Discriminative Probabilistic Models for Passage Based Retrieval

Mengqiu Wang
Computer Science Department
Stanford University
353 Serra Mall
Stanford, CA 94305
mengqiu@cs.stanford.edu

Luo Si
Department of Computer Science
Purdue University
250 N. University Street
West Lafayette, IN 47907
lsi@cs.purdue.edu

ABSTRACT

The approach of using passage-level evidence for document retrieval has shown mixed results when it is applied to a variety of test beds with different characteristics. One main reason of the inconsistent performance is that there exists no unified framework to model the evidence of individual passages within a document. This paper proposes two probabilistic models to formally model the evidence of a set of top ranked passages in a document. The first probabilistic model follows the retrieval criterion that a document is relevant if any passage in the document is relevant, and models each passage independently. The second probabilistic model goes a step further and incorporates the similarity correlations among the passages. Both models are trained in a discriminative manner. Furthermore, we present a combination approach to combine the ranked lists of document retrieval and passage-based retrieval.

An extensive set of experiments have been conducted on four different TREC test beds to show the effectiveness of the proposed discriminative probabilistic models for passage-based retrieval. The proposed algorithms are compared with a state-of-the-art document retrieval algorithm and a language model approach for passage-based retrieval. Furthermore, our combined approach has been shown to provide better results than both document retrieval and passage-based retrieval approaches.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: *Retrieval Models*

General Terms

Design, Algorithms, Experimentation

Keywords

Discriminative Models, IR, Passage Retrieval

SIGIR'08, July 20–24, 2008, Singapore.

1. INTRODUCTION

Previous research has demonstrated that using passage-level evidence can improve the accuracy of document retrieval when documents are long or span different subject areas [2, 10, 16]. However, the performance of passage-based retrieval is mixed when this approach is applied to a variety of test beds with different characteristics [12, 16].

One important reason of the inconsistent performance is that there exists no unified framework to model the evidence of individual passages within a document. Most previous research only considered evidences from the best matching passage in each document for ranking documents (e.g., [2, 11, 16]), while some other previous work used methods that would require a significant amount of tuning effort to combine the evidence of several top ranked passages (e.g., [7, 20, 23]).

To improve the performance of passage-based retrieval, this paper proposes two probabilistic models to estimate the probability of relevance of a document given the evidence of a set of top ranked passages in the document. The first probabilistic model captures the retrieval criterion that a document is relevant if any passage in the document is relevant. Since the first model estimates the probability of relevance for each passage independently, the model is called the independent passage model. On the other side, the second probabilistic model goes a step further and models the correlation among individual passages by analyzing content similarities. The second model is called the correlated passage model. Both these models are trained in a discriminative manner on a set of training queries. Furthermore, a combination approach is proposed to combine the ranked lists of document retrieval and passage-based retrieval for more accurate retrieval results. The probabilistic modeling method we proposed in this work can be classified as a density estimation method [4, 6]. To the best of our knowledge, this is the first research work that applies jointly modeling of passage evidence and discriminative training methods for passage-based retrieval.

Empirical studies have been conducted on four different TREC test beds to show the effectiveness of the proposed discriminative probabilistic models for passage-based retrieval. The new models are compared with the document retrieval algorithm and a language model approach for passage-based retrieval. Experiment results demonstrate that our first independent passage model always outperforms the language model approach of passage-based retrieval, while the second correlated passage model performs even better by consid-

ering the similarity information among passages. Both of the two proposed models are robust. In particular, the performance of the correlated model is at the same level or much better than the document retrieval method on all the test beds. Furthermore, experiments have shown that the proposed combination approach can improve the retrieval results of both the document retrieval approach and the passage-based retrieval approach.

The next section discusses related work. Section 3 describes the new probabilistic passage-based retrieval models as well as the new combined retrieval approach. Section 4 explains our experimental methodology. Section 5 presents the experimental results and the corresponding discussions. And finally section 6 gives conclusions and future work.

2. RELATED WORK

Passage retrieval has been an very attractive research direction in IR for two main reasons. The first apparent advantage of passage retrieval is that users are able to locate relevant information from returned passages much faster than if they were overloaded with long full-text documents [2, 20, 24, 22].

The second attraction of passage-based retrieval is that passage-level evidence can be used to improve effectiveness on many IR tasks. For example, Allan '95 [1] showed that using fixed window passages instead of full documents is a more effective approach for relevance feedback. MultiText retrieval system [3] also demonstrated the effectiveness of passage retrieval in a variety of tasks including query refinement and document relevance ranking. This brings us to the focus of this study — how passage-level evidence can be used to make more accurate relevance predictions for documents.

Salton et al. [20] experimented with a vector-space passage retrieval model on an encyclopedia collection and showed that retrieving paragraphs and sections instead of full-text documents yielded a significant gain on mean average precision.

UMass's system at TREC 2003 HARD track [9] experimented with passage retrieval using a language model approach. But in most cases, their passage retrieval results did not surpass document retrieval baseline.

