Utilizing Passage-Based Language Models for Document Retrieval

Michael Bendersky[1] and Oren Kurland[2]

[1] Center for Intelligent Information Retrieval, Department of Computer Science, University of Massachusetts, Amherst, MA 01003; {bemike@cs.umass.edu}
[2] Faculty of Industrial Eng. & Mgmt., Technion, Israel; {kurland@ie.technion.ac.il}

Abstract. We show that several previously proposed *passage-based* document ranking principles, along with some new ones, can be derived from the same probabilistic model. We use language models to instantiate specific algorithms, and propose a *passage language model* that integrates information from the ambient document to an extent controlled by the estimated *document homogeneity*. Several document-homogeneity measures that we propose yield passage language models that are more effective than the standard passage model for basic document retrieval and for constructing and utilizing *passage-based relevance models*; the latter outperform a document-based relevance model. We also show that the homogeneity measures are effective means for integrating document-query and passage-query similarity information for document retrieval.

Keywords: passage-based document retrieval, document homogeneity, passage language model, passage-based relevance model

1 Introduction

The ad hoc retrieval task is to rank documents in response to a query by their assumed relevance to the information need it represents. While a document can be compared as a whole to the query, it could be the case (e.g., for long and/or heterogeneous documents) that only (very few, potentially small) parts of it, i.e., *passages*, contain information pertaining to the query. Thus, researchers have proposed different approaches for utilizing passage-based information for document retrieval [1–8].

We show that some of these previously proposed passage-based document-ranking approaches can in fact be derived from the same probabilistic model. Among the methods we derive are ranking a document by the highest query-similarity score that any of its passages is assigned [2, 4, 8], and by interpolating this score with the document-query similarity score [2, 4].

We instantiate specific retrieval algorithms by using *statistical language models* [9]. In doing so, we propose a *passage language model* that incorporates information from the ambient document to an extent controlled by the estimated *document homogeneity*. Our hypothesis is that (language) models of passages in highly homogeneous documents should pull a substantial amount of information

from the ambient document; for passages in highly heterogeneous documents, minimal such information should be used.

Several document-homogeneity measures that we propose yield passage language models that are more effective than the standard passage model [8] — as experiments over TREC data attest — for basic passage-based document ranking and for constructing and utilizing *passage-based relevance models* [8]; the latter also outperform a document-based relevance model [10].

We also derive, and demonstrate the effectiveness of, a novel language-model-based algorithm that integrates, using document-homogeneity measures, the query-similarity of a document and of its passages for document ranking.

2 Retrieval Framework

In what follows we show that some previously-proposed passage-based document-ranking approaches, and some new ones, can be derived from the same model.

Notation and conventions. Throughout this section we assume that the following have been fixed: a query q, a document d, and a corpus of documents $\mathcal{C}$ ($d \in \mathcal{C}$). We use g to denote a passage, and write $g \in d$ if g is one of d's m passages. (Our algorithms are not dependent on the type of passages.) We write $p_x(\cdot)$ to denote a (smoothed) unigram language model induced from x (a document or a passage); our language model induction methods are described in Sec. 2.2.

2.1 Passage-Based Document Ranking

We rank document d in response to query q by estimating the probability $p(q|d)$ that q can be generated[3] from a model induced from d, as is common in the language modeling approach to retrieval [12, 9]. We hasten to point out, however, that our framework is not committed to any specific estimates for probabilities of the form $p(q|x)$, which we often refer to as the "query-similarity" of x.

Since passages are smaller — and hence potentially more focused — units than documents, they can potentially "help" in generating queries. Thus, assuming that *all* passages in the corpus can serve as proxies (representatives) of d for generating *any* query, and using $p(g_i|d)$ to denote the probability that passage g_i (of some document in the corpus) is chosen as a proxy of d, we can write

$$p(q|d) = \sum_{g_i} p(q|d, g_i)p(g_i|d) \ . \tag{1}$$

If we assume that d's passages are much better proxies for d than passages not in d, then we can define $\hat{p}(g_i|d) \stackrel{def}{=} \frac{p(g_i|d)}{\sum_{g_j \in d} p(g_j|d)}$ if $g_i \in d$, 0 otherwise, and

[3] While it is convenient to use the term "generate" in reference to work on language models for IR [9], we do not think of text items as literally generating the query. Furthermore, we do not we assume an underlying generative theory in contrast to Lavrenko and Croft [10], and Lavrenko [11], *inter alia*.

use it in Eq. 1 to rank d as follows:

$$Score(d) \stackrel{def}{=} \sum_{g_i \in d} p(q|d, g_i)\hat{p}(g_i|d) \ . \quad (2)$$

