External Query Expansion in the Blogosphere

Wouter Weerkamp Maarten de Rijke

ISLA, University of Amsterdam
http://ilps.science.uva.nl/

Abstract: We describe the participation of the University of Amsterdam's ILPS group in the blog track at TREC 2008. We mainly explored different ways of using external corpora to expand the original query. In the blog post retrieval task we did not succeed in improving over a simple baseline (equal weights for both the expanded and original query). Obtaining optimal weights for the original and the expanded query remains a subject of investigation. In the blog distillation task we tried to improve over our (strong) baseline using external expansion, but due to differences in the run setup, comparing these runs is hard. Compared to a simpler baseline, we see an improvement for the run using external expansion on the combination of news, Wikipedia and blog posts.

1 Introduction

We describe our participation in this year's TREC Blog track. Like last year, the blog track consists of two separate tasks: *blog post retrieval* and *blog distillation*. Besides the task of finding topically relevant blog posts, the blog post retrieval task has two further tasks: finding blog posts that contain an opinion on the given topic and determining the polarity of the opinion. To test the opinion-ranking capabilities of participants' systems, participants were asked to rerank five baseline runs based on opinionatedness, besides submitting four full opinion retrieval runs. Our main interest this year lies with the topical retrieval of both blog posts and blogs. We did not participate in the polarity determination and only submitted very basic opinion finding runs.

The remainder of this paper introduces our retrieval approaches for both tasks in Section 2, and explains the way we incorporated external sources in query modeling in Section 3. We then zoom in on the runs for both tasks and their results: post retrieval in Section 5 and blog distillation in Section 6. Finally, we conclude in Section 7.

2 Retrieval Approaches

In the blog post retrieval task we use an out-of-the-box implementation of Indri.[1] Results of previous years showed good overall performance of Indri compared to other systems and besides, it allows for easy use of query models (queries consisting of weighted terms).

In the blog distillation task we use our in-house expert retrieval model (Balog et al., 2006), which we translated to fit the task of blogger retrieval (Balog et al., 2008; Weerkamp et al., 2008). The main reason for using this model is that we believe blog distillation should be solved using a post index (as opposed to a full blog index). Although last year's blog track showed good performance of blog indexes, we stick to a post index for three reasons: (i) a post index allows for easy incremental updating, (ii) posts are a natural unit for result presentation to the user, and most importantly, (iii) only one index is needed for both post retrieval and blog distillation.

We estimate the probability of a blog *blog* generating query Q as follows:

$$P(Q|\theta_{blog}) = \prod_{t \in Q} P(t|\theta_{blog})^{n(t,Q)}. \quad (1)$$

Next, we smooth the probability of a term given a blog with the background probabilities:

$$P(t|\theta_{blog}) = (1-\lambda_{blog}) \cdot P(t|blog) + \lambda_{blog} \cdot P(t). \quad (2)$$

Finally, we estimate $P(t|blog)$ as follows:

$$P(t|blog) = \sum_{post \in blog} P(t|post, blog) \cdot P(post|blog). \quad (3)$$

We assume that the post and the blog are conditionally independent, thus $P(t|post, blog) = P(t|post)$, and approximate $P(t|post)$ with the standard maximum likelihood estimate. In Section 6 we detail our choices for estimating $p(post|blog)$.

3 Query Modeling

For both tasks we experimented with query models using external corpora. In short, we assume that documents in the

[1] http://www.lemurproject.org/indri

target collection (the blog collection) are too noisy to generate good query models based on blind relevance feedback. Instead, we use different, less noisy external corpora for expanding our original query. As much of what goes on in the blogosphere is determined by news events, we use a contemporary news corpus AQUAINT-2[2] as our external corpus. Besides this, many queries directed towards blogs and blog posts contain named entities (persons, locations, organizations, products) or general concepts (especially in blog distillation). For this we also look at Wikipedia as an external corpus, since this source contains focused information on many general concepts and named entities.

