DOCUMENT EXPANSION AND LANGUAGE MODEL RE-ESTIMATION FOR INFORMATION RETRIEVAL

BY

GARRICK SHERMAN

DISSEhATION

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Library and Information Science in the Graduate College of the University of Illinois at Urbana-Champaign, 2019

Urbana, Illinois

Doctoral Committee:

Associate Professor Jana Diesner, Chair
Professor J. Stephen Downie
Professor Ted Underwood
Associate Professor Jaime Arguello, University of North Carolina at Chapel Hill
Abstract

Document expansion is the process of augmenting the text of a document with text drawn from one or more other documents. The purpose of this expansion is to increase the size of the term sample from which document representations, such as language models, may be estimated. While document expansion has been shown to improve the effectiveness of ad-hoc document retrieval, our work differs from previous work in a variety of ways. We propose a consistent language modeling approach to document expansion of full length documents. We also explore the use of one or more external document collections as sources of data during the expansion process. Our proposed methods prove successful in improving retrieval effectiveness over baselines.

We also acknowledge that existing document expansion work, including our own, has relied on intuitive assumptions about the mechanisms by which it achieves its effects. In this thesis, we quantify aspects of document language model change resulting from expansion. We investigate the relationships between these changes and the operations of our model. In doing so, we establish evidence to support prior intuitions; specifically, we find relationships between the quality of a document’s representation, which is used to identify appropriate expansion documents, and the expansion model’s success in accurately re-estimating a language model.

Finally, recognizing the potential for further retrieval effectiveness improvement by means of selective application of our model, we investigate methods for automatically predicting whether or not to expand individual documents and, if so, which expansion collection may yield the optimal document representation. We find that, although the document expansion retrieval model has proven effective overall, accurate prediction concerning the expansion of a given document depends too heavily on predicting the document’s relevance. These findings suggest limitations to any model that may seek to optimize scoring on a per-document basis.
Acknowledgments

I would first like to thank my dissertation committee for the time and guidance they provided throughout this process. I am especially grateful to my advisor, Jana Diesner, for agreeing to take me on as her student midway through my studies, and for freely offering me her time and advice.

I would also like to acknowledge my first advisor, Miles Efron, under whom much of the work in this thesis began. I owe Miles my interest and proficiency in information retrieval research, and I appreciate the time he spent teaching and advising me. Miles is also responsible for securing the funding\footnote{Funding for this thesis came from Google in the form of unrestricted gifts in support of research.} that made my research possible, for which I am extremely appreciative.

I am indebted to my fellow student, Craig Willis, for his unending support throughout my time as a PhD student. I have learned with and from Craig, who has been consistently available to me for thoughtful counsel and open collaboration. Thank you, Craig, for your help, encouragement, and friendship.

Thank you to my parents, Ruth and Glen Sherman, for their support, encouragement, and sympathy over the years. I am extremely lucky to have a family who fostered the curiosity and intellectual engagement—and who offered the emotional support—that I have relied upon to complete my studies. Finally, thank you to my wife, Anisha Gupta, for the years of fortitude and optimism. My PhD work has shaped both of our lives profoundly, and I could not have done it without you.
# Contents

Chapter 1 Introduction ......................................................... 1  
  1.1 Motivation ........................................................................ 1  
  1.2 Contributions .................................................................... 2  
  1.3 Research questions ......................................................... 4  

Chapter 2 Background .............................................................. 6  
  2.1 Foundational IR concepts .................................................. 6  
  2.2 Document expansion in IR ................................................ 9  
  2.3 Cluster-based retrieval ...................................................... 10  
  2.4 Incorporating external collections ....................................... 12  
  2.5 Topic modeling .................................................................. 12  
  2.6 Word embedding .............................................................. 14  
  2.7 Conclusion ........................................................................ 14  

Chapter 3 Evaluation ............................................................... 16  
  3.1 Methods ........................................................................... 16  
  3.2 Data .................................................................................. 20  

Chapter 4 Document expansion using external collections .......... 23  
  4.1 Introduction ....................................................................... 23  
  4.2 Document expansion procedure ........................................ 24  
  4.3 Document expansion retrieval model ................................... 28  
  4.4 Evaluation .......................................................................... 29  
  4.5 Results .............................................................................. 31  
  4.6 Discussion ......................................................................... 44  
  4.7 Conclusions ....................................................................... 45  

Chapter 5 Document representation and re-estimation .............. 47  
  5.1 Introduction ....................................................................... 47  
  5.2 Topic term annotation dataset .......................................... 48  
  5.3 Changes to document language model quality ..................... 57  
  5.4 Representing full length documents with pseudo-queries ... 62  
  5.5 The relationship between query terms and topic terms ....... 78  
  5.6 Conclusions ....................................................................... 87  

Chapter 6 Selective document expansion .................................. 88  
  6.1 Motivation ......................................................................... 88  
  6.2 Method ............................................................................. 91  
  6.3 Features ............................................................................ 92  
  6.4 Data and evaluation ......................................................... 96  
  6.5 Results ............................................................................. 97
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.6</td>
<td>Discussion</td>
<td>101</td>
</tr>
<tr>
<td>6.7</td>
<td>Problems with the task of selective document expansion</td>
<td>105</td>
</tr>
<tr>
<td>6.8</td>
<td>Conclusions</td>
<td>110</td>
</tr>
<tr>
<td>Chapter 7</td>
<td>Conclusions</td>
<td>111</td>
</tr>
<tr>
<td>7.1</td>
<td>Revisiting research questions</td>
<td>111</td>
</tr>
<tr>
<td>7.2</td>
<td>Limitations</td>
<td>112</td>
</tr>
<tr>
<td>7.3</td>
<td>Future work</td>
<td>114</td>
</tr>
<tr>
<td>References</td>
<td></td>
<td>116</td>
</tr>
</tbody>
</table>
Chapter 1
Introduction

1.1 Motivation

Ad-hoc document retrieval requires constant grappling with the problem of sparse data: the query that a user issues is a sparse representation of some information need, and the documents to be retrieved contain relatively small amounts of data on which to base estimates of relevance. Information retrieval (IR) systems therefore benefit from any new data that may be used to estimate the relevance of documents to queries. In practice, these new data sources may take many forms. For example, in addition to the terms available in the query and document, scoring functions may benefit from temporal features (e.g. [53]), query log data from past sessions and other users (e.g. [83]), entity links to a knowledge base (e.g. [21]), etc.

IR systems may also increase their available data by augmenting the sparse text available to them. Most commonly, this takes the form of query expansion based on pseudo-relevance feedback, in which the top ranked documents are assumed to be exemplars of relevant documents and are therefore used to expand the user’s original query automatically to include terms that the user did not provide but which may be used in other relevant documents. Because the expansion procedure does not require access to relevance judgments, query expansion has been explored (with strong results [51]) that makes use of document collections that are not candidates for retrieval. Incorporating external data sources may be seen as another avenue to reduce sparsity by introducing data beyond what is available in a typical retrieval scenario.

Alternatively, some research has investigated document expansion as a means of improving the sparse data problem. Like query expansion, document expansion seeks to augment
the sample of terms available from which to estimate a language model. While documents are generally considered sparse samples, their large size relative to typical queries introduces new concerns into the expansion process. Further, document expansion and query expansion are not mutually exclusive processes; in fact, it is reasonable to believe that the two may be combined to produce results superior to either method on its own. This complementarity makes document expansion particularly valuable as a research direction, since its benefits are likely to apply to any other retrieval model that relies upon document language modeling.

While document expansion has received some attention (see Section 2.2), it has rarely been accomplished by theoretically consistent means. In particular, the identification of expansion documents is often performed heuristically despite the availability of appropriate language modeling methods for achieving the same results, as this thesis will show. Furthermore, while intuitive rationales have been proposed to explain the successes of document expansion, no studies have sought to rigorously explain the underlying mechanisms that yield the extrinsic measurement of improved retrieval effectiveness. Given the success of document expansion research under other IR paradigms, further investigation into document expansion under the language modeling approach seems likely to yield fruitful results. The success of past, related research programs serves as a strong motivation for the research conducted in this dissertation.

1.2 Contributions

This thesis begins by proposing a novel document expansion retrieval model. The model employs consistent language modeling techniques throughout, demonstrating the viability of document expansion in a language modeling context. It is also used alongside existing language model-based query expansion methods and is tested both with and without use of external data sources. Overall, the model beats baselines with statistical significance, which proves its validity, utility, and versatility. It is useful, therefore, both as a practical method ready for real world implementation and as confirmation of the suitability of language
modeling as a framework for document expansion.

While prior research has been satisfied to demonstrate the effectiveness of document expansion for IR, no systematic study of the mechanisms of document expansion has ever been attempted. This thesis therefore undertakes a careful analysis of the effects of document expansion, particularly with respect to the role of document topicality on language model change. This portion of the thesis contributes deeper understanding of the processes induced by document expansion, advancing the state of the art by delving deeper than the intuitive hypotheses suggested by all prior work on the subject. In particular, the finding that full length document pseudo-queries (described in depth in Section 4.2.2) are sufficient representations of document topicality, and the finding that topicality is not the sole factor determining successful language model improvement, are both contrary to claims made in earlier work.

In pursuit of the above, a new dataset of TREC document topic annotations was produced and made available online (see Section 5.6). Though these annotations were collected to answer the specific questions motivating this thesis, they provide novel insights into the topical makeup of some of the most well-studied documents in the field of IR. The release of this dataset is therefore itself a valuable contribution to researchers in IR and related disciplines.

Finally, the thesis proposes a method to optimize retrieval on a per-document basis by employing machine learning methods to predict the utility of expanding individual documents. Although this method is ultimately unsuccessful, its failure analysis reveals the theoretical unsoundness of per-document optimization for IR due to the need to predict document relevance. This finding adds to the theoretical understanding of model optimization in IR and should prove useful in helping future researchers to avoid similar pitfalls in optimizing retrieval models.

While the thesis is motivated by the utilization and deeper understanding of a language modeling approach for document retrieval, its findings—particularly the success of the pro-
posed retrieval model—hold practical value for industry professionals. This is not to discount its value to IR researchers, however: the proposed document expansion retrieval model is complementary to many, if not most, common retrieval models, and may be used to boost the effectiveness of future models. Further, by carefully questioning and analyzing the intuitions that guide language modeling approaches to IR, this thesis contributes deeper understandings of the importance of topicality in relevance and the role of relevance in model optimization, as well as data for future research to continue to explore these important issues.

1.3 Research questions

Specifically, the following three research questions guide this research project:

1. **How can document expansion that exclusively employs language modeling techniques be used to improve retrieval effectiveness, and can query expansion and/or use of external document collections further improve effectiveness when paired with document expansion?**

   Based on previous work in document expansion, query expansion, and work with external collections, it is reasonable to hypothesize that document expansion will work in a pure language modeling environment. However, language modeling introduces questions that are not pertinent to other retrieval models, such as how smoothing should work or how related documents should be identified. This research question will be addressed through standard IR experimentation following the Cranfield paradigm [15].

2. **What is the relationship between document language models and document expansion, and how does document expansion improve retrieval effectiveness?**

   While there are intuitive explanations for the effects of document expansion on retrieval effectiveness, this is an inherently extrinsic measure of language model change. A deeper dive into the relationship between document language models before and after
document expansion may help to explain the effects of document expansion and the methods by which it functions.

3. **Can learning algorithms be used to identify and optimize opportunities for document expansion?**

It is not reasonable to expect any single technique to improve effectiveness in all cases, and, indeed, many documents are harmed by expansion. Learning algorithms may offer a solution to automatically identify cases in which document expansion is likely to improve retrieval effectiveness. Classification techniques may also allow for certain types of optimization. In particular, this dissertation will investigate using classifiers to automatically select the source of expansion data that is most likely to improve retrieval effectiveness.
Chapter 2
Background

2.1 Foundational IR concepts

2.1.1 Probabilistic IR

The origins of probabilistic IR are generally attributed to Maron and Kuhns [60], whose work, while essentially historical in value today, introduced the notion of probability into IR—though with regard to indexing rather than retrieval. In particular, Maron & Kuhns invented the idea that documents that are more likely to satisfy the information need of a user should be presented before those that are less likely to satisfy.

This principle, known as the probability ranking principle (PRP), is foundational to modern ranked result list IR systems and underlies almost all modern IR research. Robertson stated the PRP formally: “Documents should be ranked in such a way that the probability of the user being satisfied by any given rank position is a maximum” [75]. He also defended the principle on the grounds of probability theory and decision theory, formally proving that the PRP describes the optimal strategy for ranking retrieved documents.

2.1.2 Language modeling

Prior to language modeling approaches to IR, the vector space model predominated (e.g. [79, 80, 78, 86]). This model relied on heuristic term weighting schemes, generally incorporating a term frequency (TF) component reflecting the frequency of a term in a document, an inverse document frequency (IDF) component reflecting the infrequency of documents containing the term, and a length normalization component to allow fair comparison between long and short documents. Vectors of term weights were compared based on cosine similarity. While these models were often effective, they were unwieldy and lacked theoretical rigor [26].
In response, Ponte & Croft introduced language modeling, a way of modeling text as a generative process [71]. Their work drew on prior work from the speech recognition field but adapted it to IR. While their precise model is no longer in use, it opened the door to a new type of model distinct from the vector space model and which was governed by theoretical rules derived from probability theory.

Language modeling treats queries and documents as samples of text from some underlying generative process. The most common language modeling retrieval function, query likelihood, scores documents based on the probability of generating the query text from the same process that generated the document text:

\[
\text{score}(D, Q) = P(Q|\theta_D)
\]

where \(D\) is the document, \(Q\) is the query, and \(\theta_D\) is the document language model. For simplicity, we generally assume that query term probabilities are independent, which allows us to compute \(P(Q|\theta_D)\) as a product:

\[
P(Q|\theta_D) = \prod_{i=1}^{n} P(q_i|\theta_D)
\]

Most modern IR work, including this thesis, assumes that \(\theta_D\) is a multinomial distribution over unigram terms. In order to estimate this model, we can compute the maximum likelihood estimate (MLE) of each term probability given the term frequencies observed in the document [109].

Unfortunately, the MLE underestimates the probability of documents in which any one query term does not occur. To correct for this, language modeling incorporates smoothing, which reallocates some probability mass from seen terms to unseen terms, generally by using the background collection language model to assign probabilities to words that do not appear in the document. While others exist, the most common smoothing methods are Dirichlet and Jelinek-Mercer smoothing [110]. Dirichlet smoothing works by adding pseudo-counts of
words into the document language model:

\[
P(w|D) = \frac{c(w, D) + \mu P(w|C)}{|D| + \mu}.
\]

\(c(w, D)\) gives the number of occurrences of a word in the document, \(|D|\) is the document length, \(P(w|C)\) is the term probability in the collection, and \(\mu\) is a tunable hyperparameter. Jelinek-Mercer smoothing simply interpolates the document MLE with the collection MLE according to a mixing parameter \(\lambda\):

\[
P(w|D) = (1 - \lambda)\frac{c(w, D)}{|D|} + \lambda P(w|C)
\]

Smoothing has an important role both in preventing zero-probabilities for missing words and in its IDF-like effect. Terms that appear frequently in the collection become undifferentiated after smoothing while terms that appear infrequently in the collection “survive” smoothing in that they are not strongly affected by the background model [110].

### 2.1.3 Query expansion

The query likelihood retrieval model does not allow for changes to be made to the query. However, prior work (e.g. [66]) showed that expanding a query to include more terms could dramatically improve retrieval effectiveness. Lavrenko & Croft [51] proposed a new type of language model, called a relevance model, to allow for query expansion.

Relevance models work by assuming that the query and relevant documents are all generated from the same underlying term distribution called a relevance model. Because all samples are drawn from the same distribution, we can use the query likelihood as an estimate of the probability that a document was generated by this relevance model and thereby
weight the influence of a feedback document’s terms in estimating the relevance model:

\[ P(w|\hat{\theta}_Q) = \sum_{i=1}^{k} P(w|D)P(Q|D)P(D) \]

where \( \hat{\theta}_Q \) is the estimated relevance language model, \( k \) is the number of feedback documents, and the prior probability of the document, \( P(D) \), is often assumed to be uniform. Though \( \theta_Q \) is technically a distribution over all terms in the vocabulary, for performance and efficiency reasons it is typical truncated to the \( n \) highest probability terms.

Because the query is now represented by a probability distribution and not a sample of terms, the query likelihood retrieval model is no longer applicable. Instead, most researchers use the KL-divergence retrieval model [47], which ranks documents by the KL-divergence of the document language model from the query language model:

\[ \text{score}(D, Q) = -D(\theta_Q||\theta_D) \]
\[ = -\sum_{w\in V} P(w|\theta_Q) \log \frac{P(w|\theta_Q)}{P(w|\theta_D)} \]
\[ = \sum_{w\in V} P(w|\theta_Q) \log P(w|\theta_D) \]

2.2 Document expansion in IR

Document expansion has been well studied in IR literature [23, 44, 45, 62]. Although often considered a type of smoothing [56, 92, 102], document expansion predates the probabilistic language models that necessitate smoothing. For example, Singhal & Pereira, working within the vector space model of IR, used document expansion by means of Rocchio’s formula to improve noisy document term vectors produced by imperfect speech recognition [87].

Liu & Croft proposed a method of retrieval that uses document clusters to smooth document language models [56]. In this model, called the CBDM model, each document contributes equally to the overall cluster language model. Importantly, the authors clustered
documents prior to retrieval using the $k$-means algorithm. This approach does not guarantee that a document’s assigned cluster is an optimal source of smoothing data; for example, a document situated near the boundary between two clusters is intuitively likely to be better represented by some mixture of the two clusters than by either cluster individually.

Tao et al. proposed a similar approach to Liu & Croft but placed each document at the center of its own cluster; this helps to ensure that the expansion documents are as closely related to the target document as possible [92]. Clusters are computed based on the cosine similarity between the document to be expanded and each other document in the collection. One effect of this clustering approach is that each document in the collection is a member in each other document’s “cluster” with decreasing weight for more dissimilar documents (though, for reasons of efficiency, the authors used only the top 100 most similar documents for smoothing). This is in contrast to Liu & Croft’s clustering approach in which documents are members of only one cluster and contribute equally to the cluster language model.

My research takes as its starting point that of Efron, Organisciak & Fenlon [27], who clustered very short microblog documents by issuing them as pseudo-queries. They employed a procedure closely related to relevance modeling [51] to expand the original document using those microblog documents retrieved for the pseudo-query. Their work assumed that microblog documents, due to their short length, are most appropriate for this task. We explore the application and adaptation of their work to different scenarios.

### 2.3 Cluster-based retrieval

A subject closely related to document expansion in IR research is cluster-based retrieval. This approach to document retrieval is motivated by the cluster hypothesis, which states that documents similar to one another tend to be relevant to the same queries [37, 94]. Given this assumption, cluster-based systems retrieve entire clusters of related documents rather than individual documents, with the expectation that retrieval effectiveness will be improved by essentially smoothing out the differences between documents through their
cluster memberships. This is in contrast to early work on clustering in IR, which was motivated by efficiency gains and worked by first selecting the most relevant clusters for a query and then ranking the documents within each cluster [97].

Much of the work on cluster-based retrieval employs fixed sets of clusters in which documents are assigned unchanging membership to a single cluster, e.g. [37, 18, 105]. This approach assumes that document similarities reflect similar topical makeup regardless of the relevance of the individual component documents to a given query. However, it is possible to imagine a situation in which a pair of documents are clustered together, but only one is relevant to a given information need.

To address this, many researchers turned to query-specific clustering, e.g. [35, 93, 56]. These techniques involve first retrieving a set of documents for a query and then clustering only those documents in the results set. While these approaches have proven largely successful, they represent a departure from the original cluster hypothesis. Implicit in the cluster hypothesis is the idea that related documents that may not exhibit identical levels of relevance are, nevertheless, equally relevant to a given information need. This is the most intuitive explanation for the effectiveness gains predicted by the hypothesis, since related documents that receive identical or near-identical query matching scores would not benefit, relative to traditional per-document rankings, from explicit clustering. Clustering of retrieved documents may be more appropriately thought of as reducing redundancy (e.g. as explored in [100]).

Cluster-based retrieval has historically relied on an array of hierarchical agglomerative and partitioning algorithms based on varying similarity metrics, though cosine similarity appears to have been the most popular. These clusters are generally evaluated on the basis of a representative document or the cluster centroid, with the assumption that the relevance of one part of the cluster reflects the relevance of all parts of the cluster by some transient property of relevance [104]. Interestingly, the degree to which the cluster hypothesis holds true for a collection does not necessarily reflect the change in retrieval effectiveness induced
by clustering [96], suggesting that the benefits of cluster-based retrieval are derived from something other than the accuracy of the clusters.

