A Generative Blog Post Retrieval Model that Uses Query Expansion based on External Collections

Wouter Weerkamp
w.weerkamp@uva.nl

Krisztian Balog
k.balog@uva.nl

Maarten de Rijke
mdr@science.uva.nl

ISLA, University of Amsterdam

Abstract

User generated content is characterized by short, noisy documents, with many spelling errors and unexpected language usage. To bridge the vocabulary gap between the user's information need and documents in a specific user generated content environment, the blogosphere, we apply a form of query expansion, i.e., adding and reweighing query terms. Since the blogosphere is noisy, query expansion on the collection itself is rarely effective but external, edited collections are more suitable. We propose a generative model for expanding queries using external collections in which dependencies between queries, documents, and expansion documents are explicitly modeled. Different instantiations of our model are discussed and make different (in)dependence assumptions. Results using two external collections (news and Wikipedia) show that external expansion for retrieval of user generated content is effective; besides, conditioning the external collection on the query is very beneficial, and making candidate expansion terms dependent on just the document seems sufficient.

1 Introduction

One of the grand challenges in information retrieval is to bridge the vocabulary gap between a user and her information need on the one hand and the relevant documents on the other (Baeza-Yates and Ribeiro-Neto, 1999). In the setting of blogs or other types of user generated content, bridging this gap becomes even more challenging. This has several causes: (i) the spelling errors, unusual, creative or unfocused language usage resulting from the lack of top-down rules and editors in the content creation process, and (ii) the (often) limited length of user generated documents.

Query expansion, i.e., modifying the query by adding and reweighing terms, is an often used technique to bridge the vocabulary gap. In general, query expansion helps more queries than it hurts (Balog et al., 2008b; Manning et al., 2008). However, when working with user generated content, expanding a query with terms taken from the very corpus in which one is searching tends to be less effective (Arguello et al., 2008a; Weerkamp and de Rijke, 2008b)—topic drift is a frequent phenomenon here. To be able to arrive at a richer representation of the user's information need, while avoiding topic drift resulting from query expansion against user generated content, various authors have proposed to expand the query against an external corpus, i.e., a corpus different from the target (user generated) corpus from which documents need to be retrieved.

Our aim in this paper is to define and evaluate generative models for expanding queries using external collections. We propose a retrieval framework in which dependencies between queries, documents, and expansion documents are explicitly modeled. We instantiate the framework in multiple ways by making different (in)dependence assumptions. As one of the instantiations we obtain the mixture of relevance models originally proposed by Diaz and Metzler (2006).

We address the following research questions: (i) Can we effectively apply external expansion in the retrieval of user generated content? (ii) Does conditioning the external collection on the query help improve retrieval performance? (iii) Can we obtain a good estimate of this query-dependent collection probability? (iv) Which of the collection, the query, or the document should the selection of an expansion term be dependent on? In other words, what are the strongest simplifications in terms of conditional independencies between variables that can be assumed, without hurting performance? (v) Do our models show similar behavior across topics or do we observe strong per-topic

Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, pages 1057–1065,
Suntec, Singapore, 2-7 August 2009. ©2009 ACL and AFNLP

differences between models?

The remainder of this paper is organized as follows. We discuss previous work related to query expansion and external sources in §2. Next, we introduce our retrieval framework (§3) and continue with our main contribution, external expansion models, in §4. §5 details how the components of the model can be estimated. We put our models to the test, using the experimental setup discussed in §6, and report on results in §7. We discuss our results (§8) and conclude in §9.

2 Related Work

Related work comes in two main flavors: (i) query modeling in general, and (ii) query expansion using external sources (*external expansion*). We start by shortly introducing the general ideas behind query modeling, and continue with a quick overview of work related to external expansion.

2.1 Query Modeling

Query modeling, i.e., transformations of simple keyword queries into more detailed representations of the user's information need (e.g., by assigning (different) weights to terms, expanding the query, or using phrases), is often used to bridge the vocabulary gap between the query and the document collection. Many query expansion techniques have been proposed, and they mostly fall into two categories, i.e., global analysis and local analysis. The idea of *global* analysis is to expand the query using global collection statistics based, for instance, on a co-occurrence analysis of the entire collection. Thesaurus- and dictionary-based expansion as, e.g., in Qiu and Frei (1993), also provide examples of the global approach.

Our focus in this paper is on *local* approaches to query expansion, that use the top retrieved documents as examples from which to select terms to improve the retrieval performance (Rocchio, 1971). In the setting of language modeling approaches to query expansion, the local analysis idea has been instantiated by estimating additional query language models (Lafferty and Zhai, 2003; Tao and Zhai, 2006) or relevance models (Lavrenko and Croft, 2001) from a set of feedback documents. Yan and Hauptmann (2007) explore query expansion in a multimedia setting.