For two post retrieval runs we use Lavrenko's relevance model 2 (Lavrenko and Croft, 2001) to select the top 10 terms from the top 10 external documents. After selecting weighted new terms, we construct the final query model $P(t|\theta_Q)$ by combining this new query $P(t|\hat{\theta}_Q)$ with the original query $P(t|Q)$ using:

$$P(t|\theta_Q) = \lambda_Q P(t|\hat{\theta}_Q) + (1-\lambda_Q)P(t|Q) \quad (4)$$

In two opinion retrieval runs and two blog distillation runs we use an experimental approach to query expansion. We estimate the probability of a expansion term t given the query Q and set of external corpora C:

$$P(t|Q,C) = \sum_{c \in C} \frac{P(t|c,Q) \cdot P(c|Q)}{\sum_{c' \in C} p(c'|Q)} \quad (5)$$

We estimate $p(t|c,Q)$ based on the probability of document d given the query and corpus, and the probability of term t given the document:

$$P(t|c,Q) = \sum_{D \in c; P(D|Q,c)>0} P(t|D)P(D|Q,c) \quad (6)$$

Next, we estimate $P(D|Q,c)$, the probability of document D given corpus c and query Q:

$$P(D|Q,c) = \prod_{q \in Q} P(q|D) + \frac{n(Q,D) \cdot |Q|^{-1}}{|D|} \quad (7)$$

where $n(Q,D)$ is the count of phrase Q in document D and $P(q|d) = n(q,D) \cdot |D|^{-1}$. Finally, we estimate the probability of corpus c given query Q:

$$P(c|Q) = \sum_{D \in c; P(D|Q,c)>0} \frac{P(D|Q,c)}{|D \in c; p(D|Q,c) > 0|} \quad (8)$$

4 Metrics and Significance

In this paper we report on mean average precision (MAP), precision at 5 and 10 documents (P5, P10), and mean reciprocal rank (MRR). We use the Wilcoxon signed-rank test to test for significant differences between runs. We report on significant increases (or drops) for $p < .01$ using $^{\blacktriangle}$(and $^{\blacktriangledown}$) and for $p < .05$ using $^{\triangle}$(and $^{\triangledown}$).

[2] http://trec.nist.gov/data/qa/2007_qadata/qa.07.guidelines.html#documents

5 Blog Post Retrieval

As explained in the introduction to this section, we use an out-of-the-box implementation of Indri as our retrieval system. Runs are evaluated on two topic sets: the new 2008 topics alone and the full set of 150 topics (2006–2008).

We submitted 6 runs:

(A) **`uams08n1o1`** the baseline run uses a news corpus for query expansion with $\lambda_Q = 0.5$ (i.e. equal weights to expanded and original query) and assigns priors to posts based on credibility.

(B) **`uams08n1o1sp`** identical to previous run, but with "opinionatedness prior".

(C) **`uams08class`** query expansion using both a news corpus and Wikipedia; λ_Q trained on 2006 and 2007 topics, and priors based on credibility indicators.

(D) **`uams08clspr`** identical to the previous run, but with "opinionatedness prior".

(E) **`uams08qm4it1`** query expansion following Section 3 on a news corpus and Wikipedia.

(F) **`uams08qm4it2`** identical to previous run, but with the blog post corpus as additional source.

We experiment with estimating λ_Q based on old topics: for each of the old (2006/2007) topics we know the performance of various parameter settings (weights of different corpora) in terms of MAP. We use this information in the following way: for each unseen topic t' we assing a similarity score to seen topics (t) based on overlapping documents in the result lists. Next, we multiply this overlap score by the MAP performance of each mixture setting and determine the "optimal" mixture weights this way. This method is used in runs *uams08class* and *uams08clspr*.

Four of our runs (A–D) also use credibility priors: based on a combination of 6 credibility indicators (Weerkamp and de Rijke, 2008), we estimate the prior probability of the blog post being relevant. Since all runs use the same priors, we cannot determine its effectiveness here, but it has proven successful before Weerkamp and de Rijke (2008).

Looking at opinion retrieval, we explore the use of an "opinionatedness prior". To construct this prior we use strongly opinionated terms from the OpinionFinder system[3] and calculate for each post the ratio of opinionated terms to the total number of terms. We use this prior on top of our two baseline runs *uams08n1o1* and *uams08class*, to come to runs *uams08n1o1sp* and *uams08clspr*.