2.4 Incorporating external collections

The incorporation of external collections into document retrieval is a common theme in the ad-hoc IR literature, particularly with respect to query expansion [3, 25, 54, 101, 106]. Of particular relevance to this work is that of Diaz and Metzler, who propose a mixture of relevance models [25]. Their model simply interpolates RMs built on different collections, weighting each by a query-independent quantity \( P(c) \). Though this work bears similarities, Diaz and Metzler are interested in query expansion, whereas we apply the technique as one piece in a document expansion model.

2.5 Topic modeling

Modeling the topical content of language in documents has been an important task in IR and related fields. Deerwester et al. introduced latent semantic indexing (LSI), a type of dimensionality reduction based on singular value decomposition, which they argued captured term relationships and issues of synonymy and polysemy [22].

Probabilistic LSI (pLSI), introduced by Hofmann, takes a probabilistic approach to dimensionality reduction [36]. In pLSI, documents are represented as a mixture of topics from which words are sampled:

\[
P(w, D) = P(D) \sum_z P(w|z)P(z|D)
\]

where \( z \) is a latent topic. pLSI therefore assumes that a document may consist of multiple topics with varying influence. However, pLSI requires estimation on training data and is not well defined for newly observed documents.

To correct for this, Blei et al. [6] proposed latent Dirichlet allocation (LDA), which uses
Dirichlet priors to generate topic mixture weights as a random variable. Like LSI and pLSI, LDA treats \( k \), the number of topics, as a model parameter. It is also important to note that, while the resulting models exhibit lower perplexity compared to pLSI or simpler unigram and mixture of unigram models, estimation of LDA models is generally resource-intensive.

Topic models have been shown to usefully represent the text of documents for a variety of tasks, such as IR \[102\], word sense disambiguation \[7\], and automatic document summarization \[31\], among many others.

### 2.5.1 Interpretation of topic models

As noted above, topic models have proven effective according to extrinsic evaluation of separate tasks. They are also frequently measured according to perplexity, which is essentially a measure of how “surprised” a model is to encounter observed data. While topic modeling may identify objective relationships between terms in training documents, this does not necessarily equate to success at actually capturing the semantics of text.

Chang et al. \[14\] propose word and topic intrusion as a method for measuring the semantic interpretability of topic models. Under word intrusion, users are shown a list of words ranked highly according to a topic model along with one “intruder” word with very low probability under that model. Users’ ability to identify the intruder is assumed to correspond with high semantic coherence in the model. Topic intrusion is similar, with users judging which of a set of topics (represented as word lists) is least likely to belong. These topics are chosen according to their probability of assignment to a document.

Newman et al. \[67\] study topic model coherence by testing a large number of mostly heuristic metrics against a gold standard set of human judgments about model coherence. Unlike Chang et al., Newman et al. use a simple 3-point Likert scale to measure topic model coherence. However, they report a high degree of inter-annotator agreement, suggesting that humans can fairly consistently agree on the semantic coherence of a language model.
2.6 Word embedding

Like topic models, word embeddings seek to capture the underlying relationships between terms. The most popular word embedding method, word2vec, models words as vectors of numeric weights [64, 65]. These vectors are learned from term co-occurrence using a sparse neural network approach. Essentially, each word is mapped to a set of other words that occur in close proximity, and the neural network is trained to produce this mapping automatically. The output layer is then stripped away, and the weights mapping each input term to the network’s hidden layer is used as a vector representing the term [64, 65]. These vectors quantify conceptual relationships between terms that have proven useful in several fields, including IR (e.g. [30]).

The word2vec approach to representing words as vectors can be extended to chunks of text, such as paragraphs and documents [52]. This technique, known as doc2vec, differs from word2vec by adding an input neuron that uniquely identifies each chunk of text. The model learns a numeric vector representing the document in a lower dimensional space.

2.7 Conclusion

The preceding areas of study contribute, to greater and lesser degrees, to the research presented in the remainder of this thesis. While subjects like cluster-based retrieval and topic modeling are important related areas of study, topics like query expansion, incorporating external collections, and document expansion are particularly pertinent to the research presented in this thesis.

Some of these areas of research are directly utilized in this thesis. For example, we conduct experiments using relevance models, a form of query expansion. Word2vec-style word embeddings are also directly explored as an approach for identifying related documents. Most centrally, of course, is the topic of document expansion in IR, which motivates the entirety of this thesis, from proposal of a novel document expansion retrieval model to in-
depth analysis of the internal mechanisms of the document expansion process.

Though they are not employed directly, it is also important to recognize closely related areas of study, like cluster-based retrieval and topic modeling. The former serves largely as motivation for specific document expansion approaches, including the one proposed here. That is, cluster-based retrieval operates on the belief that clusters of related documents are equally relevant to a given information need, and this belief also provides justification for using related documents to expand the representation of a document in many document expansion models.

Similarly, topic modeling, which purports to unearth coherent “topics” generating text in a collection, closely relates to both document expansion generally and to the work in this thesis concerning the topical makeup of documents specifically. In the former case, it motivates a belief that text is comprised of coherent topics that occur in multiple documents, which adds further theoretical justification for finding relevant text in expansion documents. In the latter case, it provides an important automated alternative to collecting human topic annotations, though Section 5.2.1 explains why human annotations were used instead in this thesis.

The foregoing description of related work also serves to contextualize the contributions made in the remainder of this thesis. These contributions are discussed in Section 1.2, but, briefly, comprise additions in particular to the area of document expansion: proposal of a novel document expansion retrieval model and in-depth analysis of the document expansion process as it relates to topic representation and language model change. These contributions differ from prior work in both their focus—mixing of several of these areas, such as document expansion with external collections, language modeling approaches to document expansion, etc.—and their depth, e.g. employing topic annotations to more thoroughly understand the document expansion process.
Chapter 3
Evaluation

3.1 Methods

3.1.1 IR

Methodology

The ad-hoc document retrieval components of this dissertation will fit within the Cranfield paradigm for system-oriented IR. The Cranfield paradigm is a method for evaluating retrieval models by performing experiments with set queries, documents, and relevance judgments [15]. According to Voorhees [98], the Cranfield paradigm makes the following simplifying assumptions:

- relevance “can be approximated by topical similarity”
- a single set of relevance judgments for a topic is sufficient
- all relevant documents are known

A great deal of work has gone into justifying and compensating for these simplifying assumptions, e.g. [8, 113]. While there exist valid criticisms, the Cranfield paradigm remains the basis of almost all system-oriented IR research and evaluation.

The Cranfield paradigm provides the framework for evaluating novel IR systems, but it does not dictate the design decisions inherent in creating a system. An especially important point here is the selection of model parameters; many systems perform well only when their parameters are set appropriately for the data. To tune retrieval model parameters, we employ 10-fold cross validation, in which relevance judgments are used to identify the optimal parameter setting on a training set, which is then tested on a held-out test set.
The retrieval experiments conducted as part of this dissertation will use the Indri search engine [90], created by the Lemur Project at the University of Massachusetts, in concert with the ir-tools framework∗ developed by Miles Efron, Craig Willis, and myself to facilitate experimental IR research.

Metrics

IR metrics tend to be based in some manner on the set-based metrics of recall and precision. Recall is defined as the fraction of relevant documents retrieved by the system, out of all relevant documents:

\[ R = \frac{|\{\text{rel}\} \cap \{\text{retrieved}\}|}{|\{\text{rel}\}|}. \]

Precision is the fraction of retrieved documents that are relevant:

\[ P = \frac{|\{\text{rel}\} \cap \{\text{retrieved}\}|}{|\{\text{retrieved}\}|}. \]

Since most modern retrieval systems return a ranked list of documents, set-based metrics like recall and precision are not directly applicable. Instead, modifications of these metrics have been devised to evaluate ranked lists of documents. Perhaps the most popular metric is average precision (AP), which averages the precision calculated at the rank of each relevant document in a list:

\[ AP = \frac{1}{k} \sum_{i=1}^{k} \frac{i}{r_i} \]

where \( r_i \) is the rank of the \( i \)th relevant document and \( k \) is the total number of relevant documents. When a single evaluation metric is needed for a set of queries, as is typically the case, average precision may be averaged to produce mean average precision (MAP). MAP is

∗http://github.com/uiucGSLIS/ir-tools
an important metric in the remainder of this dissertation.

Another common IR evaluation metric is normalized discounted cumulative gain (NDCG) [38]. NDCG acknowledges two points absent from MAP: first, that not all relevant documents are equally relevant; and second, that documents returned lower in the results list are less useful than those returned higher in the list. The former is addressed by cumulative gain, which refers simply to the sum of the relevance scores at a given position in the results list:

$$CG@k = \sum_{i=1}^{k} g_i$$

where $k$ is the position of a document in the results list and $g_i$ is the graded relevance of the $i$th document in the list. Cumulative gain may thought of as a numeric representation of the amount of relevant information available up to a certain point in the results list. However, there is a cost to the user when relevant information is located lower in the results list. This is encoded by discounted cumulative gain, which discounts a document’s gain based on its rank:

$$DCG@k = g_1 + \sum_{i=2}^{k} \frac{g_i}{\log_2 i}.$$  

Since the number of relevant documents and the degree of their relevance varies across queries, DCG must be normalized to allow for comparing and combining DCG scores. NDCG normalizes DCG by dividing it by the ideal DCG, IDCG, which is the optimal possible DCG achievable given the relevant documents:

$$NDCG@k = \frac{DCG@k}{IDCG@k}.$$  

NDCG@20 is used throughout this dissertation. The cutoff of twenty is the default of several common retrieval evaluation tools (e.g. trec_eval† and gdeval‡) and provides a

---

†https://trec.nist.gov/trec_eval/
‡https://trec.nist.gov/data/web/10/gdeval.pl
good indication of search result quality at the depth that a typical user or pseudo-relevance feedback algorithm is likely to reach. Use of NDCG@20 along with MAP will provide insight into the types of improvements gained by a given treatment, e.g. whether improvements occur at high or low ranks.

All effectiveness metrics will be checked for statistical significance against baseline runs. As is typical in IR research, significance is checked using paired, two-tailed $t$-tests to compare the MAP and average NDCG@20 of pairs of runs. Following the advice of [89], the randomization test (also known as the permutation test) is also employed.

### 3.1.2 Classification

Numerous machine learning algorithms exist to enable classification, among them naive Bayes [61], $k$ nearest neighbors, and support vector machines (SVMs) [39]. The goal of a classifier is to predict the class of a new instance based on its features and given a set of training instances whose classes are known. This training data indicates that classification is a type of supervised learning, meaning that a classifier must be provided a set of labeled instances, typically created manually.

As in retrieval above, classification generally requires the tuning of parameters to optimize performance on a dataset. One simple example is the setting of $k$ in the $k$ nearest neighbor algorithm. Again, cross validation is used to identify optimal parameter settings, which are tested on a smaller held-out test set. Work with classifiers is performed with the scikit-learn [69] library, which handles both cross validation and the implementation of classifiers.

There are numerous metrics that may be used in assessing the success of a classifier. The most obvious, accuracy, is defined as the number of correct predictions out of the total number of predictions. Accuracy can be a useful metric, but it is also problematic in some circumstances. For example, when a classifier predicts the majority class in all cases, its accuracy will be better than chance, but the model would have failed to identify any meaningful relationships in the data.
In the case of this dissertation, successful classification can also be measured according to its effect on the retrieval effectiveness metrics discussed in Section 3.1.1. This is because classification results are only useful if they can positively impact retrieval results. Even if a classifier achieves low accuracy or other intrinsic scores, it may nevertheless prove beneficial when applied to retrieval. The success of classifiers will therefore be assessed on the basis of whether they achieve a statistically significant improvement to retrieval effectiveness metrics compared with metrics achieved without classification.

3.1.3 Regression analysis

The work in this dissertation will also rely on well-established regression analysis methods to identify relationships between variables and predict the values of future observations. Regression analysis consists of a response variable and one or more predictors, also known as dependent and independent variables respectively. Because regression analysis makes predictions on the basis of past observations with known responses, it may be thought of as a type of supervised learning.

In this dissertation, regression models are primarily used to assess whether relationships exist between the dependent variable and the independent variable(s). These relationships can be assessed through a combination of the adjusted $R^2$ value, which measures the amount of variance in the dependent variable that is explained by variance in the independent variables, and $F$- and $t$-tests to measure the statistical significance of the overall model relationship and of the individual independent variables respectively.

3.2 Data

3.2.1 TREC datasets

This dissertation will make heavy use of data created by and used for the Text REtrieval Conference (TREC) [32]. The conference is composed of tracks, each of which is designed to
facilitate research around a specific question in IR. To enable this research, TREC organizers produce document datasets as well as topics (queries) and relevance judgments produced according to best practices; these datasets are distributed to participants and form the basis of most system-oriented IR research both for and beyond the TREC conference.

In particular, we will use the TREC datasets shown in Table 3.1. These datasets provide a good range of document types, from well-formed to messy, as well as a range of collection sizes and query difficulties. Using multiple datasets is a well-established practice in IR research; a strong retrieval model is one that generalizes well to many different types of data.

<table>
<thead>
<tr>
<th>Collection</th>
<th>Description</th>
<th>Num docs</th>
<th>Topics</th>
</tr>
</thead>
<tbody>
<tr>
<td>AP 88-89</td>
<td>AP newswire documents.</td>
<td>164,597</td>
<td>101-200</td>
</tr>
<tr>
<td>GOV2</td>
<td>Web documents.</td>
<td>25,205,179</td>
<td>701-850</td>
</tr>
<tr>
<td>Robust04</td>
<td>News documents.</td>
<td>528,155</td>
<td>301-450, 601-700</td>
</tr>
<tr>
<td>WT10g</td>
<td>Web documents.</td>
<td>1,692,096</td>
<td>451-550</td>
</tr>
</tbody>
</table>

Table 3.1: TREC datasets used in this dissertation, as well as some properties of those datasets.

3.2.2 Wikipedia

In addition to TREC data, the online encyclopedia Wikipedia\(^8\) is used for document expansion purposes. While imperfect, Wikipedia is a reasonable (and freely available) source of general purpose documents. Due to its relatively well-formed documents and broad topical coverage, Wikipedia is a popular choice for supplemental data in IR research (e.g. [3, 20, 41, 54, 67, 106])

This dissertation uses a September 2015 dump of Wikipedia. This dump contains 11,938,282 documents (Wikipedia articles). Though Wikipedia technically introduces future information by including documents written after all of the TREC collections were released, the effect is minimal and more likely to harm than help. For example, imagine that a news document is about a topic that has developmented further since the creation of the TREC

\(^8\)http://en.wikipedia.org
collections. These developments are likely to appear in related Wikipedia articles; however, since the relevance judgments matching queries to documents predate the developments, any alterations to the queries or documents to include these developments are likely to shift the language to include what was at the time of judgment irrelevant information. Rather than injecting unfair future information, Wikipedia therefore mainly serves to provide general background text.
Chapter 4
Document expansion using external collections*

4.1 Introduction

Relevance modeling is an extremely influential pseudo-relevance feedback technique in which we assume that both queries and documents are observations sampled from a relevance model (RM) [51], which is a probability distribution over terms in relevant documents. Because we do not have true relevance feedback, relevance modeling makes use of the query likelihood, $P(Q|D)$, to quantify the degree to which words in each document should contribute to the final model $R$. However, since no document is perfectly representative of its underlying generative model, we may be reasonably concerned that our estimate of $P(Q|D)$ is the result of chance. That is, there is no guarantee that $D$ is a representative sample from $R$. The quality of our RM, therefore, may benefit from a higher quality document representation than that which is estimated from the text of $D$.

Document expansion with and without external document collections is used to attempt to improve the quality of document language models. Expanded documents are expected to exhibit less random variation in term frequencies, improving probability estimates. Estimates may be further refined by expanding documents using external collections, thereby avoiding any term bias exhibited by documents in an individual collection.

Previous investigations into document expansion have tended to use only the target collection to expand documents, while this work explores the use of one or more distinct collections. Conversely, most existing work involving external corpora in ad-hoc IR has focused on query expansion; here, external collections are incorporated for purposes of document expansion.

*Much of this chapter was originally published in the Association for Computing Machinery’s SIGIR 2017 conference in Tokyo, Japan, under the title ”Document expansion using external collections.” It was coauthored with Miles Efron [85].
4.2 Document expansion procedure

4.2.1 Baseline retrieval models

The language modeling retrieval framework [47] is used throughout this thesis; specifically, we employ query likelihood (QL) and relevance modeling for ranking. These methods are discussed in detail in Sections 2.1.2 and 2.1.3. Briefly, query likelihood scores each document $D$ for a given query $Q$ on $P(Q|\theta_D)$, where $\theta_D$ is the language model (typically, as here, assumed to be a multinomial distribution over the vocabulary $V$) that generated the text of document $D$. Assuming independence among terms and a uniform distribution over documents, each document is scored by

$$P(Q|D) = \prod_{w\in Q} P(w|\theta_D)c(w,Q)$$ (4.1)

where $c(w,Q)$ is the frequency of word $w$ in $Q$. $P(w|\theta_D)$ in Eq. 4.1 is estimated following standard procedures: estimating a smoothed language model by assuming that document language models in a given collection have a Dirichlet prior distribution:

$$\hat{P}(w|\theta_D) = \frac{c(w,D) + \mu \hat{P}(w|C)}{|D| + \mu}$$ (4.2)

where $\hat{P}(w|C)$ is the maximum likelihood estimate of the probability of seeing word $w$ in a “background” collection $C$ (typically $C$ is the corpus from which $D$ is drawn), and $\mu \geq 0$ is the smoothing hyper-parameter.

Relevance modeling is a form of pseudo-relevance feedback that uses top ranked documents to estimate a language model representing documents relevant to a query [51].

Assuming uniform document prior probabilities, relevance models take the form of

$$P(w|R) = \sum_{D\in C} P(w|D)P(Q|D)$$ (4.3)
where \( P(Q|D) \) is calculated as in Eq. 4.1 and essentially weights word \( w \) in \( D \) by the query likelihood of the document. Relevance models are most efficient and robust when calculated over only the top ranked documents and limited to the top terms. These parameters are referred to as \( fbDocs \) and \( fbTerms \) respectively in Table 4.1 below.

Relevance models are prone to query drift, which is an unintended change in the topical focus of a query introduced by expansion [66]. It is therefore often desirable to linearly interpolate an RM with the original query model to improve effectiveness:

\[
P(w|Q') = (1 - \alpha)P(w|R) + \alpha P(w|Q). \tag{4.4}
\]

\( \alpha \) is a mixing parameter controlling the influence of the original query. This form of relevance model is known as “RM3.”

### 4.2.2 Expanding with document pseudo-queries

A simple approach to identifying expansion documents is to convert each target document into a pseudo-query that can be issued against the expansion collection(s). To expand a document \( D \), we begin by treating the text of \( D \) as a pseudo-query which we pose against a collection of documents \( C_E \). To transform a document into a pseudo-query we apply two transformations. First we remove all terms from \( D \) that appear in the standard Indri stoplist\(^\dagger\). Next, we prune our pseudo-query by retaining only the \( 0 < k \leq K \) most frequent words in the stopped text of \( D \), where \( K \) is the total number of unique terms in \( D \). The integer variable \( k \) is a parameter that we choose empirically. These are the non-stopwords with the highest probabilities in a maximum likelihood estimate of \( D \)'s language model and are therefore a reasonable representation of the topic of the document. Though some information may be lost with stopping, with a large enough \( k \) we hope to nevertheless capture the general topic of a document; for example, a document about Hamlet’s famous speech may not be represented by the terms “to be or not to be,” but the terms “Shakespeare,”

\[^\dagger\text{http://www.lemurproject.org/stopwords/stoplist.dft}\]
“Hamlet,” “speech,” etc. will likely represent the document’s subject sufficiently. Let $Q_D$ be the pseudo-query for $D$, consisting of the text of $D$ after our two transformations are applied.

We rank related documents, called expansion documents, by running $Q_D$ over a collection $C_E$. More formally, we rank the documents in $C_E$ against $D$ using Eq. 4.1, substituting $Q_D$ for the query and $E_i$—the $i^{th}$ expansion document—for the document. Let $\pi_i$ be the log-probability for expansion document $E_i$ with respect to $D$ given by Eq. 4.1.

We now have a ranked list of tuples $\{(E_1, \pi_1), (E_2, \pi_2), ..., (E_N, \pi_N)\}$ relating expansion document $E_i$ to $D$ with log-probability $\pi_i$. We take the top $n$ documents where $0 \leq n \leq N$. We call these top documents $E_D$ and designate them as our expansion documents for $D$.

