Study of Entity-Topic Models for OOV Proper Name Retrieval

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

Retrieving Proper Names (PNs) relevant to an audio document can improve speech recognition and content based audio-video indexing. Latent Dirichlet Allocation (LDA) topic model has been used to retrieve Out-Of-Vocabulary (OOV) PNs relevant to an audio document with good recall rates. However, retrieval of OOV PNs using LDA is affected by two issues; which we study in this paper: 1) Word Frequency Bias (less frequent OOV PNs are ranked lower); 2) Loss of Specificity (the reduced topic space representation loses lexical context). Entity-Topic models have been proposed as extensions of LDA to specifically learn relations between words, entities (PNs) and topics. We study OOV PN retrieval with Entity-Topic models and show that they are also affected by word frequency bias and loss of specificity. We evaluate our proposed methods for rare OOV PN re-ranking and lexical context re-ranking for LDA as well as for Entity-Topic models. The results show an improvement in both Recall and the Mean Average Precision.

Index Terms: proper names, OOV, topic models, LVCSR

1. Introduction

Out-Of-Vocabulary (OOV) words are inevitable when dealing with diachronic speech data for e.g. broadcast news. It has been commonly observed that majority of OOV words in news documents are Proper Names (PNs): about 57-72% [1–5]. At the same time, identifying PNs in news audio is of prime importance to enable content based indexing. In this paper, our goal is to retrieve the most relevant OOV PNs for an audio news document from a diachronic text corpus. Such a list of relevant OOV PNs can be later used to recover the target OOV PNs using phone matching [6], additional speech recognition pass [7] or spotting PNs in speech [8].

OOV word recovery and vocabulary selection have been the interest of researchers for some time. OOV word recovery techniques have used Large Vocabulary Continuous Speech Recognition (LVCSR) hypothesis to query search engines on World Wide Web (WWW) [6–8]. Vocabulary selection techniques using TF-IDF measures [13], frequency & recency of new words [14] and using linear combinations of corpora have been proposed [15–17]. In this paper we follow our approach proposed in [9] and retrieve OOV PNs using semantic and topic information. Previously, topic and semantic representations based on Latent Dirichlet Allocation (LDA) and Latent Semantic Analysis (LSA) [20] have been used to model and recognise PNs [18, 19]. These approaches use one LDA/LSA context model for each PN which restricts these approaches to model only frequent PNs. In our approach we train a global topic model and specifically address retrieval of rare OOV PNs.

In our approach proposed in [9], the In-Vocabulary (IV) words hypothesised by LVCSR are analysed for latent topics and the resulting topic space representation is used to retrieve relevant OOV PNs. With classical LDA topic model this method achieves more than 79% Recall of target OOV PNs with only top 10% of the retrieved PNs. In this paper we focus on two specific phenomena which affect the ranking of OOV PNs using LDA topic model. We refer to them as Word Frequency Bias (less frequent OOV PNs are ranked lower) and Loss of Specificity (the reduced topic space representation loses lexical context). As compared to [9], in this paper we evaluate different Entity-Topic models [10, 11] to verify if they can handle these issues. These models are extensions of LDA, designed to learn relations between words, topics and entities (or PNs).

Our evaluation shows that ranking of OOV PNs with Entity-Topic models are also affected by the same phenomena. We show that our approach to handle infrequent PNs and our lexical context model [9] used to re-score the OOV PNs retrieved with topic models, improve the Recall and Mean Average Precision (MAP) [12] for LDA as well as for Entity-Topic models.

The rest of the paper is organised as follows. The different topic models are first presented in section 2.1, followed by an overview on OOV PN retrieval using topic models in Section 2.2. Section 3 discusses about Word Frequency Bias & Loss of Specificity and presents our methods for ranking rare OOV PNs and lexical context re-ranking. Section 4 reports experimental evaluation, followed by a conclusion in Section 5.

2. OOV PN retrieval using topic models

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

2.1. Modelling PN-Topic relations

LSA [20], Probablistic LSA [21] and LDA [22] have been the most prominent methods for extracting underlying topics and semantics of a document. While LSA derives semantic spaces from word co-occurrence matrix, PLSA and LDA derive topics using probabilistic methods. We choose LDA since it is shown to outperform PLSA and LSA in document classification [22].
and word prediction [23] tasks. We further choose to evaluate Entity-Topic models [10, 11], which have been proposed as extensions of LDA to specifically model entity-topic relations. We do so in order to study if they can handle the problems faced by LDA in ranking OOV PNs (discussed in Section 3).

