An LDA-smoothed Relevance Model for Document Expansion: A Case Study for Spoken Document Retrieval

Debasis Ganguly Johannes Leveling Gareth J. F. Jones
School of Computing, Centre for Next Generation Localisation
Dublin City University, Dublin 9, Ireland
{dganguly, jleveling, gjones}@computing.dcu.ie

ABSTRACT

Document expansion (DE) in information retrieval (IR) involves modifying each document in the collection by introducing additional terms into the document. It is particularly useful to improve retrieval of short and noisy documents where the additional terms can improve the description of the document content. Existing approaches to DE assume that documents to be expanded are from a single topic. In the case of multi-topic documents this can lead to a topic bias in terms selected for DE and hence may result in poor retrieval quality due to the lack of coverage of the original document topics in the expanded document. This paper proposes a new DE technique providing a more uniform selection and weighting of DE terms from all constituent topics. We show that our proposed method significantly outperforms the most recently reported relevance model based DE method on a spoken document retrieval task for both manual and automatic speech recognition transcripts.

Categories and Subject Descriptors

H.3.3 [**INFORMATION STORAGE AND RETRIEVAL**]: Information Search and Retrieval—*Query formulation, Relevance Feedback*

Keywords

Document Expansion, Topic Modelling

1. INTRODUCTION

Document expansion (DE) in information retrieval (IR) involves expanding the contents of documents in a retrieval collection so that indexed terms of the documents better describe the contents of the document for retrieval. This can be particularly valuable for noisy and short documents. In the former case the contents may be incorrectly indexed and in the latter the terms present may not sufficiently describe the contents to support effective retrieval.

SIGIR'13, July 28–August 1, 2013, Dublin, Ireland.

An IR task that is particularly amenable to DE is spoken document retrieval (SDR) [6]. This is a challenging problem due to the vocabulary mismatch between the documents in the collection and the user query, which arises due to errors in automatic speech recognition (ASR) transcripts of the content and, particularly in the case of conversational speech, the failure to articulate details valuable for retrieval.

DE involves expanding each document in the collection by adding terms from a set of topically related documents, either from the search collection or elsewhere. We refer to this set of related documents as the *neighbourhood* of the current document. Retrieval on the expanded document index is expected to produce better retrieval results, because each expanded document is likely to contain additional beneficial indexing terms obtained from the neighbourhood.

DE faces the critical problems of finding a *good* neighbourhood for selection of the expansion terms, and selection of suitable terms from this neighbourhood. A neighbourhood comprising documents not related to the current document or additional unsuitable terms may add noise and degrade retrieval results. A typical method of computing the neighbourhood of a document is as follows. First, a set of documents is retrieved, using the current document as a query with the help of a standard retrieval method. Documents retrieved at the top ranks are then selected as the neighbourhood of the current document [6, 8]. This approach to DE can straightforwardly be applied to improve retrieval quality of images based on document metadata [5], where the document annotations are typically very short, resembling keyword queries of ad-hoc IR or web search. However, using a full document as a query may pose a problem for moderately sized documents. This is because, while standard models are well suited for short *keyword* type queries, they generally do not perform well for cases where a query and the documents in the collection are of comparable length [7].

Recently reported work on DE has shown that the relevance model (RLM) [4], which is a statistical generative model, works particularly well for expanding short documents such as tweets [2]. The tweets are usually short, focused on a topic and clean of ASR errors. The tweets are thus characteristically different from spoken documents, which may be moderately sized, multi-topical and noisier. We thus presume that RLM-based DE may not prove to be very effective for SDR, and this motivates us to devise a DE technique suitable for multi-topic spoken documents.

In this paper, we explore whether exploiting the underlying topical information of a document, D, can help in choosing a robust neighbourhood for DE with a balanced contri-

bution from all the topics in D. Such a neighbourhood can lead to a more comprehensive selection of expansion terms, thus potentially improving retrieval effectiveness in comparison to the RLM-based document expansion. Let us illustrate this with an example. Let D be a document comprising of a mixture of two topics, one pertaining to a product, say a *remote control*, and the other related to its *cost*. Consider two sets of terms from two neighbourhoods $N_1(D)$ and $N_2(D)$. Let $N_1(D) = \{pushbutton, button, tv, switch, money\}$. The first four words are related to topic-1 (*remote controls*), whereas the fifth word is related to topic-2 (*cost*). Clearly, $N_1(D)$ is biased towards topic-1. In contrast, let $N_2(D)$ be a set of terms with an even distribution of the topics, say $\{battery, button, euro, price\}$. Note that although $N_1(D)$ can add 4 terms from the first topic, it can add only 1 from the second; whereas $N_2(D)$ can add 2 terms from each topic, and hence can lead to better expansion, due to a more comprehensive coverage of topics. In our proposed method, we compute this topic-level information firstly to ensure retrieving a topic of uniform neighbourhood, and secondly to ensure a balanced selection of terms from each such topic.