Callan '94 [2] compared different passage creation methods, and tested on four TREC 1 and 2 collections using INQUERY retrieval system. For passage retrieval, the top 1 passage was considered for each query. Empirical results showed that using fixed-size window based passages was much more effective than paragraph-based passages. A significant improvement was found on the Federal Register (FR) collection (contains mostly long documents), and moderate improvement were found on two other collections. He also showed that combining document and passage retrieval scores using simple heuristic gave another 2% gain.

Zobel et al. [25] conducted similar experiments on TREC disk 1 and 2 collections using a vector-space model. Their findings were similar to Callan '94 that section-based passages degrade retrieval performance. Retrieval based on fixed-size window passages was shown to be more effective than document baseline on the FR part of the collection, but test results on the whole collection showed no significant difference.

A later study done by Xi et al. [24] re-examined fixed-size window passages on TREC 4 and 5 collections using their in-house vector-space retrieval model. Contrary to Callan '94, they did not obtain an improvement using a linear combination of passage score and whole-document score.

Hearst and Plaunt '93 [7] proposed an approach for dividing documents into passages in a way that reflects the underlying subtopic structure of the document. Their method first broke a document into small blocks of 3-5 sentences, then computed the cosine similarity of neighboring blocks based on adjusted tf-idf. Blocks were then linearly grouped into passages based on finding dramatic changes in the cosine similarity between neighbors. They conducted retrieval experiments on the Ziff subset of TIPSTER collection (containing mostly long documents) using the SMART system. And by summing up the scores of passages belonging to the same document, they obtained a substantial improvement on some of the P5-P30 measures.

Knaus, Mittendorf and Schäuble [18, 13] presented a different approach for finding passage boundaries that was based on hidden Markov model. Their preliminary experimental results favored their method over a baseline sentence retrieval model, but no comparison with full-text document retrieval was given.

Wilkinson '94 [23] designed a set of heuristic functions for assigning scores to section-based passages based on section types. In one set of experiment the top 1 passage score was used for re-ranking documents, but in a second experiment he introduced another heuristic function based on cosine similarity of passages to combine scores of passages belonging to the same document. Evaluated on a subset of the FR collection, results of the passage-based retrieval were mixed.

The work that are most closely related to ours are the work done by Kaskiel and Zobel '01 [11] and Liu and Croft '02 [16]. Kaskiel and Zobel '01 [11] used a vector-space model as their basis to compare passage retrieval results of different passage types against document retrieval on 5 different TREC collections — FR-12, FR-24, TREC24, TREC25 and WSJ. They showed consistent and significant improvements over document retrieval baseline on the longer FR collections, but the results were either negative or mixed on the shorter TREC24, TREC-45 and WSJ collections.

More recently, Liu and Croft '02 showed very similar results to Kaskiel and Zobel in their study [16]. They compared language model based passage retrieval using only the top 1 passage and document retrieval also using language model, and obtained significant improvement on FR-12 collection. But their results also showed noticeable drop on TREC-45 and AP collections. Note that our focus is passage-based retrieval algorithms that do not use query expansion and pseudo relevance feedback in order to make fair comparison with document retrieval approach that does not use query expansion and pseudo relevance feedback. Therefore, we do not compare with Liu and Croft's method based on the relevance model.

In general, results of the passage-base retrieval systems that we described were somewhat mixed. Most systems that were tested on long document collections gained improvements over document retrieval baseline, but those that were tested on other types of collections often resulted in significant drop in performance.

Another line of research that has recently received increasing attention in IR is discriminative methods for document retrieval tasks. Traditional approaches for document retrieval such as LM and BIR explicitly model the generation process of query terms from documents. However, some

modeling assumptions that are often made in these models, such as query terms are generated independently [19] from documents, are often violated in reality and thus pose limitations to these models [19, 5]. Discriminative models on the other hand, make fewer assumptions and allow arbitrary features to be incorporated into the model. Often trained to directly optimize retrieval performance [5, 19], these models have been shown to achieve significant improvements over state-of-the-art generative models in both document retrieval [5] and home-page finding tasks [19].

In this work, we propose two novel probabilistic models for passage-based retrieval which are trained discriminatively. We will describe our models and our training methods in more details in the next section.

3. PROBABILISTIC PASSAGE MODELS

This section first proposes two discriminative probabilistic models for passage-based retrieval. It then presents a combined retrieval approach that combines both passage-based retrieval and document-based retrieval.

3.1 Independent Passage Model

Our first model aims at accurately estimating the probability of relevance of each individual passage, as well as providing a probabilistic framework for combining the evidence from different passages to make a joint prediction for the document's relevance.

Given a set of n top-ranked passages $\vec{s} = \{s_1, \cdots, s_n\}$ of a document d, the relevance judgment criterion states that the document d is relevant if any one of the n passages is relevant. In other words, the relevance probability of document d is one minus the probability of all n passages being irrelevant. This can be formulated as the following

$$P(Y=1|d) = 1 - P(\vec{Z} = \vec{0}|\vec{s})) = 1 - \prod_{i=1}^{n}(P(Z_i = 0|s_i))$$

where Y and Z_i are Boolean variables that have value 1 when d and s_i are relevant, respectively, and 0 otherwise. An independent assumption is made here that whether a specific top-ranked passage s_i is relevant or not is not related to the relevance of any other top-ranked passage.