Finally, we exponentiate each $\pi_i$ and normalize our retrieval scores so they sum to one over the $n$ retained documents. Assuming a uniform prior over documents, we now have a probability distribution over our $n$ retained documents: $P(E|D)$.

Since this procedure does not depend on the query, we may compute $E_D$ once at indexing time and reuse our expansion documents across queries.

### 4.2.3 Expanding with document embeddings

More complex approaches may be used to identify expansion documents. For example, doc2vec document embeddings are an intuitively appealing approach for this purpose. Doc2vec embeddings purport to capture “deeper” relationships between documents. In other words, much as in latent semantic analysis, doc2vec’s lower dimensional document representations are theoretically able to encode document similarities in a manner that avoids problems of vocabulary mismatch to which document pseudo-queries are susceptible. In theory, two documents with wholly disjoint term sets may still appear highly similar in their embedded forms, if their terms tend to co-occur across the collection with some common set of terms. This alleged ability to link linguistically divergent but topically similar documents seems particularly useful for the purposes of document expansion since it should allow for a more
expansive set of expansion documents. By definition, document embedding methods also convert documents to (low dimensional) vector representations. This makes it natural to identify similar documents using cosine similarity, which is a well-established IR technique.

To expand documents using doc2vec embeddings, we first use the software library gensim to calculate the embeddings [74]. Multiple algorithms fall under the name “doc2vec,” but for the purposes of this work we use the distributed memory approach, which was found to be more successful across different tasks [52]. Training doc2vec algorithms is extremely time intensive, so several gensim parameter defaults were used in the interest of efficiency. Among these are the minimum word frequency—by default, gensim discards words that occur fewer than five times in the collection—and the window size between each input word and the expected output words, which is also five by default. However, we did test a range of vector sizes, using cross validation to select the empirically optimal parameter value.

We rank expansion documents in collection $C_E$ by computing the cosine similarity between each expansion document embedding $V_E$ and the target document embedding $V_D$:

$$sim(V_D, V_E) = \frac{V_D \cdot V_E}{||V_D|| \cdot ||V_E||}.$$  \hspace{1cm} (4.5)

When $C_E$ is the target collection, both $V_D$ and $V_E$ are given for all $E$ as a result of training the doc2vec embeddings. However, when $C_E$ is an external collection, only $V_E$ is known, and we must infer $V_D$ using the model trained on $C_E$. Doc2vec allows for document inference by holding constant the weights learned for words and the softmax output layer. Since the words in $D$ are known, and the weights have been optimized, we can use gradient descent to find the document weight that maximizes the probability of observing the words in $D$. This inferred vector $\hat{V}_D$ can be used in place of $V_D$.

From here, we follow the same procedure as with document pseudo-queries: take the top $n$ documents $E_D$, normalize their cosine similarity scores so that they sum to one, and use these scores as a probability distribution $P(E|D)$. 

27
As with the document pseudo-query approach, this process need only happen once at indexing time. However, the efficiency is much lower than with document pseudo-queries, since doc2vec embeddings must first be trained over the entire collection. Stable document embeddings are achieved only with a high number of training passes, or epochs. We use 50 epochs, after which the stability of document vectors did not noticeably improve.

4.3 Document expansion retrieval model

We would now like to incorporate our expansion documents into a retrieval model over documents. We assume that a query is generated by a mixture of the original document language model $\theta_D$ and language models $\theta_{E_j}$ representing the expansion documents in each corpus $C_j \in \{C_1, C_2, ..., C_n\}$. We assume that $\theta_{E_j}$ can be estimated using the text of the expansion documents $E_{D_j}$ in corpus $C_j$. This mixture model may be expressed as:

$$
\hat{P}^\lambda(Q|D) = \prod_{i=1}^{|Q|} (1 - \sum_{j=1}^n \lambda_{E_{D_j}}) P(q_i|D) + \sum_{j=1}^n \lambda_{E_{D_j}} P(q_i|E_{D_j})
$$

(4.6)

where $0 \leq \sum_{j=1}^n \lambda_{E_{D_j}} \leq 1$. $P(q_i|E_{D_j})$ is estimated in expectation:

$$
P(q_i|E_{D_j}) = \sum_{E \in E_{D_j}} P(q_i|E) P(E|D).
$$

(4.7)

Like $P(q_i|D)$, we estimate the probability of $q_i$ given expansion document $E$, $P(q_i|E)$, as a Dirichlet-smoothed query likelihood. By virtue of our expansion document scoring and normalization, we also have $P(E|D)$. This general model may be used with any number of expansion corpora.

4.3.1 Relevance modeling with expanded documents

Given our motivating intuition that document expansion allows for the more accurate estimation of document language models, we would expect that an RM computed using expanded
documents should be more accurate than a standard RM. An RM3 is therefore computed as in Eqs. 4.3 and 4.4, substituting the expanded document for the original.

4.4 Evaluation

4.4.1 Runs

For each target collection, we produce runs comprising each possible combination of expansion source and query expansion model. Expansion source refers to the collection(s) used for document expansion, while the query expansion model refers to unexpanded queries (QL) or expanded queries (RM3).

Runs are produced with expansion documents from every other target collection as well as Wikipedia. One run is also produced using both the target collection and Wikipedia in combination, which is called “Self-Wikipedia” in the results below. For each source, both the QL and RM3 variations are compared.

Stop words are removed from the query. For efficiency, we retrieve the top 1,000 documents using the default Indri QL implementation and re-rank these documents based on their expanded representations as described in Section 4.3.

4.4.2 Parameters

The parameters required for our approach, their meanings, and the values used in our experiments are shown in Table 4.1.

In general, the values of the parameters outlined in Table 4.1 were selected for efficiency reasons. For example, $k$ may equal the length of $D$; however, scoring an entire collection of documents against such a lengthy query would be prohibitively slow. The values searched were selected to allow for a range of possible settings while limiting computational complexity.

The values of $k$, $v$, $n$, $\lambda_{\varepsilon_D}$, $fbDocs$, $fbTerms$, and $\alpha$, are determined using 10-fold cross validation. In the training stage, we sweep over parameter values listed in Table 4.1, $\lambda_{\varepsilon_D}$
<table>
<thead>
<tr>
<th>Param.</th>
<th>Meaning</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$k$</td>
<td>The maximum number of document terms to use in constructing $Q_D$.</td>
<td>5, 10, 20, 50</td>
</tr>
<tr>
<td>$v$</td>
<td>The document embedding vector size used for doc2vec expansion.</td>
<td>10, 50, 100, 200, 300</td>
</tr>
<tr>
<td>$n$</td>
<td>The maximum number of expansion documents in $E_D$.</td>
<td>5, 10, 20, 50</td>
</tr>
<tr>
<td>$\lambda_{E_D}$</td>
<td>One of several related mixing parameters controlling the weights of $P(q</td>
<td>D)$ and $P(q</td>
</tr>
<tr>
<td>$\mu$</td>
<td>Used for Dirichlet smoothing of both $P(q</td>
<td>D)$ and $P(q</td>
</tr>
<tr>
<td>fbDocs</td>
<td>The number of feedback documents to use for RM3 runs.</td>
<td>10, 20, 30, 40, 50</td>
</tr>
<tr>
<td>fbTerms</td>
<td>The number of terms per document to use for RM3 runs.</td>
<td>10, 20, 30, 40, 50</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>Mixing parameter controlling the weights of the original query and relevance model for RM3 runs.</td>
<td>0.0-1.0</td>
</tr>
</tbody>
</table>

Table 4.1: Parameter settings for the document expansion procedure and retrieval model and $\alpha$ in intervals of 0.1. The concatenated results of each test fold form a complete set of topics that is used for the final reported retrieval effectiveness.

Due to combinatorial explosion, it is infeasible to search the entire parameter space for RM3 runs. While pseudo-query QL runs only need to search 176 unique parameter settings per query, an RM3 run that searched the entire parameter space would require runs for 48,400 unique parameter settings per query. For doc2vec runs, the addition of the $v$ parameter grows the potential RM3 search space to 242,000 unique parameter settings. Instead, RM3 runs search only the space defined by fbDocs, fbTerms, and $\alpha$, using the optimal settings of $k$, $n$, $\lambda_{E_D}$, and (when applicable) $v$ found for the query’s QL run. This reduces the RM3 search
to 275 unique parameter settings per query and ensures that the RM3 is built from the best document representations available.

4.5 Results

4.5.1 Pseudo-query expansion

Retrieval effectiveness of each pseudo-query run is shown in Table 4.2. Effectiveness is measured with mean average precision (MAP) and normalized discounted cumulative gain at 20 (NDCG@20) [38]. Each metric is optimized with 10-fold cross validation. Statistical significance is calculated using a paired two-tailed $t$-test with $\alpha = 0.05$. Multiple testing correction was performed using the Benjamini-Hochberg method [4] on each column, and runs that were no longer significant after correction were further marked with an asterisk. Runs were additionally tested using a paired randomization test [89]. The randomization test agreed with the $t$-test in all cases but one (WT10g expanded with Self-Wikipedia for QL MAP) and is therefore not noted in Table 4.2.

The results confirm that document expansion provides benefit over a baseline query likelihood run—no QL run performs significantly worse than the baseline, and most runs improve over the baseline QL run. The AP and Robust collections improve over the baseline in all but one case (Robust expanded with AP for NDCG@20). WT10g shows good NDCG@20 improvement as well, and its raw MAP scores appear much higher than the baseline, though they fail to achieve statistical significance. Unfortunately, the GOV2 collection proved difficult to improve. Though half of the MAP runs reach statistically significant improvement, none of the NDCG@20 runs do.

Performance of RM3 runs is more surprising with improvement over the baseline RM3 occurring more rarely compared to improvement over the baseline QL. Although two-thirds of MAP runs and 79% of NDCG@20 runs achieve raw improvement over the baseline, very few runs in both cases are statistically significant and even fewer of these survive multiple
Table 4.2: Results for document expansion QL and RM3 runs expanded with document pseudo-queries. Statistically significant improvement over the baseline at $p \leq 0.05$ is marked with $\uparrow$. Runs marked with $\ast$ are not statistically significant after multiple testing correction. Most disappointingly, two Robust MAP runs are significantly worse than the baseline RM3 run, though one of these is not significant after correction.

<table>
<thead>
<tr>
<th>Target Collection</th>
<th>Expansion Collection</th>
<th>QL</th>
<th>RM3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>MAP</td>
<td>NDCG@20</td>
</tr>
<tr>
<td>AP</td>
<td>Baseline</td>
<td>0.2337</td>
<td>0.4170</td>
</tr>
<tr>
<td></td>
<td>AP</td>
<td>0.2813$\dagger$</td>
<td>0.4567$\dagger$</td>
</tr>
<tr>
<td></td>
<td>GOV2</td>
<td>0.2716$\dagger$</td>
<td>0.4512$\dagger$</td>
</tr>
<tr>
<td></td>
<td>Robust</td>
<td>0.2764$\dagger$</td>
<td>0.4570$\dagger$</td>
</tr>
<tr>
<td></td>
<td>Self-Wikipedia</td>
<td>0.2879$\dagger$</td>
<td>0.4794$\dagger$</td>
</tr>
<tr>
<td></td>
<td>Wikipedia</td>
<td>0.2753$\dagger$</td>
<td>0.4783$\dagger$</td>
</tr>
<tr>
<td></td>
<td>WT10g</td>
<td>0.2709$\dagger$</td>
<td>0.4529$\dagger$</td>
</tr>
<tr>
<td>GOV2</td>
<td>Baseline</td>
<td>0.2697</td>
<td>0.4134</td>
</tr>
<tr>
<td></td>
<td>AP</td>
<td>0.2720</td>
<td>0.4137</td>
</tr>
<tr>
<td></td>
<td>GOV2</td>
<td>0.2797$\dagger$</td>
<td>0.4236</td>
</tr>
<tr>
<td></td>
<td>Robust</td>
<td>0.2718</td>
<td>0.4117</td>
</tr>
<tr>
<td></td>
<td>Self-Wikipedia</td>
<td>0.2784$\dagger$</td>
<td>0.4091</td>
</tr>
<tr>
<td></td>
<td>Wikipedia</td>
<td>0.2749</td>
<td>0.4170</td>
</tr>
<tr>
<td></td>
<td>WT10g</td>
<td>0.2773$\dagger$</td>
<td>0.4124</td>
</tr>
<tr>
<td>Robust</td>
<td>Baseline</td>
<td>0.2193</td>
<td>0.3835</td>
</tr>
<tr>
<td></td>
<td>AP</td>
<td>0.2281$\dagger$</td>
<td>0.3821</td>
</tr>
<tr>
<td></td>
<td>GOV2</td>
<td>0.2460$\dagger$</td>
<td>0.4163$\dagger$</td>
</tr>
<tr>
<td></td>
<td>Robust</td>
<td>0.2421$\dagger$</td>
<td>0.3982$\dagger$</td>
</tr>
<tr>
<td></td>
<td>Self-Wikipedia</td>
<td>0.2475$\dagger$</td>
<td>0.4177$\dagger$</td>
</tr>
<tr>
<td></td>
<td>Wikipedia</td>
<td>0.2409$\dagger$</td>
<td>0.4043$\dagger$</td>
</tr>
<tr>
<td></td>
<td>WT10g</td>
<td>0.2370$\dagger$</td>
<td>0.4052$\dagger$</td>
</tr>
<tr>
<td>WT10g</td>
<td>Baseline</td>
<td>0.1683</td>
<td>0.2738</td>
</tr>
<tr>
<td></td>
<td>AP</td>
<td>0.1728</td>
<td>0.2853</td>
</tr>
<tr>
<td></td>
<td>GOV2</td>
<td>0.1777</td>
<td>0.3087$\dagger$</td>
</tr>
<tr>
<td></td>
<td>Robust</td>
<td>0.1755</td>
<td>0.2962$\ast$</td>
</tr>
<tr>
<td></td>
<td>Self-Wikipedia</td>
<td>0.1856</td>
<td>0.3183$\dagger$</td>
</tr>
<tr>
<td></td>
<td>Wikipedia</td>
<td>0.1840$\dagger$</td>
<td>0.3110$\dagger$</td>
</tr>
<tr>
<td></td>
<td>WT10g</td>
<td>0.1685</td>
<td>0.2965$\dagger$</td>
</tr>
</tbody>
</table>

Wikipedia or a combination of the target collection and Wikipedia gave the largest raw effectiveness metrics in 11 of 16 cases, which is likely due to Wikipedia’s broad topical coverage. However, despite some variation, there does not appear to be an overall difference
in retrieval effectiveness based on the choice of expansion collection. This was confirmed with a series of one-way ANOVA tests, which showed no significant difference between expansion collections for any given target collection per metric.

### 4.5.2 Doc2vec expansion

<table>
<thead>
<tr>
<th>Target Collection</th>
<th>Expansion Collection</th>
<th>QL MAP</th>
<th>QL NDCG@20</th>
<th>RM3 MAP</th>
<th>RM3 NDCG@20</th>
</tr>
</thead>
<tbody>
<tr>
<td>AP</td>
<td>Baseline</td>
<td>0.2337</td>
<td>0.4170</td>
<td>0.3332</td>
<td>0.4700</td>
</tr>
<tr>
<td></td>
<td>AP</td>
<td>0.2537†</td>
<td>0.4330†*</td>
<td>0.3340</td>
<td>0.4799</td>
</tr>
<tr>
<td></td>
<td>Wikipedia</td>
<td>0.2553†</td>
<td>0.4507†*</td>
<td>0.3404†</td>
<td>0.4782</td>
</tr>
<tr>
<td>GOV2</td>
<td>Baseline</td>
<td>0.2697</td>
<td>0.4134</td>
<td>0.2883</td>
<td>0.4138</td>
</tr>
<tr>
<td></td>
<td>GOV2</td>
<td>0.2739†</td>
<td>0.4127</td>
<td>0.2889</td>
<td>0.4133</td>
</tr>
<tr>
<td></td>
<td>Wikipedia</td>
<td>0.2786†</td>
<td>0.4257</td>
<td>0.2902</td>
<td>0.4116</td>
</tr>
<tr>
<td>Robust</td>
<td>Baseline</td>
<td>0.2193</td>
<td>0.3835</td>
<td>0.2660</td>
<td>0.4002</td>
</tr>
<tr>
<td></td>
<td>Robust</td>
<td>0.2222</td>
<td>0.3807</td>
<td>0.2603†</td>
<td>0.3971</td>
</tr>
<tr>
<td></td>
<td>Wikipedia</td>
<td>0.2255†*</td>
<td>0.3954</td>
<td>0.2683</td>
<td>0.4001</td>
</tr>
<tr>
<td>WT10g</td>
<td>Baseline</td>
<td>0.1683</td>
<td>0.2738</td>
<td>0.1657</td>
<td>0.2672</td>
</tr>
<tr>
<td></td>
<td>WT10g</td>
<td>0.1725</td>
<td>0.2812</td>
<td>0.1686†*</td>
<td>0.2695</td>
</tr>
<tr>
<td></td>
<td>Wikipedia</td>
<td>0.1810</td>
<td>0.2914</td>
<td>0.1780†</td>
<td>0.2836†*</td>
</tr>
</tbody>
</table>

Table 4.3: Results for document expansion QL and RM3 runs expanded using doc2vec document embeddings. Statistically significant improvement over the baseline at $p \leq 0.05$ is marked with †. Runs marked with * are not statistically significant after multiple testing correction.

As described in Section 4.2.3, runs were also completed using doc2vec document embeddings to identify expansion documents. Because of the high computational complexity of estimating document embeddings, these runs were limited to expansion with the target collection and Wikipedia. Table 4.3 shows the results of these experiments.

Though expansion with document embeddings generally showed raw improvement over baselines, the degree of improvement often did not match that seen with document pseudo-queries. For example, both AP and Robust query likelihood runs were uniformly lower for both MAP and NDCG@20 than the lowest observed corresponding runs using pseudo-queries. Other runs achieved evaluation metrics comparable to those seen with pseudo-query
expansion, but in only one case—NDCG@20 for GOV2’s QL expanded with Wikipedia—did a doc2vec expansion run outperform all corresponding pseudo-query runs. This latter point is a slightly unfair comparison, however, since more pseudo-query runs are available than doc2vec runs.

Notably, in all but three cases, the best performing doc2vec run was expanded with Wikipedia. This echoes the performance of Wikipedia (and Self-Wikipedia) runs when pseudo-queries were used, and is again likely due to Wikipedia’s topical generality.

The disappointing overall performance of doc2vec is somewhat surprising, given the apparent applicability of this task to its intended purpose (identifying document relationships). We suggest that the problem is due to word embedding methods like doc2vec identifying relationships between similar “types” of words, rather than between words with similar topics. Zamani and Croft provide a useful example: a word2vec model might rate “safe” as being highly proximate to the query “dangerous vehicles” even though “safe” indicates the opposite topic from what was intended [108]. Our doc2vec embeddings may therefore be more likely to find proximate documents, rather than related ones.

4.5.3 Parameter sensitivity

The model described requires several parameters to be set. It is therefore important to understand the sensitivity of the model to the values of these parameters. Given the subpar performance of doc2vec runs, analysis in this section will be limited to pseudo-query runs, which are also less complex, having one fewer parameter to consider.

We first examine the impact of stoplist choice. We compare expansion document retrieval without stopping against retrieval with two separate stoplists: the Indri stoplist (as used in the experiments above) of 418 terms and the NLTK stoplist of 179 terms [5]. The two stoplists share 122 terms. We sample 100 random documents from each of the four TREC collections and compare pseudo-query term overlap as well as expansion document results overlap for each document under every combination of the three stopping alternatives.
Figure 4.1: Histograms comparing the number of terms shared between pairs of 10-word pseudo-queries under the three stopping alternatives. “Indri” and “NLTK” indicate stopping with those respective stoplists.

Figure 4.2: Histograms comparing the number of expansion documents retrieved in common between pairs of the three stopping alternatives, using the target collection as the expansion collection and retrieving ten expansion documents per pseudo-query. “Indri” and “NLTK” indicate stopping with those respective stoplists.
Table 4.4: Median term overlap between each pair of stopping alternatives. “Indri” and “NLTK” indicate stopping with those respective stoplists.

<table>
<thead>
<tr>
<th>Collection</th>
<th>Stopped/Indri</th>
<th>Stopped/NLTK</th>
<th>Indri/NLTK</th>
</tr>
</thead>
<tbody>
<tr>
<td>AP</td>
<td>2</td>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>GOV2</td>
<td>4</td>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td>Robust</td>
<td>2</td>
<td>2</td>
<td>9</td>
</tr>
<tr>
<td>WT10g</td>
<td>3</td>
<td>3</td>
<td>9</td>
</tr>
</tbody>
</table>

Table 4.5: Median expansion document results overlap between each pair of stopping alternatives. “Indri” and “NLTK” indicate stopping with those respective stoplists. Using both stoplists yields the same median overlap with unstopped results, as shown in the “Unstopped/Stopped” column.