The rest of the paper is organized as follows. In Section 2, we briefly survey the existing literature on DE. This is followed by Section 3, which describes our approach in detail. Section 4 evaluates the approach, and Section 5 summarizes the conclusions of our study.

2. RELATED WORK

All existing DE methods can be generalized to a linear combination of two term weighting functions, one for the original document terms and the other for the new terms in the document's neighbourhood. Rocchio relevance feedback has been applied for DE by shifting the current document vector closer to the centroid of the neighbouring ones [6], as shown in Equation 1. Here, D is the document to be expanded, $\{D_j^N\}_{j=1}^R$ is the set of R neighbouring documents retrieved with D as the query, and D' is the expanded document.

$$\vec{D'} = \alpha\vec{D} + \frac{1-\alpha}{R}\sum_{j=1}^{R}\vec{D_j^N} \tag{1}$$

Equation 1 reweights the terms originally present in the current document by a factor α, and reweights additional terms from the neighbourhood by the factor $(1-\alpha)/R$. Note that each document D_j^N of the neighbourhood of D in Equation 1 contributes equally to the expansion. In contrast, a method for re-weighting terms in a document from the neighbourhood by a factor proportional to its similarity with the neighbourhood, was proposed in [8], as shown in Equation 2. The proportionality factor for a document D_j^N in the neighbourhood is the ratio of its similarity with D, to the total similarity for all documents in the neighbourhood. The hypothesis is that documents in the neighbourhood retrieved at higher ranks are more similar to the current document and thus more reliable for expansion.

$$\vec{D'} = \alpha\vec{D} + (1-\alpha)\sum_{j=1}^{R}\frac{sim(\vec{D}, \vec{D_j^N})}{\sum_{k=1}^{R} sim(\vec{D}, \vec{D_k^N})}\vec{D_j^N} \tag{2}$$

This intuitive idea of using proportional similarity weights for different documents in the neighbourhood was shown to be theoretically well motivated by the relevance model (RLM) in [2]. In summary, the RLM involves computation of a model of relevance $P(w|\mathrm{R})$, given that such a model generates the pseudo-relevant documents as well as the query [4]. In the context of DE, the RLM estimates a new document model for D, denoted D', from the evidence that the new model D' generates both the current document D and its neighbourhood $\{D_j^N\}_{j=1}^R$. Figure 1a) illustrates this, assuming the current document is comprised of n terms, viz. D=(d_1, ..., d_n). The probability $P(w|D')$, approximated by the probability $P(w, d_1, \ldots, d_n)$, is given by

$$\begin{aligned} P(w, d_1, \ldots, d_n) &= \sum_{j=1}^{R} P(D_j^N) P(w, d_1, \ldots, d_n | D_j^N) \\ &= \frac{1}{R}\sum_{j=1}^{R} P(w|D_j^N) \prod_{i=1}^{n} P(d_i|D_j^N) \approx P(w|D') \end{aligned} \tag{3}$$

The language model for the expanded document D' to be used during retrieval, is recomputed by a linear combination of the relevance model estimation $P(w|D')$ (the expansion factor as computed by Equation 3), and the original unigram document model for D, as shown in Equation 4.

$$\begin{aligned} P_\alpha(w|D') &= \alpha P(w|D) + (1-\alpha) P(w|D') \\ &= \alpha P(w|D) + \frac{1-\alpha}{R}\sum_{j=1}^{R} P(w|D_j^N) \prod_{i=1}^{n} P(d_i|D_j^N) \end{aligned} \tag{4}$$

Note that Equation 4 is similar to Equation 2 because the quantity $\prod_{i=1}^{n} P(d_i|D_j^N) = P(D|D_j^N)$ acts as the proportional similarity of a neighbourhood document D_j^N with D.