We use the parametric form of logistic regression to model the probability of relevance of a passage as $P(Z|s)$:

$$P(Z=1|s) = \frac{1}{1+\exp(f(\vec{s})\vec{\theta})}$$
$$P(Z=0|s) = \frac{\exp(f(\vec{s})\vec{\theta})}{1+\exp(f(\vec{s})\vec{\theta})}$$

where $f(\vec{s})$ is a feature vector of the passage s, and $\vec{\theta}$ is the corresponding weight vector.

Given a training set consists of m queries $\{q_1, \cdots, q_m\}$, each with a set of documents $\{d_1^m, \cdots, d_k^m\}$ and their relevance judgment $\{t_1^m, t_2^m, \cdots, t_k^m\}$, we can express the conditional likelihood of the data as the following:

$$\prod_{m=1}^{M}\prod_{k=1}^{|q_m|} P(Y_k^m = 1|d_k^m)^{t_k^m} P(Y_k^m = 0|d_k^m)^{1-t_k^m}$$

where each Y_k^m is a Boolean random variable that has value 1 when d_k^m is relevant and 0 otherwise.

To train the model, we used the BFGS Quasi-Newton [8] method to estimate the parameters $\vec{\theta}$ that would maximize the conditional log likelihood of the data $\mathcal{L}$, which can be expressed as

$$\begin{aligned}\mathcal{L} &= \sum_{m=1}^{M}\sum_{k=1}^{|q_m|}[t_k^m \log(P(Y=1|d_k^m)) \\ &\quad + (1-t_k^m)\log(P(Y=0|d_k^m))] \\ &= \sum_{m=1}^{M}\sum_{k=1}^{|q_m|}\Bigg[t_k^m \log(1 - \prod_{i=1}^{|d_k^m|} P(Z_i^{d_k^m} = 0|s_i^{d_k^m})) \\ &\quad + (1-t_k^m)\log(\prod_{i=1}^{|d_k^m|} P(Z_i^{d_k^m} = 0|s_i^{d_k^m}))\Bigg]\end{aligned}$$

Since the focus of this study is on the probabilistic models rather than feature engineering, we only used two features in the independent passage model — the rank and the score of each passage that come from the baseline language model retrieval model. Despite the simplicity of these features, we were able to obtain substantial improvements over the language model approaches, as we will show in Section 4.

3.2 Correlated Passage Model

The independent passage model makes an assumption about the independence among individual passages. Although this assumption is reasonable for certain documents, such as the ones that contain short passages summarizing multiple topics, we would expect this assumption not to hold in many other cases, especially for long documents that elaborate on a single topic. To address this limitation, we take a step further in this model and exploit the correlations among individual passages. Among many different kinds of passage correlations, we focus on the correlations that are characterized by the content similarity of passages. We compute pair-wise passage similarity using the cosine similarity measure based on tf-idf vectors of the passages.

For computing the cosine similarity, we first performed a standard L2 normalization on each passage's tf-idf vector. We use $\vec{c}\{c_1, \cdots, c_n\}$ to denote an unnormalized tf-idf vector of a passage, where n is the vocabulary size. A L2-normalized vector $\vec{c'}$ is calculated as

$$\vec{c'} = \{c_1', \cdots, c_n' \mid c_i' = \frac{c_i}{\sqrt{\sum_{j=1}^{n} {c_j}^2}}, 1 \le i \le n\}$$

Then we created a document background vector by summing up the tf-idf vectors of individual passages that belong to the same document. Finally we subtracted the document background vector from each passage's vector, and use the resulting vector to calculate cosine similarity. The cosine similarity of two vector $\vec{a}$ and $\vec{b}$ (both of length n)is given as

$$\cos(\vec{a}, \vec{b}) = \frac{\sum_{i=1}^{n} a_i b_i}{\sqrt{\sum_{j=1}^{n} a_i^2}\sqrt{\sum_{j=1}^{n} b_i^2}}$$

In the independent passage model described in last section, the conditional probability of a set of passages $\vec{s} = \{s_1, \cdots, s_n\}$ having relevance judgments $\vec{v} = \{v_1, \cdots, v_n\}$ is expressed as

$$P(\vec{v}|\vec{s}) = \prod_{i=1}^{n} P(Z_i = 1|s_i)^{v_i} (P(Z_i = 0|s_i))^{1-v_i}$$