5.1 Results and Discussion

From the results in Tables 1 and 2 we have three initial observations: (i) The runs using the method for combining external corpora introduced in Section 3 (i.e., *uams08qm4it1* and *uams08qm4it2*) perform significantly worse than runs using

[3] http://www.cs.pitt.edu/mpqa/

Run	MAP	P5	P10	MRR
All topics				
uams08n1o1	0.3329	0.5987	0.5693	0.7309
uams08n1o1sp	**0.3351**$^{\blacktriangle}$	**0.6040**	**0.5687**	0.7275
uams08class	0.3297	0.5840	0.5660	**0.7377**
uams08clspr	0.3323$^{\blacktriangle}$	0.5853	0.5647	0.7349
uams08qm4it1	0.2633$^{\blacktriangledown}$	0.4747$^{\blacktriangledown}$	0.4620$^{\blacktriangledown}$	0.6007$^{\blacktriangledown}$
uams08qm4it2	0.1969$^{\blacktriangledown}$	0.3480$^{\blacktriangledown}$	0.3587$^{\blacktriangledown}$	0.4539$^{\blacktriangledown}$
2008 topics				
uams08n1o1	0.3797	0.7080	**0.6620**	**0.8052**
uams08n1o1sp	**0.3823**$^{\blacktriangle}$	**0.7120**	0.6580	**0.8052**
uams08class	0.3685	0.6680	0.6420	0.7852
uams08clspr	0.3715$^{\blacktriangle}$	0.6640	0.6400	0.7852
uams08qm4it1	0.2927$^{\blacktriangledown}$	0.5360$^{\blacktriangledown}$	0.5300$^{\blacktriangledown}$	0.6567$^{\blacktriangledown}$
uams08qm4it2	0.2122$^{\blacktriangledown}$	0.4120$^{\blacktriangledown}$	0.4120$^{\blacktriangledown}$	0.5431$^{\blacktriangledown}$

Table 1: Opinion results on the blog post retrieval task. Significance of *uams08clspr* and *uams08n1o1sp* tested against their baselines, other runs tested against the first run, *uams08n1o1*.

Run	MAP	P5	P10	MRR
All topics				
uams08n1o1	0.4350	**0.7680**	0.7480	**0.8464**
uams08n1o1sp	**0.4366**$^{\blacktriangle}$	0.7667	0.7473	0.8419
uams08class	0.4313	0.7507	**0.7493**	0.8439
uams08clspr	0.4332$^{\blacktriangle}$	0.7520	0.7473	0.8441
uams08qm4it1	0.3627$^{\blacktriangledown}$	0.6800$^{\blacktriangledown}$	0.6713$^{\blacktriangledown}$	0.7780$^{\blacktriangledown}$
uams08qm4it2	0.2745$^{\blacktriangledown}$	0.5760$^{\blacktriangledown}$	0.5740$^{\blacktriangledown}$	0.6869$^{\blacktriangledown}$
2008 topics				
uams08n1o1	0.4644	**0.8040**	**0.7620**	**0.8892**
uams08n1o1sp	**0.4661**$^{\blacktriangle}$	0.8000	**0.7620**	**0.8892**
uams08class	0.4494	0.7680	0.7480	0.8358
uams08clspr	0.4513$^{\blacktriangle}$	0.7720	0.7500	0.8408
uams08qm4it1	0.3734$^{\blacktriangledown}$	0.6720$^{\blacktriangledown}$	0.6600$^{\blacktriangledown}$	0.8052$^{\triangledown}$
uams08qm4it2	0.2606$^{\blacktriangledown}$	0.5480$^{\blacktriangledown}$	0.5380$^{\blacktriangledown}$	0.6981$^{\blacktriangledown}$

Table 2: Topical results on the blog post retrieval task. Significance of *uams08clspr* and *uams08n1o1sp* tested against their baselines, other runs tested against the first run, *uams08n1o1*.

relevance models and a linear combination of the expanded query and original query (*uams08n1o1* and *uams08class*). (ii) Looking at the runs using relevance models to construct query models (*uams08n1o1* and *uams08class*), we see that estimating the relative importance of the original query is not easy: the simple baseline approach ($\lambda = 0.5$) outperforms the slightly more advanced per-topic estimation. (iii) The runs using opinion priors (*uams08n1o1sp* and *uams08clspr*) significantly outperform their baseline counterparts in terms of MAP, not only on opinion retrieval, but also on topical retrieval.

6 Blog Distillation

Our blog distillation model allows for the estimation of the importance of individual posts to a blog, i.e., estimating association strengths between posts and their blog ($P(post|blog)$ in Eq. 3). Based on previous experiments (Weerkamp et al., 2008) and additional tests on the 2007 topics we use a combination of blog features to estimate this association strength: post length, recency, and number of comments. On top of this, we noticed that using information from the post title is an important indicator of relevance in the blog distillation task. To be able to use this information, we perform a linear combination between runs on the full post index and runs on a title-only index. This run is our baseline run, *uams08bl*.