<table>
<thead>
<tr>
<th>Collection</th>
<th>Unstopped/Stopped</th>
<th>Indri/NLTK</th>
</tr>
</thead>
<tbody>
<tr>
<td>AP</td>
<td>5</td>
<td>9</td>
</tr>
<tr>
<td>GOV2</td>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td>Robust</td>
<td>4</td>
<td>9</td>
</tr>
<tr>
<td>WT10g</td>
<td>3</td>
<td>10</td>
</tr>
</tbody>
</table>

Figure 4.1 shows the number of terms in common between each pair of stopping alternatives. In general, the figure makes clear that unstopped pseudo-queries share very few terms with stopped pseudo-queries: stopped and unstopped pseudo-queries only share a median of 2–4 out of ten terms, according to Table 4.4. In contrast, pseudo-queries stopped with the Indri stoplist are almost identical to those stopped with the NLTK stoplist, with a median of 9–10 words in common, and never fewer than six.

Different pseudo-query term composition does not necessarily imply different expansion document results, and Figure 4.2 shows that even pseudo-queries comprised of different terms often retrieve very similar results sets. However, the figure also makes clear that pseudo-queries stopped with the Indri and NLTK stoplists very rarely exhibit less than 90% overlap in expansion document results, and 100% overlap is the most common outcome for all collections. Table 4.5 reinforces this finding, with a median of 3–5 common expansion documents between unstopped pseudo-query results and pseudo-query results with either type of stopping (they yield the same medians). In contrast, the two types of stopped pseudo-queries again share a median 9–10 out of ten expansion documents.
These findings indicate two conclusions: first, that there is substantial difference between stopped and unstopped pseudo-queries, and unstopped pseudo-queries are comprised of up to 100% stopwords; second, that there is not a systematic difference between pseudo-queries stopped using one stoplist and those stopped with another. We therefore conclude that the use of the Indri stoplist is both valid and generalizable.

We now turn our attention to model parameters. Figures 4.3 and 4.4 show sweeps over the parameters $k$, $n$, and $\lambda$ for each collection. For a fixed value of each parameter, the plots show the cross validated MAP or NDCG@20 across the observed free parameter settings. A horizontal red line on each plot shows the baseline query likelihood effectiveness for comparison.

These parameter sweeps show that the values of the $k$ and $n$ parameters are largely unimportant: with few exceptions, it appears that baseline effectiveness can be surpassed at most values of these parameters.

In contrast, the model is sensitive to the value of $\lambda$. For three of the four target collections, high enough $\lambda$ values almost always result in MAP below the baseline. The remaining collection, AP, is at limited risk of falling below the baseline, but still exhibits the same peaked response to $\lambda$; that is, as in the other three collections, increased $\lambda$ improves retrieval effectiveness up to a point, after which effectiveness begins to fall off. It is noteworthy, however, that in most cases MAP does not drop below the baseline until around $\lambda = 0.9$, indicating that while the value of $\lambda$ plays an important role in maximizing retrieval effectiveness, most values of $\lambda$ are safe. GOV2 is the main exception to this finding, with MAP beginning to fall below the baseline around $\lambda = 0.4$. GOV2 NDCG@20 drops off even earlier for most expansion collections, though Wikipedia and the target collection remain above the baseline until $\lambda = 0.9$.

Collections show a preference for the Wikipedia and GOV2 collections as expansion sources. This indicates that document expansion is beneficial when it “generalizes” the language model, as might be expected using these types of “general subject” document
Figure 4.3: Parameter sweeps for query likelihood runs showing the cross validated MAP across the free parameters.
Figure 4.4: Parameter sweeps for query likelihood runs showing the cross validated NDCG@20 across the free parameters.
collections. This may also explain why, in many cases, the target collection becomes the preferred expansion source at higher values of $\lambda$: very high values of $\lambda$ may cause documents to become overly uniform in their language when more general collections are used for expansion. Wikipedia and GOV2 are also likely beneficial due to their relatively well-formed documents, which probably introduce fewer “noisy” terms (such as misspellings or non-topical content) than the target collections. Their well-formedness likely explains their advantage over the other general subject collection, WT10g.

It is also illuminating to examine the minimum observed retrieval effectiveness for each individual parameter setting as in Figures 4.5 and 4.6. Although Figures 4.3 and 4.4 show that the $n$ and $k$ parameters are generally inconsequential, Figures 4.5 and 4.6 show that it is still possible in most cases to significantly damage retrieval effectiveness at any value of $n$ and $k$. The two exceptions are AP expanded with itself, which is never lower than its baseline for either parameter, and Robust expanded with itself, which is no lower than its baseline for most values of $k$.

Interestingly, however, all collections except GOV2 with NDCG@20 guarantee some raw improvement at the lowest values of $\lambda$. Although significantly higher retrieval effectiveness scores can be achieved at higher $\lambda$ levels for most collections, this indicates that a small amount of document expansion with any expansion collection is guaranteed to match or improve over the baseline.

<table>
<thead>
<tr>
<th>Expansion Collection</th>
<th>Improved</th>
</tr>
</thead>
<tbody>
<tr>
<td>AP</td>
<td>0.91</td>
</tr>
<tr>
<td>GOV2</td>
<td>0.70</td>
</tr>
<tr>
<td>Robust</td>
<td>0.81</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>0.79</td>
</tr>
<tr>
<td>WT10g</td>
<td>0.62</td>
</tr>
</tbody>
</table>

Table 4.6: Fraction of parameter settings per expansion collection that improve retrieval effectiveness relative to the baseline for the AP collection.

Note, however, that these figures do not indicate the distribution of observed runs at each MAP value. Tables 4.6-4.9 show the overall fraction of runs that show raw improvement over
Figure 4.5: Parameter sweeps for query likelihood runs showing the minimum observed MAP.
Figure 4.6: Parameter sweeps for query likelihood runs showing the minimum observed NDCG@20.
Table 4.7: Fraction of parameter settings per expansion collection that improve retrieval effectiveness relative to the baseline for the Robust collection.

<table>
<thead>
<tr>
<th>Expansion Collection</th>
<th>Improved</th>
</tr>
</thead>
<tbody>
<tr>
<td>AP</td>
<td>0.50</td>
</tr>
<tr>
<td>GOV2</td>
<td>0.53</td>
</tr>
<tr>
<td>Robust</td>
<td>0.88</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>0.60</td>
</tr>
<tr>
<td>WT10g</td>
<td>0.47</td>
</tr>
</tbody>
</table>

Table 4.8: Fraction of parameter settings per expansion collection that improve retrieval effectiveness relative to the baseline for the WT10g collection.

<table>
<thead>
<tr>
<th>Expansion Collection</th>
<th>Improved</th>
</tr>
</thead>
<tbody>
<tr>
<td>AP</td>
<td>0.53</td>
</tr>
<tr>
<td>GOV2</td>
<td>0.60</td>
</tr>
<tr>
<td>Robust</td>
<td>0.59</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>0.64</td>
</tr>
<tr>
<td>WT10g</td>
<td>0.58</td>
</tr>
</tbody>
</table>

Although WT10g expanded with Robust shows potential for significant damage according to Figures 4.5 and 4.6, Table 4.8 shows that in fact, 59% of parameter settings for WT10g expanded with Robust result in raw improvement over the baseline. Therefore, although the potential damage done by document expansion is great in this case, the odds are that it will cause improvement. In fact, in almost all cases, the majority of parameter settings for AP, Robust, and WT10g result in improvement. GOV2 is a more difficult collection: in no cases do the majority of parameter settings improve over the baseline.

Comparing the tables with Figures 4.5 and 4.6, we can see that in most cases, expansion with the target collection is the “ safest” choice in that it is both most likely to improve over the baseline (excepting WT10g) and has the least potential damage for most parameter settings.

Overall, this analysis shows that the $\lambda$ parameter should be carefully chosen to maximize retrieval effectiveness but that suboptimal parameter selection is still likely to produce a model that improves over the baseline.
4.6 Discussion

The results indicate that our approach for document expansion based on language modeling works well even for full length documents. Earlier work suggested that only shorter documents are likely to exhibit sufficient topical coherence [27]; our work demonstrates that full length documents are similarly capable.

Although expansion using doc2vec document embeddings proved somewhat successful, overall better retrieval effectiveness was achieved with document pseudo-queries. Given these results, and taking into consideration the added computational complexity and increased number of parameters introduced by doc2vec expansion, we conclude that pseudo-query expansion is the more desirable approach. If it is true that document pseudo-queries suffer from vocabulary mismatch in identifying expansion documents, the results show that they are nevertheless more successful than the doc2vec model, which is theoretically designed to identify document similarities deeper than the observed word forms.

In contrast to query likelihood runs, query expansion using expanded documents resulted in little retrieval effectiveness improvement over the baseline. Since more accurate query likelihood scores imply that a relevance model can more accurately weight the importance of feedback documents, and given the improvement to query likelihood runs seen with document expansion—especially at higher ranks, as indicated by NDCG@20 scores—it is perhaps surprising that RM3 retrieval effectiveness was not more impacted by document expansion. We hypothesize that the small changes to query likelihood scores induced by document

<table>
<thead>
<tr>
<th>Expansion Collection</th>
<th>Improved</th>
</tr>
</thead>
<tbody>
<tr>
<td>AP</td>
<td>0.14</td>
</tr>
<tr>
<td>GOV2</td>
<td>0.45</td>
</tr>
<tr>
<td>Robust</td>
<td>0.18</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>0.27</td>
</tr>
<tr>
<td>WT10g</td>
<td>0.24</td>
</tr>
</tbody>
</table>

Table 4.9: Fraction of parameter settings per expansion collection that improve retrieval effectiveness relative to the baseline for the GOV2 collection.
expansion, while sufficient to cause beneficial re-ranking of results lists, is too small in magnitude to have a strong effect on relevance model estimation, which relies on term weights rather than document ranks. We note, however, that though the results of RM3 based on document expansion have not proved exciting, they do occasionally provide statistically significant improvement with minimal risk of significant damage to results quality.

Finally, we note that although expansion collections and target collections exhibit different relationships as discussed in Section 4.5.3, overall retrieval effectiveness metrics indicate that the choice of expansion collection is not critical to the success of the document expansion model. Nevertheless, in most cases higher raw effectiveness scores were observed when external collections were used rather than the target collection, the latter of which has been the typical choice in prior work on document expansion. The results indicate therefore that while the success of the model is not dependent on access to high quality external collections like Wikipedia, use of external collections may be beneficial to maximizing retrieval effectiveness.

4.7 Conclusions

This work is significant in its extensions of past document expansion research. Specifically, the contributions of this work are:

- Consistent and successful application of language modeling techniques during document expansion with full text documents
- The use of external collections and multiple collections to further improve retrieval effectiveness
- Examination of retrieval effectiveness with relevance models built from expanded documents

These findings indicate that document expansion yields improvements to document language model estimates. The increased effectiveness of retrieval models that rely on LM
estimates suggests that artificially augmented term samples can in fact produce more accurate models. This conclusion is further explored in Chapter 5.
Chapter 5
Document representation and re-estimation

5.1 Introduction

While document expansion has been shown to improve ad-hoc retrieval effectiveness, the precise mechanisms by which it estimates improved language models are not completely obvious. Some of the assumptions underlying the document expansion process seem unlikely. In particular, the creation of document pseudo-queries presupposes that a document’s most frequent terms will reflect a relatively coherent topical makeup. However, previous work on document expansion casts doubt on the topical uniformity of full length documents [27].

Further, the most successful topic modeling technique, latent Dirichlet allocation, assumes that documents are generated by sampling from multiple topic models; this assumption leads to lower perplexity scores when compared to mixture of unigrams approaches resembling the document pseudo-query process [6].

More broadly, document expansion raises questions about the nature of improvements to language models. Given that documents (and queries) are only sparse samples from underlying generative distributions, in what ways are those samples deficient and what are the effects of purported model improvements like document expansion? We are interested in the limits of methods like document expansion—that work by refining language model estimates—in improving retrieval effectiveness.

This chapter is designed to answer the following questions:

1. How can we quantify language model improvement?
2. Can full length documents be adequately represented by truncated language models, i.e., pseudo-queries?
3. In what ways are language model re-estimation techniques unlikely to improve retrieval
These questions are predicated on notions of document topicality; they are all concerned with aspects of language model sufficiency and quality, which, in the context of systems-oriented IR research, are best assessed with respect to topicality. It is therefore extremely valuable in answering these questions to have access to data that indicates the topical makeup of documents. In Section 5.2, we discuss the collection of this data. In Sections 5.3–5.5, we use this data to attempt to answer the research questions posed above.

5.2 Topic term annotation dataset

5.2.1 Annotation task description

Figure 5.1: An example of the interface design. The document is shown on one side with topic term choices shown on the right. Annotators read the document, select the best words, and then submit their decision with the “Submit” button. In case annotators have an issue loading a document, they are also provided a skip button.
In order to collect data reflecting document topicality, we asked human annotators to interpret document subject matter and make decisions about how it may be best represented. Though topic modeling approaches like LDA provide a tantalizing automated alternative to the onerous annotation collection process, automatically discovered topics are not always semantically coherent (as discussed in Section 2.5.1), and the choice of \( k \), the number of topics to model, would likely have a profound effect on the results of this research. In contrast, human annotations allow us to directly identify the terms that constitute a coherent topical description of the document.

To collect these annotations, each annotator was presented with a document, which he or she was asked to read. In addition, the participant was presented with a list of terms that may or may not describe the document, generated according to a process described in Section 5.2.2. After reading the document, the participant selected some subset of terms that he felt best describe the document, which we will call the “topic terms,” up to ten terms per document. This procedure was repeated with many documents to ensure an adequate sample.

Seven study participants were recruited from among the University of Illinois at Urbana-Champaign School of Information Sciences current and former graduate students as well as law school students at the University of Michigan known personally by the author. Requirements for participation were English proficiency and access to a web-capable computer. Participants were paid $15.00 in Amazon.com gift certificates per hour of work for a maximum of five hours of work on the project. All annotators worked the full five hours.

The procedure was completed through a simple web-based system created for this project that randomly assigned documents to study participants, provided an interface for document annotation, and stored data. A screenshot of the interface is shown in Figure 5.1. Annotators were required to attend an introductory meeting to explain the task and interface and to answer any questions. Because the system was internet accessible, annotators were permitted to complete the remainder of the task from whatever location they chose using any web-
capable computer.

As part of this initial meeting, study participants were provided broad guidelines to help them understand and complete the task. The guidelines were designed to be suggestive, rather than strict, in recognition of the subjectivity involved in interpreting topicality. In addition to the task description, the guidelines directed annotators not to follow hyperlinks nor to use sources outside the document itself in making term selections. Participants were also instructed to select only one form of each word (e.g., to select either the singular or the plural, but not both) to maximize the breadth of topic coverage. Participants were asked to complete practice annotations to ensure their complete understanding of the system and the task.

In order to assess inter-annotator agreement, all participants were required to annotate the same fifteen documents, which were randomly chosen from the pool of documents. These shared documents were the first fifteen annotated by all participants, and were the only documents that they could not elect to skip.

In addition, to assess the quality of annotations, documents were randomly injected with requests for annotators to select specific terms. These terms were sampled with likelihood inverse to their document language model probabilities to attempt to ensure that they would not be selected by chance as part of the annotation process. Participants were advised that these requests would be present in documents.

5.2.2 Data

Participants were presented with documents selected at random, without replacement, from the TREC collections used for document expansion experiments (see Section 3.2.1). Although unjudged documents are assumed to be nonrelevant in most IR applications, only judged documents were sampled, since knowing their relevance to a query provides valuable information in assessing the effects of language model change. Relevant and nonrelevant documents were sampled with equal probability initially, but after removing documents longer
than 1,000 words in length to ensure a sufficient number of annotations would be produced, the ratio of relevant to nonrelevant documents fell to approximately 40:60. This squares with prior observations that the probability that a document is relevant tends to increase as document length increases [86, 88].

An important component of the annotation interface was the set of words presented to participants as choices for topic term selection. The goal in sampling terms was to select them from a variety of sources that were likely to describe the document well while including terms that do not necessarily appear in the document itself. Providing a choice of terms, rather than allowing annotators to write in terms themselves, was intended to increase the chances that terms absent from the document text would be selected, while also increasing the likelihood that participants would select a variety of terms rather than focusing in on specific topics appearing within the document. Terms were sampled from several sources:

1. The original document language model
2. The Wikipedia expansion document language model ($P(w|E_{D_j})$ from Section 4.3)
3. The target collection expansion document language model
4. Relevance models constructed for each query to which the document was judged relevant

In each case, terms were sampled according to their probabilities in the underlying distribution. For example, each term sampled from an expansion document language model was sampled with probability $P(w|E_{D_j})$. However, two limitations were placed on the terms sampled: first, terms were sampled without replacement across the language models for a given document, i.e., each subsequent distribution in the list above was sampled excluding any terms that had been sampled from an earlier distribution; and second, no stopwords or numbers were sampled from any distribution. On average, approximately 50 topic term choices were presented to participants for each document, though the precise number of terms varied somewhat.
Figure 5.2: (a) Per-document term selection frequencies across all annotators, not including term selections required for quality checks. (b) The number of terms selected per document by each participant.

5.2.3 Annotation summary

<table>
<thead>
<tr>
<th>Annotator</th>
<th># Docs</th>
<th># Terms</th>
<th>Avg. Terms/Doc</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>40</td>
<td>268</td>
<td>6.70</td>
</tr>
<tr>
<td>B</td>
<td>165</td>
<td>1,486</td>
<td>9.01</td>
</tr>
<tr>
<td>C</td>
<td>111</td>
<td>1,067</td>
<td>9.61</td>
</tr>
<tr>
<td>D</td>
<td>90</td>
<td>895</td>
<td>9.94</td>
</tr>
<tr>
<td>E</td>
<td>140</td>
<td>1,396</td>
<td>9.97</td>
</tr>
<tr>
<td>F</td>
<td>147</td>
<td>1,166</td>
<td>7.93</td>
</tr>
<tr>
<td>G</td>
<td>116</td>
<td>950</td>
<td>8.19</td>
</tr>
</tbody>
</table>

Table 5.1: The total number of documents annotated and terms selected by each participant, and the average number of terms selected per document.

Seven annotators participated in the study. To protect their privacy, they are identified with the letters A through G. Table 5.1 shows the total number of documents, total number of terms, and average number of terms per document annotated by each participant. Participants completed a total of 809 annotations, 105 of which belong to the shared set of 15 inter-annotator agreement documents. Since document annotations—not annotators—are the focus of our work, we were satisfied with the amount of data collected. Nevertheless, it is important to ensure that annotations produced by different annotators are approximately interchangeable.

In the majority of cases, annotators opted to select the full ten terms for each document,
as can be seen in Figure 5.2(a). (Note that, due to an oversight, participants who failed
to select the required quality check term were allowed to select an additional topic term,
which explains the small number of documents annotated with eleven terms.) Annotator
A is a clear outlier here, having annotated less than half as many documents as the next
fewest annotations, and annotating only about 30% as many total terms. Annotator A also
selected by far the least average number of terms per document, and appears to have been
much more likely to select fewer terms than other participants, as shown in Figure 5.2(b).

Analysis of variance (ANOVA) was used to test whether participants differed significantly
in their annotation habits. The number of terms selected per document was found to differ
with statistical significance between study participants \((F(6, 802) = 60.59, p < 0.01)\). The
post hoc Tukey’s HSD test further showed that annotators C, D, and E did not differ
from each other significantly, nor did Annotators F and G. All other pairs of participants
annotated documents at significantly different rates. These statistics show that participants’
annotation behavior differed, but does not necessarily imply any problematic conclusions.

### 5.2.4 Quality checks

<table>
<thead>
<tr>
<th>Annotator</th>
<th># Shown</th>
<th>% Complete</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>11</td>
<td>54.55</td>
</tr>
<tr>
<td>B</td>
<td>46</td>
<td>65.22</td>
</tr>
<tr>
<td>C</td>
<td>23</td>
<td>78.26</td>
</tr>
<tr>
<td>D</td>
<td>24</td>
<td>100.00</td>
</tr>
<tr>
<td>E</td>
<td>35</td>
<td>74.29</td>
</tr>
<tr>
<td>F</td>
<td>38</td>
<td>78.95</td>
</tr>
<tr>
<td>G</td>
<td>27</td>
<td>81.48</td>
</tr>
</tbody>
</table>

Table 5.2: Summary statistics for quality check completion.