Balog et al. (2008b) compare methods for sampling expansion terms to support query-dependent and query-independent query expansion; the latter is motivated by the wish to increase "aspect recall" and attempts to uncover aspects of the information need not captured by the query. Kurland et al. (2005) also try to uncover multiple aspects of a query, and to that they provide an iterative "pseudo-query" generation technique, using cluster-based language models. The notion of "aspect recall" is mentioned in (Buckley, 2004; Harman and Buckley, 2004) and identified as one of the main reasons of failure of the current information retrieval systems. Even though we acknowledge the possibilities of our approach in improving aspect recall, by introducing aspects mainly covered by the external collection being used, we are currently unable to test this assumption.

2.2 External Expansion

The use of external collections for query expansion has a long history, see, e.g., (Kwok et al., 2001; Sakai, 2002). Diaz and Metzler (2006) were the first to give a systematic account of query expansion using an external corpus in a language modeling setting, to improve the estimation of relevance models. As will become clear in §4, Diaz and Metzler's approach is an instantiation of our general model for external expansion.

Typical query expansion techniques, such as pseudo-relevance feedback, using a blog or blog post corpus do not provide significant performance improvements and often dramatically hurt performance. For this reason, query expansion using external corpora has been a popular technique at the TREC Blog track (Ounis et al., 2007). For blog post retrieval, several TREC participants have experimented with expansion against external corpora, usually a news corpus, Wikipedia, the web, or a mixture of these (Zhang and Yu, 2007; Java et al., 2007; Ernsting et al., 2008). For the blog finding task introduced in 2007, TREC participants again used expansion against an external corpus, usually Wikipedia (Elsas et al., 2008a; Ernsting et al., 2008; Balog et al., 2008a; Fautsch and Savoy, 2008; Arguello et al., 2008b). The motivation underlying most of these approaches is to improve the estimation of the query representation, often trying to make up for the unedited nature of the corpus from which posts or blogs need to be retrieved. Elsas et al. (2008b) go a step further and develop a query expansion technique using the links in Wikipedia.

Finally, Weerkamp and de Rijke (2008b) study external expansion in the setting of blog retrieval to uncover additional perspectives of a given topic. We are driven by the same motivation, but where they considered rank-based result combinations and simple mixtures of query models, we take a more principled and structured approach, and develop four versions of a generative model for query expansion using external collections.

3 Retrieval Framework

We work in the setting of generative language models. Here, one usually assumes that a document's relevance is correlated with query likelihood (Ponte and Croft, 1998; Miller et al., 1999; Hiemstra, 2001). Within the language modeling approach, one builds a language model from each document, and ranks documents based on the probability of the document model generating the query. The particulars of the language modeling approach have been discussed extensively in the literature (see, e.g., Balog et al. (2008b)) and will not be repeated here. Our final formula for ranking documents given a query is based on Eq. 1:

$$\begin{aligned}\log P(D|Q) \propto \\ \log P(D) + \sum_{t \in Q} P(t|\theta_Q) \log P(t|\theta_D) \quad (1)\end{aligned}$$

Here, we see the prior probability of a document being relevant, $P(D)$ (which is independent of the query Q), the probability of a term t for a given query model, θ_Q, and the probability of observing the term t given the document model, θ_D. Our main interest lies in in obtaining a better estimate of $P(t|\theta_Q)$. To this end, we take the query model to be a linear combination of the maximum-likelihood query estimate $P(t|Q)$ and an expanded query model $P(t|\hat{Q})$:

$$P(t|\theta_Q) = \lambda_Q \cdot P(t|Q) + (1 - \lambda_Q) \cdot P(t|\hat{Q}) \quad (2)$$

In the next section we introduce our models for estimating $p(t|\hat{Q})$, i.e., query expansion using (multiple) external collections.

4 Query Modeling Approach

Our goal is to build an expanded query model that combines evidence from multiple external collections. We estimate the probability of a term t in the expanded query $\hat{Q}$ using a mixture of collection-specific query expansion models.

$$P(t|\hat{Q}) = \textstyle\sum_{c \in C} P(t|Q,c) \cdot P(c|Q), \quad (3)$$

where C is the set of document collections.

To estimate the probability of a term given the query and the collection, $P(t|Q,c)$, we compute the expectation over the documents in the collection c:

$$P(t|Q,c) = \sum_{D \in c} P(t|Q,c,D) \cdot P(D|Q,c). \quad (4)$$

Substituting Eq. 4 back into Eq. 3 we get

$$\begin{aligned}&P(t|\hat{Q}) = \quad (5)\\ &\sum_{c \in C} P(c|Q) \cdot \sum_{D \in c} P(t|Q,c,D) \cdot P(D|Q,c).\end{aligned}$$

This, then, is our query model for combining evidence from multiple sources.

The following subsections introduce four instances of the general external expansion model (EEM) we proposed in this section; each of the instances differ in independence assumptions:

- EEM1 (§4.1) assumes collection c to be independent of query Q and document D jointly, and document D individually, but keeps the dependence on Q and of t and Q on D.
- EEM2 (§4.2) assumes that term t and collection c are conditionally independent, given document D and query Q; moreover, D and Q are independent given c but the dependence of t and Q on D is kept.
- EEM3 (§4.3) assumes that expansion term t and original query Q are independent given document D.
- On top of EEM3, EEM4 (§4.4) makes one more assumption, viz. the dependence of collection c on query Q.