We again experiment with expansion on external corpora using the novel method introduced in Section 3. In run *uams08nw* we use the news corpus and Wikipedia, in run *uams08pnw* we also use the post index as external corpus. The difference with the baseline is that we do not use the combination with the title-only index: for this submission we would like to look at the influence of the query expansion and scores of the two runs (using query expansion and the title-only run) are in a very different range, calling for other, more suitable ways of combining these scores.

The final run we submitted, *uams08nonr* is a highly experimental run: an important aspect of the blog distillation task is to return not just blogs that mention this topic, but mention it quite often. In that sense, we do not only want to determine the relevance of the blog for a given topic, but also the non-relevance for that topic (i.e. relevant regarding different topics). We tried to estimate this by looking at the performance of blogs on the 2007 topics and use this as indicator of non-relevance (assuming the 2008 topics are different from the 2007 topics); the relevance score of a blog (Eq. 1) is divided by the average relevance score of that blog on all 2007 topics. A blog with a high relevance score and low relevance scores on other topics will get a score (and rank) boost.

Summarizing, we submitted the following runs, and added one extra baseline run to our results table: We submitted 6 runs:

(A) **`uams08bl`** $P(post|blog)$ based on number of comments, post length, and recency; combination of title+body run and title-only run.

(B) **`baseline`** identical to previous run, but without the combination with a title-only run.

(C) **`uams08nw`** identical to previous run, but with query expansion following Section 3 on a news corpus and Wikipedia.

(D) **`uams08pnw`** identical to previous run, but with the blog post corpus as additional external corpus.

(E) **`uams08nonr`** ratio of relevance to non-relevance of a blog.

Run	MAP	P5	P10	MRR
baseline	0.2567	0.4480	0.4180	**0.7298**
uams08bl	**0.2638**$^{\triangle}$	**0.4600**	**0.4200**	0.7294
uams08nonr	0.0257$^{\blacktriangledown}$	0.1000$^{\blacktriangledown}$	0.0900$^{\blacktriangledown}$	0.2393$^{\blacktriangledown}$
uams08nw	0.2489	0.4080	0.3660	0.6515
uams08pnw	0.2620	0.4080	0.3900	0.6303$^{\triangledown}$

Table 3: Results on the blog distillation task. Significance tested against *baseline*.

6.1 Results and Discussion

The results of our submitted runs, plus the evaluation of one additional run are presented in Table 3. The results show some interesting things: (i) The experimental run using "non-relevance" fails completely, indicating we need different ways of incorporating this notion of non-relevance. (ii) Our baseline (*uams08bl*) is a pretty strong baseline and cannot be beaten by the other runs (except on MRR by *baseline*). (iii) Query expansion can improve over the absolute baseline in terms of MAP, but still performs less than the combination with the titles.

7 Conclusions

In this year's participation in the blog track we mainly explored different ways of using external corpora to expand the original query. In the blog post retrieval task we did not succeed in improving over a simple baseline (equal weights for both the expanded and original query) and we need a thorough analysis to find out why this did not work. For the same task, further investigation is needed to determine the effectiveness of the credibility priors and to see what happens when the opinion prior is applied.

In the blog distillation task we tried to improve over our (strong) baseline using external expansion. Since this baseline also uses information from the title explicitly, it is hard to determine why the expanded runs do not improve over the baseline. Compared to a baseline without the title component, we see an improvement for the run using expansion on the combination of news, Wikipedia and blog posts. For this task, further research into the combination of title and full post components is needed, as well as the combination with expanded queries. The run that tried to capture non-relevance of a blog failed, but exploring this area further could lead to significant improvements over a baseline that looks only at "relevance."

Finally, looking at the two tasks combined, we see that query expansion on the blog distillation task is much more effective than on the blog post retrieval task. Further analysis is needed to find out why this difference occurs.

8 Acknowledgments

This research was 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 220-80-001, 017.001.190, 640.001.501, 640.002.501, 612.066.512, 612.061.814, 612.061.815, and 640.004.802.

9 References

Balog, K., Azzopardi, L., and de Rijke, M. (2006). Formal models for expert finding in enterprise corpora. In *SIGIR'06*.

Balog, K., de Rijke, M., and Weerkamp, W. (2008). Bloggers as experts. In *SIGIR '08*.

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

Weerkamp, W., Balog, K., and de Rijke, M. (2008). Finding key bloggers, one post at a time. In *ECAI 2008*.

Weerkamp, W. and de Rijke, M. (2008). Credibility improves topical blog post retrieval. In *HLT/NAACL '08*.