Enhancing Relevance Models with Adaptive Passage Retrieval

Xiaoyan Li[1] and Zhigang Zhu[2]

[1] Department of Computer Science, Mount Holyoke College, South Hadley,
MA 01075, USA
xli@mtholyoke.edu
[2] Department of Computer Science, CUNY City College, New York, NY 10031, USA
zzhu@ccny.cuny.edu

Abstract. Passage retrieval and pseudo relevance feedback/query expansion have been reported as two effective means for improving document retrieval in literature. Relevance models, while improving retrieval in most cases, hurts performance on some heterogeneous collections. Previous research has shown that combining passage-level evidence with pseudo relevance feedback brings added benefits. In this paper, we study passage retrieval with relevance models in the language-modeling framework for document retrieval. An *adaptive passage retrieval* approach is proposed to document ranking based on the best passage of a document given a query. The proposed passage ranking method is applied to two relevance-based language models: the Lavrenko-Croft relevance model and our *robust relevance model*. Experiments are carried out with three query sets on three different collections from TREC. Our experimental results show that combining adaptive passage retrieval with relevance models (particularly the robust relevance model) consistently outperforms solely applying relevance models on full-length document retrieval.

Keywords: Relevance models, passage retrieval, language modeling.

1 Introduction

Language modeling approach is a successful alternative to traditional retrieval models for text retrieval. The language modeling framework was first introduced by Ponte and Croft [19], followed by many research activities related to this framework since then [1, 3, 4, 8, 10-12, 14-18, 20, 21, 23]. For example, query expansion techniques [3,11,12,17,18,21,23], pseudo-relevance feedback [4,11,12,17,18,21,23], parameter estimation methods [10], multi-word features [20], passage segmentations [16] and time constraints [14] have been proposed to improve the language modeling frameworks. Among them, query expansion with pseudo feedback can increase retrieval performance significantly [11,18,23]. It assumes a few top ranked documents retrieved with the original query to be relevant and uses them to generate a richer query model.

However, two major problems remain unsolved in the query expansion techniques. First, the performance of a significant number of queries decreases when query

C. Macdonald et al. (Eds.): ECIR 2008, LNCS 4956, pp. 463–471, 2008.

expansion techniques are applied on some collections. Second, existing query expansion techniques are very sensitive to the number of documents used for pseudo feedback. Most approaches usually achieved the best performance when about 30 documents are used for pseudo feedback. As the number of feedback documents increases beyond 30, retrieval performance drops quickly. In our recent work [15], a robust relevance model is proposed based on a study of features that affected retrieval performance. These features included key words from original queries, relevance ranks of documents from the first round retrieval, and common words in the background data collection. The robust relevance model seamlessly incorporated these features into the relevance-based language model in [11] and further improved the performance and robustness of the model. The three features were also used in a recent work by Tao and Zhai [22] with regularized mixture models.

The robust relevance model and the regularized mixture model greatly ease the second problem, i.e. sensitivity of the retrieval performance to the number of documents used for pseudo feedback. However, the solution to the first problem is only partially. As we have reported in [15], the performance of the robust relevance model outperformed the Lavrenko-Croft relevance model and the simple query likelihood model on four test query sets, but it underperformed the simple query likelihood model on a query set against a subset of the TREC terabyte collection.

Passage retrieval is another effective means to improve document retrieval [5,6,7,16]. Particularly in [16], it was incorporated into the language modeling framework via various approaches. However, a major concern of passage retrieval in the language modeling framework is that it hurts retrieval performance on some collections, although it can provide comparable results and sometimes significant improvements over full-length document retrieval on collection with long and multi-topic documents. Therefore, one important research issue for both relevance models and passage retrieval is when and how to apply relevance models and passage retrieval for better retrieval performance.

In this paper, an *adaptive* passage retrieval approach is proposed to document ranking based on the *best passage* of a document given a query. The best passage of a document is the passage with the highest relevance score with respect to the query. The size of the best passage varies from document to document and from query to query. The best passage of a document can be a passage of the smallest window size considered or the document itself depends on whether it has the highest relevance score among all available passages. This adaptive passage selection is applied to two relevance-based language models: the Lavrenko-Croft relevance model [11] and our robust relevance model [15]. Experiments are carried out with three query sets on three different collections from TREC, including the ones that caused under-performance in the robust relevance model [15] and the fixed-size passage retrieval approach [16]. Our experimental results show that combining adaptive passage retrieval with relevance models consistently outperforms solely applying relevance models on full-length document retrieval. It indicates that passage-level evidence, if used appropriately, can be incorporated in relevance models to achieve better performance in terms of mean average precision, especially in the case of the robust relevance model.