4.1 External Expansion Model 1 (EEM1)

Under this model we assume collection c to be independent of query Q and document D jointly, and document D individually, but keep the dependence on Q. We rewrite $P(t|Q,c)$ as follows:

$$\begin{aligned}&P(t|Q,c)\\ &\quad = \sum_{D \in c} P(t|Q,D) \cdot P(t|c) \cdot P(D|Q)\\ &\quad = \sum_{D \in c} \frac{P(t,Q|D)}{P(Q|D)} \cdot P(t|c) \cdot \frac{P(Q|D)P(D)}{P(Q)}\\ &\quad \propto \sum_{D \in c} P(t,Q|D) \cdot P(t|c) \cdot P(D) \quad (6)\end{aligned}$$

Note that we drop $P(Q)$ from the equation as it does not influence the ranking of terms for a given query Q. Further, $P(D)$ is the prior probability of a document, regardless of the collection it appears in (as we assumed D to be independent of c). We assume $P(D)$ to be uniform, leading to the following equation for ranking expansion terms:

$$P(t|\hat{Q}) \propto \sum_{c \in C} P(t|c) \cdot P(c|Q) \cdot \sum_{D \in c} P(t,Q|D). \quad (7)$$

In this model we capture the probability of the expansion term given the collection ($P(t|c)$). This allows us to assign less weight to terms that are less meaningful in the external collection.

4.2 External Expansion Model 2 (EEM2)

Here, we assume that term t and collection c are conditionally independent, given document D and query Q: $P(t|Q,c,D) = P(t|Q,D)$. This leaves us with the following:

$$\begin{aligned} P(t|Q,D) &= \frac{P(t,Q,D)}{P(Q,D)} \\ &= \frac{P(t,Q|D) \cdot P(D)}{P(Q|D) \cdot P(D)} \\ &= \frac{P(t,Q|D)}{P(Q|D)} \end{aligned} \quad (8)$$

Next, we assume document D and query Q to be independent given collection c: $P(D|Q,c) = P(D|c)$. Substituting our choices into Eq. 4 gives us our second way of estimating $P(t|Q,c)$:

$$P(t|Q,c) = \sum_{D \in c} \frac{P(t,Q|D)}{P(Q|D)} \cdot P(D|c) \quad (9)$$

Finally, we put our choices so far together, and implement Eq. 9 in Eq. 3, yielding our final term ranking equation:

$$P(t|\hat{Q}) \propto \sum_{c \in C} P(c|Q) \cdot \sum_{D \in c} \frac{P(t,Q|D)}{P(Q|D)} \cdot P(D|c). \quad (10)$$

4.3 External Expansion Model 3 (EEM3)

Here we assume that expansion term t and both collection c and original query Q are independent given document D. Hence, we set $P(t|Q,c,D) = P(t|D)$. Then

$$\begin{aligned} P(t|Q,c) &= \sum_{D \in c} P(t|D) \cdot P(D|Q,c) \\ &= \sum_{D \in c} P(t|D) \cdot \frac{P(Q|D,c) \cdot P(D|c)}{P(Q|c)} \\ &\propto \sum_{D \in c} P(t|D) \cdot P(Q|D,c) \cdot P(D|c) \end{aligned}$$

We dropped $P(Q|c)$ as it does not influence the ranking of terms for a given query Q. Assuming independence of Q and c given D, we obtain

$$P(t|Q,c) \propto \sum_{D \in c} P(D|c) \cdot P(t|D) \cdot P(Q|D)$$

so

$$P(t|\hat{Q}) \propto \sum_{c \in C} P(c|Q) \cdot \sum_{D \in c} P(D|c) \cdot P(t|D) \cdot P(Q|D).$$

We follow Lavrenko and Croft (2001) and assume that $P(D|c) = \frac{1}{|\mathcal{R}_c|}$, the size of the set of top ranked documents in c (denoted by $\mathcal{R}_c$), finally arriving at

$$P(t|\hat{Q}) \propto \sum_{c \in C} \frac{P(c|Q)}{|\mathcal{R}_c|} \cdot \sum_{D \in \mathcal{R}_c} P(t|D) \cdot P(Q|D). \quad (11)$$

4.4 External Expansion Model 4 (EEM4)

In this fourth model we start from EEM3 and drop the assumption that c depends on the query Q, i.e., $P(c|Q) = P(c)$, obtaining

$$P(t|\hat{Q}) \propto \sum_{c \in C} \frac{P(c)}{|\mathcal{R}_c|} \cdot \sum_{D \in \mathcal{R}_c} P(t|D) \cdot P(Q|D). \quad (12)$$

Eq. 12 is in fact the "mixture of relevance models" external expansion model proposed by Diaz and Metzler (2006). The fundamental difference between EEM1, EEM2, EEM3 on the one hand and EEM4 on the other is that EEM4 assumes independence between c and Q (thus $P(c|Q)$ is set to $P(c)$). That is, the importance of the external collection is independent of the query. How reasonable is this choice? Mishne and de Rijke (2006) examined queries submitted to a blog search engine and found many to be either news-related *context* queries (that aim to track mentions of a named entity) or *concept* queries (that seek posts about a general topic). For context queries such as *cheney hunting* (TREC topic 867) a news collection is likely to offer different (relevant) aspects of the topic, whereas for a concept query such as *jihad* (TREC topic 878) a knowledge source such as Wikipedia seems an appropriate source of terms that capture aspects of the topic. These observations suggest the collection should depend on the query.