2.1. LDA topic model

For modelling LDA topics on a corpus of (D) text documents, topic vocabulary size (\( N_w \)), the number of topics (\( T \)) and Dirichlet priors (\( \alpha \), \( \beta \)) are first chosen [24]. Topic model parameters \( \theta \) and \( \phi \) are then estimated using Gibbs sampling algorithm [24]. \( \theta = [\theta_d]_{D \times T} \) is the topic distribution for each document \( d \), and \( \phi = [\phi_t]_{N_w \times T} \) is the topic distribution to words from the vocabulary, both across \( T \) topics.

2.1.2. Entity-Topic models

Entity-Topic models were proposed to specifically model entity-topic relations and were shown to perform slightly better than LDA for entity prediction tasks [10]. Figure 1 shows the graphical representation of the Entity-Topic models proposed in [10]. In the figure, \( \tilde{z} \) denotes the latent entity topic which generates an entity \( \tilde{w} \) based on the entity topic distribution \( \phi \), whereas \( z \) denotes the latent word topic which generates a word \( w \) based on the word topic distribution \( \phi \). It can be seen that the entity topic models share structural similarity with LDA, except that in Entity-Topic models there is a separate hierarchy for generation of words and entities. SwitchLDA has a switch variable \( x \) which controls the generation of words and entities. In Conditionally Independent LDA (CI-LDA) and SwitchLDA, the document specific topic distribution \( \theta_d \) generates both word and entity topics. Whereas in the Correspondence LDA’s (CorrLDA1 and CorrLDA2), the document specific topic distribution \( \theta_d \) generate the word topics and then word topics are used to generate entities. CorrLDA2 has additional hierarchy which allows different number of word and entity topics (\( T \) and \( \tilde{T} \)). A more detailed description of the variables and the generative/sampling process for these models is available in [10].

To use these models for our task of OOV PN retrieval we divide the topic model vocabulary into a set of \( N_w \) words & PNs in the LVCSR vocabulary and a set of \( N_v \) OOV PNs (PNs out of LVCSR vocabulary). Additionally, we tried further variations of the Entity-Topic models by changing the hierarchy of word topics and entity topics. For example in the CorrLDA1 model [10] word topic \( \tilde{z}_i \) is first sampled from the document topic distribution \( \theta \), and entity topic \( z_j \) is then sampled uniformly from the word topics. Instead of this, entity topic \( z_j \) can be first sampled from the document topic distribution \( \theta \) and the word topic can then be sampled uniformly from entity topic. We refer to this variation of CorrLDA1 as CorrLDA1-Flipped (CorrLDA1-F). A similar variation of the CorrLDA2 model has been proposed as Entity Centred Topic Model (ECTM) [11]. The motivation for using CorrLDA1-F and ECTM is that since these models learn entity centric topics, they may be useful in retrieving OOV PNs from the document topic. Thus we study performance of 7 different topic models for OOV PN retrieval: classic LDA, CI-LDA, SwitchLDA, CorrLDA1 and CorrLDA2, ECTM and the CorrLDA1-F model discussed above.

2.2. OOV PN retrieval

Let us denote the LVCSR word hypothesis by \( h \) and OOV PNs in diachronic corpus (or topic model vocabulary) by \( \tilde{v} \). To retrieve OOV PNs we calculate \( p(\tilde{v}|h) \) for each \( \tilde{v} \) and then use it as a score to rank OOV PNs relevant to \( h \). The latent topic mixture of \( h \), i.e. \( p(\tilde{v}|h) \), is inferred by sampling the topic assignments for words in \( h \) with the word-topic distribution \( \phi \) learned during training [24]. Then the likelihood of an OOV PN (\( \tilde{v} \)) in the diachronic corpus is calculated as:

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

where \( p(\tilde{v}|t) \) is from \( \phi \) for LDA or \( \tilde{\phi} \) for Entity-Topic models.

