How Diachronic Text Corpora Affect Context based Retrieval of OOV Proper Names for Audio News

Imran Sheikh, Irina Illina, Dominique Fohr
Université de Lorraine, LORIA, UMR 7503, Vandoeuvre-lès-Nancy, F-54506, France
Inria, Villers-lès-Nancy, F-54600, France
CNRS, LORIA, UMR 7503, Vandoeuvre-lès-Nancy, F-54506, France
{imran.sheikh, irina.illina, dominique.fohr}@loria.fr

Abstract

Out-Of-Vocabulary (OOV) words missed by Large Vocabulary Continuous Speech Recognition (LVCSR) systems can be recovered with the help of topic and semantic context of the OOV words captured from a diachronic text corpus. In this paper we investigate how the choice of documents for the diachronic text corpora affects the retrieval of OOV Proper Names (PNs) relevant to an audio document. We first present our diachronic French broadcast news datasets, which highlight the motivation of our study on OOV PNs. Then the effect of using diachronic text data from different sources and a different time span is analysed. With OOV PN retrieval experiments on French broadcast news videos, we conclude that a diachronic corpus with text from different sources leads to better retrieval performance than one relying on text from single source or from a longer time span.

Keywords: oov, proper names, lvcsr, diachronic corpus

1. Introduction

News content is diachronic in nature and is characterised by different topics which change with time, leading to a change in the linguistic content and vocabulary. As a result, a typical problem faced by *Large Vocabulary Continuous Speech Recognition* (LVCSR) systems processing diachronic audio news is *Out-Of-Vocabulary* (OOV) words. In previous works it has been observed that the majority of OOV words are *Proper Names* (PNs). PN percentage in OOV words been reported as: 56% by Qin (2013), 66% by Parada (2011), 57.6% by Palmer (2005), 70% by Allauzen (2005), 72% by Bechet (2000). On the other hand, PNs in audio news are of prime importance for content based indexing and browsing applications. Our work is closely related to recognition and recovery of OOV PNs i.e., PNs which appear in diachronic audio but are not present in the LVCSR vocabulary and cannot be recognised by the LVCSR system.

To recognise OOV PNs in a test audio document, we rely on new PNs extracted from collections of diachronic text news from the internet, referred to as *a diachronic corpus*. Instead of using all the new PNs to recover the *target OOV PNs*[1], it would be efficient to incorporate only those new PNs which are contextually related. We refer to this task as *retrieval of OOV PNs relevant to an audio document*. To achieve this, we use the topic and lexical context of the audio document and OOV PNs. In our previous works, we have presented different methods exploiting topic context to retrieve OOV PNs (Sheikh et al., 2015a) and we have further discussed topic model variations and methods to handle the less frequent OOV PNs (Sheikh et al., 2015b). The topic and lexical context of the OOV PNs are modelled and derived from the diachronic corpus and therefore it is important to study the selection of documents for the diachronic corpus. In this paper we try to investigate some characteristics of the diachronic corpus which can affect the performance of retrieval of OOV PNs. For instance we analyse the effect on the retrieval performance when: (a) the diachronic corpus is built with text from different originating sources (b) a diachronic corpus from a single source is supplemented with text resources for the less frequent OOV PNs (c) the diachronic corpus timeline extends beyond test dataset timeline. Following our earlier works we will focus on retrieval of OOV PNs in French broadcast news videos.

The rest of the paper is organised as follows. In Section 2. we briefly discuss the approach that we have adopted in this paper to retrieve OOV PNs relevant to an audio document. In Section 3. we present diachronic French broadcast news datasets used in our experiments and in Section 4. we present the different configurations used in our study. Section 5. presents the experiment setup, followed by a discussion in Section 6.

2. OOV PN Retrieval using Topic Models

As mentioned earlier, our task is to retrieve OOV PNs relevant to an audio document. To achieve this, we rely on new PNs extracted from collections of diachronic text news from the internet, which we refer as a diachronic corpus. Topic models are trained using the diachronic text corpus as a training corpus, in order to learn relations between words, latent topics and OOV PNs. During test, the audio news document is transcribed by an LVCSR system with a standard vocabulary. *In-Vocabulary* IV words (including IV PNs) hypothesised by the LVCSR are then input to the topic model to infer the topic distribution of the test audio document. From the test document topic distribution we infer a list of most relevant OOV PNs.