We now re-define this term to be

$$\begin{aligned} P(\vec{v}|\vec{s}) =& \frac{1}{\mathcal{Z}} \exp\left(\sum_{i}^{n} \log\left(P(Z_i = 1|s_i)^{v_i} P(Z_i = 0|s_i)^{1-v_i} \right) \right. \\ & \left. + \frac{\alpha}{|\vec{s}|} \sum_{i,j(i<j)} g(\omega_{ij})(v_i v_j) \right) \\ =& \frac{1}{Z} \exp\left(\sum_{i}^{n} (1 - v_i)(f(\vec{s_i})\vec{\theta}) - \log(1 + \exp(f(\vec{s_i})\vec{\theta})) \right. \\ & \left. + \frac{\alpha}{|\vec{v}|} \sum_{i,j(i<j)} g(\omega_{ij})(v_i v_j) \right) \end{aligned}$$

where ω_{ij} denotes the correlation (cosine similarity) between passage s_i and s_j; α is a parameter for controlling the weight of the correlation term; $g(\omega)$ is a normalizing function defined as

$$g(\omega) = \begin{cases} 0, & \text{if } \omega < t \\ \frac{1}{1-t}(\omega - t), & \text{otherwise} \end{cases}$$

where t is a threshold parameter and $\mathcal{Z}$ is a partition function defined as $\mathcal{Z} = \sum_{gen(\vec{v'})} P(\vec{v'}|\vec{s})$. Here we use the notation $gen(\vec{v'})$ to denote a generation function that enumerates over all possible values of the zero-one vector v' exactly once.

The first operand of the plus operator models the independent passage relevance, while the second operand models the passage correlations. Notice that when α is set to 0, the model becomes exactly the same as the independent passage model.

For training, we first used the independent passage model to obtain the $\vec{\theta}$ values, then we used 10x10 depth-4 grid-search method to choose the α and t values by directly optimizing mean average precision (MAP) on the training set. We also experimented with gradient-based search methods to optimize the conditional probability instead of MAP. The resulting objective function is very bumpy, yielding many bad local minimas, and the results of gradient-based search are not as good as the grid search. Since this model is a generalization of the independent passage model, it is guaranteed to improve MAP on training set over the first model. Despite the greedy nature of the grid-search method, we observed good generalization performance on the final test set, as we will show in Section 4.

3.3 Combination Model

Previous research [2] has demonstrated that combining passage-level evidence with document-level evidence often yields better retrieval results than any one model alone. We propose an approach for systematically combining passage retrieval results with document retrieval results.

Our approach is similar to the CombMNZ method described in Lee '97 [15]. We first take the top n ranked lists of document retrieval and passage retrieval. Then for each document, we count the number of times it occurred in the two rank lists (denoted as c), and linearly interpolate the two retrieval scores to produce the combined score (we use d and p to denote document and passage retrieval scores, and s for combined score). The formulae we used is

$$s = (\beta * p + (1 - \beta) * d) * c$$

We used a greedy search procedure similar to the one described in Section 3.2 to find parameters n and β that maximize MAP on training set. And finally the combined scores were used to re-rank documents on the testing set.

4. EXPERIMENTS

4.1 Data Set

We evaluated our models on four TREC data sets: the Associated Press newswire 1998-1990 (AP) with TREC topics 51-150, the Federal Register 1988 and 1989 (FR-12) with TREC topics 51-150, all the data on TREC disk 4 and 5 (TREC-45) with TREC topics 351-450, and the Wall Street Journal 1986-1989 (WSJ) with TREC topics 51-150. The FR-12 collection is the smallest in size, containing less than 1/5, 1/10 and 1/4 as many documents as the AP, TREC-45 and WSJ collections, respectively. But the documents in FR-12 are much longer on average, and have a larger variance in length than the documents in AP and TREC-45. The TREC-45 collection represents a heterogeneous collection, including materials from 5 different document sources. Queries are taken from the "title" fields of TREC topics. Document-level relevance judgments come from the judged pool of TREC participating teams. Queries without relevant documents in the judged pool are removed from the query set. Table 1 and Table 2 gives detailed statistics of the collections and query sets. Previous work on passage-based retrieval has shown negative or mixed results on AP, TREC-45 and WSJ collections [11, 16].

For training and testing, we split the 100 queries used for AP, TREC-45 and WSJ sets into the first half and the second half, and then we carried out 2-fold cross-validations. In each fold, no test data was seen at training time for all our models. For the FR-12 collection, since Liu and Croft '02 [16] used only queries 51-100 in their experiments, we added queries 101-150 for training, while leaving out queries 51-100 for testing.

4.2 Experiment Setup

In all our experiments, both query and documents were stemmed using Krovetz stemmer. All punctuations in the queries were replaced with spaces, and no acronym expansion or replacement were used. Stopwords were removed based on the standard INQUERY 418 words stoplist [14].

In order to compare with generative models based on language model, we used Indri[1] retrieval engine [17] to retrieve top 1000 passages for each query. And all 1000 passages of each query from Indri's passage retrieval were used as inputs to our models for training and testing. Passages were created

[1]Indri is a state-of-the-art retrieval engine that combines the merits of language model and inference network. Strictly speaking, it is not a purely LM-based system, but we think it is a very strong generative baseline to compare against. Descriptions of Indri's retrieval model can be found here: *http://ciir.cs.umass.edu/ǫńmetzler/indriretmodel.html*.