EEM3 and EEM4 assume that expansion term t and original query Q are independent given document D. This may or may not be too strong an assumption. Models EEM1 and EEM2 also make independence assumptions, but weaker ones.

5 Estimating Components

The models introduced above offer us several choices in estimating the main components. Below we detail how we estimate (i) $P(c|Q)$, the importance of a collection for a given query, (ii) $P(t|c)$, the *un*importance of a term for an external collection, (iii) $P(Q|D)$, the relevance of a document in the external collection for a given query, and (iv) $P(t,Q|D)$, the likelihood of a term co-occurring with the query, given a document.

5.1 Importance of a Collection

Represented as $P(c|Q)$ in our models, the importance of an external collection depends on the query; how we can estimate this term? We consider three alternatives, in terms of (i) query clarity, (ii) coherence and (iii) query-likelihood, using documents in that collection.

First, query clarity measures the structure of a set of documents based on the assumption that a small number of topical terms will have unusually large probabilities (Cronen-Townsend et al., 2002). We compute the query clarity of the top ranked documents in a given collection c:

$$clarity(Q,c) = \sum_t P(t|Q) \cdot \log \frac{P(t|Q)}{P(t|\mathcal{R}_c)}$$

Finally, we normalize $clarity(Q,c)$ over all collections, and set $P(c|Q) \propto \frac{clarity(Q,c)}{\sum_{c' \in C} clarity(Q,c')}$.

Second, a measure called "coherence score" is defined by He et al. (2008). It is the fraction of "coherent" pairs of documents in a given set of documents, where a coherent document pair is one whose similarity exceeds a threshold. The coherence of the top ranked documents $\mathcal{R}_c$ is:

$$Co(\mathcal{R}_c) = \frac{\sum_{i \neq j \in \{1,\ldots,|\mathcal{R}_c|\}} \delta(d_i,d_j)}{|\mathcal{R}_c|(|\mathcal{R}_c|-1)},$$

where $\delta(d_i,d_j)$ is 1 in case of a similar pair (computed using cosine similarity), and 0 otherwise. Finally, we set $P(c|Q) \propto \frac{Co(\mathcal{R}_c)}{\sum_{c' \in C} Co(\mathcal{R}_{c'})}$.

Third, we compute the conditional probability of the collection using Bayes' theorem. We observe that $P(c|Q) \propto P(Q|c)$ (omitting $P(Q)$ as it will not influence the ranking and $P(c)$ which we take to be uniform). Further, for the sake of simplicity, we assume that all documents within c are equally important. Then, $P(Q|c)$ is estimated as

$$P(Q|c) = \frac{1}{|c|} \cdot \sum_{D \in c} P(Q|D) \qquad (13)$$

where $P(Q|D)$ is estimated as described in §5.3, and $|c|$ is the number of documents in c.

5.2 Unimportance of a Term

Rather than simply estimating the importance of a term for a given query, we also estimate the *un*importance of a term for a collection; i.e., we assign lower probability to terms that are common in that collection. Here, we take a straightforward approach in estimating this, and define $P(t|c) = 1 - \frac{n(t,c)}{\sum_{t'} n(t',c)}$.

5.3 Likelihood of a Query

We need an estimate of the probability of a query given a document, $P(Q|D)$. We do so by using Hauff et al. (2008)'s refinement of term dependencies in the query as proposed by Metzler and Croft (2005).

5.4 Likelihood of a Term

Estimating the likelihood of observing both the query and a term for a given document $P(t,Q|D)$ is done in a similar way to estimating $P(Q|D)$, but now for t, Q in stead of Q.

6 Experimental Setup

In his section we detail our experimental setup: the (external) collections we use, the topic sets and relevance judgements available, and the significance testing we perform.

6.1 Collections and Topics

We make use of three collections: (i) a collection of user generated documents (blog posts), (ii) a news collection, and (iii) an online knowledge source. The blog post collection is the TREC Blog06 collection (Ounis et al., 2007), which contains 3.2 million blog posts from 100,000 blogs monitored for a period of 11 weeks, from December 2005 to March 2006; all posts from this period have been stored as HTML files. Our news collection is the AQUAINT-2 collection (AQUAINT-2, 2007), from which we selected news articles that appeared in the period covered by the blog collection, leaving us with about 150,000 news articles. Finally, we use a dump of the English Wikipedia from August 2007 as our online knowledge source; this dump contains just over 3.8 million encyclopedia articles.