To estimate $p(q|d, g_i)$, we integrate $p(q|d)$ and $p(q|g_i)$ based on the assumed *homogeneity* of d: the more homogeneous d is assumed to be, the higher the impact it has as a "whole" on generating q. Specifically, we use the estimate[4] $h^{[\mathcal{M}]}(d)p(q|d) + (1 - h^{[\mathcal{M}]}(d))p(q|g_i)$, where $h^{[\mathcal{M}]}(d)$ assigns a value in $[0, 1]$ to d by homogeneity model $\mathcal{M}$. (Higher values correspond to higher estimates of homogeneity; we present document-homogeneity measures in Sec. 2.3.) Using some probability algebra (and the fact that $\sum_{g_i \in d} \hat{p}(g_i|d) = 1$), Eq. 2 then becomes

$$Score(d) \stackrel{def}{=} h^{[\mathcal{M}]}(d)p(q|d) + (1 - h^{[\mathcal{M}]}(d)) \sum_{g_i \in d} p(q|g_i)\hat{p}(g_i|d) \ , \quad (3)$$

with more weight put on the "match" of d as a whole to the query as d is considered more homogeneous.

If we consider d to be highly heterogeneous and consequently set $h^{[\mathcal{M}]}(d)$ to 0, and in addition use the relative importance (manually) attributed to g_i as a surrogate for $\hat{p}(g_i|d)$, Eq. 3 is then a previously proposed ranking approach for (semi-)structured documents [4]; if a uniform distribution is used for $\hat{p}(g_i|d)$, instead, we score d by the mean "query-similarity" of its constituent passages, which yields poor retrieval performance that supports our premise from Sec. 1 about long (and heterogeneous) documents.

Alternatively, we can bound Eq. 3 by

$$Score_{inter-max}(d) \stackrel{def}{=} h^{[\mathcal{M}]}(d)p(q|d) + (1 - h^{[\mathcal{M}]}(d)) \max_{g_i \in d} p(q|g_i) \ . \quad (4)$$

This scoring function is a generalized form of approaches that interpolate the document-query similarity score and the maximum query-similarity score assigned to any of its passages using fixed weights [14, 2, 15, 4]; hence, such methods (implicitly) assume that all documents are homogeneous to the same extent. Furthermore, note that assuming that d is highly homogeneous and setting $h^{[\mathcal{M}]}(d) = 1$ results in a standard document-based ranking approach; on the other hand, assuming d is highly heterogeneous and setting $h^{[\mathcal{M}]}(d) = 0$ yields a commonly-used approach that scores d by the maximum query-similarity measured for any of its passages [2, 7, 4, 8]:

$$Score_{max}(d) \stackrel{def}{=} \max_{g_i \in d} p(q|g_i) \ . \quad (5)$$

2.2 Language-Model-Based Algorithms

Following standard practice in work on language models for IR [9], we estimate $p(q|d)$ and $p(q|g_i)$ using the unigram language models induced from d

[4] This is reminiscent of some work on cluster-based retrieval [13].

and g_i, i.e., $p_d(q)$ and $p_{g_i}(q)$, respectively. Then, Eq. 4 yields the novel **Interpolated Max-Scoring Passage** algorithm, which scores d by $h^{[\mathcal{M}]}(d)p_d(q) + (1 - h^{[\mathcal{M}]}(d)) \max_{g_i \in d} p_{g_i}(q)$. Using language models in Eq. 5 yields the **Max-Scoring Passage** algorithm, which scores d by $\max_{g_i \in d} p_{g_i}(q)$ as was proposed by Liu and Croft [8].

Language Model Induction. We use $\widetilde{p}_x^{MLE}(w)$ to denote the maximum likelihood estimate (MLE) of term w with respect to text (or text collection) x, and smooth it using corpus statistics to get the standard (basic) language model [16]:

$$\widetilde{p}_x^{[basic]}(w) = (1 - \lambda_{\mathcal{C}})\widetilde{p}_x^{MLE}(w) + \lambda_{\mathcal{C}}\widetilde{p}_{\mathcal{C}}^{MLE}(w) \ ; \tag{6}$$

$\lambda_{\mathcal{C}}$ is a free parameter.

We extend the estimate just described to a sequence of terms $w_1 w_2 \cdots w_n$ by using the unigram-language-model term-independence assumption

$$p_x^{[basic]}(w_1 w_2 \cdots w_n) \stackrel{def}{=} \prod_{j=1}^{n} \widetilde{p}_x^{[basic]}(w_j) \ . \tag{7}$$

Passage Language Model. Using $p_{g_i}^{[basic]}(q)$ in the above-described algorithms implies that document d is so heterogeneous that in estimating the "match" of each of its passages with the query we do not consider any information from d, except for that in the passage itself.