3. Ranking of OOV PNs by topic models

In this section we highlight two specific phenomena which affect the ranking of OOV PNs using LDA topic model. We refer to them as Word Frequency Bias and Loss of Specificity. We present a simple explanation to these phenomena using Equation (1) and some special cases of the topic distribution/mixture in the test document.

---

[1] These tasks are more similar to our task than general text document retrieval in which LSA is more dominant. Also we have verified in another set of experiments that LDA outperforms LSA for our task.

[2] The topic model vocabulary can be divided in two more ways, (a) words and PNs, as in the original work [10]; (b) forming three categories i.e. IV PNs, OOV PNs and other words. We tried these variations, however our approach discussed above (to separate OOV PNs from IV words) gives relatively better or similar results on our dataset.
3.1. Word frequency bias

Let us consider a special case where the test document contains a single topic: in other words, only one of the $T$ topic dimensions, let’s say $t_i$, in the topic mixture $p(t_i|v)$ is dominant (a value close to 1, and others to 0). When Equation (1) is used for ranking OOV PNs ($\tilde{v}$), it will simply order them by increasing value of $p(\tilde{v}|t_i)$. But since $p(\tilde{v}|t_i)$ depends on the frequency of occurrence of $\tilde{v}$, OOV PNs observed only few times in the entire diachronic corpus (termed as rare OOV PNs) will be ranked down in the list. In a more realistic scenario, Equation (1) acts as a weighted combination, with the weights coming from test document topic mixture, and it still tends to be biased against rare OOV PNs. To address this, we propose rare OOV PN re-ranking [9], where only for rare OOV PNs, we update Equation (1) to include $p(\tilde{v}|t)$ and $p(t|\cdot)$ normalisation as follows:

$$p(\tilde{v}|h) \approx \frac{C_h^1 \sum_{t=1}^{T} p(\tilde{v}|t) p(t|h)}{\sqrt{\sum_{t=1}^{T} p(\tilde{v}|t)^2} \sqrt{\sum_{t=1}^{T} p(t|h)^2}}$$

where $C_h^1$ is a scaling factor to scale the scores to that of frequent OOV PNs, obtained using Equations (1). The reasoning behind this normalisation is that Equation (1) can be seen as a dot product between topic space vector representations of the test document $h$ and an OOV PN $\tilde{v}$, whereas Equation (2) resembles cosine similarity which is weakly affected by vector magnitude (related to word frequency in this case). A similar word frequency effect was observed in the study of word associations using semantic spaces [23].

3.2. Loss of specificity

Let us consider again the case where the test document topic mixture has a single dominant topic dimension, resulting into a ranking of OOV PNs ($\tilde{v}$) ordered by increasing value of $p(\tilde{v}|t_i)$. It is possible that this topic (e.g. football) has some sub-topics (e.g. world cup, European club matches, etc.). This level of specificity is lost in the (reduced) topic space representation but is still contained in word level representations (e.g. team names). It must be noted that increasing number and granularity of topics is not always feasible. Thus topics alone are not discriminant enough for ranking OOV PNs and they must be enhanced with lexical context information. Even in text information retrieval, LDA model needs to be combined with document-specific word distributions to capture both general topic as well as document specific information [25].

For our task, we address this problem by re-ranking with a lexical context model, proposed in [9]. In this re-ranking, a topic model is first used to choose top-$N$ (topic relevant) OOV PNs, and then lexical context model is used to re-rank OOV PNs in the top-$N$ list. The idea is to improve the scores of OOV PNs which are related by topic as well as by lexical context. Given the PN lexical context model, trained on the diachronic corpus, Gibbs sampling is used to infer the best OOV PN which generates each word in $h$. The word counts for each top-$N$ OOV PN are then used to re-score the OOV PN as follows:

$$P_N(\tilde{v}|h) \approx p(\tilde{v}|h) + C_h^2 \sum_{k} N_{\tilde{v}w}^{h} + \alpha_o \frac{N_{\tilde{v}w}^{h}}{N_{wh} + \alpha_o N}$$

where, $N_{\tilde{v}w}^{h}$ is the number of times a word $w$ in $h$ is assigned to top-$N$ OOV PN $\tilde{v}$; $p(\tilde{v}|t)$ is the score from topic model, $N_{wh}$ is number of words in $h$, $\alpha_o$ is the model prior and $C_h^2$ is a scaling factor to combine topic and lexical model scores.