During 2006–2008, the TRECBlog06 collection has been used for the topical blog post retrieval task (Weerkamp and de Rijke, 2008a) at the TREC Blog track (Ounis et al., 2007): to retrieve posts about a given topic. For every year, 50 topics were developed, consisting of a title field, description, and narrative; we use only the title field, and ignore the other available information. For all 150 topics relevance judgements are available.

6.2 Metrics and Significance

We report on the standard IR metrics Mean Average Precision (MAP), precision at 5 and 10 documents (P5, P10), and the Mean Reciprocal Rank (MRR). To determine whether or not differences between runs are significant, we use a two-tailed paired t-test, and report on significant differences for $\alpha = .05$ ($^\vartriangle$ and $^\triangledown$) and $\alpha = .01$ ($^\blacktriangle$ and $^\blacktriangledown$).

7 Results

We first discuss the parameter tuning for our four EEM models in Section 7.1. We then report on the results of applying these settings to obtain our retrieval results on the blog post retrieval task. Section 7.2 reports on these results. We follow with a closer look in Section 8.

7.1 Parameters

Our model has one explicit parameter, and one more or less implicit parameter. The obvious parameter is λ_Q, used in Eq. 2, but also the number of terms to include in the final query model makes a difference. For training of the parameters we use two TREC topic sets to train and test on the held-out topic set. From the training we conclude that the following parameter settings work best across all topics: (EEM1) $\lambda_Q = 0.6$, 30 terms; (EEM2) $\lambda_Q = 0.6$, 40 terms; (EEM3 and EEM4) $\lambda_Q = 0.5$, 30 terms. In the remainder of this section, results for our models are reported using these parameter settings.

7.2 Retrieval Results

As a baseline we use an approach without external query expansion, viz. Eq. 1. In Table 1 we list the results on the topical blog post finding task of (i) our baseline, and (ii) our model (instantiated by EEM1, EEM2, EEM3, and EEM4). For all models that contain the query-dependent collection probability ($P(c|Q)$) we report on multiple ways of estimating this: (i) uniform, (ii) best global mixture (independent of the query, obtained by a sweep over collection probabilities), (iii) coherence, (iv) query clarity, (v) $P(Q|c)$, and (vi) using an oracle for which optimal settings were obtained by the same sweep as (ii). Note that methods (i) and (ii) are not query dependent; for EEM3 we do not mention (ii) since it equals (i). Finally, for EEM4 we only have a query-independent component, $P(c)$: the best performance here is obtained using equal weights for both collections.

model	$P(c\|Q)$	MAP	P5	P10	MRR
Baseline		0.3815	0.6813	0.6760	0.7643
EEM1	uniform	0.3976$^\blacktriangle$	0.7213$^\blacktriangle$	0.7080$^\blacktriangle$	0.7998
	0.8N/0.2W	0.3992	0.7227	0.7107	0.7988
	coherence	0.3976	0.7187	0.7060	0.7976
	query clarity	0.3970	0.7187	0.7093	0.7929
	$P(Q\|c)$	0.3983	0.7267	0.7093	0.7951
	oracle	0.4126$^\blacktriangle$	0.7387$^\vartriangle$	0.7320$^\blacktriangle$	0.8252$^\vartriangle$
EEM2	uniform	0.3885$^\blacktriangle$	0.7053$^\vartriangle$	0.6967$^\vartriangle$	0.7706
	0.9N/0.1W	0.3895	0.7133	0.6953	0.7736
	coherence	0.3890	0.7093	0.7020	0.7740
	query clarity	0.3872	0.7067	0.6953	0.7745
	$P(Q\|c)$	0.3883	0.7107	0.6967	0.7717
	oracle	0.3995$^\blacktriangle$	0.7253$^\blacktriangle$	0.7167$^\blacktriangle$	0.7856
EEM3	uniform	0.4048$^\blacktriangle$	0.7187$^\vartriangle$	0.7207$^\blacktriangle$	0.8261$^\blacktriangle$
	coherence	0.4058	0.7253	0.7187	0.8306
	query clarity	0.4033	0.7253	0.7173	0.8228
	$P(Q\|c)$	0.3998	0.7253	0.7100	0.8133
	oracle	**0.4194**$^\blacktriangle$	**0.7493**$^\blacktriangle$	**0.7353**$^\blacktriangle$	**0.8413**
EEM4	0.5N/0.5W	0.4048$^\blacktriangle$	0.7187$^\vartriangle$	0.7207$^\blacktriangle$	0.8261$^\blacktriangle$

Table 1: Results for all model instances on all topics (i.e., 2006, 2007, and 2008); aN/bW stands for the weights assigned to the news (a) and Wikipedia corpora (b). Significance is tested between (i) each uniform run and the baseline, and (ii) each other setting and its uniform counterpart.