Table 1: Statistics of test collections (top 3 rows adopted from Liu & Croft 2002)

Collection	# of Docs	Size	Average # of Words/Doc	Std Dev. of Doc Length	Contents
AP	242,918	0.73 Gb	272.3	132.72	Associated Press newswire 1988-90 (from TREC disk 1-3)
FR-12	45,820	0.47 Gb	873.9	2514.16	Federal Register 1988-89 (from TREC disk 1-2)
TREC-45	556,077	2.13 Gb	305.3	775.78	The Financial Times 1991-94, Federal Register 1994, Congressional Record 1993, Foreign Broadcast Information Service, the LA Times
WSJ	173,252	0.51 Gb	390.7	435.67	Wall Street Journal (1986, 1987, 1988, 1989)

Table 2: Query set statistics. For FR-collection, queries 101-150 were used for training and 51-100 were used for testing. For AP, TREC-45 and WSJ collections, we performed 2-fold cross-validations using the half-half splits on the query sets.

		# of Queries			# of Relevant Documents			
		with	Average			Per Query (for queries with rel. docs)		
Collection	Queries	Relevant Docs	Query Length	Std. Dev. of Query Length	Total	Avg	Min	Max
AP	TREC topics 51-100	49	3.8	2.00	11946	243.8	2	1142
	TREC topics 101-150	50	4.9	1.91	9883	197.7	11	847
FR-12	TREC topics 51-100	21	4.1	2.14	502	23.9	1	118
	TREC topics 101-150	33	4.7	1.83	406	12.3	1	103
TREC-45	TREC topics 301-350	50	2.7	0.80	4611	92.2	3	474
	TREC topics 351-400	50	2.5	0.7	4674	93.5	7	361
WSJ	TREC topics 51-100	50	3.8	1.99	6228	124.56	3	591
	TREC topics 101-150	50	4.9	1.91	4556	91.12	2	338

using half-overlapping windows of size 50, and we took the top 3 ranked passages of each document as input to our models. We also used Indri as the baseline document retrieval system for comparison. In both Indri's passage retrieval and document retrieval, we used Jelinek-Mercer smoothing method (linear smoothing) with the lambda parameter set to 0.5.

5. RESULTS

An extensive set of experiments were conducted on 3 test beds to address the following questions:

1) How good are the proposed passage-based retrieval algorithms compared with other passage-based retrieval algorithms? Experiments are conducted to compare the two new discriminative probabilistic models with the language model based algorithm for passage-based retrieval [16].

2) Whether the proposed passage-based retrieval algorithms are robust compared with document retrieval algorithm? Experiments are conducted to compare our passage-based retrieval algorithms with the document retrieval algorithm described in Section 4.2.

3) Whether the combined retrieval approach can further improve the accuracy of both passage-based retrieval and document retrieval? Experiments are conducted to compare our approach of combining the language model based document retrieval with our discriminative correlated passage-based retrieval, to the language model based passage-retrieval and document retrieval approaches.

To provide more detailed information, we did statistical significance test to compare the results. Sanderson and Zobel '05 [21] showed in their recent paper that t-test is more reliable than sign test or Wilcoxon test, and mean average precision (MAP) is more reliable than precision at rank 10. Inspired by their findings, we report our results mainly using MAP and report significance test results of non-directional paired t-test.

5.1 Our Models vs. Language Model on Passage Retrieval

From Table 3 we can see that on all four collections, all of our models consistently out-performed language model passage retrieval. On AP, TREC-45 and WSJ collections, the improvements over LM-based passage retrieval from both our independent model and correlated model were found to be statistically significant with large margin. Two-tailed paired t-tests on comparing the correlated model with the LM passage model showed p-values of 1.85e-10 on AP, 9.9e-4 on TREC-45 and 3.37e-11 on WSJ.

On FR-12 data set, our correlated model also made a substantial improvement over the LM-based passage retrieval (over 3 percentage points and over 13% relative improvement in MAP), but the difference was not found to be statistically significant. However, it is worth noting that the number of queries with relevant documents on the testing set of FR-12 collection is smaller (21), as compared to 99 and 100 queries on the AP and TREC-45 sets (we performed

Table 3: Comparison of mean average precision with Indri LM-based passage retrieval using half-overlapping window of size 50 on FR-12, AP, TREC-45 and WSJ datasets. *LM-psg-lin-0.5* denotes Indri LM with linear smoothing parameter set to 0.5. *Inde* denotes our independent passage model, *Corr* denotes our correlated passage model. $\%^{+}_{-}$*LM* denotes relative percentage change over LM-psg-lin-0.5. Best results on each collection are highlighted. The †symbol indicates statistical significance at 0.95 confidence interval.