[1] Ideally new PNs extracted from collections of diachronic text news are OOV PNs with respect to the LVCSR vocabulary. However, all new PNs are not present in the test set audio documents. Hence we use the term *target OOV PNs* to refer to the OOV PNs actually present in the test set audio documents. The term 'OOV PNs' will be used to refer to the new PNs.

Table 1: Diachronic French broadcast news datasets

	L'Express	*Le Figaro*	*L'Express +*	*L'Express*	*Euronews*	*Euronews*
			Le Figaro			
	(LX)	(FIG)	(LX+FIG)	(LX-18m)	(Dev)	(Test)
Type of Documents	Text	Text	Text	Text	Text	Video
Time Period	Jan - Jun 2014	Jan - Jun 2014	Jan - Jun 2014	Jul 2013 - Dec 2014	Jan - Jun 2014	Jan - Jun 2014
Number of Documents[1]	45K	59K	104K	142K	3.1K	3K
Vocabulary Size (unigrams)[2]	150K	140K	180K	270K	42K	45K
Corpus Size (approx. word count)	24M	18M	42M	70M	550K	700K
Number of PN unigrams[2]	57K	51K	80K	104K	12K	11K
Total PN count	1.45M	1.3M	2.7M	4.2M	54K	42K
Number of OOV unigrams[3]	12.4K	11.9K	24.4K	37.1K	4.9K	4.3K
Documents with OOV[3]	32.3K	36.4K	73K	109K	2.25K	2.2K
Total OOV count[3]	141K	142K	320K	509K	9.1K	8K
Number of OOV PN unigrams[3]	9.3K	8.8K	18.4K	28.2K	3.4K	3.1K
Documents with OOV PN[3]	26.5K	30K	61.3K	93.5K	1.9K	1.9K
Total OOV PN count[3]	107K	103K	243K	388K	6.9K	6.2K

[1]K denotes *Thousand* and M denotes *Million*, [2]unigrams occurring less than two times are ignored
[3]unigrams occurring in less than three documents ignored, documents with more than 20 and less than 500 terms
Note: (a) OOV statistics are post filtering steps discussed in Section 5.1. (b) [2] and [3] do not apply to *Euronews*

Following our approach in (Sheikh et al., 2015a; Sheikh et al., 2015b), Latent Dirichlet Allocation (LDA) (Blei et al., 2003) is used to model topics in the diachronic corpus. For modelling the LDA topics, we fix a reduced vocabulary, the number of topics T and symmetric Dirichlet priors. Topic model parameters θ and ϕ are estimated using the Gibbs sampling algorithm (Griffiths and Steyvers, 2004). LDA captures the topical context of the OOV PNs and enables retrieval of OOV PNs relevant to a test document. Let us denote the LVCSR hypothesis of a test document by h and OOV PNs in diachronic corpus by $\tilde{v}_x$. 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. With 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. Given $p(\tilde{v}_x|t) = \phi_{vt}$, the likelihood of an OOV PN ($\tilde{v}_x$) can be calculated as: $p(\tilde{v}_x|h) = \sum_{t=1}^{T} p(\tilde{v}_x|t)\ p(t|h)$.

3. Diachronic Broadcast News Datasets

Table 1 presents realistic diachronic broadcast news datasets which will be used in our study. These datasets also highlight the motivation for handling OOV PNs. The datasets are collected from three sources: (a) French newspaper *L'Express* (http://www.lexpress.fr/) (b) French newspaper *Le Figaro* (http://www.lefigaro.fr/) (c) the French website of the *Euronews* (http://fr.euronews.com/) television channel. The *L'Express* and *Le Figaro* datasets contain text news whereas the Euronews dataset contains text news as well as news videos and their approximate text transcriptions. In our study the *L'Express* and *Le Figaro* datasets will be used as diachronic corpora to train topic models, in order to infer the OOV PNs relevant to Euronews (Test) videos. The *Euronews* text articles (marked as Dev) will be the development set.