A few observations. First, our baseline performs well above the median for all three years (2006–2008). Second, in each of its four instances our model for query expansion against external corpora improves over the baseline. Third, we see that it is safe to assume that a term is dependent only on the document from which it is sampled (EEM1 vs. EEM2 vs. EEM3). EEM3 makes the strongest assumptions about terms in this respect, yet it performs best. Fourth, capturing the dependence of the collection on the query helps, as we can see from the significant improvements of the "oracle" runs over their "uniform" counterparts. However, we do not have a good method yet for automatically estimating this dependence, as is clear from the insignificant differences between the runs labeled "coherence," "query clarity," "$P(Q|c)$" and the run labeled "uniform."

8 Discussion

Rather than providing a pairwise comparison of all runs listed in the previous section, we consider two pairwise comparisons—between (an instantion of) our model and the baseline, and between two instantiations of our model—and highlight phenomena that we also observed in other pairwise comparisons. Based on this discussion, we also consider a combination of approaches.

8.1 EEM1 vs. the Baseline

We zoom in on EEM1 and make a per-topic comparison against the baseline. First of all, we observe behavior typical for all query expansion methods: some topics are helped, some are not affected, and some are hurt by the use of EEM1; see Figure 1, top row. Specifically, 27 topics show a slight drop in AP (maximum drop is 0.043 AP), 3 topics do not change (as no expansion terms are identified) and the remainder of the topics (120) improve in AP. The maximum increase in AP is 0.5231 (+304%) for topic 949 (*ford bell*); Topics 887 (*world trade organization*, +87%), 1032 (*I walk the line*, +63%), 865 (*basque*, +53%), and 1014 (*tax break for hybrid automobiles*, +50%) also show large improvements. The largest drop (-20% AP) is for topic 1043 (*a million little pieces*, a controversial memoir that was in the news during the time coverd by the blog crawl); because we do not do phrase or entity recognition in the query, but apply stopword removal, it is reduced to *million pieces* which introduced a lot of topic drift.

Let us examine the "collection preference" of topics: 35 had a clear preference for Wikipedia, 32 topics for news, and the remainder (83 topics) required a mixture of both collections. First, we look at topics that require equal weights for both collections; topic 880 (*natalie portman*, +21% AP) concerns a celebrity with a large Wikipedia biography, as well as news coverage due to new movie releases during the period covered by the blog crawl. Topic 923 (*challenger*, +7% AP) asks for information on the space shuttle that exploded during its launch; the 20th anniversary of this event was commemorated during the period covered by the crawl and therefore it is newsworthy as well as present in Wikipedia (due to its historic impact). Finally, topic 869 (*muhammad cartoon*, +20% AP) deals with the controversy surrounding the publication of cartoons featuring Muhammad: besides its obvious news impact, this event is extensively discussed in multiple Wikipedia articles.

As to topics that have a preference for Wikipedia, we see some very general ones (as is to be expected): Topic 942 (*lawful access*, +30% AP) on the government accessing personal files; Topic 1011 (*chipotle restaurant*, +13% AP) on information concerning the Chipotle restaurants; Topic 938 (*plug awards*, +21% AP) talks about an award show. Although this last topic could be expected to have a clear preference for expansion terms from the news corpus, the awards were not handed out during the period covered by the news collection and, hence, full weight is given to Wikipedia.

At the other end of the scale, topics that show a preference for the news collection are topic 1042 (*david irving*, +28% AP), who was on trial during the period of the crawl for denying the Holocaust and received a lot of media attention. Further examples include Topic 906 (*davos*, +20% AP), which asks for information on the annual world economic forum meeting in Davos in January, something typically related to news, and topic 949 (*ford bell*, +304% AP), which seeks information on Ford Bell, Senate candidate at the start of 2006.

8.2 EEM1 vs. EEM3

Next we turn to a comparison between EEM1 and EEM3. Theoretically, the main difference between these two instantiations of our general model is that EEM3 makes much stronger simplifying indepence assumptions than EEM1. In Figure 1 we compare the two, not only against the baseline, but, more interestingly, also in terms of the difference in performance brought about by switching from uniform estimation of $P(c|Q)$ to oracle estimation. Most topics gain in AP when going from the uniform distribution to the oracle setting. This happens for both models, EEM1 and EEM3, leading to less topics decreasing in AP over the baseline (the right part of the plots) and more topics increasing (the left part). A second observation is that both gains and losses are higher for EEM3 than for EEM1.

Zooming in on the differences between EEM1 and EEM3, we compare the two in the same way, now using EEM3 as "baseline" (Figure 2). We observe that EEM3 performs better than EEM1 in 87

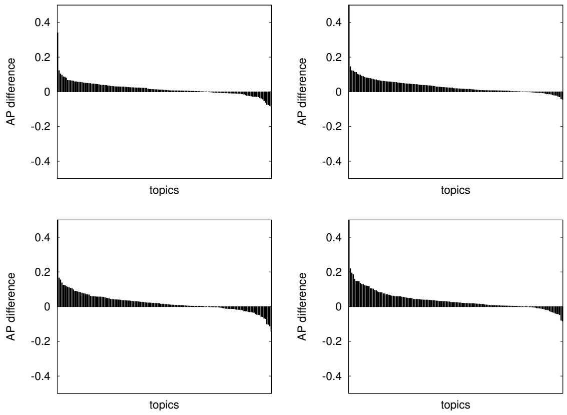

Figure 1: Per-topic AP differences between the baseline and (Top): EEM1 and (Bottom): EEM3, for (Left): uniform $P(c|Q)$ and (Right): oracle.