Some past work on question answering, and passage and XML retrieval [17–22] uses a passage language model that exploits information from the ambient document to the same fixed extent for all passages and documents. In contrast, here we suggest to use the document estimated homogeneity to control the amount of reliance on document information. (Recall that homogeneity measures are used in the Interpolated Max-Scoring Passage algorithm for fusion of similarity scores.) Hence, for $g \in d$ we define the passage language model

$$\widetilde{p}_g^{[\mathcal{M}]}(w) \stackrel{def}{=} \lambda_{psg}(g)\widetilde{p}_g^{MLE}(w) + \lambda_{doc}(d)\widetilde{p}_d^{MLE}(w) + \lambda_{\mathcal{C}}\widetilde{p}_{\mathcal{C}}^{MLE}(w) \ ; \tag{8}$$

we fix $\lambda_{\mathcal{C}}$ to some value, and set $\lambda_{doc}(d) = (1 - \lambda_{\mathcal{C}})h^{[\mathcal{M}]}(d)$ and $\lambda_{psg}(g) = 1 - \lambda_{\mathcal{C}} - \lambda_{doc}(d)$ to have a valid probability distribution. We then extend this estimate to sequences as we did at the above

$$p_g^{[\mathcal{M}]}(w_1 w_2 \cdots w_n) \stackrel{def}{=} \prod_{j=1}^{n} \widetilde{p}_g^{[\mathcal{M}]}(w_j) \ . \tag{9}$$

Setting $h^{[\mathcal{M}]}(d) = 0$ — considering d to be highly heterogeneous — we get the standard passage language model from Eq. 7. On the other hand, assuming d is highly homogeneous and setting $h^{[\mathcal{M}]}(d) = 1$ results in representing each of d's passages with d's standard language model from Eq. 7; note that in this case the Max-Scoring Passage algorithm amounts to a standard document-based language model retrieval approach.

2.3 Document Homogeneity

We now consider a few simple models $\mathcal{M}$ for estimating document homogeneity. We define functions $h^{[\mathcal{M}]} : \mathcal{C} \rightarrow [0, 1]$ with higher values corresponding to (assumed) higher levels of homogeneity.

Long documents are often considered as more heterogeneous than shorter ones. We thus define the normalized length-based measure

$$h^{[length]}(d) \stackrel{def}{=} 1 - \frac{\log |d| - \min_{d_i \in \mathcal{C}} \log |d_i|}{\max_{d_i \in \mathcal{C}} \log |d_i| - \min_{d_i \in \mathcal{C}} \log |d_i|} \ ,$$

where $|d_j|$ is the number of terms in d_j.[5]

The length-based measure does not handle the case of short heterogeneous documents. We can alternatively say that d is homogeneous if its term distribution is concentrated around a small number of terms [23]. To model this idea, we use the *entropy* of $d's$ unsmoothed language model and normalize it with respect to the maximum possible entropy of *any* document with the same length as that of d (i.e., $\log |d|$):[6]

$$h^{[ent]}(d) \stackrel{def}{=} 1 + \frac{\sum_{w' \in d} \widetilde{p}_d^{MLE}(w') \log(\widetilde{p}_d^{MLE}(w'))}{\log |d|} \ .$$

Both homogeneity measures just described are based on the document as a whole and do not explicitly estimate the variety among its passages. We can assume, for example, that the more similar the passages of a document are to each other, the more homogeneous the document is. Alternatively, a document with passages highly similar to the document as a whole might be considered homogeneous. Assigning d's passages with unique IDs, and denoting the tf.idf[7] vector-space representation of text x as $\boldsymbol{x}$, we can define these homogeneity notions using the functions $h^{[interPsg]}(d)$ and $h^{[docPsg]}(d)$, respectively, where

$$h^{[interPsg]}(d) \stackrel{def}{=} \begin{cases} \frac{2}{m(m-1)} \sum_{i<j; g_i, g_j \in d} \cos(\boldsymbol{g_i}, \boldsymbol{g_j}) & \text{if } m > 1 \ , \\ 1 & \text{otherwise} \ ; \end{cases}$$

$$h^{[docPsg]}(d) \stackrel{def}{=} \frac{1}{m} \sum_{g_i \in d} \cos(\boldsymbol{d}, \boldsymbol{g_i}) \ .$$

[5] Normalizing the length with respect to documents in several corpora (including the ambient corpus) yields very similar retrieval performance to that resulting from normalization with respect to documents in the ambient corpus alone.

[6] $Entropy(d) \stackrel{def}{=} -\sum_{w' \in d} \widetilde{p}_d^{MLE}(w') \log(\widetilde{p}_d^{MLE}(w'))$; higher values correspond to (assumed) lower levels of homogeneity. A document d with all terms different from each other has the maximum entropy ($\log |d|$) with respect to documents of length $|d|$. If $|d| = 1$, we set $h^{[ent]}(d)$ to 1.