TreeTagger[2] (Schmid, 1994) is used to automatically tag PNs in the text. The words and PNs which occur in the lexicon of our *Automatic News Transcription System* (ANTS) (Illina et al., 2004) are tagged as IV, and the remaining PNs are tagged as OOV. The ANTS lexicon is based on news articles appearing in French newspaper *LeMonde* until 2008 and contains about 123K unique words. As shown in Table 1, 72% (3.1K out of 4.3K) of OOV words in the *Euronews* video dataset are PNs and about 63% (1.9K out of 3K) of the videos contain OOV PNs. An important statistic *target OOV PN coverage* is not shown in Table 1. We use the term target OOV PN coverage to refer to the percentage of OOV PNs in *Euronews* videos which can be recovered with a given diachronic corpus. The target OOV PN coverage for each of the diachronic text corpus is as follows: 42% for LX, 40% for FIG, 52% for LX+FIG combined and 54% for LX-18m. We can observe that although LX-18m captures more new PNs (28.2K) as compared to those by LX+FIG (18.4K), the OOV PN coverage of the two differs only 2% absolute.

4. Configurations of the Training Diachronic Corpus

The topic context of OOV PNs, which enables retrieval of the relevant OOV PNs, is learned from a diachronic text corpus. We would like to study the effect of selection of documents for the training diachronic corpus. In particular we study the following configurations of the diachronic corpus.

(A) Documents containing OOV PNs[3] and from the same time period as the test set, e.g. *L'Express* documents

[2]http://www.cis.uni-muenchen.de/ schmid/tools/TreeTagger/

[3]including documents not containing OOV PN did not give significant improvement in the retrieval performance

containing OOV PNs and corresponding to 6 months of the *Euronews* video test set.

(B) Documents coming from two different originating sources, e.g. *L'Express* and *Le Figaro*.

(C) Documents from a time period extending beyond the timeline of the test set (e.g. *L'Express* documents from 18 months for the *Euronews* video test set).

(D) Documents with OOV PNs are collected from one source and then for the less frequent OOV PNs in this collection new documents are additionally collected from another source. As discussed in our previous works (Sheikh et al., 2015a; Sheikh et al., 2015b), retrieval of less frequent OOV PNs has a poor performance because there is not enough data to learn their topic distribution. This problem of reduced representation of less frequent OOV PNs motivates us to study this configuration.

5. Experiment Setup

5.1. Corpora Setup

The datasets presented in Table 1 will be used for our experiments. The *L'Express* and *Le Figaro* datasets will be used as a diachronic corpus to train the topic models. Audio news extracted from the *Euronews* video dataset will be the test set, whereas the *Euronews* text articles (marked as Dev) will be the development set. The different configurations discussed in Section 4. will be studied with the *L'Express* and *Le Figaro* datasets. Throughout the experiments and discussions, LX, FIG and LX+FIG denote diachronic corpora with documents from *L'Express*, *Le Figaro* (both from Jan 2014 - Jun 2014) and a combination of the two, respectively. These correspond to configurations A and B of Section 4. LX-18m denotes a diachronic corpus with documents of *L'Express* from Jul 2013 - Dec 2014 and corresponds to configuration C of Section 4. LX+rFIG denotes documents of LX (*L'Express*, Jan 2014 - Jun 2014) supplemented with documents from FIG (*Le Figaro*, Jan 2014 - Jun 2014) which contain OOV PNs occurring less than 10 times in LX. This corresponds to configuration D of Section 4. Target OOV PN coverage for LX+rFIG is 49%.

5.2. LVCSR Processing

The ANTS (Illina et al., 2004) 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[4] available from *Euronews*.