The rest of the paper is structured as follows. In Section 2, we give a brief overview of the two relevance-based language models used in this paper. Section 3 describes our approach to combining the adaptive passage retrieval with the relevance models. Section 4 provides experimental results, compared to baseline results of full-length document retrieval. Section 5 summarizes the paper with conclusions and some future work.

2 Relevance Models

2.1 The Lavrenko-Croft Relevance Model

Lavrenko and Croft's relevance-based language model [5] is a model-based query expansion approach in the language-modeling framework [18]. A relevance model is a distribution of words in the relevant class for a query. Both the query and its relevant documents are treated as random samples from an underlying relevance model R, as shown in Figure 1. Once the relevance model is estimated, the *KL-divergence* between the relevance model (of a query and its relevant documents) and the language model of a document can be used to rank the document. Documents with smaller divergence are considered more relevant thus have higher ranks. Equations (1) and (2) are the formulas used in [5] and in this paper for approximating a relevance model for a query:

$$P_o(w \mid R) \approx \frac{P(w, q_1 ... q_k)}{P(q_1 ... q_k)} \tag{1}$$

$$P(w, q_1 ... q_k) = \sum_{D \in M} P(D) P(w \mid D) \prod_{i=1}^{k} P(q_i \mid D) \tag{2}$$

where $P_o(w \mid R)$ stands for the relevance model of the query and its relevant documents, in which $P(w, q_1...q_k)$ stands for the total probability of observing the word w together with query words $q_1...q_k.$ A number of top ranked documents (say N) returned with a query likelihood language model are used to estimate the relevance model. In Equation (2), M is the set of the N top ranked documents used for estimating the relevance model for a query (together with its relevant documents). $P(D)$ is the prior probability to select the corresponding document language model D for generating the total probability in Equation (2). In the original relevance model approach, a uniform distribution was used for the prior.

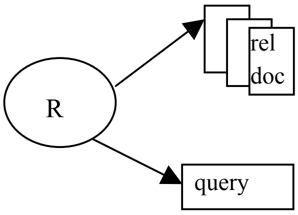

Fig. 1. The Lavrenko-croft relevance model

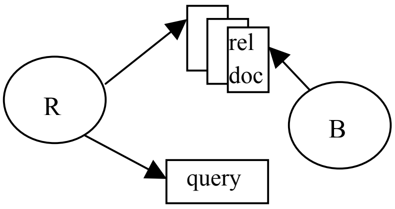

Fig. 2. Our robust relevance model

2.2 Our Robust Relevance Model

Based on the Lavrenko-Croft relevance model approach, we have proposed a robust relevance model to further improve retrieval performance and robustness [15]. In the robust relevance model, Queries are random samples from the underlying relevance model R, and relevant documents are sampled from both the underlying relevance model R and a background language model B, as shown in Figure 2.

Three significant changes were made to the original relevance model in order to estimate a more accurate relevance model for a query: *treating the original query as a special document, introducing rank-related prior,* and *discounting common words*. The robust relevance model seamlessly incorporated these three features into the original relevance-based language model. Equations (3), (4) and (5) are the formulas used in [15] and also in this paper for approximating a relevance model for a query:

$$P_{new}(w, q_1 \ldots q_k) = \sum_{D \in S} P(D) P(w \mid D) \prod_{i=1}^{k} P(q_i \mid D) \tag{3}$$

$$P(D) = \frac{1}{Z_1} * \frac{\alpha + |D|}{\beta + Rank(D)}, \quad Z_1 = \sum_{D \in S} \frac{\alpha + |D|}{\beta + Rank\ (D)} \tag{4}$$

$$P_{new}(w \mid R) = \frac{1}{Z_2} \frac{P_{new}(w, q_1 \ldots q_k)}{\gamma + P(w \mid B)}, \quad Z_2 = \sum_{w \in V} \frac{P_{new}(w, q_1 \ldots q_k)}{\gamma + P(w \mid B)} \tag{5}$$

Unlike the set M including only top N documents' models in equation (2) for the Lavrenko-Croft relevance model, the robust relevance model treats the original query as a special document: the set S in equation (3) includes both the query model and the document models for the top N documents.