DE has thus evolved over time, starting with a simple vector space approach [6], moving onto applying different confidence measures for different documents in the neighbourhood [8], and establishing this intuitive notion from a theoretical perspective [2]. None of these methods, however, has attempted to utilize the topical information of the current document and its neighbourhood to achieve a uniform expansion of concepts for all different aspects of the current document. Our proposed model seeks to achieve this uniform balanced expansion of concepts within a document.

3. LDA-SMOOTHED RELEVANCE MODEL

Our proposed model for DE uses the RLM-based model as a framework. We begin this section with a discussion on how the RLM-based model can be extended, and then describe the technical details of the extended model.

Motivation. Referring back to Equation 3, we can see that the RLM relies on the probability of co-occurrences of a new word, $(P(w|D_j^N))$, with that of an existing word,

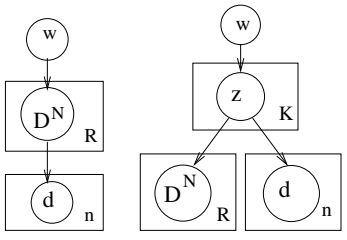


Figure 1: Plate diagrams of document expansion for a) RLM (left) and b) LDA-smoothed RLM (right).

$(P(d_i|D_j^N))$, in the neighbourhood of the current document. Highly co-occurring terms contribute more to the expanded document model. However, such co-occurrences, if computed at the level of whole documents, can lead to the problem of unbalanced selection of expansion terms For example, referring back to Section 1, the *remote control* related terms such as *pushbutton*, *button* etc. may have a higher degree of co-occurrence with terms in D, than the *cost* related words such as *money*, *euro* etc., whose contributions are thus down weighted. However, if co-occurrences are computed at the level of topics, it may be the case that the top co-occurring terms for the first topic are *battery* and *button*, whereas for the second topic, these could be *money* and *price*. This will lead to a better expansion, because it gives an even chance to the document to be retrieved against a query related to the production cost. These topic-level co-occurrences can be computed with the help of a topic-modelling approach, which we describe next.

LDA for topical co-occurrences. Latent Dirichlet allocation (LDA) has been shown to model the underlying topics of a collection of documents effectively [1]. Furthermore, it has been shown that term generation probabilities of a document, marginalized over a set of topics, improves ad-hoc IR effectiveness [9]. When applied in the context of RLM-based DE, this marginalization with topics leads to a computation of co-occurrences at the level of topics, rather than computing it over the whole document in the RLM. Thus, it can lead to a more uniform contribution from each topic, rather than being biased towards a few in the estimated values of $P(w|D')$. We can thus assume that both the neighbourhood and the current document are generated from a set of K latent topics, as shown in Figure 1b). Marginalizing Equation 3 over the K topic nodes yields Equation 5. We use Equation 5 to estimate the term weights for the expanded document model D'.

$$
\begin{aligned}
P(w|D') &= \sum_{j=1}^{R} P(D_j^N) P(w, d_1, \ldots, d_n | D_j^N) \\
&= \frac{1}{R} \sum_{j=1}^{R} \sum_{k=1}^{K} P(w|z_k) P(z_k|D_j^N) \prod_{i=1}^{n} \sum_{k=1}^{K} P(d_i|z_k) P(z_k|D_j^N) \\
&= \frac{1}{R} \sum_{j=1}^{R} P_{LDA}(w|D_j^N, \boldsymbol{\theta}, \boldsymbol{\phi}) \prod_{i=1}^{n} P_{LDA}(d_i|D_j^N, \boldsymbol{\theta}, \boldsymbol{\phi})
\end{aligned}
\tag{5}
$$

Implementation details. We perform off-line LDA inference over a collection of documents with a pre-configured number of topics, K, to get the output matrices $\boldsymbol{\theta}$ (document-topic) and $\boldsymbol{\phi}$ (word-topic) mappings. The optimal value of K is determined empirically. A subset of rows (those corresponding to the neighbourhood document indices for the current document) of these matrices are then used to compute the probabilities $P_{LDA}(w|D_j^N, \boldsymbol{\theta}, \boldsymbol{\phi})$ (LDA smoothing for the neighbourhood documents) and $P_{LDA}(d_i|D_j^N, \boldsymbol{\theta}, \boldsymbol{\phi})$ (LDA smoothing for the current document), as shown in Equation 5. As a final step, we substitute the value of $P(w|D')$ in Equation 4 to compute the linearly interpolated expansion model, similar to [2]. It is worth mentioning here that our method of DE, shown in Equation 5, is different from running RLM on top of LDA-smoothed documents models, an approach named *RM+LBDM* in [9]. In our approach, both the query (the current document) and the pseudo-relevant documents (the neighbourhood) are LDA-smoothed, instead of smoothing only the pseudo-relevant documents, as in *RM+LBDM*.