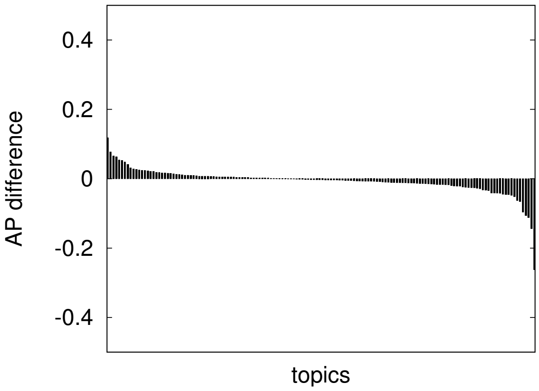

Figure 2: Per-topic AP differences between EEM3 and EEM1 in the oracle setting.

cases, while EEM1 performs better for 60 topics. Topics 1041 (*federal shield law*, 47% AP), 1028 (*oregon death with dignity act*, 32% AP), and 1032 (*I walk the line*, 32% AP) have the highest difference in favor of EEM3; Topics 877 (*sonic food industry*, 139% AP), 1013 (*iceland european union*, 25% AP), and 1002 (*wikipedia primary source*, 23% AP) are helped most by EEM1. Overall, EEM3 performs significantly better than EEM1 in terms of MAP (for $\alpha = .05$), but not in terms of the early precision metrics (P5, P10, and MRR).

8.3 Combining Our Approaches

One observation to come out of §8.1 and 8.2 is that different topics prefer not only different external expansion corpora but also different external expansion methods. To examine this phenomemon, we created an articificial run by taking, for every topic, the best performing model (with settings optimized for the topic). Twelve topics preferred the baseline, 37 EEM1, 20 EEM2, and 81 EEM3. The articifical run produced the following results: MAP 0.4280, P5 0.7600, P10 0.7480, and MRR 0.8452; the differences in MAP and P10 between this run and EEM3 are significant for $\alpha = .01$. We leave it as future work to (learn to) predict for a given topic, which approach to use, thus refining ongoing work on query difficulty prediction.

9 Conclusions

We explored the use of external corpora for query expansion in a user generated content setting. We introduced a general external expansion model, which offers various modeling choices, and instantiated it based on different (in)dependence assumptions, leaving us with four instances.

Query expansion using external collection is effective for retrieval in a user generated content setting. Furthermore, conditioning the collection on the query is beneficial for retrieval performance, but estimating this component remains difficult. Dropping the dependencies between terms and collection and terms and query leads to better performance. Finally, the best model is topic-dependent: constructing an artificial run based on the best model per topic achieves significant better results than any of the individual models.

Future work focuses on two themes: (i) topic-dependent model selection and (ii) improved estimates of components. As to (i), we first want to determine whether a query should be expanded, and next select the appropriate expansion model. For (ii), we need better estimates of $P(Q|c)$; one aspect that could be included is taking $P(c)$ into account in the query-likelihood estimate of $P(Q|c)$. One can make this dependent on the task at hand (blog post retrieval vs. blog feed search). Another possibility is to look at solutions used in distributed IR. Finally, we can also include the estimation of $P(D|c)$, the importance of a document in the collection.

Acknowledgements

We thank our reviewers for their valuable feedback. This research is supported by the DuOMAn project carried out within the STEVIN programme which is funded by the Dutch and Flemish Governments (http://www.stevin-tst.org) under project number STE-09-12, and by the Netherlands Organisation for Scientific Research (NWO) under project numbers 017.001.190, 640.001.501, 640.002.501, 612.066.512, 612.061.814, 612.061.815, 640.004.802.

References

AQUAINT-2 (2007). URL: `http://trec.nist.gov/data/qa/2007_qadata/qa.07.guidelines.html#documents`.

Arguello, J., Elsas, J., Callan, J., and Carbonell, J. (2008a). Document representation and query expansion models for blog recommendation. In *Proceedings of ICWSM 2008*.

Arguello, J., Elsas, J. L., Callan, J., and Carbonell, J. G. (2008b). Document representation and query expansion models for blog recommendation. In *Proc. of the 2nd Intl. Conf. on Weblogs and Social Media (ICWSM)*.

Baeza-Yates, R. and Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. ACM.

Balog, K., Meij, E., Weerkamp, W., He, J., and de Rijke, M. (2008a). The University of Amsterdam at TREC 2008: Blog, Enterprise, and Relevance Feedback. In *TREC 2008 Working Notes*.

Balog, K., Weerkamp, W., and de Rijke, M. (2008b). A few examples go a long way: constructing query models from elaborate query formulations. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 371–378, New York, NY, USA. ACM.

Buckley, C. (2004). Why current IR engines fail. In *SIGIR '04*, pages 584–585.