[7] Modeling these two homogeneity notions using the KL divergence between language models yields substantially-inferior retrieval performance to that of using the proposed vector space representation with the cosine measure.

3 Related Work

There is a large body of work on utilizing (different types of) passages for document retrieval [1–8]. We showed in Sec. 2 that several of these methods can be derived and generalized from the same model.

Utilizing passage language models is a recurring theme in question answering [24, 25, 18], sentence and passage retrieval [26, 20, 22], document retrieval [3, 6, 8] and XML retrieval [19, 21]. As mentioned in Sec. 2.2, some prior work [17–22] smooth the passage (sentence) model with its ambient document's statistics, by using interpolation with fixed weights. We present in Section 4.1 the relative merits of our approach of using document homogeneity measures for controlling the reliance on document statistics.

Liu and Croft's work [8] most resembles ours in that they use the Max-Scoring Passage algorithm with the basic passage model from Eq. 7; they also use a *passage-based relevance model* [10] to rank documents. We demonstrate the merits in using their methods with our passage language model in Sec. 4.

4 Evaluation

We conducted our experiments on the following four TREC corpora:

corpus	# of docs	avg. length	queries	disk(s)
FR12	45,820	935	51-100	1,2
LA+FR45	186,501	317	401-450	4,5
WSJ	173,252	263	151-200	1-2
AP89	84,678	264	1-50	1

FR12, which was used in work on passage-based document retrieval [2, 8], and LA+FR45, which is a challenging benchmark [27], contain documents that are longer on average (and often considered more heterogeneous) than those in WSJ and AP89.

We used the Lemur toolkit (www.lemurproject.org) to run our experiments. We applied basic tokenization and Porter stemming, and removed INQUERY stopwords. We used titles of TREC topics as queries.

To evaluate retrieval performance, we use the mean average (non-interpolated) precision (MAP) at 1000, and the precision of the top 10 documents (p@10). We determine statistically significant differences in performance using the two-tailed Wilcoxon test at the 95% confidence level.

Passages. While there are several passage types we can use [7], our focus is on the general validity of our retrieval algorithms and language-model induction techniques. Therefore, we use *half overlapping fixed-length windows* (of 150 and 50 terms[8]) as passages and mark them *prior* to retrieval time. Such passages are computationally convenient to use and were shown to be effective for document retrieval [2], specifically, in the language model framework [8].

[8] Passages of 25 terms yield degraded performance as in some previous reports [2, 8].

Table 1. Performance numbers of the Max-Scoring Passage algorithm (MSP) with either the basic passage language model (MSPbase) or our passage language model (MSP[$\mathcal{M}$]) that utilizes homogeneity model $\mathcal{M}$. Document-based language-model (DOCbase) retrieval performance is presented for reference. Boldface: best result per colomn; underline: best performance for a corpus per evaluation measure. d and p mark statistically significant differences with DOCbase and MSPbase, respectively.

	FR12				LA+FR45			
	PsgSize 150		PsgSize 50		PsgSize 150		PsgSize 50	
	MAP	p@10	MAP	p@10	MAP	p@10	MAP	p@10
DOCbase	22.0	13.3	22.0	13.3	22.7	26.4	22.7	**26.4**
MSPbase	28.4	14.8	30.1^d	14.8	21.9	25.5	21.7	25.7
MSP[length]	**29.6**^{d}	15.7	**31.8**^d_p	15.7	23.1$_p$	27.5	**23.6**$_p$	26.0
MSP[ent]	29.3^d	**16.2**	30.1^d	**16.2**	22.2	26.2	21.8	26.0
MSP[interPsg]	29.1^d	15.7	30.7^d	**16.2**	22.8$_p$	26.6	21.9	25.3
MSP[docPsg]	29.3^d	**16.2**	31.0^d	15.7	**23.2**$^{\mathbf{d}}$	**27.9**	23.0	25.5

	WSJ				AP89			
	PsgSize 50		PsgSize 150		PsgSize 50		PsgSize 150	
	MAP	p@10	MAP	p@10	MAP	p@10	MAP	p@10
DOCbase	28.4	39.6	28.4	39.6	**20.0**	**24.1**	**20.0**	24.1
MSPbase	28.8	41.8	26.1^d	40.4	18.8^d	23.0	17.7^d	22.4
MSP[length]	**29.3**^{d}	**43.0**^{d}	29.0$_p$	**44.8**^d_p	19.3$_p$	23.7	18.7$_p$	**24.6**
MSP[ent]	**29.3**$_p$	41.6	27.9$_p$	41.8	19.1$_p$	22.8	18.2^d_p	22.6
MSP[interPsg]	29.2^d	42.4^d	28.2$_p$	43.2$_p$	19.5$_p$	23.7	18.4^d_p	23.9
MSP[docPsg]	29.1^d	42.6^d	**29.2**$_p$	**44.8**^d_p	19.8$_p$	23.3	19.1$_p$	**24.6**

4.1 Experimental Results

Passage Language Model. To study the performance of our passage language model independently of score-integration (as performed by Interpolated Max-Scoring Passage), we use it in the Max-Scoring Passage algorithm, which was previously studied with the basic passage model [8].