Collection	LM-psg lin-0.5	Inde	$\%^{+}_{-}$LM	Corr	$\%^{+}_{-}$LM
FR-12	0.2751	0.2870	+4.32	0.3110	+13.05
AP	0.1785	0.2063	+15.57†	0.2084	+16.75†
TREC-45	0.1749	0.1860	+6.35†	0.1870	+6.92†
WSJ	0.2043	0.2331	+14.10†	0.2345	+14.78†

Table 4: Comparison of mean average precision with Indri LM document retrieval on FR-12, AP, TREC-45 and WSJ datasets. *LM-doc-lin-0.5* denotes Indri LM document retrieval with linear smoothing parameter set to 0.5. *Inde* denotes our independent passage model, *Corr* denotes our correlated passage model. $\%^{+}_{-}$*Doc* denotes relative percentage change over LM-doc-lin-0.5. Best results on each collection are highlighted. The †symbol indicates statistical significance at 0.95 confidence interval.

Collection	LM-doc lin-0.5	Inde	$\%^{+}_{-}$Doc	Corr	$\%^{+}_{-}$Doc
FR-12	0.2144	0.2870	+33.86†	0.3110	+45.06†
AP	0.2088	0.2063	-1.20	0.2084	-0.19
TREC-45	0.1891	0.1860	-1.64	0.1870	-1.11
WSJ	0.2188	0.2331	+6.54†	0.2345	+7.18†

2-fold cross-validations on these two sets, and therefore we have full test results for all queries, see Table 2). Therefore the power of the significance test on FR-collection is weaker, giving rise to a higher chance of *type II* errors [21]. Table 6 gives more detailed statistics of different model's performance on FR-12 collection, and thus helps us to make a better comparison. We noticed from the table that our correlated model improves over LM-base passage retrieval at all recall levels and on every other measure. The improvements are most salient in *P5-P20* and the top 5 recall levels of the 11-point precisions. We think that improvements on these measures are particularly important, perhaps more so than MAP, because the top documents are most likely to have an impact on user's perception of retrieval quality.

We can also see from Table 3 that our correlated model is consistently better or at leaset as good as the independent model on all four data sets. This result indicates that incorporating passage similarity correlations into our model helped us to make more accurate predictions of document relevance than treating passages solely independently. The parameter values of the learned correlated model are given in Table 7 to aid reproducing our results. Note that for the WSJ corpus, when training on topics 101-150 the passage correlation is not found helpful, and therefore the α weight is set to 0.0, which means we back off to the independent model.

Table 5: Comparison of combination model with Indri LM-based document retrieval and passage retrieval. *LM-doc-lin-0.5* denotes Indri LM document retrieval with linear smoothing parameter set to 0.5, and *LM-psg-lin-0.5* denotes Indri LM passage retrieval with the same setting. *Combo* denotes our combination model. Best results on each collection are highlighted. The †symbol indicates statistical significance at 0.95 confidence interval.

Collection	LM-doc -lin-0.5	LM-psg -lin-0.5	Combo	$\%^{+}_{-}$doc	$\%^{+}_{-}$psg
FR-12	0.2144	0.2751	0.3123	+45.67†	+13.52
AP	0.2088	0.1785	0.2167	+3.78†	+21.40†
TREC-45	0.1891	0.1749	0.1978	+4.60†	+13.09†
WSJ	0.2188	0.2043	0.2431	+11.11†	+18.99†

5.2 Comparison with Document Retrieval

On the FR-12 dataset, both our independent model and correlated model out-performed document retrieval algorithms with very significant margin, as shown in the FR-12 row of Table 4. A more thorough analysis of several different measures such as R-precision, P5-P10 and 11-point precision showed that our models retrieved 17.2% more relevant documents that were not retrieved by document retrieval, and precision on almost all recall levels were significantly better than document retrieval.

Our two models also significantly outperformed the document retrieval model on the WSJ dataset. On AP and TREC-45 datasets, the retrieval performance obtained by our models were extremely close to the performance of document retrieval algorithms (see Table 4), with relative percentage changes ranging from 0.19% to 1.64%, and p-value of t-statistics as high as 0.9376 (the higher the p-value, the less likely that the differences are statistically significant).

When we compared the LM-based passage retrieval against LM-based document retrieval (see Table 5), although LM-based passage retrieval showed a substantial improvement on long document collection (FR-12), its performance was significantly worse than document retrieval on the other three shorter collection (-14.51% reduction in MAP on AP, -7.50% on TREC-45, and -6.67% on WSJ, with t-test p-values 1.9e-05, 9.3e-03 and 1.2e-02). In contrast, our independent model and correlated model achieved higher improvements on FR-12 and WSJ, and performed at the same level as document retrieval on AP and TREC-45 set with statistically insignificant differences.

Our results confirm the deficiency of LM-based passage-retrieval models on shorter collections, which was also observed in Liu and Croft '02 [16], and also demonstrated the robustness of our proposed models.

5.3 Combination Model Results

The combination model that we proposed was able to combine the merits of our passage retrieval models and document retrieval, and achieved the best performance on all four collections (see Table 5). The improvement of combi-

Table 6: Detailed comparison with passage retrieval using half-overlapping window size of 50 on FR-12. *LM-psg-lin-0.5* denotes Indri LM with linear smoothing parameter set to 0.5. *Inde* denotes our independent passage model, *Corr* denotes our correlated passage model. $\%^{+}_{-}LM$ denotes relative percentage change over LM-psg-lin-0.5. The total number of relevant docs over all queries on FR-12 is 502. Rel-ret is the total number of rel docs retrieved; R-prec is precision after R (= rel) retrieved docs; ircl-prn lists the 11-point precisions.