The robust relevance model also introduces a rank-related prior. In equation (4), $|D|$ denotes the length of document *D* or the length of the query – the special document. *Rank(D)* denotes the rank of document *D* in the ranked list of documents returned by using the basic query likelihood language model. The rank of the query is set to 0 so that it has the highest rank among all the documents used for relevance model approximation. Z_1 is the normalization factor that makes the sum of the priors to 1. Parameters α and β are used to control how much a document's length and its rank affect the prior of the document, respectively.

Finally, the robust relevance model discounts common words in the background data collection. In equation (5), P_{new} $(w \mid R)$ denotes the probability of word *w* in the new relevance model. P $(w \mid B)$ denotes the probability of word *w* in the background language model B. γ is the parameter for discounting the probability of a word in the new relevance model by its probability in the background language model. Z_2 is the normalization factor that makes the sum of the probabilities of words in the new relevance model to 1. The best values for parameters α , β and γ reported in [15] are used in our experiments in this paper.

3 Combining Passage Retrieval with Relevance Models

Passage retrieval can be applied in the language modeling framework. Various approaches were proposed in [16] to implement passage retrieval in the language modeling environment. In the context of relevance models, three different methods R1, R2 and R3 were developed in [16], and the method R1 was reported as the best candidate. Different types of passages including half-overlapped window and arbitrary passage with fixed or variable lengths were also tried for passage retrieval.

In our paper, as baselines, we use the method R1 with fixed-size half-overlapped windows to retrieve relevance documents. Half-overlapped windows of 150, 350 and 500 words are considered in our experiments.

Given a window size, documents are first broken into half-overlapped passages. A language model is then built for each passage. At query time, a simple query likelihood language model implemented in LEMUR [13] is used to retrieve top passages. In the case of the Lavrenko-Croft relevance model, the top retrieved passages are assumed relevant and used to build a relevance model for the query. In the case of our robust relevance model, both the top passages and the query itself are used to build a relevance model. Once the relevance model is built, a *KL divergence score* is computed between each passage model and the relevance model. The KL divergence score is then used for ranking passages. Documents finally are ranked based on the score of their best passage.

However, the problem with a fixed-size window approach is that the best performance is achieved with different window size on different collections. Therefore, it is not clear how to preset a window size at query time on a new collection that is previously unseen. To solve this problem, we propose an *adaptive* passage retrieval approach in this paper. Documents are ranked based on their *best passage*. The best query in this context is the passage that can represent a document better than other passages with respect to a query. We observe that the size of a best passage can vary from document to document and query from query, because documents may discuss multiple topics, have a different focus, and by authors with different writing styles. There are various approaches that can be used to locate the best passage of a document. In this paper, we choose a very simple but efficient way to find the best passage of a document to improve retrieval performance for a given query.

We simply take the retrieval results from relevance models on full-length documents retrieval, fixed-size passages on relevance models with several preset window sizes, for example 150, 350 and 500, respectively. With the four result files, each document has four scores: full-length document score, the highest score among all 150-word-long passages, the highest score among all 350-word-long passages, and the highest score among all 500-word-long passages. We take the highest score among the four scores as the score of the best passage of the document. Documents are then ranked according to the score of their best passages. Note that the size of the best passage varies from document to document. The best passage of a document can be a passage of the smallest window size considered (say 150 in this paper) or the full-length document itself depends on whether it has the highest relevance score among all available passages of variable size. Therefore, the adaptive passage retrieval approach combines the best of the full document and fixed-size passage retrieval results.

4 Experiments and Results

We have carried out experiments with three TREC query sets on three data collections. All experiments were performed with the Lemur toolkit [13]. The Krovetz stemmer [9] was used for stemming and the standard stopword list of LEMUR was used to remove about 420 common terms. Top 30 documents or passages are used to estimate a relevance model for a query when using relevance model approaches. Parameters α , β and γ in Equations (3)-(5) are set to the same values as used in [15].