4. EVALUATION

This section describes the SDR test collection data, the experimental settings and the results.

4.1 SDR Test Collection

The SDR test collection used for our experiments is derived from the AMI corpus[1], comprising of 100 hours of recorded planned meetings transcribed both manually and using automatic speech recognition (ASR). For our experiments, we use both the manual and ASR transcripts to assess the impact of ASR errors and the effectiveness of our method on both transcript types. The query set consists of text extracted from 25 PowerPoint slides which are supplied as part of the AMI dataset, the objective of the retrieval task being to locate the spoken content relevant to the topic of a given query slide. The relevant content is manually labelled across the spoken contents for each query [3].

The average length of each meeting in terms of spoken time duration is 1998 secs, with a range between 306.4 secs and 5297.84 secs. Retrieving these very long meeting documents for a given query slide does not benefit a searcher, whose objective is to precisely locate the relevant time range(s) within a meeting for each query slide. Thus, to get more meaningful retrieval units, we segmented the documents in AMI corpus into chunks of fixed duration (180 secs), because previous research using this test collection based on the AMI corpus reports that a fixed time-based segmentation of 180 secs produced the best retrieval results [3]. The average length of these segments[2] is 99.49 words. These segments on average are much shorter than the TREC-8 documents (311.94) words, thus justifying the need for DE; but longer than tweets (21.92) [2], justifying the necessity of LDA smoothing to achieve a uniform selection of terms from each topic with the segments. In fact, one of the reasons for not using a tweet corpus for our experiements is that a) short uni-topical tweets are unsuitable to test the topical co-occurrence hypothesis for expansion of these mid-range length documents, and b) for tweets, temporal evidence is an important criterion for choosing a good expansion neighbourhood, which we do not pursue here.

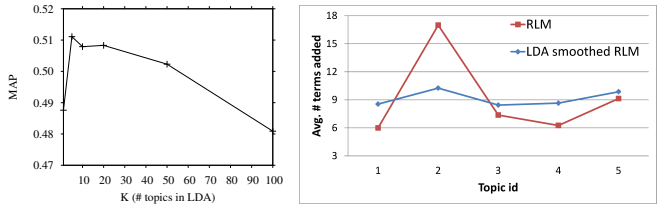
4.2 Experimental Settings

Baselines. Our proposed DE method, DE_{LDA} is compared against the following baselines: i) retrieval without DE, DE_{NO}; and ii) retrieval after expanding documents by RLM [2], DE_{RLM}. We do not compare our results against approaches reported in [6, 8], since we have shown in Section 2 that these have the same mathematical form as [2].

Parameters. The parameters in our experiments were empirically optimized on the AMI test set of 25 queries. A common parameter for both DE_{RLM} and DE_{LDA}, is α (see Equation 4). This was empirically optimized to 0.6. This optimal value of α on the AMI test set is identical to the values reported in [8, 2], although these used different test collections. Another common parameter is the number of neighbourhood documents to be used for expansion, viz.

[1] `http://www.amiproject.org/`

[2] *segment* and *document* are used interchangeably henceforth



(a) (b)

Figure 2: a) Effect of varying K in DE_{LDA} and b) avg. #terms added per topic, on ASR documents.

R (see Equations 3 and 5). The value of R was varied in the range of $[5, 50]$, and was set to the optimal value of 20 in both cases. An additional parameter for DE_{LDA} is the number of topics, K. To find its optimal value, we report the MAP for different values of K on the ASR data in the range of $[1, 100]$ in Figure 2a. Note that the leftmost point in Figure 2a, i.e. K=1, is identical to the method DE_{RLM}. The figure shows that the optimal value of K is 5. The intuitive reason for optimality at a small value of K is that the meeting documents are more homogeneous in content than news articles, for which, values of K as high as 800, are optimal [9]. Instead of optimizing the value of K separately on the manual transcripts, we used the optimal value of K as obtained for ASR.

Metrics. To evaluate the retrieval effectiveness, in addition to standard metrics such as MAP, recall etc., we calculate the mean average segment precision (MASP) [3]. Unlike the binary valued relevance of MAP, relevance in MASP can be a real number in $[0, 1]$ according to the amount of relevant content present in a segment, and in addition to the rank of relevant content, MASP also considers the difference between the starting time of the retrieved segment and that of the relevant content [3].