As discussed in Section 5.2.1, annotators were required to complete occasional “quality
checks” to ensure that their choice of topic terms was based on full consideration of the
document text. These quality checks took the form of instructions randomly injected into
document text that told annotators to select a specific term as part of the set of topic terms
Among the participants, only Annotator D successfully completed all of the quality checks presented. However, most other annotators completed a strong majority of quality checks; on average, they completed approximately 76% of quality checks.

Figure 5.3(a) shows the percentage of quality checks successfully completed across collections. Quality checks for documents from the AP and Robust collections were much more likely to be completed, likely because documents from these collections suffer from minimal formatting issues due to the simple structure of news documents. In contrast, the web documents constituting the GOV2 and WT10g collections are more likely to suffer from formatting issues causing quality checks to become relatively hidden.

Although we hypothesized that annotators might be less likely to notice quality checks in longer documents, there is no indication that this is the case: a two-tailed $t$-test comparing the lengths of documents whose quality checks were and were not completed shows no statistically significant difference ($p = 0.9365$).

Completion rates did vary across annotators, as shown in Figure 5.3(b). However, these factors are confounded by the rates at which quality checks were presented; for example, neither Annotator A nor G completed any quality checks for WT10g documents, but both were presented only two quality checks from that collection. Ultimately, the data suggests that all annotators appeared to attempt to complete quality checks, so we have chosen not
to discard any annotations on the basis thereof.

5.2.5 Inter-annotator agreement

The annotation task required participants to form subjective interpretations of document topicality. Because the remainder of the chapter uses these interpretations as a type of ground truth for the topical nature of documents, it is important to have a sense of inter-annotator agreement.

Fleiss’ $\kappa$ is a chance-corrected measure of inter-annotator agreement [28]. It extends Cohen’s $\kappa$ [16] to allow for more than two annotators. $\kappa$ statistics are defined as the amount of agreement observed above that expected by chance:

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e}$$

where $\bar{P}$ is the observed proportion of annotations in which the annotators agree and $\bar{P}_e$ is the expected mean proportion of agreement if annotators made their selections randomly. $\kappa$ therefore gives the proportion of annotations observed to be in agreement above the proportion expected to be made by chance.

In a typical inter-annotator agreement evaluation, it is assumed that each annotator makes one and only one selection (or “rating”) per annotated item (or “subject”). Because this task required participants to select multiple terms per document and did not require participants to select equal numbers of terms, the document does not qualify as the subject of annotation. Instead, we calculate an individual $\kappa$ score for each document, treating terms as the subject with binary ratings select/reject. That is, we measure the extent of agreement among annotators for each document by comparing, for each topic term choice presented to them, whether they selected the term or chose not to select it. Although participants were limited to ten selections and may have agreed more had they been allowed to select additional terms, the ten term limit was put in place to ensure that the terms selected were
the most descriptive of the document topic(s), and participants’ decisions in the face of this limit are an important indication of term importance.

<table>
<thead>
<tr>
<th>Document</th>
<th>Unstemmed $\kappa$</th>
<th>Stemmed $\kappa$</th>
</tr>
</thead>
<tbody>
<tr>
<td>AP880721-0095</td>
<td>0.3345</td>
<td>0.5750</td>
</tr>
<tr>
<td>AP880917-0104</td>
<td>0.4902</td>
<td>0.6797</td>
</tr>
<tr>
<td>AP890308-0220</td>
<td>0.5739</td>
<td>0.6183</td>
</tr>
<tr>
<td>AP890424-0284</td>
<td>0.2253</td>
<td>0.6526</td>
</tr>
<tr>
<td>FBIS3-39240</td>
<td>0.4969</td>
<td>0.5718</td>
</tr>
<tr>
<td>FBIS3-40185</td>
<td>0.3392</td>
<td>0.5219</td>
</tr>
<tr>
<td>FR940318-0-00101</td>
<td>0.5031</td>
<td>0.6096</td>
</tr>
<tr>
<td>FT941-10100</td>
<td>0.6062</td>
<td>0.6793</td>
</tr>
<tr>
<td>GX010-11-11242343</td>
<td>0.4209</td>
<td>0.3038</td>
</tr>
<tr>
<td>GX011-80-1074652</td>
<td>0.3528</td>
<td>0.5437</td>
</tr>
<tr>
<td>WTX016-B14-3</td>
<td>0.4102</td>
<td>0.6507</td>
</tr>
<tr>
<td>WTX018-B20-194</td>
<td>0.5490</td>
<td>0.6623</td>
</tr>
<tr>
<td>WTX026-B08-107</td>
<td>0.3578</td>
<td>0.2985</td>
</tr>
<tr>
<td>WTX099-B24-354</td>
<td>0.6214</td>
<td>0.7041</td>
</tr>
<tr>
<td>WTX101-B11-76</td>
<td>0.5035</td>
<td>0.6111</td>
</tr>
</tbody>
</table>

Table 5.3: Fleiss’ $\kappa$ scores for each document, treating per-term selection/rejection as the possible ratings to compare.

Per-document Fleiss’ $\kappa$ scores for the pool of documents shared among annotators are shown in Table 5.3. Note that all annotators were provided the same choice of topic terms. Annotators achieve reasonably strong agreement for most documents; mean $\kappa$ across all documents is 0.4523, and median $\kappa$ is 0.4902. By stemming terms, apparent disagreement due only to word form differences can be reduced*. After applying the so-called Porter2 stemming algorithm (often called the “Snowball” stemmer following the eponymous stemming programming language) [72, 73, 5], mean $\kappa$ improves to 0.5788 and median $\kappa$ to 0.6111. While interpretations of $\kappa$ vary, one frequently cited paper calls these levels of agreement “moderate” to “substantial” [48]. We consider these results to be quite good, given the subjectivity involved.

*Though annotators selected unstemmed terms (since interpretation of stems can be difficult), they were advised to select only a single form of each word in anticipation of later stemming. The purpose of stemming is to recover equivalence between multiple forms of a word—to reconcile syntactically different tokens with near-identical semantic content. Given the overall accuracy of modern stemming algorithms, the great majority of agreement added by calculating $\kappa$ on the basis of stems is therefore the result of resolving superficial differences in annotator selections.
5.3 Changes to document language model quality

Document expansion’s impact on retrieval effectiveness is the result of document language model re-estimation. By quantifying the quality of the re-estimated LM, it becomes possible to measure which characteristics of documents and of the document expansion process lead to improved language model estimates.

Like the concept of relevance in IR more generally, the notion of language model quality is complex and difficult to measure. To operationalize LM quality, we can assume that a good language model is one that emphasizes the topic(s) expressed by the document. A positive change in language model quality is therefore one that increases the emphasis on topic terms. Using this definition of LM improvement, there are several potential measurements of language model change. In this work, we will consider the following quantities.

Note that, in this and subsequent sections, analysis often makes reference to expanded document language models. Although the choice of collection used to expand document LMs can affect retrieval effectiveness, for simplicity in this analysis we will use Wikipedia due to its strong overall effectiveness (see Section 4.5).

5.3.1 Topic term likelihood change

The change in document LM topic term likelihood is a straightforward metric of language model change. If document expansion is beneficial—and results from Chapter 4 show that it is, overall—then we might reasonably expect topic term likelihood to increase as a result of expansion. Topic term likelihood can be calculated analogously to query likelihood. That is, we treat terms as independent and calculate the likelihood according to Eq. 4.2 (when the document is not expanded) and Eq. 4.6 (when the document is expanded). We then simply subtract the original likelihood from the expanded likelihood to produce the change in topic term likelihood resulting from expansion.

Figure 5.4 shows the distribution of topic term likelihood changes across annotated documents. As is visible, many documents are adversely affected by expansion. The proportion
of documents that show improved topic term likelihood, 59.7%, corresponds very closely with
the proportion of documents whose rank changed in the correct direction (i.e., increased for
relevant documents, decreased for nonrelevant documents), 56.3%. A two-tailed $t$-test shows
that the likelihoods of topic terms are nevertheless higher on average as a result of expansion
($p < 0.01$) as we would expect given overall retrieval effectiveness gains. Importantly, an-
other $t$-test shows that the average topic term likelihood change is not significantly different
between relevant and nonrelevant documents ($p = 0.1640$). This is what we would expect,
since topic terms reflect the topic of the document regardless of whether that topic is relevant
to a specific query.

Reassuringly, it can be seen in Figure 5.5 that although annotators sometimes made
different term selections, the change in term likelihoods as a result of expansion is consistent
across annotators. This finding is confirmed by one-way ANOVA ($F(6,802) = 1.6824,$
$p = 0.1223$).
5.3.2 Query likelihood change

Query terms are an appropriate basis for evaluating language model change because, when paired with relevance judgments, they can indicate the topic of a document. Since only judged documents were sampled for annotation, relevance judgments can be leveraged to measure language model change based on query likelihood. When a document is relevant to a query, we can reasonably expect its query likelihood to increase as a result of beneficial language model changes, as from document expansion. Query likelihood is calculated according to Eq. 4.2. As with topic term likelihood change, we can simply subtract the original document query likelihood from the expanded document query likelihood to calculate the QL change.

Unsurprisingly, Figure 5.6 shows that the query likelihood of relevant documents is higher under both the original (“target”) and expanded language models. Surprisingly, Figure 5.7 indicates that even most nonrelevant documents show an increase in query likelihood as a result of expansion. This figure further suggests that the query likelihood increase of relevant documents tends to be greater than that of nonrelevant documents, as might be expected. This is confirmed with a two-tailed \( t \)-test \( (p < 0.05) \).
Figure 5.6: The query likelihood for nonrelevant and relevant documents under the target and expanded language models.

Figure 5.7: The difference in query log likelihood between the expanded document language model and the baseline document language model, by relevance.
Query likelihood change shows moderate correlation with topic term likelihood change. Kendall’s $\tau$ correlation of the two quantities is 0.1721 ($p < 0.01$) for relevant documents and 0.1033 ($p < 0.05$) for nonrelevant documents. The higher relevant document correlation makes sense, since a document that is relevant to a query is generally on the topic of the query, and increased topic term likelihood should therefore correspond to increased query likelihood. In contrast, the likelihood of topic terms may increase in a nonrelevant document without necessarily corresponding to a decrease in query likelihood. The correlations therefore indicate that while the two measures may “get at” a similar concept—document language model improvement—they do so without being redundant.

5.3.3 Expansion document coherence

A slightly different approach to measuring language model quality, expansion document coherence uses the average pairwise cosine similarity among the expansion documents (as retrieved by the pseudo-query) to quantify the extent to which expansion documents are about the same topic. The idea in this case is that a more coherent set of expansion documents is a signal that the expanded document language model will have shifted in a specific,
clear direction. In this sense, the intuition is similar to that behind query clarity [19] or other query performance prediction cohesion scores [13, 24, 95]. Although this metric requires a greater leap in intuition, we include it for consideration because it is a language model improvement metric that can be calculated without reference to human annotations, making it much more inexpensive and versatile than the preceding two metrics.

Further, there is evidence to support the claim that expansion document coherence measures language model improvement. The metric correlates fairly well with topic term likelihood change: their Kendall’s $\tau$ correlation is 0.2541 ($p < 0.01$). It also correlates with query likelihood change: $\tau = 0.1845$ ($p < 0.01$) for relevant documents, $\tau = 0.0821$ ($p < 0.01$) for nonrelevant documents. As before, these correlations are strong enough to reassure us that all three quantities measure something similar (the quality of the document language model change) without being redundant.

5.4 Representing full length documents with pseudo-queries

Section 4.5 showed that document pseudo-queries were an effective approach to identifying related documents to incorporate into the expansion process, called “expansion documents.” To understand the source of their success and where there is room for improvement, it is important to understand how well document pseudo-queries are able to capture the main topic(s) of a document and the extent to which the quality of pseudo-queries relates to positive language model change as a result of document expansion.

This analysis is closely related to methods of query performance prediction (QPP), and, in fact, query performance predictors will be considered as metrics of pseudo-query quality in Section 5.4.2. The goal of QPP is to find measurements of the query and/or results list that predict the quality of the results list. A good predictor is one whose predictions correlate with true retrieval effectiveness metrics like average precision [13].

The task here is similar in that we want to identify characteristics of pseudo-queries that correlate with retrieval of high quality expansion documents. However, unlike in query
performance prediction, we cannot evaluate the quality of the expansion documents using traditional retrieval evaluation metrics, because there are no judgments assessing the relevance of expansion documents to the target document. Instead, we employ the metrics of language model change discussed in Section 5.3 to represent expansion document quality. This decision is premised on the idea that high quality expansion documents result in high quality language model changes—a reasonable premise, since the only change between the baseline and expanded document runs is the incorporation of expansion documents into document representations.

There are many possible methods of quantifying pseudo-query quality. We now explain several metrics intended to quantify pseudo-query quality and analyze their relationships to document language model improvement. We calculate this relationship using Kendall’s $\tau$ correlation, a non-parametric correlation coefficient that measures the association of two variables by comparing the ranking of their values [42]. Kendall’s $\tau$ can take values between -1 and 1, with -1 indicating perfect disagreement and 1 indicating perfect agreement. We also investigate the predictive power of combinations of pseudo-query quality metrics by employing multiple linear regression models, assessed by adjusted $R^2$.

5.4.1 Topical pseudo-query quality

**Topic term/pseudo-query term overlap**

One straightforward metric for quantifying the topical quality of pseudo-queries is the recall of topic terms in the pseudo-query. Presumably, the more topic terms appear in the pseudo-query, the better the pseudo-query has captured the primary subject matter of the document.

An alternative to term recall is the average precision (AP) of topic terms within the document pseudo-query. This metric differentiates between cases in which the same number of topic terms appear in the pseudo-query, but with differing prominence. That is, a pseudo-query that weights topic terms more highly (relative to other terms) might have the same recall as, but a higher average precision than, another pseudo-query that weights the same
Table 5.4: Results comparing two term overlap metrics with various metrics of language model improvement. The Kendall’s $\tau$ correlation between that metric and the LM change metric is reported with its statistical significance: * indicates $p < 0.05$, ** indicates $p < 0.01$. Terms less highly. A high AP of topic terms would indicate that they are prominent within the document pseudo-query, which is presumably desirable.

The definitions of recall and average precision can be found in Section 3.1.1. In calculating both metrics, terms are stemmed to minimize the effect of differing word forms.

Kendall’s $\tau$ correlations quantifying the association between term overlap metrics and language model improvement metrics are shown in Table 5.4. Language model change shows limited to no correlation with the two metrics of term overlap. Significant positive correlations are observed for relevant document QL change and expansion document coherence. However, the small magnitudes of the correlations show that these relationships are limited.

In general, it is surprising how little association there is between the quality of the pseudo-query, as measured by topic term overlap, and the quality of the language model change. In the following section, we propose an alternative measure of pseudo-query topical quality that better elucidates this association.

**Topic term/pseudo-query result overlap**

There are several reasons why term overlap metrics may not suffice for quantifying pseudo-query quality. For example, even when the two term sets strongly overlap, the pseudo-query may accord too little relative importance to the topic terms relative to other terms in the pseudo-query. This is partly addressed by average precision, but AP is a rank-based metric,
<table>
<thead>
<tr>
<th>LM change</th>
<th>Kendall’s $\tau$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topic term likelihood change</td>
<td>0.2034**</td>
</tr>
<tr>
<td>QL change (Nonrel. docs)</td>
<td>0.1193**</td>
</tr>
<tr>
<td>QL change (Rel. docs)</td>
<td>0.1148**</td>
</tr>
<tr>
<td>Expansion doc coherence</td>
<td>0.2113**</td>
</tr>
</tbody>
</table>

Table 5.5: Results comparing overlap of pseudo-query and topic query results, as measured by Jaccard similarity, with various metrics of language model change. The Kendall’s $\tau$ correlation between that metric and each LM change metric is reported with its statistical significance: * indicates $p < 0.05$, ** indicates $p < 0.01$.

meaning that even a pseudo-query with high topic term AP may be undermined by extreme term weights. In the other direction, a pseudo-query may fail to include many topic terms and still capture the main topic of the document, i.e., it may exhibit vocabulary mismatch. If one topic term is particularly discriminative, a pseudo-query containing that term and no other topic terms may also still be able to effectively capture the primary topic of the document.

A slightly different approach that may help account for these issues is to measure the overlap between the documents retrieved with a pseudo-query (i.e., the expansion documents) and those retrieved by using the topic terms as a keyword query. This more extrinsic measure of topic term/pseudo-query similarity may clarify the quality of the pseudo-query by more directly measuring the pseudo-query’s ability to identify topical expansion documents.

Jaccard similarity is an appropriate measure of results overlap. It is defined as:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

where $|A \cap B|$ is the number of items in the intersection of sets $A$ and $B$ and all other quantities are analogous. In this case, sets $A$ and $B$ would contain the documents retrieved for the pseudo-query and the “topic query.”

Table 5.5 shows the Kendall’s $\tau$ correlations between result overlap and various language
Table 5.6: A comparison of disjoint Wikipedia results sets for the pseudo-query and topic term query of document GX269-69-7323852. The document exhibited high overall language model similarity between the two results sets.

model change metrics. In contrast to term overlap metrics, results overlap appears to correlate more strongly with language model change, though correlations are still not as strong as we might expect. The weaker, approximately matching query likelihood correlations will be discussed in more detail below.

**Topic term/pseudo-query result similarity**

Sources of noise remain in the results overlap metric described above. For example, it is possible for two sets of results to be disjoint but cover very similar subject matter. One such comparison is shown in Table 5.6. The titles of the retrieved Wikipedia documents suggest that the subject matter of the competing results sets will be extremely similar; however, they happen to have selected distinct document sets.

We propose a final measure of pseudo-query topicality that may help alleviate these issues by comparing the *language* used in pseudo-query results and that of topic query results. Since the topic terms are the gold standard, pseudo-query results that use language similar to the topic query results likely indicate a higher quality pseudo-query.

An efficient way of estimating the language model of each set of retrieved documents is to concatenate them into a single large “pseudo-document” (not to be confused with the
Table 5.7: Results comparing the cosine similarity of pseudo-query and topic query results pseudo-documents with various metrics of language model change. The Kendall’s τ correlation between that metric and the LM change metric is reported with its statistical significance: * indicates $p < 0.05$, ** indicates $p < 0.01$.

Table 5.7 summarizes the correlations between results similarity and various language model changes. In general, correlations show an increase compared to their counterparts in Table 5.5. Though all correlations are significant, the correlations with topic term likelihood change and expansion document coherence are strongest, indicating moderate levels of positive correlation between pseudo-query quality and document language model improvement. See below for discussion of the query likelihood results.

**Query likelihood change**

Intuitively, we expected for the correlation between pseudo-query quality and query likelihood change to approximately match those of other language model improvement metrics, at least when limited to relevant documents. The rationale is as follows. First, a high quality pseudo-query should reflect the topic(s) of the document: when the document is relevant to a query, we would expect the pseudo-query to reflect the topic expressed by that query; when the document is nonrelevant, we have no such expectation. Second, we would expect a pseudo-query that reflects the topic of a query to retrieve expansion documents also on the
topic of that query. The query-oriented expansion documents should therefore boost query likelihood in the final expanded document language model.

However, the findings discussed in the previous two sections have indicated that query likelihood change does not correlate with pseudo-query quality in the expected manner: not only are QL change correlations much lower than other metrics of language model improvement, but, unexpectedly, the correlation coefficients of relevant and nonrelevant documents also tend to be much closer than anticipated. Despite these findings, there is some evidence to suggest that the description above may have merit. We observe that the piecemeal relationships relating pseudo-query quality to query likelihood change hold true at each step, as outlined below, and we suggest that the low correlations reported in the previous sections are indicative not of a lack of association, but of the extremely noisy phenomenon we are attempting to measure.

<table>
<thead>
<tr>
<th>P.-Q. quality metric</th>
<th>P.-Q./query sim.</th>
<th>Rel.</th>
<th>Kendall’s $\tau$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Results Jaccard sim.</td>
<td>Results precision</td>
<td>Nonrel.</td>
<td>0.0365</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Rel.</td>
<td>0.1949**</td>
</tr>
<tr>
<td></td>
<td>Query weight percent.</td>
<td>Nonrel.</td>
<td>0.0642*</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Rel.</td>
<td>0.1664**</td>
</tr>
<tr>
<td>Results cosine sim.</td>
<td>Results precision</td>
<td>Nonrel.</td>
<td>0.0608*</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Rel.</td>
<td>0.1674**</td>
</tr>
<tr>
<td></td>
<td>Query weight percent.</td>
<td>Nonrel.</td>
<td>0.0317</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Rel.</td>
<td>0.1376**</td>
</tr>
</tbody>
</table>

Table 5.8: Kendall’s $\tau$ correlations between pseudo-query quality metrics and measures of pseudo-query/query similarity for relevant and nonrelevant documents. Statistical significance at $p < 0.05$ (*) and $p < 0.01$ (**) is shown. The results show that relevant documents consistently achieve higher correlations and greater significance than nonrelevant documents.