4. Experiments and results

4.1. Experiment setup

To evaluate the entity topic models we use the Euronews corpus. This corpus consists of French news videos and articles collected from Euronews (http://fr.euronews.com). We used the text news corresponding to the period 01/01/2014 - 31/05/2014, from this corpus as our diachronic corpus. The diachronic corpus vocabulary is filtered by removing PNs occurring only once, non PN words occurring less than 4 times, and using a stop-list of common french words and non content words which do not carry any topic information. The filtered vocabulary has 8155 PNs and 8732 words. Of these, the words and PNs which occur in the lexicon of our Automatic News Transcription System (ANTS) [26] are tagged as IV and the remaining (2418) PNs are tagged as OOV PNs. (ANTS lexicon is based on news articles until 2008 from French newspaper LeMonde.)

For modelling topics and lexical context we used 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, and our second test set (TestSet-II) is the LVCSR transcriptions of these videos. 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, 38% of the OOV PNs have occurred only 5 or lesser times in the training set. Similarly, the number of OOV PNs to be retrieved for TestSet-II is 220.

For modelling topics on diachronic corpus, we choose symmetric weak priors (for $\alpha, \beta, \tilde{\beta}, \gamma$). We experimented with different values and chose ones which gave best results on our test set. Similarly, we tried with different number of word topics $T$ and entity topics $\tilde{T}$ (in the range 20-200), the best performance is obtained with about 50 topics for all topic models. Scaling factors $C_h^1$ and $C_h^2$, used in Equation (2) and Equation (3) respectively, are also set empirically.

4.2. Results

Figure 2 shows the performance of different topic models for retrieval of OOV PNs on TestSet-I of text news articles. For each of the graphs in Figure 2, X-axis represents the number of OOV PNs retrieved from the diachronic corpus. Y-axis represents recall of the target OOV PNs. Figure 2 (a) shows OOV PN recall for the LDA topic model, CI-LDA and SwitchLDA Entity-Topic models. Figure 2 (b) shows recall for the CorrLDA1, CorrLDA2, CorrLDA1-F and ECTM Entity-Topic models. It can be seen that although the Entity-Topic models are targeted for capturing entity-topic relationships, they give recall comparable to LDA for our task of OOV PN retrieval. For top 10% retrieved OOV PNs, the topic models give a recall of 74-78%, except for CorrLDA1-F and ECTM which have a recall of 68% and 63% respectively. Our observation is that, since the total count of entities (OOV PNs in our task) is quite small compared to words, the topics learned by these models are not optimal. And hence the associations between words and entities via topic space are not as good as LDA and other entity topic models.

To highlight the issue of Word Frequency Bias, Figure 3 shows a distribution of ranks of target OOV PNs in TestSet-I versus their frequency in training corpus. This graph is for CorrLDA1 (the best performing Entity-Topic model) before (top)
and after (bottom) the rare OOV PN and lexical context re-rankings. It is evident from this graph that our proposed re-rankings improve the overall ranks of OOV PNs. To summarise Table 1 compares OOV PN retrieval performance of the topic models on TestSet-I of text news articles, in terms of Recall and Mean Average Precision (MAP) [12] obtained with top 10% of retrieved OOV PNs. Table 1 shows the gain with different re-ranking methods. No denotes no re-ranking, Lex denotes lexical context re-ranking (with \( N=100 \)) cf. Section 3.2, Freq denotes re-ranking for rare OOV PNs (occurring less than 5 times in training data) cf. Section 3.1 and Freq+Lex denotes combination of both lexical context and rare OOV PN re-ranking. It can be seen that for each of the topic models our re-ranking methods help to improve MAP and recall (upto 12% for CorrLDA1).

Figure 2: OOV PN Recall for TestSet-I.

Table 1: OOV PN Recall and Mean Average Precision (MAP) at 10% OOV PNs, for TestSet-I.