4.1 Data

We used three query sets on three document collections in our experiments. (1). Queries 51 to 150 on a homogeneous collection AP88_90. AP88_90 includes newswires from Associated Express 1988, 1989 and 1990. It is a collection of short documents. This was used in [16] where fixed-size passage retrieval hurt the relevance retrieval performance. (2). Queries 101 to 150 on a heterogeneous collection AP&FR collection, which includes the Associated Press data set (AP88 and AP89) and the FR88&89 collection. We created this collection to test the performance of our approach to such a heterogeneous data collection. (3). Queries 701 to 750 on a sub-collection of the TREC Terabyte data set on which the robust relevance model [15] had some problem. To construct the subset, the top-ranked 10,000 documents for each of the 50 queries that were retrieved using the basic query likelihood language model were selected. The subset has 466,724 unique web documents and is about 2% of the entire terabyte collection [2]. This collection is by nature a more heterogeneous collection with web documents, blogs, emails as well as news articles. The statistics of AP88_89 collection, AP&FR collection, and the subset of terabyte collection are shown in Table 1. Table 2 summaries the information about the three sets of queries used in our experiments and relevant documents on the corresponding three document collections. The queries are taken from TREC topics and only title field are used in our experiments. The queries are on average 3 or 4 words long, and the number of relevant documents per query varies across collections.

Table 1. Statistics of the three document collections

Collection Statistics	AP88_90	AP&FR	Terabyte (GOV2)
# of documents	242,918	210,417	466,724
# of terms	61,975,608	83,936,199	958,740,730
# of unique terms	255,617	362,886	3,637,433
Average Length of documents	255	398	2,054
Average frequency of terms	242	231	264

Table 2. Information of the three query sets (N1: # of queries with relevant documents; N2: total # of relevant documents; N3: average # of relevant doc. per query)

Collections	Queries (title only)	N1	N2	N3
AP88_90	TREC topics 51-150	99	21829	220.5
AP&FR	TREC topics 101-150	50	5,211	104.2
Terabyte	TREC topics 701-750	49	10,617	216.7

Table 3. Performance comparison of passage retrieval + relevance models (RM: Lavrenko-Croft relevance model; RRM: our robust relevance model)

Datasets	Methods	FullDoc	P150	P350	P500	BestP
AP88_90	RM	0.2779	0.2677	0.2747	0.2771	0.2844
	RRM	0.2821	0.2655	0.2800	0.2822	0.2882
AP&FR	RM	0.2696	0.2799	0.2720	0.272	0.2761
	RRM	0.2724	0.3084	0.3093	0.3106	0.3113
Terabyte	RM	0.1872	0.2026	0.2119	0.2067	0.2202
	RRM	0.2361	0.2256	0.2448	0.2376	0.2528

4.2 Experimental Results

We have carried out experiments on three query sets. In each query set, the four baselines - full document retrieval (FullDoc) and three fixed-size passage retrieval baselines with three different window sizes (P150, P350 and P500), and the adaptive passage retrieval method (BestP), is applied to the Lavrenko-Croft relevance model (RM) and our robust relevance model (RRM), respectively. Mean average precision is used for performance evaluation. Three different window sizes in the fixed-size passage retrieval baselines are 150, 350 and 500. The "best" passage of a document in the proposed adaptive passage retrieval approach (BestP) is the passage with the highest KL divergence score among all passages of different sizes (150, 350, 500 and full-length document). The performance of three query sets in terms of mean average precision is given in Table 3. The following observations can be obtained based on the experimental results.

(1) Combining adaptive passage retrieval with the two relevance models consistently outperforms solely applying the relevance models on full-length document retrieval on all the three collections. Robust relevance model with fixed-size passages also gives better performance than full-length document retrieval on all three collections. But original relevance model with fixed-size passage achieves outperforms full-length document retrieval only on two of the collections.

(2) The adaptive passage retrieval consistently provides the best performance than the full-length document retrieval and the fixed-size passage retrieval, when using the two relevance models, on all three collections. The only exception is for queries 101-150 with the original relevance model, where the best performance was achieved when the passage size is fixed to 150. However, the adaptive passage retrieval method ranked the second best, and is very close to the first best.

(3) Better performance is achieved when the robust relevance model is used. This is true for all the four baselines as well as the adaptive passage retrieval approach. The performance is always the best when combining the adaptive passage retrieval with the robust relevance model.

5 Conclusions and Future Work

In this paper, we study how to better combine passage retrieval with relevance models in the language modeling framework for better retrieval performance. Three main conclusions have been drawn from the experimental results. First, combining passage retrieval with relevance models consistently outperforms relevance models on full-length document retrieval in terms of mean average precision on document retrieval. Second, the proposed adaptive passage retrieval approach for identifying best passage gives better performance than using passages of fixed sizes. Third, the robust relevance model uniformly outperforms the original relevance models, especially when combining with passage-level evidence.