To justify our first assumption above, that a high quality pseudo-query should reflect the topic of a query to which the document is relevant, we compare the Kendall’s $\tau$ correlations between measures of pseudo-query quality and two measures of pseudo-query/query similarity. To measure pseudo-query quality, we select Jaccard similarity of results and cosine similarity of result pseudo-documents—the two metrics that best correlated with language model improvement. To measure pseudo-query/query similarity, we first compute the pre-
cision of the top ten pseudo-query results from the target collection, which we call “results precision.” This metric is a convenient method of determining whether the pseudo-query is able to capture the topic expressed by the query. Since this metric can be used only when relevance judgments are available, we compare it against one that can be calculated by a fully independent system: the total fraction of pseudo-query weight contributed by non-stopword query terms. As in earlier sections, we use Wikipedia as the queried collection since it is unbiased in its contents.

Findings are shown in Table 5.8. Metrics of pseudo-query quality consistently correlate at a much higher rate with metrics of pseudo-query/query similarity for relevant documents than for nonrelevant documents. While correlations for nonrelevant documents are occasionally statistically significant, all relevant documents show statistical significance (and at much greater levels). We also observe that this pattern of much higher correlations for relevant documents than nonrelevant ones holds for both the relevance judgment dependent and independent metrics. Overall, these findings support the hypothesis that a high quality pseudo-query will capture the topic of a query to which the document is relevant.

<table>
<thead>
<tr>
<th>Pseudo-query/query sim.</th>
<th>Rel.</th>
<th>Kendall’s $\tau$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Results precision</td>
<td>Nonrel.</td>
<td>0.2178**</td>
</tr>
<tr>
<td></td>
<td>Rel.</td>
<td>0.2819**</td>
</tr>
<tr>
<td>Query weight percent.</td>
<td>Nonrel.</td>
<td>0.1172**</td>
</tr>
<tr>
<td></td>
<td>Rel.</td>
<td>0.2188**</td>
</tr>
</tbody>
</table>

Table 5.9: Kendall’s $\tau$ correlations between pseudo-query/query similarity metrics and query likelihood increase from the original to the expanded document language models. Statistical significance at $p < 0.05$ (*) and $p < 0.01$ (**) is shown.

There is also evidence to support our second assumption, that pseudo-queries that are more inclusive of the query tend to result in greater increases in query likelihood in the expanded document. Table 5.9 shows the Kendall’s $\tau$ correlations between pseudo-query/query similarity metrics and query likelihood change. In general, the results indicate that significant correlations exist between these quantities. This makes sense, since a pseudo-query that reflects the topic of a query is likely to retrieve expansion documents that are also on
that topic, thus boosting the topic’s prominence in the expanded language model.

Interestingly, in both cases relevant documents show a stronger correlation than nonrelevant documents; the difference is particularly pronounced using the query weight percentage of the pseudo-query. This outcome is surprising since we might expect that a pseudo-query that is highly similar to a query would tend to emphasize query terms regardless of whether that is correct, i.e., regardless of the document’s relevance. We hypothesize that relevant document pseudo-queries tend to appropriately weight query terms as a whole better than nonrelevant pseudo-queries. The latter may disproportionately emphasize certain query terms while underweighting others. As a result, similar total measures of pseudo-query/query similarity may not be equally representative of pseudo-query/query similarity, leading to the mismatched correlation coefficients for relevant and nonrelevant documents.

We have shown that clear relationships exist between pseudo-query quality and pseudo-query/query similarity as well as between pseudo-query/query similarity and query likelihood change. While we believe the evidence clearly supports these relationships, it is nevertheless the case that the magnitude of the correlations at each step are moderate at best. As the following section discusses, these weaker correlation coefficients are reflective of the inherent noise involved in quantifying these concepts. For example, two pseudo-queries with perfect results precision can (and almost certainly would) result in differing amounts of query likelihood change, since the precise language used in relevant documents varies. In fact, though Figure 5.7 shows that QL almost always increases, the QL of documents with perfect pseudo-query results precision does nevertheless occasionally decrease, possibly as a result of multiple topics occurring in long expansion documents. Pseudo-query quality metrics suffer from noise, too, as the following section discusses.

In general, we believe this discussion has shown that there is evidence to support a relationship between pseudo-query quality and query likelihood change, but that this relationship is difficult to detect as a result of noisy metrics. Perhaps these issues can be addressed in future work.
Combining pseudo-query quality metrics

<table>
<thead>
<tr>
<th>Topic term</th>
<th>Pseudo-query term</th>
<th>Pseudo-query weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>bodyguard</td>
<td>bodyguard</td>
<td>0.1630</td>
</tr>
<tr>
<td>gegeo</td>
<td>gegeo</td>
<td>0.1630</td>
</tr>
<tr>
<td>odai</td>
<td>odai</td>
<td>0.1630</td>
</tr>
<tr>
<td>hussein</td>
<td>hussein</td>
<td>0.1220</td>
</tr>
<tr>
<td>president</td>
<td>president</td>
<td>0.0816</td>
</tr>
<tr>
<td>son</td>
<td>son</td>
<td>0.0816</td>
</tr>
<tr>
<td>investigated</td>
<td>firing</td>
<td>0.0612</td>
</tr>
<tr>
<td>killing</td>
<td>justice</td>
<td>0.0612</td>
</tr>
<tr>
<td>saddam</td>
<td>presidential</td>
<td>0.0612</td>
</tr>
<tr>
<td>shooting</td>
<td>shooting</td>
<td>0.0408</td>
</tr>
</tbody>
</table>

Table 5.10: A comparison of topic terms and pseudo-query terms as chosen by Annotator G for document AP881121-0106. Bolded rows do not show term overlap.

<table>
<thead>
<tr>
<th>Topic terms results</th>
<th>Pseudo-query results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uday_Hussein</td>
<td>Uday_Hussein</td>
</tr>
<tr>
<td>The_Devil’s_Double</td>
<td>Kamel_Hana_Gegeo</td>
</tr>
<tr>
<td>Kamel_Hana_Gegeo</td>
<td>The_Devil’s_Double</td>
</tr>
<tr>
<td>House_of_Saddam</td>
<td>Sajida_Talfah</td>
</tr>
<tr>
<td>Adnan_Khairallah</td>
<td>Adnan_Khairallah</td>
</tr>
<tr>
<td>Sajida_Talfah</td>
<td>Samira_Shahbandar</td>
</tr>
<tr>
<td>Samira_Shahbandar</td>
<td>House_of_Saddam</td>
</tr>
<tr>
<td>Qusay_Hussein</td>
<td>Suzanne_Mubarak</td>
</tr>
<tr>
<td>Operation_Red_Dawn</td>
<td>Odai_Hussein</td>
</tr>
<tr>
<td>CIA_activities_in_Iraq</td>
<td>Kamel_Gegeo</td>
</tr>
</tbody>
</table>

Table 5.11: A comparison of Wikipedia results sets for the topic term query and pseudo-query of document AP881121-0106. The document exhibited low overall language model similarity between the results sets, despite high results overlap.

Measures of pseudo-query quality are inherently noisy. Even when terms and results overlap to a high degree, influential dissimilar results can cause language model dissimilarity. An example of this problem is shown in Tables 5.10 and 5.11. In the case presented there, the document scored highly on topic term/pseudo-query term and results overlap, but the results showed an unusually low cosine similarity between the results pseudo-documents. Cases like this demonstrate that no single measure presented in this section can fully capture topical pseudo-query quality, nor are these measures capable of fully representing the relationship.
Table 5.12: Pearson’s $r$ correlations of explanatory variables.

<table>
<thead>
<tr>
<th>Response</th>
<th>Topic Term AP</th>
<th>Topic Term Recall</th>
<th>Results Jaccard</th>
<th>Results Cosine</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topic Term AP</td>
<td>1.0000</td>
<td>0.8562</td>
<td>0.4173</td>
<td>0.3224</td>
</tr>
<tr>
<td>Topic Term Recall</td>
<td>1.0000</td>
<td>1.0000</td>
<td>0.3847</td>
<td>0.3531</td>
</tr>
<tr>
<td>Results Jaccard</td>
<td></td>
<td>1.0000</td>
<td>0.4482</td>
<td></td>
</tr>
<tr>
<td>Results Cosine</td>
<td></td>
<td></td>
<td>1.0000</td>
<td></td>
</tr>
</tbody>
</table>

Table 5.13: Linear regression model results. The first four columns show the variable coefficients. Bolded values indicate statistical significance at $\alpha = 0.05$. The model $p$-value gives the significance of the regression model as a whole using an $F$-test.

between the pseudo-query and the topic terms.

Nevertheless, we can establish that a relationship does exist by estimating linear regression models using language model improvement metrics as the response variable and pseudo-query quality metrics as explanatory variables. First, though, it is important to check whether the explanatory variables correlate highly with each other. These correlations are shown in Table 5.12. Though the variables tend to correlate somewhat, the highest correlation by far is between topic term average precision and topic term recall, which is unsurprising since they both attempt to measure term overlap. We choose to discard topic term AP because topic term recall is a more interpretable metric.

Linear regression models using the remaining three explanatory variables are shown in Table 5.13. While all regression models are significant at $\alpha = 0.05$, the adjusted $R^2$ values show that the expansion document coherence is best fit by its regression line. Both topic term likelihood and expansion document coherence are explained by most of the measures of pseudo-query quality, but expansion document coherence’s $R^2$ is almost twice that of topic term likelihood. This may suggest that expansion document coherence is the “clearest”
metric proposed. The poor QL change models, at least, are likely the result of noise, as was previously discussed.

5.4.2 Other measures of pseudo-query quality

In the above, we have attempted to measure pseudo-query quality with reference to topicality. Unsurprisingly, but importantly, we have found that the extent to which pseudo-queries capture document topicality correlates to the extent of language model improvement in the expanded document. While confirmation of the importance of pseudo-query topicality is important, other measures of pseudo-query quality are also identifiable. We briefly examine simple, established query performance prediction metrics as alternative measures of pseudo-query quality. Many query performance prediction metrics exist; many of them correlate with one another strongly (e.g. average inverse document frequency and average inverse collection term frequency [70]), and some involve complex and/or computationally costly calculations. Because our goal is only to explore the relationship between query difficulty and language model improvement, highly precise predictors are unnecessary. We therefore limit our exploration to metrics that are simple and inexpensive to calculate.

The simplest predictor we calculate is the average collection query similarity (average SCQ) [111]. SCQ of a query term is a straightforward measure of the term’s prominence in the collection:

\[
SCQ(w) = [1 + \log(c(w, C))] \cdot IDF(w, C)
\]

where \(c(w, C)\) is the count of term \(w\) in collection \(C\) and \(IDF(w, C)\) is the inverse document frequency of \(w\) in \(C\). We can combine the SCQ scores of each query term into a single score in several ways; we choose the average SCQ for simplicity. The intuition underlying this predictor is that the more similar a query is to a collection, the more likely the collection is to contain useful documents. Note, however, that we discard term weights in calculating
Another predictor considered is the simple clarity score (SCS) [112], a simplified form of the classic clarity score [19], which measures the KL-divergence between the query language model and the background collection language model. Whereas the traditional clarity score uses a relevance model [51] to better represent the query LM, the SCS simply takes the query as is, and is therefore much more computationally inexpensive:

\[ SCS(Q) = \sum_{w \in Q} P(w|Q) \log \frac{P(w|Q)}{P(w|C)}. \]

With a typical keyword query, \( P(w|Q) = \frac{1}{|Q|} \), i.e., term probabilities in the query are uniform. However, terms in the pseudo-query are weighted, and these weights can be used as \( P(w|Q) \). Pseudo-queries are therefore well-suited to SCS.

Finally, we consider weighted information gain (WIG) [112], which operates by comparing the scores of top ranked documents against the overall collection query score:

\[ WIG(Q) = \frac{1}{k} \sum_{D \in F} \sum_{w \in Q} \lambda(w) \log \frac{P(w|D)}{P(w|C)} \]

where \( F \) is the set of top-ranked documents for query \( Q \). \( P(w|D) \) is calculated using Dirichlet-smoothed query likelihood, while \( P(w|C) \) is calculated as the simple maximum likelihood estimate. \( \lambda(w) \) is related to term types in the Markov random field retrieval model [63]; for our purposes, it is simply \( \frac{1}{\sqrt{|Q|}} \). WIG’s premise is similar to that of clarity, but compares retrieval scores rather than language models.

Kendall’s \( \tau \) correlations comparing these query performance predictors against metrics of language model change are shown in Table 5.14. In general, average SCQ appears to relate minimally with language model improvement. SCS is even worse, although interestingly we find an unusually significant negative correlation between SCS and expansion document coherence. Indeed, SCS is weakly but significantly correlated negatively with topic term likelihood increase as well. In contrast to traditional QPP studies, this suggests that docu-
### Table 5.14: Kendall’s τ correlations comparing three query performance prediction metrics against language model improvement metrics. Statistical significance at $p < 0.05$ (*) and $p < 0.01$ (**) is shown.

<table>
<thead>
<tr>
<th>LM change metric</th>
<th>QPP metric</th>
<th>Kendalls τ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topic term likelihood change</td>
<td>Avg. SCQ, SCS, WIG</td>
<td>0.0718**, -0.0566*, 0.3105**</td>
</tr>
<tr>
<td>QL change (nonrel.)</td>
<td>Avg. SCQ, SCS, WIG</td>
<td>0.0989**, 0.0523*, 0.1345**</td>
</tr>
<tr>
<td>QL change (rel.)</td>
<td>Avg. SCQ, SCS, WIG</td>
<td>0.0444, 0.0317, 0.1626**</td>
</tr>
<tr>
<td>Expansion doc coherence</td>
<td>Avg. SCQ, SCS, WIG</td>
<td>0.0197*, -0.1886**, 0.2642**</td>
</tr>
</tbody>
</table>

Table 5.15: Correlations among query performance predictors.

<table>
<thead>
<tr>
<th></th>
<th>Avg. SCQ</th>
<th>SCS</th>
<th>WIG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avg. SCQ</td>
<td>1.0000</td>
<td>0.6354</td>
<td>0.3964</td>
</tr>
<tr>
<td>SCS</td>
<td>1.0000</td>
<td>0.2113</td>
<td></td>
</tr>
<tr>
<td>WIG</td>
<td>1.0000</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

As with topical pseudo-query quality metrics, we can combine query performance predictors to predict language model change. Checking correlations among predictors (see Table 5.15), we find average SCQ and SCS to be moderately correlated. Given average SCQ’s poor $\tau$ coefficients with LM improvement metrics, we opt to remove it from the regression model. Results of the model are shown in Table 5.16. Contrasting with Table 5.13, we find...
Table 5.16: Linear regression models predicting language model improvement using query performance predictors.

<table>
<thead>
<tr>
<th></th>
<th>Topic term likeli.</th>
<th>QL (nonrel)</th>
<th>QL (rel)</th>
<th>Exp. cohere.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>$1.7020 \cdot 10^{-4}$</td>
<td>$-6.2980 \cdot 10^{-5}$</td>
<td>$9.3110 \cdot 10^{-5}$</td>
<td>$1.0248$</td>
</tr>
<tr>
<td>SCS</td>
<td>$-6.1910 \cdot 10^{-5}$</td>
<td>$-2.3290 \cdot 10^{-6}$</td>
<td>$-2.0250 \cdot 10^{-5}$</td>
<td>$-0.0489$</td>
</tr>
<tr>
<td>WIG</td>
<td>$1.5630 \cdot 10^{-4}$</td>
<td>$4.9690 \cdot 10^{-5}$</td>
<td>$6.8800 \cdot 10^{-5}$</td>
<td>$0.0582$</td>
</tr>
<tr>
<td>Adj. $R^2$</td>
<td>$0.2106$</td>
<td>$0.0406$</td>
<td>$0.0487$</td>
<td>$0.3065$</td>
</tr>
<tr>
<td>Model $p$-val.</td>
<td>$&lt; 0.0100$</td>
<td>$&lt; 0.0100$</td>
<td>$&lt; 0.0100$</td>
<td>$&lt; 0.0100$</td>
</tr>
</tbody>
</table>

Table 5.17: Linear regression models predicting language model improvement using query performance predictors and topical pseudo-query quality metrics.

<table>
<thead>
<tr>
<th></th>
<th>Topic term likeli.</th>
<th>QL (nonrel)</th>
<th>QL (rel)</th>
<th>Exp. cohere.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>$-1.3490 \cdot 10^{-3}$</td>
<td>$1.6950 \cdot 10^{-5}$</td>
<td>$-5.3080 \cdot 10^{-4}$</td>
<td>$0.0273$</td>
</tr>
<tr>
<td>Term overlap</td>
<td>$-3.2010 \cdot 10^{-4}$</td>
<td>$-3.3120 \cdot 10^{-5}$</td>
<td>$6.1350 \cdot 10^{-5}$</td>
<td>$-0.0836$</td>
</tr>
<tr>
<td>Result overlap</td>
<td>$9.0610 \cdot 10^{-5}$</td>
<td>$6.1200 \cdot 10^{-5}$</td>
<td>$-1.9260 \cdot 10^{-5}$</td>
<td>$-0.0138$</td>
</tr>
<tr>
<td>Result sim.</td>
<td>$1.6980 \cdot 10^{-3}$</td>
<td>$-9.4790 \cdot 10^{-5}$</td>
<td>$6.8670 \cdot 10^{-4}$</td>
<td>$0.9846$</td>
</tr>
<tr>
<td>SCS</td>
<td>$-5.5890 \cdot 10^{-5}$</td>
<td>$1.6530 \cdot 10^{-6}$</td>
<td>$-2.2170 \cdot 10^{-5}$</td>
<td>$-0.0357$</td>
</tr>
<tr>
<td>WIG</td>
<td>$1.3420 \cdot 10^{-4}$</td>
<td>$4.4710 \cdot 10^{-5}$</td>
<td>$5.2830 \cdot 10^{-5}$</td>
<td>$0.0515$</td>
</tr>
<tr>
<td>Adj. $R^2$</td>
<td>$0.2607$</td>
<td>$0.0420$</td>
<td>$0.0357$</td>
<td>$0.4097$</td>
</tr>
<tr>
<td>Model $p$-val.</td>
<td>$&lt; 0.0100$</td>
<td>$&lt; 0.0100$</td>
<td>$&lt; 0.0100$</td>
<td>$&lt; 0.0100$</td>
</tr>
</tbody>
</table>

that query performance predictors are actually better at predicting LM improvement than are topical pseudo-query quality metrics.

Better still are linear models that combine query performance predictors with topical pseudo-query quality metrics, as shown in Table 5.17. The increase in adjusted $R^2$ displayed by these “full” linear models, as well as the consistent significance of both QPP and topical variables, reveal that these two types of measurements complement one another in predicting pseudo-query quality.

### 5.4.3 Discussion

The success of document expansion rests largely on the success of the document pseudo-query in retrieving useful expansion documents. Previous work on expansion of microblog documents assumed that pseudo-query viability was tied to topical coherence; in other words, justification for microblog pseudo-queries derived from the “hypothesis that most short texts
discuss only a single topic” [27].

The results presented in Chapter 4 have already proven that pseudo-queries work well for full length documents. This section has shown that the success of pseudo-queries is indeed related to their ability to capture the topic of a document. But if the suppositions of [27] are correct, and full length documents do not share the topical coherence of microblog documents, our findings indicate that pseudo-queries are nevertheless able to capture documents’ multifaceted topical makeup, and that the more they do so, the better their effects on document language model change.

However, QPP measures demonstrate that there is more to pseudo-query quality than simply topical pertinence. For example, it is easy to imagine a scenario in which a highly topical pseudo-query is nevertheless unable to differentiate relevant expansion documents from nonrelevant ones, particularly if the topic of the document itself is somehow overly broad, or unreasonably specific, relative to the collection as a whole. Query performance predictors therefore demonstrate the need not only for appropriate pseudo-query topicality, but also for pseudo-query suitability—meaning a pseudo-query of appropriate intelligibility for retrieval from the collection.

The question explored in this section—whether pseudo-query quality is associated with document language model improvement—is a noisy one. There are not objective measures of pseudo-query quality or language model quality, which means the metrics used in this section are proxies for quality, necessarily built on intuitions and subjective interpretations. Given the amount of subjectivity required by the analysis, the adjusted $R^2$ values of linear models, which may be considered low in some situations, should be interpreted here as strong evidence of a relationship between pseudo-query quality, with reference to topicality and intelligibility, and document language model improvement.