Specifically, let $MSP[\mathcal{M}]$ denote the implementation of Max-Scoring Passage with our passage model $p_g^{[\mathcal{M}]}(\cdot)$, and $MSPbase$ denote its implementation with the basic passage model $p_g^{[basic]}(\cdot)$ [8]. Since our passage model leverages information from the ambient document, we also use as a reference comparison a standard document-based language-model retrieval arpproach ($DOCbase$) that scores document d by $p_d^{[basic]}(q)$.

All tested algorithms incorporate a single free parameter $\lambda_{\mathcal{C}}$, which controls the extent of corpus-based smoothing. We fix $\lambda_{\mathcal{C}}$ to 0.5, because this results in (near) optimal (MAP) performance for *both* our reference comparisons ($MSPbase$ and $DOCbase$) with respect to values in $\{0.1, 0.2, \ldots, 0.9\}$.[9]

We present the performance numbers in Table 1. Our first observation is that the Max-Scoring Passage algorithm is consistently more effective (many times to a statistically significant degree) when utilizing our new passage language model ($MSP[\mathcal{M}]$) than when using the basic passage language model ($MSPbase$).

We can also see in Table 1 that the most effective homogeneity measures for inducing our passage model are $length$ — demonstrating its correlation with

[9] Similar relative-performance patterns are observed for $\lambda_{\mathcal{C}} = 0.3$.

heterogeneity — and *docPsg*; the latter measures the similarity between a document and its passages, and is thus directly related to the balance we want to control of using document-based vs. passage-based information. Furthermore, $MSP[length]$ and $MSP[docPsg]$ yield performance that is superior to that of document-based retrieval ($DOCbase$) in many of the relevant comparisons, espeically for FR12 and WSJ. For AP89, however, document-based retrieval is superior (in terms of MAP) to using (any) passage-based information, possibly due to the high homogeneity of the documents.

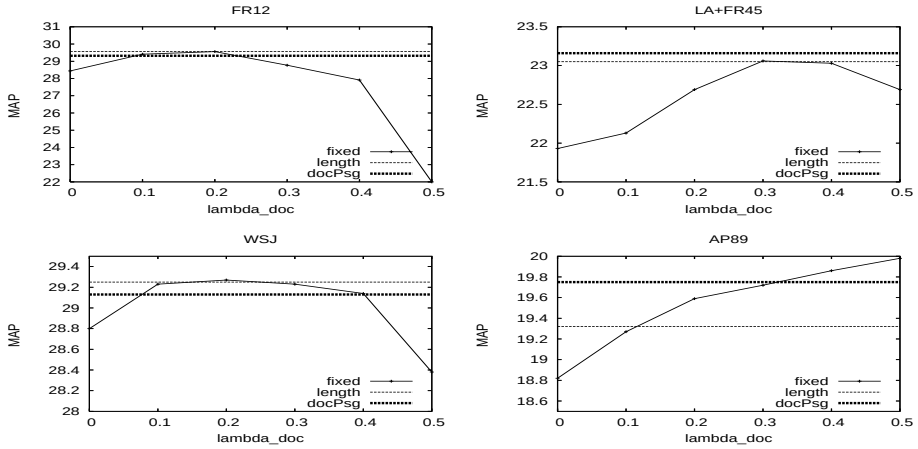

Fig. 1. The MAP performance curve of Max-Scoring Passage (PsgSize=150) when setting $\lambda_{doc}(d)$ (see Eq. 8) to the same *fixed* value in $\{0, 0.1, \ldots, 0.5\}$ for *all* documents. (0 and 0.5 correspond to $MSPbase$ and $DOCbase$, respectively.) The performance of using the homogeneity measures *length* and *docPsg* is plotted for comparison with thin and thick horizontal lines, respectively. Note: figures are not to the same scale.

Further Analysis. Our passage model incorporates information from the ambient document to an extent controlled by the estimated document homogeneity. We now study the alternative of fixing the reliance on document information to the same extent for all documents and passages, as proposed in some past work [17, 18, 20, 22]. We do so by fixing $\lambda_{doc}(d)$ in Eq. 8 to a value in $\{0, 0.1, \ldots, 0.5\}$. (Recall that $\lambda_{doc}(d) = (1 - \lambda_{\mathcal{C}})h^{[\mathcal{M}]}(d)$ and $\lambda_{\mathcal{C}} = 0.5$; also, setting $\lambda_{doc}(d)$ to 0 and 0.5 corresponds to $MSPbase$ and $DOCbase$, respectively.) We depict the resultant MAP performance curve (for passages of 150 terms) of the Max-Scoring Passage algorithm in Fig. 1. We plot for comparison the performance of using our best-performing homogeneity measures *length* and *docPsg*.