In the current experiments, for testing the ideas of the adaptive passage sizes, we only used a few typical document sizes that have been tested empirically in literature. As a future work, the approach proposed by Jiang and Zhai [5] for identifying variable-length passages using HMMs could be used. As another future work, new approaches to query expansion techniques need to be developed for retrieval on heterogeneous collections (e.g., the Terabyte collection), which may include web documents, blogs, emails as well as news articles. In this case, incorporating selective query expansion techniques, such as Cronen-Townsend et al's work in [3], and features like metadata into relevance models may be helpful.

Acknowledgments. The first author is supported by Mount Holyoke College and second author by NSF (Grant No. CNS-0551598).

References

[1] Abdul-Jaleel, N., et al.: UMASS at TREC2004. In: Thirteen Text Retrieval Conference Notebook (2004)

[2] Clarke, C., Craswell, N., Soboroff, I.: Overview of the TREC 2004 terabyte track. In: Thirteen Text Retrieval Conference Notebook (2004)

[3] Cronen-Townsend, S., Zhou, Y., Croft, W.B.: A framework for selective query expansion. In: Proc. 13th Int. Conf. on Information and Knowledge Management, pp. 236–237 (2004)

[4] Hiemstra, D.: Using language models for information retrieval. PhD thesis, University of Twente (2001)

[5] Jiang, J., Zhai, C.: UIUC in HARD 2004 – Passage Retrieval using HMMs. In: Thirteen Text Retrieval Conference Notebook (2004)

[6] Kaszkiel, M., Zobel, J.: Passage retrieval revisited. In: Proc. 20th ACM-SIGIR Conf. on Research and Development in Information Retrieval, pp. 178–185 (1997)

[7] Kaszkiel, M., Zobel, J.: Effective ranking with arbitrary passages. Journal of the American Society for Information Science and Technology 52(4), 344–364 (2001)

[8] Kraaij, W., Westerveld, T., Hiemstra, D.: The importance of prior probabilities for entry page search. In: 25th ACM SIGIR Conf. on Research and Development in Information Retrieval, pp. 27–34 (2002)

[9] Krovetz, R.: Viewing morphology as an inference process. In: Proc. 16th ACM SIGIR Conf. on Research and Development in Information Retrieval, pp. 191–202 (1993)

[10] Lafferty, J., Zhai, C.: Document language models, query models, and risk minimization for information retrieval. In: 24th ACM SIGIR Conf. on Research and Development in Information Retrieval, pp. 111–119 (2001)

[11] Lavrenko, V., Croft, W.B.: Relevance-based language models. In: Proc. 24th ACM SIGIR Conf. on Research and Development in Information Retrieval, pp. 120–127 (2001)

[12] Lavrenko, V., Croft, W.B.: Relevance models in information retrieval. In: Croft, W.B., Lafferty, J. (eds.) Language modeling for information retrieval, pp. 11–56. Kluwer Academic Publishers, Dordrecht (2003)

[13] LEMUR, http://www-2.cs.cmu.edu/~lemur/3.1/doc.html

[14] Li, X., Croft, W.B.: Time-based language models. In: Proc. 12th Int. Conf. on Information and Knowledge Management, pp. 469–475 (2003)

[15] Li, X.: A new robust relevance model in the language model framework, Information Processing and Management (2007), doi:10.1016/j.ipm.2007.07.005

[16] Liu, X., Croft, W.B.: Passage retrieval based on language models. In: Proc. 11th Int. Conf. on Information and Knowledge Management, pp. 375–382 (2002)

[17] Miller, D.H., Leek, T., Schwartz, R.: A hidden Markov model information retrieval system. In: Proc. 22nd ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 214–221 (1999)

[18] Ponte, J.: A Language modeling approach to information retrieval. PhD thesis, UMass-Amherst (1998)

[19] Ponte, J., Croft, W.B.: A language modeling approach to information retrieval. In: Proc. 21st ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 275–281 (1998)

[20] Song, F., Croft, W.B.: A general language model for information retrieval. In: Proc. 22nd ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 279–280 (1999)

[21] Tao, T., Zhai, C.: A two-stage mixture model for pseudo feedback. In: Proc.27th ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 486–487 (2004)

[22] Tao, T., Zhai, C.: Regularized estimation of mixture models for robust pseudo-relevance feedback. In: Proc. 29th ACM SIGIR Conf. on Research and Development in Information Retrieval, pp. 162–169 (2006)

[23] Zhai, C., Lafferty, J.: Model-based feedback in the language modeling approach to information retrieval. In: Proc. 10th Int. Conf. on Information and Knowledge Management, pp. 403–410 (2001)