These findings may not be surprising—a reasonable investigator would likely have hypothesized that pseudo-query quality, as measured with respect to topicality and intelligibility, was tied to document language model improvement. However, we have not only confirmed
this likely hypothesis—a valuable finding in its own right—but also shown that pseudo-query topicality is an appropriate quality measure even of full length documents with presumably multiplex subject matter. We have further shown that pseudo-query quality can be quantified according to multiple axes: topicality and intelligibility. Finally, our ability to predict language model improvement on the basis of pseudo-query quality holds potential for more than the establishment of relationships that we have achieved here; it also opens the door to automated feedback about the success of a document expansion process. We explore such an idea in the next chapter.

5.5 The relationship between query terms and topic terms

Topic terms may reasonably be interpreted as a query representing an information need that the annotator believes the document satisfies. The annotations collected for this project therefore provide an implicit relevance judgment, indicating that the document is relevant to an information need that may be expressed by the terms selected. This interpretation of relevance has precedent in Lavrenko’s generative theory of relevance, in which queries (and documents) are seen as random term samples representing an information need [49, 50].

TREC data also provides an explicit, alternative source of relevance judgments. Because every document annotated for this work was sampled from the pool of judged documents, each document is guaranteed to have at least one associated TREC relevance judgment. In actuality, approximately 25% of annotated documents were judged by TREC assessors with respect to multiple queries. The maximum observed number of query relevance judgments for a single document was eight; no document was judged relevant to more than two queries.

In general, we would expect explicit and implicit relevance judgments to agree. This assumption appears to be supported by the substantial difference in the average recall of query terms in the topic term sets of nonrelevant and relevant documents. Comparing stemmed query and topic terms, we find that the mean recall of non-stopword query terms in topic terms is 0.1701 for nonrelevant documents, compared to 0.3798 for relevant documents;
a two-tailed $t$-test indicates that this difference is significant with $p < 0.01$. The exact distributions are shown in Figure 5.9. Since implicit relevance judgments are indicated by the amount of term overlap (greater overlap indicating relevance, less overlap indicating nonrelevance), the association between magnitude of overlap and explicit relevance supports the idea that explicit and implicit relevance judgments typically agree.

It is therefore illuminating to examine the cases in which the implicit topic term relevance judgment appears to contradict the explicit TREC relevance judgment. These cases will help to understand ways in which reasonable modifications to retrieval functions, such as re-estimation of document language models, are unlikely to prove capable of increasing retrieval effectiveness. This is because disagreement between implicit and explicit relevance judgments suggests that deeper semantic issues may be at play than a standard term-based retrieval model is able to capture.

In the remainder of this section, we will describe situations in which topic term annotators and TREC assessors appear to disagree about the topicality of documents. In doing so, we hope to illuminate areas in which typical language modeling techniques like document
expansion are likely to continue to fall short. As in Section 5.4, analysis in this section will use stemmed terms whenever term overlap is measured. Doing so minimizes the misleading effects of word form mismatch.

5.5.1 TREC relevance without term overlap

Annotators selected no query terms as topic terms for approximately 33% of explicitly relevant documents. Since TREC relevance judgments indicate that these documents are about the subject expressed by the query, these cases suggest a fundamental disagreement between topic term annotators and TREC annotators about the topical makeup of these documents. While data was not collected about annotators’ decision making processes, we may attempt to understand why they may have chosen not to select query terms.

In some cases, this supposed disagreement may be the fault of the term sampling scheme used to provide topic term choices to annotators. About one in ten explicitly relevant documents did not offer query terms as options. The absence of query terms among topic term choices indicates that query terms did not appear, or appeared with very small probability,
across the term sources described in Section 5.2.2.

This conclusion is supported by Figure 5.10, which shows that query terms absent from the topic term choice set were almost always also absent from the text of the document. These terms evidently also do not appear with any frequency in related documents (i.e., expansion documents and pseudo-feedback documents) since these sources also contributed to the sample of terms offered as choices. These absent query terms therefore suffer from extreme mismatch in these documents; given that topic term choices were sampled from sources used for query and document expansion, it seems reasonable to conclude that typical language modeling approaches for reducing vocabulary mismatch are unlikely to benefit these outlying cases.

A more common scenario was for annotators to be presented terms from relevant queries as topic term choices but to decide not to select them. Figure 5.11(a) shows the distribution of counts of rejected query terms in annotated documents; it is contrasted with Figure 5.11(b), which shows the same distribution for selected query terms. From these figures, it can be seen that rejected query terms were much more likely to be absent from the document text than were present query terms. This bias in favor of terms appearing in the document text appears to explain a substantial fraction of query term rejections.

In fact, selection of query terms as topic terms appears to relate more generally to the prominence of the terms in the document text. Comparing Figures 5.11(a) and 5.11(b), it is
clear that query terms are much more likely to be selected when they appear more frequently in the document they describe. Unsurprisingly, a two-tailed $t$-test indicates that the mean frequency of rejected terms in documents (1.66) is significantly lower than that for selected terms (4.19) with $p < 0.01$. Clearly, this bias in favor of selecting prominent terms from the document is responsible for a large number of the rejected query terms.

While this bias does not reflect ideal annotation practices, it is important to remember that the rejected query terms were offered as topic term options despite their absence or dearth from the document text. Annotators therefore presumably considered, and ultimately decided against, selecting these terms. It is therefore important to remember that the decision not to select these query terms does not necessarily indicate rejection of the query terms’ topical pertinence; rather, given that around 60% of these document annotations used the maximum allowed number of terms (ten) it is likely the case that annotators often found other terms that they felt better described the topic of the document than did the excluded query terms.

This still leaves many documents where annotators did not maximize their topic term allowance, yet rejected query terms. While the true reasons behind each annotator’s choices cannot be known, manual examination of these term/document pairs reveals a few common cases. First, some of these terms are “instructional” terms that communicate a command to the search engine, rather than the desired topic of information to be retrieved. For example,
the word “information” was rejected for at least one document relevant to TREC query 495, which instructs the search engine to “find information” about the 1920s; since this term is not topical within the query, it is unsurprising that it was not selected to describe the document. Another reason for rejection of these query terms may be oversight. For example, Annotator G neglected to select the query term “Scotland” when annotating document WTX075-B17-172, which discusses locations and attractions throughout the country, but which uses the term “Scotland” only once, within the web page’s title element and therefore separate from the main contents of the document; it is possible that Annotator G did not realize which country the document described.

Most rejections, however, appear to indicate more fundamental differences between the document’s relevance and its topicality. For example, Annotator A chose not to annotate document WTX020-B20-176 with the query term “steroid,” although the document was judged relevant to query 509, “steroids; what does it do to your body.” The document, which pertains to yeast in the human body, mentions steroids only once, during a discussion of environmental factors contributing to yeast proliferation: “Examples include stress, inadequate nutrition, pollution, steroids, antibiotics and hormones in the meat we eat, sugar rich foods and refined carbohydrates.” We include the entire quote here to demonstrate how inconsequential the discussion of steroids is to the main topic of the document. Nevertheless, while a reasonable annotator would likely agree with Annotator A’s decision to reject “steroids” as a topic term, it also seems reasonable to conclude that the document is relevant to the information need expressed by the query, in that it provides an example of the effects of steroids on the body. Another example: document FBIS4-5631 is judged relevant to query 321, “women in parliaments.” Yet the document contains only a single mention of women—or parliaments—among its discussion of a government shake-up in the Marshall Islands: “Konou, the only woman in parliament, has been health minister since last year.” Annotator G, understandably, did not select the term “parliaments” nor the term “women” (nor “woman”) to describe a document concerned primarily with changes to
the foreign and education minister positions. Still, the document does discuss the state of women in a specific parliament, and is therefore arguably relevant to the query.

Technically, most theories of relevance would treat the preceding examples as instances of “topical” relevance; certainly, they do not qualify as “cognitive,” “psychological,” “situational,” or other commonly identified types of relevance [33, 81, 82]. Theoreticians equate topicality to “aboutness.” It would be difficult, however, to argue that these documents are “about” the queries, at least in the usual sense of the word “about.” Perhaps it is more accurate to claim that the documents include information about the queries. This nuance brings to mind Cooper’s definition of topical relevance, which is based on sentences and therefore exists “on a sub-document level” [17, 50].

This “sub-document level” of relevance is more appropriate for the documents described above, but it is a problematic level to operate at where TREC collections and document retrieval are concerned. This is of course because traditional TREC collections operate at the level of the document—it is the unit of both relevance judgments and of retrieval. Retrieval functions have therefore developed to assess the relevance of the document as a whole to the query; typically, the more a document appears to be about the query (i.e., the more it includes query terms) the higher its retrieval score. Accordingly, the two example documents discussed above are both ranked quite low in standard query likelihood retrievals. Traditional language modeling approaches to IR may have limited ability to adequately represent these types of sub-document level topics, though approaches like positional language models [58, 59] may help. Document expansion is unlikely to correct for many of these situations.

These deficiencies would not be deficiencies in a passage retrieval scenario [77, 10, 68, 40]. But in the traditional document retrieval scenario, for which the collections analyzed here are used, they point to a more fundamental mismatch between TREC’s stated goal to retrieve relevant documents and its actual operationalization of that goal, which is to find documents containing relevant information. This subtle distinction is critically important in the design of retrieval algorithms, and treating TREC relevance judgments as though
they assess document topicality as a whole is likely an impediment to improved retrieval effectiveness. (Of course, this presupposes that retrieving documents that include relevant information at high ranks is beneficial, as TREC judgments assert. It seems reasonable to claim, in contrast, that these documents should be ranked lower than those that are about the query. This question is outside the scope of this work.)

We have identified several major contributors to mismatch between query terms and topic terms in explicitly relevant documents:

- Extreme vocabulary mismatch between query terms and document text (including the text of related documents) leads to an absence of query terms from topic term options
- Annotator bias favors selection of terms appearing prominently in document text; absent or infrequent query terms are therefore less likely to be selected
- Query relevance does not necessarily imply query “aboutness,” leading annotators to disregard query terms that do not reasonably reflect the topic of the document

Annotator bias is an unfortunate side effect of the annotation process, but the remaining points suggest fundamental problems for language model re-estimation methods like document expansion. Both indicate that query terms are too far removed from document topicality for current methods to compensate, and while future methods may be able to compensate for extreme vocabulary mismatch and sub-document manifestations of relevance, understanding these limitations of current language model re-estimation techniques is an important step towards the development of those future methods.

5.5.2 Term overlap without TREC relevance

Much simpler to consider are documents annotated with query terms by topic term annotators, but judged to be nonrelevant by TREC assessors. It is difficult to imagine scenarios in which a document is best described by query terms yet is not relevant to the information need expressed by those query terms.
Manual analysis of explicitly nonrelevant documents that were annotated with query terms suggests that the problem is generally one of query representation. For example, Annotator B annotated document AP890525-0147 with all of the terms in query 192, “oil spill cleanup.” This is a reasonable description of the document’s subject matter, yet the document was judged nonrelevant by TREC assessors. Though we cannot determine with certainty the reason for this relevance judgment, it is likely due to the criteria described in the narrative form of the query, which states: “The mere mention of cleanup efforts without identifying the method or chemical used is not relevant.” Indeed, the document in question, while on the topic of “oil spill cleanup” (the title form of the query) does not specifically describe the methods used to accomplish a cleanup.

Other examples are plentiful. Documents relevant to query 493, “retirement,” are required to “list retirement communities . . . in the U.S. and Canada.” Query 395, “tourism,” explains in its narrative form that “only documents which detail an actual [economic] increase are relevant.” Query 689, “family-planning aid,” refers to international aid provided by the U.S., and documents about domestic funding are explicitly nonrelevant.

Use of the title form of TREC queries is a common practice, but these examples show that they do not always fully reflect the information need underlying them and in relation to which relevance judgments are made. Improvements to document language models are unlikely to benefit these cases; instead, new approaches to retrieval that take into account the description or narrative forms of queries—or new relevance judgments made on the basis of the title forms—are likely required. Fortunately, these situations have proven atypical among topic term annotations: only about 2.7% of the annotations made for explicitly nonrelevant documents contain all query terms (and, among these, nearly half are one-word queries); only around 5.5% contain more than half of the query terms.
5.6 Conclusions

The data collected for this chapter has been analyzed with an eye toward our document expansion retrieval model and language model re-estimation more generally. It has resulted in the following specific contributions:

- Measures of document language model quality have been proposed, based in part on the annotations collected
- Improvements to document language model quality have been linked to the quality of document pseudo-queries, suggesting that further retrieval effectiveness gains can likely be derived from improved document pseudo-queries
- Deficiencies of language model re-estimation techniques have been uncovered relating to vocabulary mismatch and differences between document aboutness and the inclusion of relevant information
- TREC topic query insufficiencies have been highlighted that standard language modeling techniques are unlikely to overcome

In addition, the data collected has potential application beyond our purposes here. For this reason, the data has been made available online [84].† The creation of this dataset is itself a valuable and significant contribution of this work.

†https://doi.org/10.13012/B2IDB-9822674_V1
Chapter 6
Selective document expansion

6.1 Motivation

The document expansion model described in Chapter 4 requires several parameters to be tuned, such as the number of terms included in the document pseudo-query, the number of expansion documents retrieved, and the mixing weight controlling the interpolation of the original and expansion language models. Cross validation was used to determine the values of these parameters that optimize, among the options specified in our search space, the overall retrieval effectiveness metrics of interest.

However, the approach taken in that chapter assumes that a single setting of these parameters will be found to optimize the effectiveness metric across all queries and documents. Analysis of the document expansion runs shows that, as might be expected, the optimal overall parameter setting is not equivalent to the optimal parameter setting on a per-document basis.

For example, we observe that among the top 1,000 documents retrieved for each query in the baseline Dirichlet-smoothed query likelihood runs reported in Section 4.5.1, 50.4% of relevant documents actually drop in rank as a result of document expansion with Wikipedia, while 41.9% of known nonrelevant documents incorrectly rise in rank. This means that for a large proportion of documents, the mixing parameter value that is best overall is actually worse than one that assigns no weight to the expansion data. The percentage of relevant and nonrelevant documents in each collection that are penalized by expansion with Wikipedia are detailed in Table 6.1.

As has been discussed previously, document expansion may be thought of as a form of smoothing. It is reasonable to believe that different documents may require different
amounts of smoothing; for example, shorter documents may require more smoothing than longer documents, as might documents comprised of fewer unique terms due to problems of sparsity. According to Zhai & Lafferty, smoothing methods that adapt to the needs of individual documents, as in Dirichlet smoothing, are desirable for improved estimation of the document language model [110].

Given that uniform document expansion has an adverse effect on the ranks of a sizable portion of documents, and given that document expansion is a form of smoothing, it is therefore reasonable and consistent to conclude that retrieval effectiveness would improve by selectively expanding documents only when doing so is likely to improve the retrieval effectiveness metric.

Oracle runs, which give the model’s highest possible effectiveness when the optimal expansion collection is always selected, are shown in Tables 6.2 and 6.3. They demonstrate that there is a great deal of room for improvement, suggesting that even limited success at predicting the optimal expansion collection could yield significant retrieval effectiveness

Table 6.1: The percentage of nonrelevant and relevant documents that were damaged by document expansion with Wikipedia (i.e., the percentage of nonrelevant documents whose rank increased, and the percentage of relevant documents whose rank decreased, as a result expansion).

<table>
<thead>
<tr>
<th></th>
<th>AP</th>
<th>GOV2</th>
<th>Robust</th>
<th>WT10g</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nonrelevant</td>
<td>43.20%</td>
<td>39.95%</td>
<td>41.94%</td>
<td>43.19%</td>
</tr>
<tr>
<td>Relevant</td>
<td>38.04%</td>
<td>52.99%</td>
<td>52.37%</td>
<td>52.59%</td>
</tr>
</tbody>
</table>

Table 6.2: Mean average precision of each target collection selecting the optimal expansion collection for each document on the basis of rank or score. “Highest MAP Expansion” refers to the maximum MAP observed for a run where documents are uniformly expanded with a single collection, as reported in Section 4.5.
Table 6.3: NDCG@20 of each target collection selecting the optimal expansion collection for each document on the basis of rank or score. “Highest NDCG@20 Expansion” refers to the maximum NDCG@20 observed for a run where documents are uniformly expanded with a single collection, as reported in Section 4.5.

<table>
<thead>
<tr>
<th></th>
<th>AP</th>
<th>GOV2</th>
<th>Robust</th>
<th>WT10g</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baselines</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unexpanded Baseline</td>
<td>0.4170</td>
<td>0.4134</td>
<td>0.3835</td>
<td>0.2738</td>
</tr>
<tr>
<td>Highest NDCG@20 Expansion</td>
<td>0.4794</td>
<td>0.4236</td>
<td>0.4177</td>
<td>0.3183</td>
</tr>
<tr>
<td>Multiclass</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rank</td>
<td>0.6766</td>
<td>0.5311</td>
<td>0.5785</td>
<td>0.4496</td>
</tr>
<tr>
<td>Score</td>
<td>0.8191</td>
<td>0.6050</td>
<td>0.6645</td>
<td>0.5524</td>
</tr>
<tr>
<td>Binary class</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rank</td>
<td>0.5822</td>
<td>0.5079</td>
<td>0.5159</td>
<td>0.4164</td>
</tr>
<tr>
<td>Score</td>
<td>0.6659</td>
<td>0.5813</td>
<td>0.5584</td>
<td>0.4780</td>
</tr>
</tbody>
</table>

improvement. These tables report both multiclass oracles, in which the optimal of the five candidate expansion collections is selected for each document, and binary class oracles, in which the given target collection is either expanded with Wikipedia or left unexpanded. Though the binary class oracles are consistently lower than their corresponding multiclass oracles, they nevertheless show substantial room for improvement over either baseline. We are therefore motivated to explore both the multiclass and binary class cases.

We are further motivated by research into selective query expansion, which has achieved at least limited improvement in retrieval effectiveness [57, 34, 107, 1]. Though query expansion and document expansion are not identical, the techniques are similar enough that prior success in query expansion may suggest the possibility of success for document expansion. However, it is important to note that improvements resulting from selective query expansion have been characterized as “marginal” [13], indicating that successful selective document expansion may be difficult to achieve.

This chapter presents a classification approach to expansion collection selection. In the remainder of this chapter, we describe the method used, define features, and present and analyze results.
6.2 Method

Note that in Tables 6.2 and 6.3, effectiveness is improved much more by selecting the expansion collection with the optimal document score, rather than rank. The rank of any one document is dependent on all other documents, which means that an expansion collection may be optimal for a given document only due to the effect of expanding all documents with this collection; since the proposed technique varies the expansion collection per-document, ranks become volatile. In contrast, the optimal score is defined independently of other documents: the optimal score for a relevant document is the highest score available, while the optimal score for a nonrelevant document is the lowest score available. We therefore assign collection labels based on document scores, rather than ranks, in much the same way that retrieval models assign scores rather than ranks.

Following the example of [2], we train one logistic regression model per candidate expansion collection $C_i$ and select the collection whose classifier returns the highest confidence, $P_{C_i}(Y = 1|D)$. This is a natural approach for collection selection, because each feature is calculated with respect to an individual collection. While other approaches exist (e.g. [103]), this per-collection classifier approach is also beneficial in that it allows us to switch seamlessly to predicting whether or not to expand a given document using a single collection, which may be useful when multiple expansion collections are unavailable.

In contrast to a typical multiclass problem, in which the label of a given instance is always one of the $k$ classes, this approach requires us to label documents using binary classes for each of the candidate collections. Our class labels should naturally indicate whether expansion with a given collection is desirable or not. One simple solution is to compare the document’s expanded retrieval score against its baseline score: when the expanded score is preferred, the label is positive; when the baseline score is preferred, the label is negative. Although this approach makes strong assumptions about the nature of score changes in expanded documents, we are reassured by the oracle runs that this approach may nevertheless yield substantial retrieval effectiveness improvements. We therefore employ this method of labeling...
in the following work.

6.3 Features

Features are intended to measure the “appropriateness” of a collection as a source of data to use in expanding a document. While many possible features exist, computational efficiency is a factor, since features must be computed for every document retrieved. Features that involve sums or products over the entire vocabulary or over all terms in one or more documents are not considered.

6.3.1 Query-independent features

Query-independent features are desirable for their efficiency, because they can be computed at indexing time. In the case of this task, these features generally relate the document or its pseudo-query to the candidate expansion collection. All but the final feature in this section are established query performance predictors that have been shown to correlate with retrieval effectiveness. Because retrieval effectiveness in this case corresponds to retrieval of appropriate expansion documents, these predictors are a logical tool for our purposes.