We can see in Fig. 1 that using homogeneity measures improves performance over a poor choice of a fixed $\lambda_{doc}(d)$; furthermore, for FR12, LA+FR45 and WSJ, the measures yield performance that is sometimes better than the best

performance obtained by using some fixed $\lambda_{doc}(d)$, and always better than that of using either passage-only information or document-only information (see the end points of the curves). Many of the performance improvements posted by the homogeneity measures over a fixed $\lambda_{doc}(d)$ are also statistically significant, e.g., $MSP[length]$ and $MSP[docPsg]$'s performance is better to a statistically significant degree than setting $\lambda_{doc}(d)$ to (i) 0 for LA+FR45 and AP89, (ii) 0.5 for FR12 and WSJ, and (iii) $\{0.1, 0.2\}$ for AP89.

Table 2. Performance numbers of a *passage-based relevance model* [8]. We use either the originally suggested basic passage language model (relPsgBase) or our passage language model (relPsg[$\mathcal{M}$]). Document-based relevance-model performance is presented for reference (relDoc). Best result in a colomn is boldfaced, and best result for a corpus (per evaluation measure) is underlined; statistically significant differences with relDoc and relPsgBase are marked with d and p, respectively.

	FR12				LA+FR45			
	PsgSize 150		PsgSize 50		PsgSize 150		PsgSize 50	
	MAP	p@10	MAP	p@10	MAP	p@10	MAP	p@10
relDoc	10.7	9.1	10.7	9.1	20.7	23.8	20.7	23.8
relPsgBase	**31.7**^{d}	14.3^d	31.1^d	16.2^d	**22.4**	26.0	21.9	24.7
relPsg[length]	28.0^d	14.8^d	30.7^d	**18.1**^{d}	21.8$_p$	**26.6**	**23.3**^d_p	25.3
relPsg[docPsg]	26.9^d	**15.7**^{d}	**34.2**^{d}	**18.1**^{d}	20.4$_p$	25.1	22.8^d_p	**25.7**

	WSJ				AP89			
	PsgSize 150		PsgSize 50		PsgSize 150		PsgSize 50	
	MAP	p@10	MAP	p@10	MAP	p@10	MAP	p@10
relDoc	33.9	48.4	33.9	48.4	25.6	28.5	**25.6**	28.5
relPsgBase	34.5	47.2	34.0	45.0	24.1	29.8	22.2	25.9
relPsg[length]	35.4^d	50.0	37.5^d_p	49.0$_p$	25.1	**30.4**	24.3$_p$	30.0$_p$
relPsg[docPsg]	**35.9**^{d}	**50.2**	**37.6**^d_p	**50.2**$_p$	**25.7**	29.8	25.1$_p$	**31.3**$_p$

Relevance Model. The most effective *passage-based relevance model* approach for ranking documents that was suggested by Liu and Croft [8] is to construct a relevance model [10] only from passages and use it to rank documents. We compare their original implementation $relPsgBase$, which utilizes the basic passage model, to an implementation $relPsg[\mathcal{M}]$, which utilizes our passage language model $p_g^{[\mathcal{M}]}(\cdot)$. We also use a document-based relevance model ($relDoc$) [10] as a reference comparison.

We optimize the performance of *each* of our reference comparisons ($relPsgBase$ and $relDoc$) with respect to the number of top-retrieved elements (i.e., passages or documents) and the number of terms used for constructing the relevance models; specifically, we select these parameters' values from $\{25, 50, 75, 100, 250, 500\}$ — i.e., 36 parameter settings — so as to optimize MAP performance. We set $\lambda_{\mathcal{C}} = 0.5$ (as at the above) except for estimating top-retrieved elements' language models for constructing relevance models, wherein we set $\lambda_{\mathcal{C}} = 0.2$ following past recommendations [10]. Our $relPsg[\mathcal{M}]$ ($\mathcal{M} \in \{length, docPsg\}$) algorithms use the parameter values selected for the $relPsgBase$ reference comparison; therefore, their performance is not necessarily the optimal one they can achieve.

Table 2 shows that in most of the relevant comparisons using our passage language model yields passage-based relevance models ($relPsg[\mathcal{M}]$) that outperform both the original implementation ($relPsgBase$) [8] — which utilizes the basic passage model — and the document-based relevance model ($relDoc$). (Note, for example, that underlined numbers that constitute the best performance for a corpus per evaluation metric appear only in $relPsg[\mathcal{M}]$ rows.) In many cases, the performance differences are also statistically significant.