5.3. LDA Topic Models

For training topic models, diachronic corpus words are lemmatised and filtered by removing PNs and non PN words occurring less than 3 times. Additionally a stoplist of common words and non content words which do not carry any topic-related information is applied. Moreover, a POS based filter is employed to retain words tagged as PN, noun, adjective, verb and acronym. PNs not present in the ANTS LVCSR lexicon are tagged as OOV PNs. An LDA topic model is trained with the filtered corpus vocabulary.
We trained 100, 200, 300, 400 and 500 topics for each of the configurations and first evaluated the performance on our *Euronews* (text) development set. Figure 1 shows the effect of number of topics on the OOV PN retrieval performance. The comparison in Figure 1 is in terms of *Recall@100* and *MAP@100* (Manning et al., 2008), which represent the *Recall* and *Mean Average Precision* (MAP) values obtained by considering the top 100 retrieved OOV PNs. Higher number of topics are more favourable for larger a corpus and we can observe the same from Figure 1. LX-18m and LX+FIG continue to give better performance with increased number of topics but for LX and FIG the recall does not improve beyond 300 topics. The MAP however still improves for LX and FIG (we will discuss more about it in Section 6.1.). As each diachronic corpus configuration will perform its best for a certain number of topics, choosing a fixed number of topics for analysis of corpora of different sizes may not be appropriate. However, since our study is focused on effects of corpus selection on OOV PN retrieval, we will continue our analysis on the test set with each diachronic corpus modelled with 300 topics. We will refer back to Figure 1 to support relevant observations whenever required.

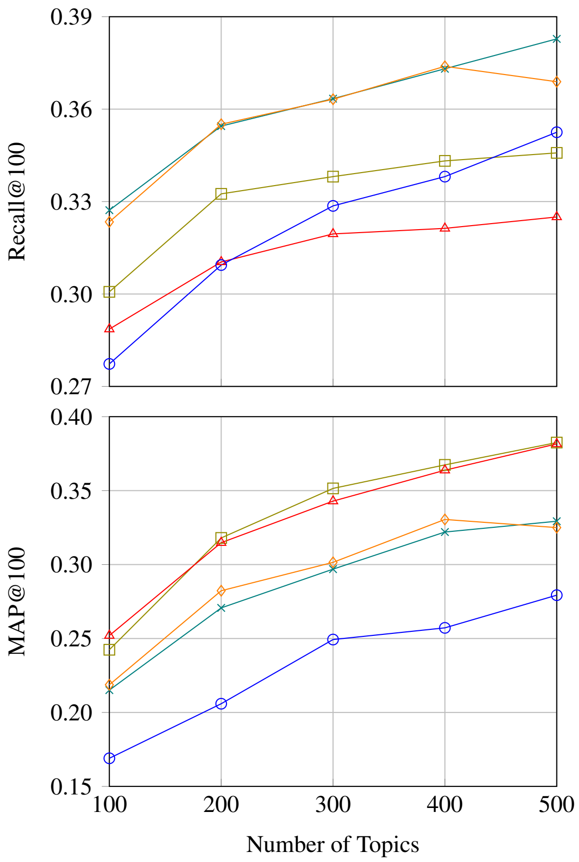

Figure 1: Recall@100 and MAP@100 for OOV PN retrieval on *Euronews* text (Dev set) with different number of LDA topics trained on the diachronic corpora. (—□— LX, —△— FIG, —×— LX+FIG, —◇— LX+rFIG, —○— LX-18m)

[4] these are approximate transcriptions; on a set of 10 manually transcribed audio files we obtained a WER of 33.8%

6. Discussion

We present Figure 2 to analyse the effects of diachronic corpus configurations A, B, C and D discussed in Section 4. Figure 2 shows a graph of recall and MAP of retrieval of OOV PNs. The performance on both reference (left) and LVCSR (right) transcriptions of the *Euronews* video test set are shown. The X-axes represent the number (N) of top-N retrieved OOV PNs. Y-axes represent recall (top) and MAP (bottom) of the target OOV PNs. For comparison, retrieval results obtained with 300 topics are shown.

6.1. What Recall and MAP indicate

It is necessary to understand the importance and differences for the recall and MAP curves. After retrieval of the relevant OOV PNs, the top-N relevant OOV PNs are to be used for recognition and recovery of the target OOV PNs. To recover the target OOV PNs one can use phone matching (Pan et al., 2005), or additional speech recognition pass (Oger et al., 2008); or spotting PNs in speech (Parada et al., 2010; Sheikh et al., 2015c). 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* (i.e. N in the top-N choice) will be the same whether the ranks of the retrieved OOV PNs are closer to 1 or to N. The MAP value is a direct function of the retrieval ranks. It can be seen that the MAP for LX is best but it does not have the best recall rates. LX has a smaller number of OOV PNs to choose from and it makes smaller retrieval errors but at the same time gives lesser coverage of the target OOV PNs. We can also observe that the recall values for the reference and LVCSR transcriptions appear close but the MAP values capture the differences in ranks. Similarly, in Figure 1 the recall values for LX and FIG do not improve beyond 300 topics but the MAP values increase due to the improvement in ranks of the target OOV PNs within the top 100 retrieved OOV PNs.