<table>
<thead>
<tr>
<th>Model</th>
<th>No Recall</th>
<th>MAP</th>
<th>No Recall</th>
<th>MAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>LDA</td>
<td>0.78</td>
<td>0.26</td>
<td>0.88</td>
<td>0.25</td>
</tr>
<tr>
<td>CI-LDA</td>
<td>0.76</td>
<td>0.24</td>
<td>0.86</td>
<td>0.24</td>
</tr>
<tr>
<td>SwitchLDA</td>
<td>0.75</td>
<td>0.22</td>
<td>0.86</td>
<td>0.22</td>
</tr>
<tr>
<td>CorrLDA1</td>
<td>0.77</td>
<td>0.24</td>
<td>0.89</td>
<td>0.23</td>
</tr>
<tr>
<td>CorrLDA1-F</td>
<td>0.68</td>
<td>0.24</td>
<td>0.72</td>
<td>0.24</td>
</tr>
<tr>
<td>CorrLDA2</td>
<td>0.74</td>
<td>0.25</td>
<td>0.84</td>
<td>0.24</td>
</tr>
<tr>
<td>ECTM</td>
<td>0.63</td>
<td>0.20</td>
<td>0.67</td>
<td>0.20</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>Freq</th>
<th>Recall</th>
<th>MAP</th>
<th>Freq</th>
<th>Recall</th>
<th>MAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>LDA</td>
<td>0.78</td>
<td>0.31</td>
<td>0.88</td>
<td>0.30</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CI-LDA</td>
<td>0.76</td>
<td>0.30</td>
<td>0.85</td>
<td>0.28</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SwitchLDA</td>
<td>0.75</td>
<td>0.27</td>
<td>0.86</td>
<td>0.25</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CorrLDA1</td>
<td>0.77</td>
<td>0.27</td>
<td>0.89</td>
<td>0.28</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CorrLDA1-F</td>
<td>0.68</td>
<td>0.27</td>
<td>0.73</td>
<td>0.27</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CorrLDA2</td>
<td>0.74</td>
<td>0.29</td>
<td>0.84</td>
<td>0.29</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ECTM</td>
<td>0.63</td>
<td>0.22</td>
<td>0.66</td>
<td>0.22</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2 shows performance on TestSet-II of news audio. LVCSR transcripts of the news audio are obtained using ANTS [26], with 46% Word Error Rate (WER) compared to the manual transcripts. Recall on LVCSR transcripts is less affected as the retrieval method relies on the topic mixture inferred from the LVCSR hypothesis, which smoothes out the LVCSR errors.

Figure 3: Rank-Frequency Distribution for CorrLDA1 before (top) and after (bottom) rare OOV PN and lexical re-rankings.

Table 2: OOV PN Recall on TestSet-II at 10% OOV PNs.

<table>
<thead>
<tr>
<th>Model</th>
<th>Manual Transcripts</th>
<th>LVCSR Transcripts</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No Freq</td>
<td>Lex</td>
</tr>
<tr>
<td>LDA</td>
<td>0.79</td>
<td>0.86</td>
</tr>
<tr>
<td>CI-LDA</td>
<td>0.78</td>
<td>0.86</td>
</tr>
<tr>
<td>SwitchLDA</td>
<td>0.79</td>
<td>0.90</td>
</tr>
<tr>
<td>CorrLDA1</td>
<td>0.79</td>
<td>0.89</td>
</tr>
<tr>
<td>CorrLDA1-F</td>
<td>0.70</td>
<td>0.75</td>
</tr>
<tr>
<td>CorrLDA2</td>
<td>0.78</td>
<td>0.85</td>
</tr>
<tr>
<td>ECTM</td>
<td>0.66</td>
<td>0.70</td>
</tr>
</tbody>
</table>

5. Conclusion

We discussed two specific phenomena, Word Frequency Bias and Loss of Specificity, which affect the retrieval of OOV PNs using LDA topic model. Entity-Topic models, which are extensions of LDA designed to learn relations between words-topics-PNs, were studied to verify if they can handle these issues. Experimental evaluation shows that their performance does not outperform LDA and that they face same problems. We showed that our proposed methods of rare OOV PN and lexical context re-ranking improve the Recall and MAP for LDA as well for the Entity-Topic models. Upto 11% (absolute) improvement in Recall on LVCSR transcripts is less affected as the retrieval method relies on the topic mixture inferred from the LVCSR hypothesis, which smoothes out the LVCSR errors.

6. Acknowledgements

This work is funded by the ContNomina project supported by the French national Research Agency (ANR) under the contact ANR-12-BS02-0009.
7. References


