(t|d^{'})]_T$ for each $d^{'}$ is available from $\theta_{d'}$ estimated during training. Topic mixture $[p(t|h)]_T$ for h is inferred and likelihood of $\tilde{v}_x$ is calculated as:

$$\begin{aligned} p(\tilde{v}_x|h) &\approx \max_{d'}\{CosSim(h, d^{'})\} \\ &= \max_{d'}\left\{\frac{\sum_{t=1}^{T}\; p(t|h)\; p(t|d^{'})}{\sqrt{\sum_{t=1}^{T} p(t|h)^2}\,\sqrt{\sum_{t=1}^{T} p(t|d^{'})^2}}\right\} \end{aligned} \tag{2}$$

where $CosSim(h, d^{'})$ is the cosine similarity between test and diachronic document in topic space (for our task, cosine similarity gives better performance than KL divergence and Hellinger distance measures which quantify similarity between probability distributions). The main idea behind this technique is to associate each OOV PN with several topic distributions, each of which is derived from the documents in the diachronic corpus in which the OOV PN was observed. While this technique gives the best retrieval ranks to OOV PNs, it requires iterating through the diachronic corpus.

3.2.3. Method III: Using PN-Word Associations

This proposed method relies on topic based associations between words in test document (h) and OOV PNs. It does not require explicit inference of topic mixture of h, and simply takes the word-topic distributions from ϕ. Although this simplification skips the hierarchical generative process of LDA, it exploits the separation of words in topic space. This method can work with online LVCSR decoding. Denoting words in h by $\{w_i\}_{i=1}^{N_{vh}}$, where N_{vh} is number of words in h,

$$\begin{aligned} p(\tilde{v}_x|h) = p(\tilde{v}_x|\{w_i\}_{i=1}^{N_{vh}}) &\approx \prod_{i=1}^{N_{vh}} p(\tilde{v}_x|w_i) \\ &= \prod_{i=1}^{N_{vh}} \sum_{t=1}^{T}\; p(\tilde{v}_x|t)\; p(t|w_i) \end{aligned} \tag{3}$$

3.3. Retrieving Rare/Less-Frequent OOV PNs

Methods I and *III* rely on $p(\tilde{v}_x|t)$ which is biased against OOV PNs observed only few times in the diachronic corpus. As a result, these methods give lower retrieval ranks to the rare OOV PNs. Hence, *only for* rare OOV PNs, we update Equations (1) and (3) to include $p(\cdot|t)$ and $p(t|\cdot)$ normalisation, and factors C_h^1 and C_h^3 to scale the scores to that of frequent OOV PNs obtained using Equations (1) and (3).

$$p(\tilde{v}_x|h) \approx \frac{C_h^1 \quad \sum_{t=1}^{T} p(\tilde{v}_x|t)\; p(t|h)}{\sqrt{\sum_{t=1}^{T} p(\tilde{v}_x|t)^2}\,\sqrt{\sum_{t=1}^{T} p(t|h)^2}} \tag{4}$$

$$p(\tilde{v}_x|h) \approx \prod_{i=1}^{N_{vh}} \frac{C_h^3 \quad \sum_{t=1}^{T} p(\tilde{v}_x|t)\; p(t|w_i)}{\sqrt{\sum_{t=1}^{T} p(\tilde{v}_x|t)^2}\,\sqrt{\sum_{t=1}^{T} p(t|w_i)^2}} \tag{5}$$

4. RE-RANKING WITH LEXICAL CONTEXT

A problem with ranking PNs using topic models is that for a test document h if topic t is prominent (i.e. $p(t|h)$ is high) then all the PNs which have high $p(\tilde{v}_x|t)$ take higher ranks. For instance, the diachronic news corpus from the period of the 2014 Football World Cup leads to sports topic dominated by football. As a result, the topic models tend to give higher scores to football PNs for documents on related sports topic. Increasing the number and granularity of topics is not a feasible solution when diachronic corpus is not large enough. Thus topics alone may not be discriminant enough for ranking OOV PNs. This problem can affect rankings of both rare and frequent OOV PNs. To address this we propose a lexical context model to re-rank OOV PNs. It models lexical distribution of PN-word co-occurrences within and across documents.

4.1. Lexical Context Model

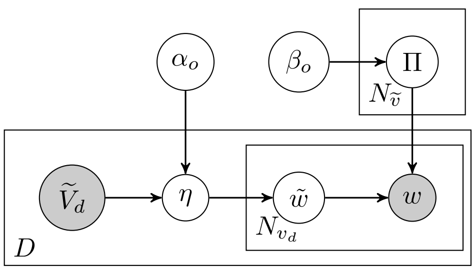


Fig. 1. Proper name lexical context model.