6.2. Effects of different Diachronic Corpora Configurations

By analysing the recall and MAP curves of Figure 2 we can draw the following conclusions.

- Expanding the time period of a diachronic corpus (as in LX-18m) gives better target OOV PN coverage, but not the best recall rates. Additionally it gives a low MAP. Such corpora can be possibly exploited by training larger number of topics or by employing better retrieval methods (Sheikh et al., 2015c; Sheikh et al., 2016).

- Adding new documents corresponding to less frequent OOV PNs is not effective: performance of LX+rFIG is similar to that of LX+FIG. We found that adding documents containing less frequent OOV PNs in LX+rFIG leads to inclusion of more than 60% of FIG documents. The additional data not only increases data for learning topic representation of less frequent OOV PNs but also comes with additional less frequent OOV PNs and more instances of frequent OOV PNs. Further analysis of the ranks of the less frequent OOV PNs obtained with LX, LX+FIG and LX+rFIG showed that the ranks with LX+rFIG are better with respect to LX but similar to that with LX+FIG.

- Using documents of the same time period and from multiple sources (LX+FIG) gives a good balance of recall, MAP and target OOV PN coverage.

7. Conclusion

Diachronic text corpora are essential for recovery of OOV words and proper names missed by LVCSR systems. We performed a study to analyse the selection of text documents in the diachronic corpus used for retrieval of OOV proper names. French broadcast news videos from a particular time period were used as the diachronic test set and text articles from different French news websites formed the diachronic text corpora. With OOV proper name retrieval based on topic context, we can conclude that (a) text from a longer time span can give increased coverage of OOV proper names (b) but a corpus with text from different sources leads to better retrieval performance than relying on text from a single source, even if it corresponds to a longer time span (c) and less frequent OOV proper names need improvement in retrieval methods (Sheikh et al., 2016) and not just additional training data.

Acknowledgment

This work is funded by the ContNomina project supported by the French National Research Agency (ANR) under the contract ANR-12-BS02-0009.

8. Bibliographical References

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

Béchet, F., Nasr, A., and Genet, F. (2000). Tagging unknown proper names using decision trees. In *38th Annual Meeting on Association for Computational Linguistics*, pages 77–84, PA, USA.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022.

Griffiths, T. L. and Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101:5228–5235.

Illina, I., Fohr, D., Mella, O., and Cerisara, C. (2004). The Automatic News Transcription System: ANTS some Real Time experiments. In *INTERSPEECH*, pages 377–380.

Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, UK.

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

Palmer, D. and Ostendorf, M. (2005). Improving out-of-vocabulary name resolution. *Computer Speech & Language*, 19:107 – 128.

