Towards Robust Query Expansion: Model Selection In The Language Modeling Framework

Mattan Winaver, Oren Kurland and Carmel Domshlak
Faculty of Industrial Engineering and Management, Technion – Israel institute of technology
mattanw@tx.technion.ac.il, kurland@ie.technion.ac.il, dcarmel@ie.technion.ac.il

ABSTRACT

We propose a language-model-based approach for addressing the performance *robustness* problem — with respect to free-parameters' values — of pseudo-feedback-based query-expansion methods. Given a query, we create a set of language models representing different forms of its expansion by varying the parameters' values of some expansion method; then, we select a single model using criteria originally proposed for evaluating the performance of using the original query, or for deciding whether to employ expansion at all. Experimental results show that these criteria are highly effective in selecting *relevance language models* that are not only significantly more effective than poor performing ones, but that also yield performance that is almost indistinguishable from that of manually optimized relevance models.

Categories and Subject Descriptors: H.3.3 [Information Search and Retrieval]: Retrieval Models

General Terms: Algorithms, Experimentation

Keywords: query expansion, language models, query clarity, model selection, query drift

1. INTRODUCTION

Automatic *query expansion* based on *pseudo feedback* [12] is an effective approach for addressing issues rising from the use of short queries. The idea is to construct a *query model* using terms from both the original query and documents that are the highest-ranked ones by some initial search.

However, since typically the documents from the initial search are not all relevant, and do not necessarily exhibit all *query aspects* [4], the constructed query model is often far from being "optimal". Indeed, the resultant retrieval performance is often highly sensitive to the choice of the number of documents and the number of terms used for defining the model. Furthermore, for some queries, retrieval based on the original query results in performance superior to that resulting from the use of an expansion model.

Thus, recent research on improving the *robustness* of expansion methods has focused on either predicting whether a given expansion will be more effective for retrieval than the original query [2, 7], or on improving the performance robustness of specific expansion methods [10, 13].

Inspired by work on combining multiple, mainly boolean-based, query representations [3], we propose a new approach for improving the robustness of automatic expansion methods. Given a query, we (i) create a set of models that represent different forms of its expansion, (ii) estimate which model will yield the best retrieval performance, and (iii) use the chosen model for retrieval.

The set of models considered for a query can be derived, for instance, by using different expansion methods, and/or by varying the values of the free parameters that these methods incorporate. Our estimates of which model will perform best utilize measures that were proposed to either predict the resultant performance of using the *original* query [6], or to decide whether the query should be expanded at all [7]; specifically, we use the language modeling framework [5] and adapt the *query clarity* [6] and distance between retrieved document lists [7] measures, respectively.

Our experimental results show that using these selection criteria over sets of *relevance models* [9] results in performance that is significantly better than that of the poor-performing relevance models, and is in many cases almost as good as that of a manually optimized one.

2. MODEL-SELECTION CRITERIA

Throughout this section we assume that the following have been fixed: a corpus $\mathcal{C}$ of documents, a query q, and a set of language models $\mathcal{M} = \{M_1, \ldots, M_n\}$ that represent "expanded forms" of q.

Our first model-selection criterion, denoted clr, is based on the *query clarity* measure [6] — we choose the model in $\mathcal{M}$ that *maximizes* the KL divergence $D(p(\cdot|M_i)||p(\cdot|\mathcal{C})) = \sum_w p(w|M_i) \log \frac{p(w|M_i)}{p(w|\mathcal{C})}$, where $p(w|\mathcal{C})$ is the maximum likelihood estimate of term w with respect to $\mathcal{C}$. Note that in the original proposal of query clarity [6], the divergence is calculated for a single expansion model, and is shown to correlate with the retrieval performance of using the original query q. In contrast, here we assume that the divergence is correlated with the performance of using the expansion model *itself.*

The selection criterion just described does not, however, handle the *query drift* problem [11] — an expansion model can be "clear" (i.e., distant from the corpus model), and yet represent topic(s) different than those exhibited by q. We therefore experiment with the $ndrift$ selection criterion that utilizes an approach originally suggested for deciding whether to expand a query at all [7]. Specifically, for each (expanded) query model M_i, we construct a (unigram) language model θ_{M_i} by (i) constructing the unsmoothed unigram language models of the top-100 retrieved documents in a search performed over $\mathcal{C}$ using the KL-retrieval method [8] with M_i as the query model, and (ii) smoothing the uni-

SIGIR'07, July 23–27, 2007, Amsterdam, The Netherlands.
ACM 978-1-59593-597-7/07/0007.

	queries	disks
AP89	1-50	1
AP88+89	101-150	1,2
WSJ	101-150	1,2
SJMN	51-150	3
TREC8	401-450	4,5 – CR

	Relevance Model				Interpolated Relevance Model			
	worst	*best*	*clr*	*ndrift*	*worst*	*best*	*clr*	*ndrift*
AP89	17.5	24.36^w	22.56^w_b	23.44^w_b	19.7	24.76^w	23.24^w_b	23.89^w_b
AP88+89	20.26	29.35^w	27.65^w	28.62^w	26.61	30.39^w	29.44^w	30.36^w
WSJ	15.21	28.2^w	26.57^w_b	27.65^w	23.25	29.63^w	28.11^w_b	29.11^w
SJMN	17.22	23.96^w	21.97^w_b	23.5^w	19.78	24.58^w	22.78^w_b	24.16^w
TREC8	17.6	24.22^w	22.19^w_b	24.01^w	24.23	27.17^w	26.05^w_b	27.07^w

Figure 1: The left table presents the details of the TREC corpora used for experiments. The right table depicts the MAP performance numbers of the (interpolated) relevance model that was chosen from a set of candidates either manually (*worst*, *best*) or by using a model-selection criterion (*clr*, *ndrift*). Statistically significant differences with *worst* and *best* are marked with w and b respectively.

form interpolation of these 100 language models with the Jelinek-Mercer approach wherein the smoothing parameter is set to 0.2. We then choose the M_j for which $D(\theta_{M_q}||\theta_{M_j})$ is *minimal*, where M_q is an unsmoothed unigram language model constructed from q, and θ_{M_q} is constructed as above using the top-100 retrieved documents from a search with M_q as the query model.