Figure 1 shows the graphical representation of our proposed PN lexical context model. It shares similarity with the smoothed LDA model [15]. For each word in a diachronic document d, a PN $\tilde{w}_i$ is first sampled from document specific PN distribution η, i.e. $\tilde{w}_i \sim Multi(\eta_d)$. Then the corresponding word w_i is sampled from the PN specific word distribution Π, i.e. $w_i \sim Multi(\Pi_{\tilde{w}_i})$. $\widetilde{V}_d$ represents PNs specific to d and α_o, β_o are Dirichlet priors to η, Π respectively. $N_{\tilde{v}}$ is total number of PNs and N_{v_d} is document specific word count. For each document, words are assigned to a PN based on the learnt PN specific word distribution Π as well as the document specific PN distribution η. We use Gibbs sampling to estimate the PN assignments; the sampling equation being:

$$p(\tilde{w}_i = \tilde{v}|w_i = v, \tilde{w}_{-i}, w_{-i}, \widetilde{V}_d) \propto \frac{C^{V\tilde{V}}_{v\tilde{v}} + \beta_o}{\sum_k C^{V\tilde{V}}_{v_k\tilde{v}} + \beta_o N_v} \frac{C^{D\tilde{V}}_{d\tilde{v}} + \alpha_o}{\sum_m C^{D\tilde{V}}_{d\tilde{v}_m} + \alpha_o N_{\tilde{v}}} \quad (6)$$

where $\tilde{w}_i = \tilde{v}$ implies that the i^{th} term in a document is assigned PN $\tilde{v}$. $C^{V\tilde{V}}_{v\tilde{v}}$ is the number of times word v is assigned to PN $\tilde{v}$ and $C^{D\tilde{V}}_{d\tilde{v}}$ is the number of words in d assigned to PN $\tilde{v}$. $\tilde{w}_{-i}$ and w_{-i} represent all PN and word assignments not including the i^{th} word. Equation (6) is derived by marginalising out Π and η, which can be later estimated from the first and second parts of RHS of Equation (6) respectively.

4.2. OOV PN Re-Ranking

Lexical context is used only to re-rank OOV PNs. The topic model is first used to choose top-*N* (topic relevant) OOV PNs, and then the lexical context model to re-rank OOV PNs in the top-*N* list. We use Gibbs sampling to infer the best OOV PN assignments to each word in h using a modified equation:

$$p(\tilde{w}_i = \tilde{v}_x|w_i = v, \tilde{w}_{-i}, w_{-i}, \widetilde{V}_d) \propto \frac{N^{V\tilde{V}}_{v\tilde{v}_x} + C^{V\tilde{V}}_{v\tilde{v}_x} + \beta_o}{\sum_k \left(N^{V\tilde{V}}_{v_k\tilde{v}_x} + C^{V\tilde{V}}_{v_k\tilde{v}_x}\right) + \beta_o N_v} p(\tilde{v}_x|h) \quad (7)$$

where $N^{V\tilde{V}}_{v\tilde{v}_x}$ is the number of times term w_i in h is word v and assigned to top-*N* OOV PN $\tilde{v}_x$; $C^{V\tilde{V}}_{v\tilde{v}}$ is the count saved from training. The top-*N* OOV PNs are then re-ranked using:

$$P_N(\tilde{v}_x|h) \approx p(\tilde{v}_x|h) + C^L_h \frac{\sum_k N^{V\tilde{V}}_{v_k\tilde{v}_x} + \alpha_o}{N_{vh} + \alpha_o N} \quad (8)$$

where, C^L_h is a scaling factor to combine topic and lexical model scores of top-*N* OOV PNs.