Passage-based Retrieval Results on FR-12					
	LM-psg -lin-0.5	Inde	$\%^{+}_{-}$LM	Corr	$\%^{+}_{-}$LM
rel-ret	275	280	+1.8	280	+1.8
R-prec	0.2690	0.2659	-1.2	0.3142	+16.8
recip-rank	0.3491	0.3972	+13.8	0.4268	+22.3
P5	0.1429	0.1714	+19.9	0.1714	+19.9
P10	0.1333	0.1571	+17.9	0.1571	+17.9
P15	0.1206	0.1460	+21.1	0.1429	+18.5
P20	0.1190	0.1310	+10.1	0.1333	+12.0
P30	0.1159	0.1222	+5.4	0.1222	+5.4
ircl-prn					
0.00	0.3880	0.4252	+9.6	0.4483	+15.5
0.10	0.3784	0.4011	+6.0	0.4251	+12.3
0.20	0.3434	0.3754	+9.3	0.4003	+16.6
0.30	0.3339	0.3579	+7.2	0.3817	+14.3
0.40	0.2811	0.2977	+5.9	0.3205	+14.0
0.50	0.2771	0.2841	+2.5	0.3069	+10.8
0.60	0.2578	0.2643	+2.5	0.2858	+10.9
0.70	0.2560	0.2552	-0.3	0.2797	+9.3
0.80	0.2166	0.2116	-2.3	0.2354	+8.7
0.90	0.2087	0.1957	-6.2	0.2199	+5.4
1.00	0.1947	0.1788	-8.2	0.2025	+4.0
map	0.2144	0.2870	+4.3	0.3110	+13.1

nation model over document retrieval was statistically significant on all four data sets, with p-values 0.045 on FR, 0.029 on AP, 0.038 on TREC-45 and 3.8e-06 on WSJ. The results were also significantly better than LM-based passage retrieval on all collections except for FR-12. However, as we explained in Section 5.1, due to the on the small number of queries on the testing portion of the FR-12 collection, the power of the significance test on FR-collection is weaker.

To better illustrate why our combination model worked well, we draw the β and n values that were learnt from each training set in Table 8. As we can see from the table, the combination model learned to choose very small β values for collections TREC-45 and AP in the process of optimizing MAP on training set, and therefore assigning more weights to the document retrieval scores in the final combination scores. On the other hand, the model learnt a much larger β value for collection FR-12, putting more weight on passage retrieval scores. The learned model corresponds well to our knowledge about the relative strength and weakness of document retrieval and passage retrieval on these collections, and was able to adapt to different collections and to make decisions better than any one model alone.

5.4 Overall Comparison

Table 7: Parameter values of the correlated passage model, learnt from each training set

Collection	Training Set	α	t
FR-12	TREC topics 101-150	4.8	0.298
AP	TREC topics 51-100	3.5	0.730
	TREC topics 101-150	3.9	0.850
TREC-45	TREC topics 301-350	3.0	0.695
	TREC topics 351-400	4.3	0.899
WSJ	TREC topics 51-100	6.0	0.420
	TREC topics 101-150	0.0	-

Table 8: Parameter values of the combination model, learnt from each training set

Collection	Training Set	β	n
FR-12	TREC topics 101-150	0.7184	500
AP	TREC topics 51-100	0.0312	700
	TREC topics 101-150	0.0328	1000
TREC-45	TREC topics 301-350	0.0272	2000
	TREC topics 351-400	0.0220	1100
WSJ	TREC topics 51-100	0.0584	800
	TREC topics 101-150	0.1024	800

In summary, our independent passage model achieved 33+% relative improvement over the state-of-the-art Indri document retrieval algorithm on long document collections (FR-12) and maintained a level of performance better or at least as good as the document retrieval algorithm on short (AP), medium length (WSJ) and heterogeneous (TREC-45) collections. By modeling passage correlations in the correlated passage model, we further improved retrieval results on all four collections. And finally, through combining passage-level and document-level evidences in our combination model, we obtained even further improvements consistently across all collections, and out-performed the document retrieval algorithms on all four collections.

Our models also gave consistent and significant gains over state-of-the-art LM-based passage retrieval approaches. Empirical results demonstrated that our method overcomes the inconsistency problem found in LM-based passage retrieval approaches and our models were very robust across different test beds with diverse characteristics.

6. CONCLUSIONS AND FUTURE WORK

It is an intuitive idea to utilize passage-level evidence for document retrieval. Passage-based retrieval has shown promising results in collections where the documents are long or span different subject areas. However, most prior passage-based retrieval algorithms were not robust enough when they were tested on collections with different types of corpus characteristics. Many previous passage-based retrieval algorithms only considered the evidence of the best matching passage in each document for ranking available documents, while some other previous work used heuristics to combine the evidence of several top ranked passages.