4.3 Results

Table 1 shows the retrieval effectiveness obtained on the manual and ASR transcripts by the different approaches. The first observation is that the retrieval effectiveness with DE is higher than the baseline without DE, i.e. DE_{NO}, both for the manual and ASR transcripts. The second observation is that our approach to DE using LDA smoothing, significantly[3] outperforms the standard RLM-based DE method, again on both manual and ASR transcripts. Note that DE_{LDA} achieves both higher recall and higher P@10. Moreover, the MASP results show that DE_{LDA} is able to retrieve segments with more overlapping relevant content at higher ranks, in comparison to DE_{RLM}. The retrieval effectiveness is better on the manual transcript in comparison to the ASR, as expected. Counterintuitively, the P@10 value on ASR is slightly higher than its manual counterpart. We conjencture that this is due to the fact that the manual transcript documents, due to the absence of incorrectly recognized words, should have a higher number of topics than the ASR documents. In fact, setting K to 20 on the manual transcripts increases P@10 to 0.6480, which is higher than the value of P@10 for ASR transcript reported in Table 1.

To test the hypothesis that DE_{LDA} is less biased towards a particular topic while selecting expansion terms, we plot the distribution of expansion terms over the set of topics, averaged over all documents of the ASR collection in Figure 2b. To plot the topic assignments for DE_{RLM}, we simply used the topic mappings from DE_{LDA}. Figure 2b shows that DE_{LDA} indeed results in a uniform selection of terms from each topic as evident by a flatter graph, in comparison to the sharp peaks and valleys of DE_{RLM}.

[3]Measured by Wilcoxon test with 95% confidence measure

Doc. Type	Expansion Method	Evaluation Metrics				
		MAP	Recall	P@10	R-Prec	MASP
ASR	DE_{NO}	0.4718	0.9353	0.6160	0.4774	0.2457
	DE_{RLM}	0.4876	0.9533	0.5880	0.4887	0.2501
	DE_{LDA}	**0.5111**	**0.9615**	**0.6400**	**0.5030**	**0.2659**
Manual	DE_{NO}	0.5105	0.9509	0.6280	0.4792	0.2630
	DE_{RLM}	0.5129	0.9689	0.6120	0.4950	0.2654
	DE_{LDA}	**0.5347**	**0.9705**	**0.6360**	**0.5087**	**0.2790**

Table 1: Retrieval results using different document expansion methods on the AMI dataset.

5. CONCLUSIONS

We proposed a novel DE method to ameliorate the vocabulary mismatch problem of short and noisy spoken documents by extending RLM-based DE with topic-based LDA smoothing. This results in a more uniform selection of expansion terms for each topic in the current document in comparison to the RLM-based method, which may be biased towards certain topics. It is shown empirically that our proposed method significantly outperforms the RLM-based DE on a spoken document retrieval task.

In future, we would like to investigate the effect of dynamically varying the value of K per document.

Acknowledgments

This research is supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the CNGL project.

6. REFERENCES

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003.

[2] M. Efron, P. Organisciak, and K. Fenlon. Improving retrieval of short texts through document expansion. In *Proceedings of the SIGIR '12*, pages 911–920, 2012.

[3] M. Eskevich, W. Magdy, and G. J. F. Jones. New metrics for meaningful evaluation of informally structured speech retrieval. In *Proceedings of ECIR '12*, pages 170–181, 2012.

[4] V. Lavrenko and B. W. Croft. Relevance based language models. In *SIGIR 2001*, pages 120–127. ACM, 2001.

[5] J. Min, J. Leveling, D. Zhou, and G. J. F. Jones. Document expansion for image retrieval. In *RIAO*, pages 65–71, 2010.

[6] A. Singhal and F. C. N. Pereira. Document expansion for speech retrieval. In *Proceedings of the SIGIR '99*, pages 34–41, 1999.

[7] T. Takaki, A. Fujii, and T. Ishikawa. Associative document retrieval by query subtopic analysis and its application to invalidity patent search. In *Proceedings of CIKM '04*, pages 399–405, 2004.

[8] T. Tao, X. Wang, Q. Mei, and C. Zhai. Language model information retrieval with document expansion. In *Proceedings of the HLT-NAACL*, 2006.

[9] X. Wei and W. B. Croft. LDA-based document models for ad-hoc retrieval. In *Proceedings of SIGIR '06*, pages 178–185. ACM, 2006.