Once a query model M is selected from $\mathcal{M}$ by either of the two criteria just described, we can rank documents in the corpus using the KL-retrieval approach [8] with M.

3. EXPERIMENTS

To derive the set of query models $\mathcal{M}$, we use either a *relevance model* [9], or an *interpolated relevance model* [1] — the latter interpolates the former with the original query model. We set the number of top-retrieved documents and the number of terms, for constructing the (interpolated) relevance models, to values in $\{5, 10, 20, 30, 40, 50, 75, 100\}$ and $\{50, 100, 500, 1000\}$ respectively[1]. Thus, for each query we get two sets of 32 (interpolated) relevance models. For each set we identify the *best* and *worst* performing models in terms of MAP as measured over all tested queries[2].

We ran our experiments on TREC corpora (details in Figure 1) using the Lemur toolkit (www.lemurproject.org). We used topics' titles as queries, applied the Porter stemmer and removed INQUERY stopwords. Statistical significance was determined using the two-sided Wilcoxon test at the 95% confidence level.

Figure 1 depicts the MAP performance results.[3] Our first observation is that, indeed, a poor selection of a relevance model results in performance that is much inferior to that of a manually optimized one. (Note that *best* outperforms *worst* to a statistically significant degree in all cases.)

We also see in Figure 1 that both selection criteria result in performance that is always significantly better than that of *worst*. Furthermore, the performance of *ndrift* is in most cases statistically indistinguishable from that of *best*.

While the resultant performance of using *ndrift* is always better than that resulting from using *clr*, we note that the former requires running all the different query models while the latter only requires running the selected model.

4. CONCLUSION

We showed that using language-model-based selection criteria for choosing a query model from a set of candidates results in performance that is significantly better than that of a poor-performing query model and is almost (statistically speaking) as good as that of a manually optimized one.

Acknowledgments We thank the anonymous reviewers for their comments. The paper is based upon work done in part while the second author was at Cornell University and upon work supported in part by the National Science Foundation under grant no. IIS-0329064. Any opinions, findings, and conclusions or recommendations expressed are those of the authors and do not necessarily reflect the views of any sponsoring institutions or the U.S. government.

5. REFERENCES

[1] N. Abdul-Jaleel, J. Allan, W. B. Croft, F. Diaz, L. Larkey, X. Li, M. D. Smucker, and C. Wade. UMASS at TREC 2004 — novelty and hard. In *Proceedings of TREC-13*, 2004.

[2] G. Amati, C. Carpineto, and G. Romano. Query difficulty, robustness, and selective application of query expansion. In *Proceedings of ECIR*, pages 127–137, 2004.

[3] N. J. Belkin, C. Cook, W. B. Croft, and J. P. Callan. The effect of multiple query representations on information retrieval system performance. In *Proceedings of SIGIR*, pages 339–346, 1993.

[4] C. Buckley. Why current IR engines fail. In *Proceedings of SIGIR*, pages 584–585, 2004. Poster.

[5] W. B. Croft and J. Lafferty, editors. *Language Modeling for Information Retrieval.* Number 13 in Information Retrieval Book Series. Kluwer, 2003.

[6] S. Cronen-Townsend, Y. Zhou, and W. B. Croft. Predicting query performance. In *Proceedings of SIGIR*, pages 299–306, 2002.

[7] S. Cronen-Townsend, Y. Zhou, and W. B. Croft. A language modeling framework for selective query expansion. Technical Report IR-338, Center for Intelligent Information Retrieval, University of Massachusetts, 2004.

[8] J. D. Lafferty and C. Zhai. Document language models, query models, and risk minimization for information retrieval. In *Proceedings of SIGIR*, pages 111–119, 2001.

[9] V. Lavrenko and W. B. Croft. Relevance-based language models. In *Proceedings of SIGIR*, pages 120–127, 2001.

[10] X. Li and W. B. Croft. Improving the robustness of relevance-based language models. Technical Report IR-401, Center for Intelligent Information Retrieval, University of Massachusetts, 2005.

[11] M. Mitra, A. Singhal, and C. Buckley. Improving automatic query expansion. In *Proceedings of SIGIR*, pages 206–214, 1998.

[12] I. Ruthven and M. Lalmas. A survey on the use of relevance feedback for information access systems. *Knowledge Engineering Review*, 18(2):95–145, 2003.

[13] T. Tao and C. Zhai. Regularized estimation of mixture models for robust pseudo-relevance feedback. In *Proceedings of SIGIR*, pages 162–169, 2006.

[1] For constructing the relevance models, the value of the smoothing parameter of the top-retrieved documents' language models is set to 0.2.

[2] The interpolation parameter of *all* interpolated relevance models, which controls the reliance on the original query model, is fixed to the (manually optimized) value chosen for the *best* model.

[3] Similar performance patterns are obtained with respect to precision@10. The actual performance numbers are omitted due to space considerations.