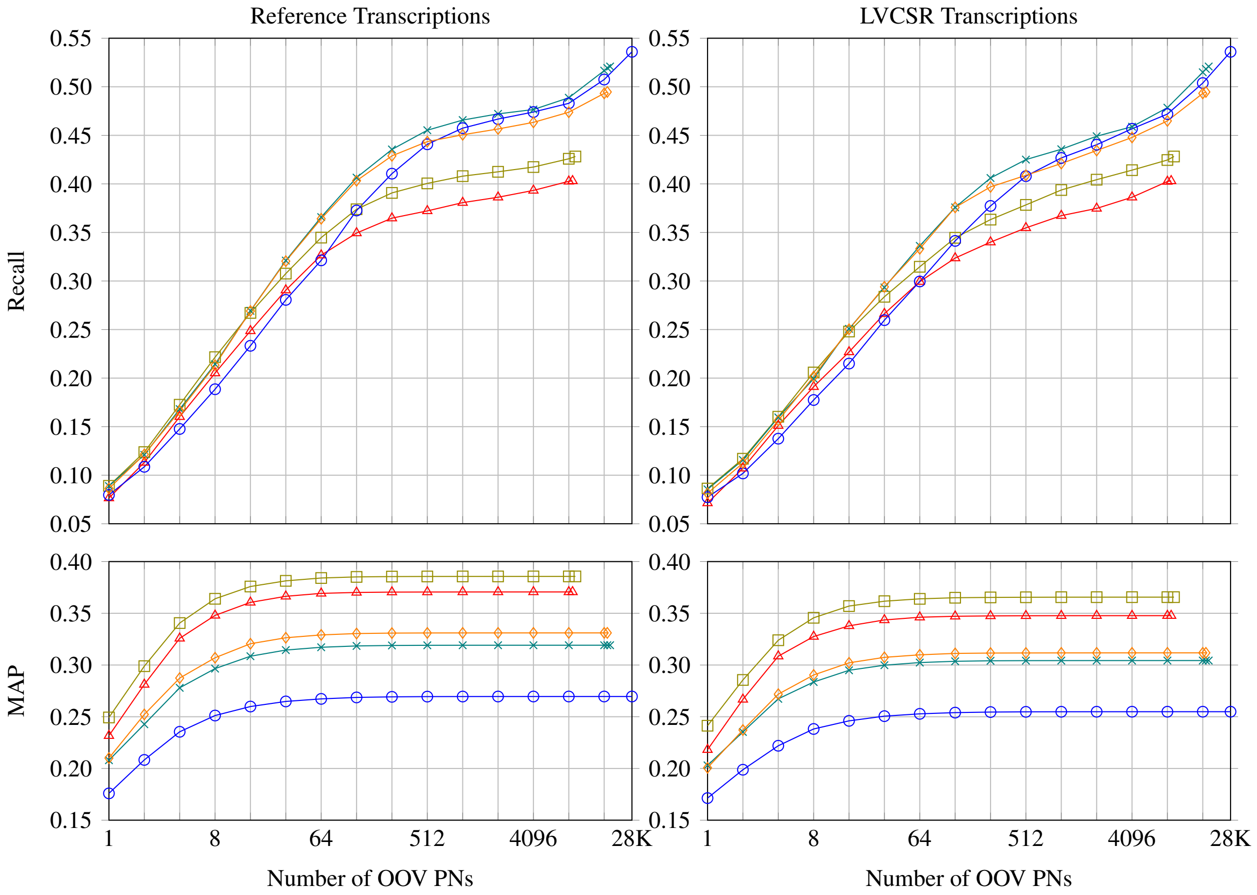

Figure 2: Recall and MAP for OOV PN retrieval on *Euronews* news video test set with different diachronic corpora for training LDA topic model. (—□— LX, —△— FIG, —×— LX+FIG, —◇— LX+rFIG, —○— LX-18m)

Pan, Y.-C., Liu, Y.-Y., and Lee, L.-S. (2005). Named entity recognition from spoken documents using global evidences and external knowledge sources with applications on mandarin chinese. In *IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 296–301.

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

Parada, C., Dredze, M., and Jelinek, F. (2011). OOV sensitive named-entity recognition in speech. In *INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association*, pages 2085–2088.

Qin, L. (2013). *Learning Out-of-Vocabulary Words in Automatic Speech Recognition*. Ph.D. thesis, Language Technologies Institute, School of Computer Science, Carnegie Mellon University.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, UK.

Sheikh, I., Illina, I., Fohr, D., and Linares, G. (2015a). OOV proper name retrieval using topic and lexical context models. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5291–5295.

Sheikh, I. A., Illina, I., and Fohr, D. (2015b). Study of entity-topic models for OOV proper name retrieval. In *INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association*, pages 1344–1348.

Sheikh, I. A., Illina, I., Fohr, D., and Linarès, G. (2015c). Learning to retrieve out-of-vocabulary words in speech recognition. *CoRR*, abs/1511.05389.

Sheikh, I., Illina, I., Fohr, D., and Linarès, G. (2016). Document level semantic context for retrieving OOV proper names. In *(To Appear) 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.