5. EXPERIMENTS AND RESULTS

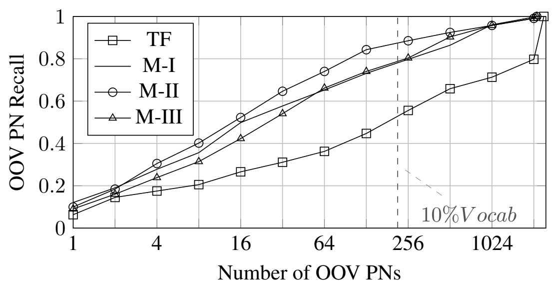
Our approach is evaluated on the *Euronews* corpus. This corpus consists of French news videos and articles collected from Euronews (http://fr.euronews.com). From this corpus, the text news corresponding to the period 01/01/2014 - 31/05/2014 is chosen as our diachronic corpus. (Automatically choosing a temporal subset for a diachronic corpus, or retrieving OOV PNs from a larger diachronic corpus using temporal context, is another interesting problem which can be addressed using topic models, but is beyond the scope of this paper.) Diachronic corpus vocabulary was filtered by removing PNs occurring only once, non PN words occurring less than 4 times, and using a stoplist of common French words and non content words which do not carry any topic-related information. The filtered vocabulary has 8155 PNs and 8732 words. The words and PNs which occur in the lexicon of our *Automatic News Transcription System* (ANTS) [24] are tagged as IV and the remaining (2418) PNs are tagged as OOV PNs. (ANTS lexicon is based on news articles up to 2008 from French newspaper *LeMonde*.) 64% of OOV words in Euronews corpus are PNs and these OOV PNs comprise 0.52% of the corpus.

Topics and lexical context are modelled with a *training set* of 3850 news articles from dates 01-29 of each month. Our *first test set* (TestSet-I) is 170 news articles from dates 30-31 of each month. Out of 170 articles, 120 have corresponding news videos. Our *second test set* (TestSet-II) is the automatic and manual transcriptions of these videos. For TestSet-I and for the manual transcriptions of TestSet-II, the OOV PNs are removed and the remaining test document is used as input. The total number of OOV PNs to be retrieved for TestSet-I, obtained by counting unique OOV PNs per document, is

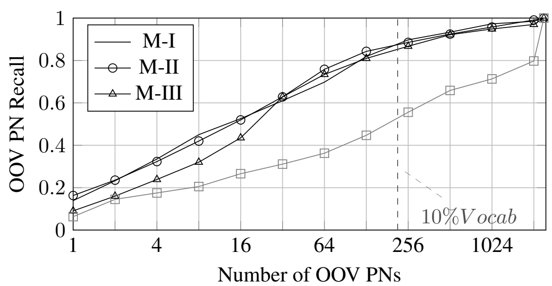
331. Out of 331, 126 (38%) of the OOV PNs have occurred 5 times or fewer in the training set. Similarly, the number of OOV PNs to be retrieved for TestSet-II is 220.

We tried different number of topics in our experiments, the best performance being obtained for 50 topics. Scaling factors C_h^1, C_h^3 and C_h^L are also set empirically.

5.1. Performance of OOV PN Retrieval



(a) Topic Models vs Term Frequency based.



(b) Topic Models + Lexical & Rare OOV PN Re-ranking.

Fig. 2. Recall for OOV PN Retrieval Methods.

Figure 2 shows the performance of the proposed methods for retrieval of OOV PNs on TestSet-I of text news. For each of the graphs in Figure 2, X-axis represents the number of OOV PNs selected from the diachronic corpus by various methods. Y-axis represents recall of the target OOV PNs. The methods discussed in Section 3.2 are denoted as M-I, M-II and M-III respectively. TF is the *Term Frequency* baseline, in which the top most frequent OOV PNs in diachronic corpus are selected. Figure 2 (a) shows OOV PN recall using the methods discussed in Section 3.2, whereas Figure 2 (b) shows recall after lexical context re-ranking (with N=100) and rare OOV PN re-ranking as discussed in Section 4 and Section 3.3 respectively. For vocabulary selection, the proposed methods clearly outperform Term Frequency based selection. OOV PN retrieval with proposed methods can recover about 78% of the target OOV PNs by adding only 10% of PNs from diachronic corpus. Proposed re-ranking methods further improve the recall curves by 8-10% (absolute) at top 10% OOV PN vocabulary (operating point chosen by us to compare further results).

Table 1 compares OOV PN retrieval performance of the methods on TestSet-I of text news articles in terms of Recall (R) and Mean Average Precision (MAP) [25] obtained with top 10% of the retrieved OOV PNs. The gain with different re-ranking methods is shown, where *None* denotes no re-ranking, *Lex* denotes lexical context re-ranking as discussed in Section 4, *Freq* denotes re-ranking for rare OOV PNs as discussed in Section 3.3 and *Lex+Freq* denotes a combination of the two. Rare OOV PN re-ranking is not required for M-II; and no re-ranking applied to TF. As shown for each of the methods, lexical context re-ranking helps to improve MAP and rare OOV PN re-ranking improves recall by 3-10% due to retrieval of rare OOV PNs. *Methods I* and *II*, which are based on topic inference, give better results compared to *Method III*, which can work with online LVCSR decoding.