Interpolated Max-Scoring Passage. The algorithm scores document d by interpolation (governed by the homogeneity-based interpolation weight $h^{[\mathcal{M}]}(d)$) of the document-based language model score ($DOCbase$) with the score assigned by Max-Scoring Passage. (See Sec. 2.2.) To focus on this score integration, rather than combine it with information integration at the language model level[10], which we explored at the above, we use the basic passage language model $p_g^{[basic]}(\cdot)$; the Max-Scoring Passage implementation is then the $MSPbase$ defined above.

In Fig. 2 we present the MAP performance of Interpolated Max-Scoring Passage (with passages of 150 terms). We either use $h^{[\mathcal{M}]}(d)$ with the $length$ and $docPsg$ homogeneity measures[11], or set $h^{[\mathcal{M}]}(d)$ to a fixed value in $\{0, 0.1, \ldots, 1\}$ for *all* documents (0 and 1 correspond to $MSPbase$ and $DOCbase$, respectively), which echoes some past work [2, 4].

We see in Fig. 2 that homogeneity measures yield performance that is (i) better than that of several fixed values of $h^{[\mathcal{M}]}(d)$, (ii) always better than the worse performing among $MSPbase$ and $DOCbase$ (see the end points of the curves), and (iii) sometimes (e.g., for FR12 and WSJ) better than the best performance attained by using some fixed $h^{[\mathcal{M}]}(d)$ for all documents[12]. Many of the improvements obtained by our homogeneity measures over a fixed $h^{[\mathcal{M}]}(d)$ are also statistically significant, e.g., $length$ is significantly better than setting $h^{[\mathcal{M}]}(d)$ to (i) 0 for LA+FR45, WSJ and AP89, (ii) 0.9 for FR12, (iii) $\{0.1, \ldots, 0.4\}$ for LA+FR45, and (iv) $\{0.1, 0.3\}$ for WSJ.

5 Conclusions

We derived some previously-proposed and new passage-based document-ranking approaches from the same model. We proposed an effective *passage language model* that incorporates information from the ambient document to an extent controlled by the estimated *document homogeneity*. Our homogeneity measures are also effective for integrating document and passage query-similarity information for document retrieval.

[10] Experiments show that such combination yields additional performance gains.

[11] ent and $interPsg$ yield inferior performance to that of $length$ and $docPsg$, and are omitted to avoid cluttering of the figure.

[12] These observations also hold if Dirichlet-smoothed [16] language models are used for both passages and documents.

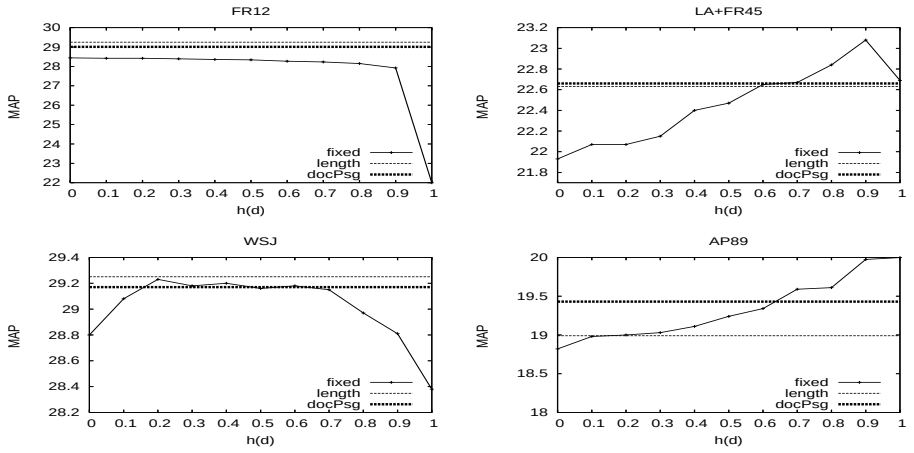

Fig. 2. The MAP performance curve of the Interpolated Max-Scoring Passage algorithm (PsgSize=150) when setting $h^{[\mathcal{M}]}(d)$ (see Eq. 4) to the same *fixed* value in $\{0, 0.1, \ldots, 1\}$ for *all* documents. (0 and 1 correspond to $MSPbase$ and $DOCbase$, respectively.) We also plot the performance of setting $\mathcal{M}$ to $length$ and $docPsg$ with thin and thick horizontal lines, respectively. Note: figures are not to the same scale.