In this paper, we have proposed a novel probabilistic framework for formally modeling the evidence of individual passages in a document. Our first probabilistic model captures the retrieval criterion that a document is relevant if any pas-

sage of the document is relevant and models individual passages independently. The second probabilistic model goes a step further and takes into account the content similarities among passages. The proposed probabilistic models of passage-based retrieval are trained in a discriminative manner. Furthermore, we have presented an approach to combine document-level and passage-level evidence.

The proposed models were evaluated on four TREC collections with diverse characteristics. Our models achieved consistent and significant improvements over state-of-the-art language model approaches to passage retrieval. Whereas previous approaches to passage retrieval have shown mixed results when tested on collections of different characteristics, we have demonstrated that our proposed models are much more robust and performed consistently across collections. Furthermore, our proposed combination approach has demonstrated promising results on all four collections, with significant improvements over both document retrieval and passage-based retrieval.

There are several possibilities to extend the research in this paper. For example, a more sophisticated query-specific combination approach may provide more accurate results, which automatically adjusts the weights on passage-based retrieval and document retrieval with respect to the characteristics of user queries.

7. ACKNOWLEDGMENTS

This research was partially supported by the NSF grant IIS-0749462 and a research grant from the State of Indiana. Any opinions, findings, conclusions, or recommendations expressed in this paper are the authors', and do not necessarily reflect those of the sponsor.

8. REFERENCES

[1] J. Allan. Relevance feedback with too much data. In *Proc. of SIGIR*, 1995.

[2] J. P. Callan. Passage-level evidence in document retrieval. In *Proc. of SIGIR*, 1994.

[3] G. V. Cormack, C. L. A. Clarke, C. R. Palmer, and S. S. L. To. Passage-based query refinement. *IPM*, 36(1):133–153, 1999.

[4] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley Interscience, 2nd edition, 2002.

[5] J. Gao, H. Qi, X. Xia, and J.-Y. Nie. Linear discriminant model for information retrieval. In *Proc. of SIGIR*, 2005.

[6] D. J. Harper. *Relevance Feedback in Document Retrieval Systems: An Evaluation of Probabilistic Strategies*. PhD thesis, Cambridge University, 1980.

[7] M. A. Hearst and C. Plaunt. Subtopic structuring for full-length document access. In *Proc. of SIGIR*, 1993.

[8] J. J. E. Dennis and R. B. Schnabel. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Soc for Industrial and Applied Math, 1996.

[9] N. A. Jaleel, A. Corrada-Emmanuel, Q. Li, X. Liu, C. Wade, and J. Allan. Umass at trec 2003: Hard and qa. In *Proc. of TREC-12*, 2003.

[10] M. Kaskiel and J. Zobel. Passage retrieval revisited. In *Proc. of SIGIR*, 1997.

[11] M. Kaskiel and J. Zobel. Effective ranking with arbitrary passages. *JASIS*, 52(4):344–364, 2001.

[12] K. Kise, M. Junker, A. Dengel, and K. Matsumoto. Experimental evaluation of passage-based document retrieval. In *Proceedings of the 6th International Conference on Document Analysis and Recognition*, 2001.

[13] D. Knaus, E. Mittendorf, and P. Schauble. Improving a basic retrieval method by links and passage level evidence. In *Proc. of TREC-3*, 1994.

[14] V. Lavrenko and B. Croft. Relevance based language models. In *Proc. of SIGIR*, 2001.

[15] J. H. Lee. Analyses of multiple evidence combination. In *Proc. of SIGIR*, 1997.

[16] X. Liu and B. Croft. Passage retrieval based on language models. In *Proc. of CIKM*, 2002.

[17] D. Metzler and B. Croft. Combining the language model and inference network approaches to retrieval. *IPM Special Issue on Bayesian Networks and Information Retrieval*, 40(5):735–750, 2004.

[18] E. Mittendorf and P. Schäuble. Document and passage retrieval based on hidden markov models. In *Proc. of SIGIR*, 1994.

[19] R. Nallapati. Discriminative models for information retrieval. In *Proc. of SIGIR*, 2004.

[20] G. Salton, J. Allan, and C. Buckley. Approaches to passage retrieval in full text information systems. In *Proc. of SIGIR*, 1993.

[21] M. Sanderson and J. Zobel. Information retrieval system evaluation: Effort, sensitivity, and reliability. In *Proc. of SIGIR*, 2005.

[22] C. Wade and J. Allan. Passage retrieval and evaluation. Technical report, CIIR, University of Massachusetts, 2005.

[23] R. Wilkinson. Effective retrieval of structured documents. In *Proc. of SIGIR*, 1994.

[24] W. Xi, R. Xu-Rong, C. S. Khoo, and E.-P. Lim. Incorporating window-based passage-level evidence in document retrieval. *JIS*, 27(2):73–80, 2001.

[25] J. Zobel, A. Moffat, R. Wilkinson, and R. Sacks-Davis. Efficient retrieval of partial documents. *IPM*, 31(3):361–377, 1995.