Table 1. Recall (R) and Mean Average Precision (MAP) of OOV PN Retrieval Methods at top 10% OOV PNs.

Re-Ranking	TF		M-II	
	R	MAP	R	MAP
None	0.54	0.15	**0.87**	0.27
Lex	-	-	**0.87**	**0.31**

Re-Ranking	M-I		M-III	
	R	MAP	R	MAP
None	0.78	0.26	0.79	0.21
Lex	0.78	**0.31**	0.79	**0.22**
Freq	**0.88**	0.25	**0.82**	0.21
Lex+Freq	**0.88**	**0.30**	**0.82**	**0.22**

Table 2. OOV PN Recall on news videos (at 10% OOV PNs).

	Re-Ranking	M-I	M-II	M-III
Manual	*None*	0.79	0.87	0.82
	Lex+Freq	0.86	0.87	0.82
LVCSR	*None*	0.79	0.87	0.75
	Lex+Freq	0.86	0.87	0.76

Table 2 shows performance on TestSet-II of audio news. LVCSR transcripts of the audio news are obtained using ANTS [24], with 46% Word Error Rate (WER). Performance of M-I and M-II is not affected for LVCSR transcripts as they rely on the topic mixture inferred on the test document. Whereas M-III directly operates on LVCSR hypothesised words and so its recall is lower than manual transcripts.

6. CONCLUSION

We proposed new methods to retrieve OOV PNs relevant to an audio news document by using probabilistic topic models. These methods recover 75-87% of OOV PNs by adding only 10% of PNs from a diachronic corpus. *Methods I* and *II*, which are based on topic inference, give better results compared to *Method III* which directly relies on LVCSR hypothesised words, but can work with online LVCSR decoding. We addressed retrieval of rare OOV PNs, which further improves the recall. And our proposed lexical context model improves the mean average precision of OOV PN retrieval. For vocabulary selection, our proposed methods outperform Term Frequency based selection.

7. REFERENCES

[1] 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.

[2] D. Palmer and M. Ostendorf, “Improving out-of-vocabulary name resolution,” *Computer Speech & Language*, vol. 19, no. 1, pp. 107 – 128, 2005.

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

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

[5] F. Béchet, A. Nasr, and F. Genet, “Tagging unknown proper names using decision trees,” in *38th Annual Meeting on Association for Computational Linguistics*, PA, USA, 2000, pp. 77–84.

[6] 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.

[7] 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.

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

[9] 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 ICASSP*, 2010, pp. 5410–5413.

[10] N. Bertoldi and M. Federico, “Lexicon adaptation for broadcast news transcription,” in *ISCA Adaptation Methods for Speech Recognition*, 2001, pp. 187–190.

[11] C. Martins, A. Texeira, and J. Neto, “Dynamic vocabulary adaptation for a daily and real-time broadcast news transcription system,” in *Spoken Language Technology Workshop, IEEE*, 2006, pp. 146–149.

[12] A. Allauzen and J.-L. Gauvain, “Diachronic vocabulary adaptation for broadcast news transcription,” in *ISCA INTERSPEECH*, 2005, pp. 1305–1308.

[13] W. Wang, “Techniques for effective vocabulary selection,” in *EUROSPEECH*, 2003, pp. 245–248.

[14] 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.

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

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

[17] 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.

[18] 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.

[19] A. Sarma and D. D. Palmer, “Context-based speech recognition error detection and correction,” in *HLT-NAACL*, 2004, pp. 85–88.

[20] I. Illina, D. Fohr, and G. Linarès, “Proper Name Retrieval from Diachronic Documents for Automatic Speech transcription using Lexical and Temporal Context,” in *In ISCA/IEEE Workshop on Speech, Language and Audio in Multimedia (SLAM)*, 2014, pp. 29–33.

[21] T. Hofmann, “Probabilistic latent semantic analysis,” in *Uncertainty in Artificial Intelligence*, 1999, pp. 289–296.

[22] T. L. Griffiths, J. B. Tenenbaum, and M. Steyvers, “Topics in semantic representation,” *Psychological Review*, vol. 114, p. 2007, 2007.

[23] T. L. Griffiths and M. Steyvers, “Finding scientific topics,” *Proceedings of the National Academy of Sciences*, vol. 101, no. suppl 1, pp. 5228–5235, 2004.

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

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