Acknowledgments. We thank the anonymous reviewers for their helpful comments. This paper is based upon work done in part while the first author was at the Technion and the second author was at Cornell University, and upon work supported in part by the Center for Intelligent Information Retrieval and by the National Science Foundation under grant no. IIS-0329064. Any opinions, findings and conclusions or recommendations expressed in this material are the authors' and do not necessarily reflect those of the sponsoring institutions.

References

1. Salton, G., Allan, J., Buckley, C.: Approaches to passage retrieval in full text information systems. In: Proceedings of SIGIR. (1993) 49–58
2. Callan, J.P.: Passage-level evidence in document retrieval. In: Proceedings of SIGIR. (1994) 302–310
3. Mittendorf, E., Schäuble, P.: Document and passage retrieval based on hidden Markov models. In: Proceedings of SIGIR. (1994) 318–327
4. Wilkinson, R.: Effective retrieval of structured documents. In: Proceedings of SIGIR. (1994) 311–317
5. Kaszkiel, M., Zobel, J.: Passage retrieval revisited. In: Proceedings of SIGIR. (1997) 178–185
6. Denoyer, L., Zaragoza, H., Gallinari, P.: HMM-based passage models for document classification and ranking. In: Proceedings of ECIR. (2001) 126–135
7. Kaszkiel, M., Zobel, J.: Effective ranking with arbitrary passages. Journal of the American Society for Information Science **52**(4) (November 2001) 344–364

8. Liu, X., Croft, W.B.: Passage retrieval based on language models. In: Proceedings of the 11th International Conference on Information and Knowledge Managment (CIKM). (2002) 375–382
9. Croft, W.B., Lafferty, J., eds.: Language Modeling for Information Retrieval. Number 13 in Information Retrieval Book Series. Kluwer (2003)
10. Lavrenko, V., Croft, W.B.: Relevance-based language models. In: Proceedings of SIGIR. (2001) 120–127
11. Lavrenko, V.: A Generative Theory of Relevance. PhD thesis, University of Massachusetts Amherst (2004)
12. Ponte, J.M., Croft, W.B.: A language modeling approach to information retrieval. In: Proceedings of SIGIR. (1998) 275–281
13. Kurland, O., Lee, L.: Corpus structure, language models, and ad hoc information retrieval. In: Proceedings of SIGIR. (2004) 194–201
14. Buckley, C., Salton, G., Allan, J., Singhal, A.: Automatic query expansion using SMART: TREC3. In: Proceedings of of the Third Text Retrieval Conference (TREC-3). (1994) 69–80
15. Cai, D., Yu, S., Wen, J.R., Ma, W.Y.: Block-based web search. In: Proceedings of SIGIR. (2004) 456–463
16. Zhai, C., Lafferty, J.D.: A study of smoothing methods for language models applied to ad hoc information retrieval. In: Proceedings of SIGIR. (2001) 334–342
17. Abdul-Jaleel, N., Allan, J., Croft, W.B., Diaz, F., Larkey, L., Li, X., Smucker, M.D., Wade, C.: UMASS at TREC 2004 — novelty and hard. In: Proceedings of the Thirteenth Text Retrieval Conference (TREC-13). (2004)
18. Hussain, M.: Language modeling based passage retrieval for question answering systems. Master's thesis, Saarland University (2004)
19. Ogilvie, P., Callan, J.: Hierarchical language models for XML component retrieval. In: Proceedings of INEX. (2004)
20. Murdock, V., Croft, W.B.: A translation model for sentence retrieval. In: Proceedings of HLT/EMNLP. (2005) 684–695
21. Sigurbjörnsson, B., Kamps, J.: The effect of structured queries and selective indexing on XML retrieval. In: Proceedings of INEX. (2005) 104–118
22. Wade, C., Allan, J.: Passage retrieval and evaluation. Technical Report IR-396, Center for Intelligent Information Retrieval (CIIR), University of Massachusetts (2005)
23. Kurland, O., Lee, L.: PageRank without hyperlinks: Structural re-ranking using links induced by language models. In: Proceedings of SIGIR. (2005) 306–313
24. Corrada-Emmanuel, A., Croft, W.B., Murdock, V.: Answer passage retrieval for question answering. Technical Report IR-283, Center for Intelligent Information Retrieval, University of Massachusetts (2003)
25. Zhang, D., Lee, W.S.: A language modeling approach to passage question answering. In: Proceedings of the Twelfth Text Retrieval Conference (TREC-12). (2004) 489–495
26. Jiang, J., Zhai, C.: UIUC in HARD 2004 — passage retrieval using HMMs. In: Proceedings of the Thirteenth Text Retrieval Conference (TREC-13). (2004)
27. Kurland, O., Lee, L., Domshlak, C.: Better than the real thing? Iterative pseudo-query processing using cluster-based language models. In: Proceedings of SIGIR. (2005) 19–26