Cronen-Townsend, S., Zhou, Y., and Croft, W. B. (2002). Predicting query performance. In *SIGIR02*, pages 299–306.

Diaz, F. and Metzler, D. (2006). Improving the estimation of relevance models using large external corpora. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 154–161, New York, NY, USA. ACM.

Elsas, J., Arguello, J., Callan, J., and Carbonell, J. (2008a). Retrieval and feedback models for blog distillation. In *The Sixteenth Text REtrieval Conference (TREC 2007) Proceedings*.

Elsas, J. L., Arguello, J., Callan, J., and Carbonell, J. G. (2008b). Retrieval and feedback models for blog feed search. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 347–354, New York, NY, USA. ACM.

Ernsting, B., Weerkamp, W., and de Rijke, M. (2008). Language modeling approaches to blog post and feed finding. In *The Sixteenth Text REtrieval Conference (TREC 2007) Proceedings*.

Fautsch, C. and Savoy, J. (2008). UniNE at TREC 2008: Fact and Opinion Retrieval in the Blogsphere. In *TREC 2008 Working Notes*.

Harman, D. and Buckley, C. (2004). The NRRC reliable information access (RIA) workshop. In *SIGIR '04*, pages 528–529.

Hauff, C., Murdock, V., and Baeza-Yates, R. (2008). Improved query difficulty prediction for the web. In *CIKM '08: Proceedings of the seventeenth ACM conference on Conference on information and knowledge management*, pages 439–448.

He, J., Larson, M., and de Rijke, M. (2008). Using coherence-based measures to predict query difficulty. In *30th European Conference on Information Retrieval (ECIR 2008)*, page 689694. Springer, Springer.

Hiemstra, D. (2001). *Using Language Models for Information Retrieval*. PhD thesis, University of Twente.

Java, A., Kolari, P., Finin, T., Joshi, A., and Martineau, J. (2007). The blogvox opinion retrieval system. In *The Fifteenth Text REtrieval Conference (TREC 2006) Proceedings*.

Kurland, O., Lee, L., and Domshlak, C. (2005). Better than the real thing?: Iterative pseudo-query processing using cluster-based language models. In *SIGIR '05*, pages 19–26.

Kwok, K. L., Grunfeld, L., Dinstl, N., and Chan, M. (2001). TREC-9 cross language, web and question-answering track experiments using PIRCS. In *TREC-9 Proceedings*.

Lafferty, J. and Zhai, C. (2003). Probabilistic relevance models based on document and query generation. In *Language Modeling for Information Retrieval*, Kluwer International Series on Information Retrieval. Springer.

Lavrenko, V. and Croft, W. B. (2001). Relevance based language models. In *SIGIR '01*, pages 120–127.

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

Metzler, D. and Croft, W. B. (2005). A markov random field model for term dependencies. In *SIGIR '05*, pages 472–479, New York, NY, USA. ACM.

Miller, D., Leek, T., and Schwartz, R. (1999). A hidden Markov model information retrieval system. In *SIGIR '99*, pages 214–221.

Mishne, G. and de Rijke, M. (2006). A study of blog search. In Lalmas, M., MacFarlane, A., Rüger, S., Tombros, A., Tsikrika, T., and Yavlinsky, A., editors, *Advances in Information Retrieval: Proceedings 28th European Conference on IR Research (ECIR 2006)*, volume 3936 of *LNCS*, pages 289–301. Springer.

Ounis, I., Macdonald, C., de Rijke, M., Mishne, G., and Soboroff, I. (2007). Overview of the TREC 2006 Blog Track. In *The Fifteenth Text Retrieval Conference (TREC 2006)*. NIST.

Ponte, J. M. and Croft, W. B. (1998). A language modeling approach to information retrieval. In *SIGIR '98*, pages 275–281.

Qiu, Y. and Frei, H.-P. (1993). Concept based query expansion. In *SIGIR '93*, pages 160–169.

Rocchio, J. (1971). Relevance feedback in information retrieval. In *The SMART Retrieval System: Experiments in Automatic Document Processing*. Prentice Hall.

Sakai, T. (2002). The use of external text data in cross-language information retrieval based on machine translation. In *Proceedings IEEE SMC 2002*.

Tao, T. and Zhai, C. (2006). Regularized estimation of mixture models for robust pseudo-relevance feedback. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 162–169, New York, NY, USA. ACM.

Weerkamp, W. and de Rijke, M. (2008a). Credibility improves topical blog post retrieval. In *ACL-08: HLT*, pages 923–931.

Weerkamp, W. and de Rijke, M. (2008b). Looking at things differently: Exploring perspective recall for informal text retrieval. In *8th Dutch-Belgian Information Retrieval Workshop (DIR 2008)*, pages 93–100.

Yan, R. and Hauptmann, A. (2007). Query expansion using probabilistic local feedback with application to multimedia retrieval. In *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 361–370, New York, NY, USA. ACM.

Zhang, W. and Yu, C. (2007). UIC at TREC 2006 Blog Track. In *The Fifteenth Text REtrieval Conference (TREC 2006) Proceedings*.