Pseudo-query clarity

Pseudo-query clarity is the simple clarity score (SCS) [112] of the pseudo-query against the expansion collection. SCS was explored as a measure of pseudo-query quality in Section 5.4.2. SCS is defined as:

\[ SCS(Q) = \sum_{w \in Q} P(w|Q) \log \frac{P(w|Q)}{P(w|C)}. \]

\( P(w|Q) \) is simply the relative weight of term \( w \) in pseudo-query \( Q \).
Weighted Information Gain

Like SCS, weighted information gain (WIG) [112] was explored in Section 5.4.2 as a measure of pseudo-query quality. It is defined as:

$$WIG(Q) = \frac{1}{k} \sum_{D \in F} \sum_{w \in Q} \lambda(w) \log \frac{P(w|D)}{P(w|C)}.$$ 

where $F$ is the set of top-ranked expansion documents for pseudo-query $Q$ and $\lambda(w) = \frac{1}{\sqrt{|Q|}}$.

Average IDF

A simpler metric is average inverse document frequency [70]. Since IDF is intuitively a measure of a term’s ability to differentiate relevant from nonrelevant documents, a pseudo-query with higher average IDF is one that is able, on average, to differentiate appropriate expansion documents from inappropriate ones.

Pseudo-query/expansion document similarity

We also compute the cosine similarity between the pseudo-query and the pseudo-document formed by concatenating the expansion documents. This metric is intended to measure the extent to which the expansion documents remain on-topic, where the topic is that expressed by the document pseudo-query.

6.3.2 Query-dependent features

Query-dependent features are more costly to compute than query-independent features, since they must be computed at query time. However, since our definition of expansion collection appropriateness (i.e., the class label) is derived from the query likelihood score of the expanded document and query relevance, it is reasonable to expect that information about the query will improve classifier success.
Figure 6.1: Boxplots comparing the Jaccard similarity of query results and expansion documents for nonrelevant and relevant documents. The figure shows that, although disjoint sets of expansion documents and query results are most common for both nonrelevant and relevant documents, relevant documents are much more likely to achieve a higher degree of overlap.

**Jaccard similarity of query results and expansion documents**

The labels our classifiers are attempting to predict are a factor of score and relevance. We suggest that a comparison between query results (from the candidate expansion collection) and expansion documents is one way to provide our classifiers with information that may indicate relevance.

Following from Section 5.4.1, in which we compared topic term results against pseudo-query results, we here compute the Jaccard similarity between query results and pseudo-query results (i.e., expansion documents).

A two-tailed $t$-test comparing relevant and nonrelevant documents suggests that our motivating intuition for this metric is correct: the mean Jaccard similarity of nonrelevant documents, 0.0109, is less than that of relevant documents, 0.0662, with statistical significance ($p < 0.01$). This difference is also visible in Figure 6.1. The figure shows that, although most expansion document sets do not overlap with query results, the expansion documents of relevant target documents are more likely to achieve some overlap.
Relevance model similarity to expansion documents

In a similar vein to the above metric, we can compute the cosine similarity between the relevance model (RM) [51] computed for the query and the expansion documents’ pseudo-document. While computing a unique relevance model for each candidate expansion collection would be possible, and perhaps desirable, it would also be extremely costly. An RM computed on the target collection should adequately capture the query language model and ensures that each expansion pseudo-document is compared against the same query representation, simplifying comparison. We compute RM1 models from ten documents and truncate to the top 20 terms.

As before, a two-tailed t-test suggests that cosine similarity is greater on average for relevant documents (0.3440) than nonrelevant documents (0.1677), with statistical significance ($p < 0.01$). The similarities of relevant and nonrelevant documents are visualized in Figure 6.2.
Collection query likelihood

We also consider the query likelihood of each expansion collection. This metric is independent of the document and its pseudo-query. Instead, it seeks to capture the intuition that some collections may be more appropriate for a query in general. We calculate collection QL using add-one smoothing to prevent zero-counts of terms.

6.4 Data and evaluation

For each target collection, we build five individual logistic regression models: one per candidate expansion collection. The target collections under consideration are those used in Chapter 4 (TREC collections AP, GOV2, Robust, and WT10g) to allow for direct comparison of retrieval effectiveness. As in Chapter 4, each of these collections, as well as Wikipedia, is also considered as a candidate expansion collection for each of the target collections.

To evaluate classifier success, we measure not only accuracy but, more importantly, the mean average precision (MAP) and normalized discounted cumulative gain at rank 20 (NDCG@20) of query results scored using the predicted expansion collections. Whereas accuracy can be computed with any random sample of test data, evaluating based on MAP and NDCG@20 requires us to form tests sets cohesively. We therefore use grouped 10-fold cross validation for model evaluation. Documents are grouped according to query, so that each target collection is comprised of \( n \) groups where \( n \) is the number of TREC queries for that collection. Grouped cross validation ensures that members of a group are not split across train and test sets; in our case, this ensures that all documents for a given query are treated as part of the same hold-out validation fold, enabling straightforward MAP and NDCG@20 calculation.

For example, we consider 100 queries with respect to the AP collection. Grouping by query, 10-fold cross validation tests ten queries per fold. Because the test queries are complete (i.e., all of the documents constituting the query are part of the test fold), we can compute
MAP and NDCG@20 across these ten queries. Each query appears in only one test fold. We report the average MAP and NDCG@20 across folds in Section 6.5 below.

Each logistic regression model predicts whether a given document is optimally expanded (with a particular expansion collection) or left unmodified. We therefore evaluate accuracy, MAP, and NDCG@20 for each of the individual “binary” models. We also evaluate these metrics using the “multiclass” approach, in which one of the five expansion collections is chosen based on the model with the highest confidence.

6.5 Results

<table>
<thead>
<tr>
<th>Expansion collection</th>
<th>Method</th>
<th>Accuracy</th>
<th>MAP</th>
<th>NDCG@20</th>
</tr>
</thead>
<tbody>
<tr>
<td>AP</td>
<td>Baseline</td>
<td>0.2428</td>
<td>0.2337</td>
<td>0.4170</td>
</tr>
<tr>
<td></td>
<td>Uniform</td>
<td>0.7572</td>
<td>0.2813</td>
<td>0.4567</td>
</tr>
<tr>
<td></td>
<td>Predicted</td>
<td>0.7882</td>
<td>0.2558</td>
<td>0.4486</td>
</tr>
<tr>
<td></td>
<td>Predicted.weight</td>
<td>0.4978</td>
<td>0.2550</td>
<td>0.4430</td>
</tr>
<tr>
<td>GOV2</td>
<td>Baseline</td>
<td>0.5170</td>
<td>0.2337</td>
<td>0.4170</td>
</tr>
<tr>
<td></td>
<td>Uniform</td>
<td>0.4830</td>
<td>0.2716</td>
<td>0.4512</td>
</tr>
<tr>
<td></td>
<td>Predicted</td>
<td>0.6372</td>
<td>0.2364</td>
<td>0.4256</td>
</tr>
<tr>
<td></td>
<td>Predicted.weight</td>
<td>0.4372</td>
<td>0.2573</td>
<td>0.4469</td>
</tr>
<tr>
<td>Robust</td>
<td>Baseline</td>
<td>0.4309</td>
<td>0.2337</td>
<td>0.4170</td>
</tr>
<tr>
<td></td>
<td>Uniform</td>
<td>0.5691</td>
<td>0.2764</td>
<td>0.4570</td>
</tr>
<tr>
<td></td>
<td>Predicted</td>
<td>0.6691</td>
<td>0.2509</td>
<td>0.4453</td>
</tr>
<tr>
<td></td>
<td>Predicted.weight</td>
<td>0.4672</td>
<td>0.2454</td>
<td>0.4309</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>Baseline</td>
<td>0.4129</td>
<td>0.2337</td>
<td>0.4170</td>
</tr>
<tr>
<td></td>
<td>Uniform</td>
<td>0.5871</td>
<td>0.2753</td>
<td>0.4783</td>
</tr>
<tr>
<td></td>
<td>Predicted</td>
<td>0.6863</td>
<td>0.2363</td>
<td>0.4348</td>
</tr>
<tr>
<td></td>
<td>Predicted.weight</td>
<td>0.4698</td>
<td>0.2580</td>
<td>0.4455</td>
</tr>
<tr>
<td>WT10g</td>
<td>Baseline</td>
<td>0.4774</td>
<td>0.2337</td>
<td>0.4170</td>
</tr>
<tr>
<td></td>
<td>Uniform</td>
<td>0.5226</td>
<td>0.2709</td>
<td>0.4529</td>
</tr>
<tr>
<td></td>
<td>Predicted</td>
<td>0.6453</td>
<td>0.2405</td>
<td>0.4282</td>
</tr>
<tr>
<td></td>
<td>Predicted.weight</td>
<td>0.4686</td>
<td>0.2546</td>
<td>0.4374</td>
</tr>
</tbody>
</table>

Table 6.4: Selective document expansion results for the AP collection.

Results for the various models under consideration are shown in Tables 6.4–6.8.

Tables 6.4–6.7 report the results of binary classification models deciding whether each document should be expanded with the specified expansion collection or left unmodified.
<table>
<thead>
<tr>
<th>Expansion collection</th>
<th>Method</th>
<th>Accuracy</th>
<th>MAP</th>
<th>NDCG@20</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Baseline</td>
<td>0.2986</td>
<td>0.2697</td>
<td>0.4134</td>
</tr>
<tr>
<td></td>
<td>Uniform</td>
<td>0.7014</td>
<td>0.2720</td>
<td>0.4137</td>
</tr>
<tr>
<td></td>
<td>Predicted</td>
<td>0.7044</td>
<td>0.2712</td>
<td>0.4125</td>
</tr>
<tr>
<td></td>
<td>Predicted.weight</td>
<td>0.4571</td>
<td>0.2698</td>
<td>0.4093</td>
</tr>
<tr>
<td>GOV2</td>
<td>Baseline</td>
<td>0.3424</td>
<td>0.2697</td>
<td>0.4134</td>
</tr>
<tr>
<td></td>
<td>Uniform</td>
<td>0.6576</td>
<td>0.2797</td>
<td>0.4236</td>
</tr>
<tr>
<td></td>
<td>Predicted</td>
<td>0.7029</td>
<td>0.2717</td>
<td>0.4154</td>
</tr>
<tr>
<td></td>
<td>Predicted.weight</td>
<td>0.4147</td>
<td>0.2671</td>
<td>0.4091</td>
</tr>
<tr>
<td>Robust</td>
<td>Baseline</td>
<td>0.2782</td>
<td>0.2697</td>
<td>0.4134</td>
</tr>
<tr>
<td></td>
<td>Uniform</td>
<td>0.7218</td>
<td>0.2718</td>
<td>0.4117</td>
</tr>
<tr>
<td></td>
<td>Predicted</td>
<td>0.7235</td>
<td>0.2706</td>
<td>0.4123</td>
</tr>
<tr>
<td></td>
<td>Predicted.weight</td>
<td>0.5035</td>
<td>0.2699</td>
<td>0.4095</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>Baseline</td>
<td>0.3405</td>
<td>0.2697</td>
<td>0.4134</td>
</tr>
<tr>
<td></td>
<td>Uniform</td>
<td>0.6595</td>
<td>0.2749</td>
<td>0.4170</td>
</tr>
<tr>
<td></td>
<td>Predicted</td>
<td>0.6891</td>
<td>0.2586</td>
<td>0.4123</td>
</tr>
<tr>
<td></td>
<td>Predicted.weight</td>
<td>0.4469</td>
<td>0.2615</td>
<td>0.4052</td>
</tr>
<tr>
<td>WT10g</td>
<td>Baseline</td>
<td>0.3236</td>
<td>0.2697</td>
<td>0.4134</td>
</tr>
<tr>
<td></td>
<td>Uniform</td>
<td>0.6764</td>
<td>0.2773</td>
<td>0.4124</td>
</tr>
<tr>
<td></td>
<td>Predicted</td>
<td>0.6948</td>
<td>0.2699</td>
<td>0.4142</td>
</tr>
<tr>
<td></td>
<td>Predicted.weight</td>
<td>0.4134</td>
<td>0.2685</td>
<td>0.3987</td>
</tr>
</tbody>
</table>

Table 6.5: Selective document expansion results for the GOV2 collection.

Several types of runs are reported for each candidate expansion collection:

- **Baseline**: Runs in which no documents are expanded, i.e., query likelihood runs. Note that Baseline MAP and NDCG@20 are always the same, but accuracy changes by expansion collection because classifier labels vary depending on the scores of documents expanded with each collection.

- **Uniform**: Runs in which all documents are expanded with the given expansion collection. These correspond to runs reported in Table 4.2 in Chapter 4.

- **Predicted**: The results of expanding documents according to model predictions. Accuracy, MAP, and NDCG@20 are computed for the hold-out set in each fold of cross validation and are then averaged across folds to produce the values reported here.

- **Predicted.weight**: The same as predicted, but with training instances weighted in the model’s cost function. These weights are based on a combination of relevance (since
Table 6.6: Selective document expansion results for the Robust collection.

so few documents in each collection are relevant, models tend to favor nonrelevant documents; see Section 6.6) and score change (within relevance groups, the direction of score changes are imbalanced; e.g. documents of a certain relevance may overwhelmingly benefit from expansion, in which case models may incorrectly learn to expand all documents of that relevance). Documents are weighted inversely to the proportion of the data comprised by their particular relevance/score change combination.

Table 6.8 gives the results of multiclass selective expansion. In this case, we select the expansion collection whose corresponding model reports the highest confidence among candidate expansion collections. For example, if our target collection is AP, for a given document we consider all five models reported in Table 6.4. The collection corresponding to the model with the highest confidence is selected for that document. Predicted and predicted.weight correspond to the models described above. Maximum uniform refers to the
<table>
<thead>
<tr>
<th>Expansion collection</th>
<th>Method</th>
<th>Accuracy</th>
<th>MAP</th>
<th>NDCG@20</th>
</tr>
</thead>
<tbody>
<tr>
<td>AP</td>
<td>Baseline</td>
<td>0.1718</td>
<td>0.1683</td>
<td>0.2738</td>
</tr>
<tr>
<td></td>
<td>Uniform</td>
<td>0.8282</td>
<td>0.1728</td>
<td>0.2853</td>
</tr>
<tr>
<td></td>
<td>Predicted</td>
<td>0.8315</td>
<td>0.1730</td>
<td>0.2935</td>
</tr>
<tr>
<td></td>
<td>Predicted.weight</td>
<td>0.4708</td>
<td>0.1716</td>
<td>0.2825</td>
</tr>
<tr>
<td>GOV2</td>
<td>Baseline</td>
<td>0.2433</td>
<td>0.1683</td>
<td>0.2738</td>
</tr>
<tr>
<td></td>
<td>Uniform</td>
<td>0.7567</td>
<td>0.1777</td>
<td>0.3087</td>
</tr>
<tr>
<td></td>
<td>Predicted</td>
<td>0.7726</td>
<td>0.1683</td>
<td>0.2841</td>
</tr>
<tr>
<td></td>
<td>Predicted.weight</td>
<td>0.5145</td>
<td>0.1766</td>
<td>0.2938</td>
</tr>
<tr>
<td>Robust</td>
<td>Baseline</td>
<td>0.1543</td>
<td>0.1683</td>
<td>0.2738</td>
</tr>
<tr>
<td></td>
<td>Uniform</td>
<td>0.8457</td>
<td>0.1755</td>
<td>0.2962</td>
</tr>
<tr>
<td></td>
<td>Predicted</td>
<td>0.8463</td>
<td>0.1766</td>
<td>0.2947</td>
</tr>
<tr>
<td></td>
<td>Predicted.weight</td>
<td>0.6433</td>
<td>0.1734</td>
<td>0.2837</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>Baseline</td>
<td>0.2423</td>
<td>0.1683</td>
<td>0.2738</td>
</tr>
<tr>
<td></td>
<td>Uniform</td>
<td>0.7577</td>
<td>0.1840</td>
<td>0.3110</td>
</tr>
<tr>
<td></td>
<td>Predicted</td>
<td>0.7950</td>
<td>0.1742</td>
<td>0.2946</td>
</tr>
<tr>
<td></td>
<td>Predicted.weight</td>
<td>0.4854</td>
<td>0.1769</td>
<td>0.2914</td>
</tr>
<tr>
<td>WT10g</td>
<td>Baseline</td>
<td>0.2671</td>
<td>0.1683</td>
<td>0.2738</td>
</tr>
<tr>
<td></td>
<td>Uniform</td>
<td>0.7329</td>
<td>0.1685</td>
<td>0.2965</td>
</tr>
<tr>
<td></td>
<td>Predicted</td>
<td>0.7710</td>
<td>0.1708</td>
<td>0.2811</td>
</tr>
<tr>
<td></td>
<td>Predicted.weight</td>
<td>0.5541</td>
<td>0.1678</td>
<td>0.2814</td>
</tr>
</tbody>
</table>

Table 6.7: Selective document expansion results for the WT10g collection.

greatest values of MAP and NDCG@20 achievable with uniform expansion, i.e., expanding all documents with a single collection.

The binary prediction results show that, while it is fairly common for selective expansion runs to outperform unexpanded baselines, they do not outperform uniform expansion runs. (A few runs, such as MAP of predicted for WT10g expanded with itself, show apparent slight improvement over the uniform run; however, these are generally an artifact of averaging and do not reflect true improvement.) There is no apparent pattern of difference between unweighted and weighted models when it comes to retrieval effectiveness, although accuracy is always substantially lower for weighted runs than unweighted runs. This is discussed further in Section 6.6. Without the ability to beat uniform expansion, it is difficult to justify building these models.

The multiclass prediction results are similarly negative, although as a rule they achieve
<table>
<thead>
<tr>
<th>Target collection</th>
<th>Method</th>
<th>Accuracy</th>
<th>MAP</th>
<th>NDCG@20</th>
</tr>
</thead>
<tbody>
<tr>
<td>AP</td>
<td>Maximum uniform</td>
<td>0.4631</td>
<td>0.2813</td>
<td>0.4783</td>
</tr>
<tr>
<td></td>
<td>Predict</td>
<td>0.4415</td>
<td>0.2767</td>
<td>0.4724</td>
</tr>
<tr>
<td></td>
<td>Predict.weight</td>
<td>0.4014</td>
<td>0.2825</td>
<td>0.4816</td>
</tr>
<tr>
<td>GOV2</td>
<td>Maximum uniform</td>
<td>0.1864</td>
<td>0.2797</td>
<td>0.4236</td>
</tr>
<tr>
<td></td>
<td>Predict</td>
<td>0.1654</td>
<td>0.2693</td>
<td>0.4114</td>
</tr>
<tr>
<td></td>
<td>Predict.weight</td>
<td>0.1818</td>
<td>0.2772</td>
<td>0.4233</td>
</tr>
<tr>
<td>Robust</td>
<td>Maximum uniform</td>
<td>0.1100</td>
<td>0.2460</td>
<td>0.4163</td>
</tr>
<tr>
<td></td>
<td>Predict</td>
<td>0.4453</td>
<td>0.2297</td>
<td>0.3907</td>
</tr>
<tr>
<td></td>
<td>Predict.weight</td>
<td>0.4074</td>
<td>0.2471</td>
<td>0.4119</td>
</tr>
<tr>
<td>WT10g</td>
<td>Maximum uniform</td>
<td>0.4549</td>
<td>0.1840</td>
<td>0.3110</td>
</tr>
<tr>
<td></td>
<td>Predict</td>
<td>0.3115</td>
<td>0.1722</td>
<td>0.2820</td>
</tr>
<tr>
<td></td>
<td>Predict.weight</td>
<td>0.3350</td>
<td>0.1736</td>
<td>0.3007</td>
</tr>
</tbody>
</table>

Table 6.8: Selective document expansion results for multiclass models. For each target collection, each model predicts the optimal choice of expansion collection from among the five candidate collections (the four TREC collections plus Wikipedia). “Maximum uniform” refers to the highest MAP or NDCG@20 achieved using uniform expansion. Accuracy of “maximum uniform” runs is based on MAP.

evaluation scores closer to the uniform runs. We also note that, in contrast to the binary prediction runs, weighted models consistently achieve higher retrieval effectiveness scores.

### 6.6 Discussion

The question of weighting instances in the logistic regression cost function is one that reveals a great deal about the inner workings of these models. Table 6.9 shows the accuracies of unweighted and weighted models with Wikipedia as the expansion collection. (Throughout this section, we will compare against only Wikipedia for simplicity of discussion; excluded analysis has shown that the conclusions based on Wikipedia hold true for other candidate expansion collections.)

As the table shows, unweighted models achieve much greater accuracy on nonrelevant documents than on relevant ones. This is unsurprising, since only around 5–10% of the classified documents are relevant for each collection. This low proportion of relevant documents means that models can achieve high accuracy by simply learning to treat all documents as
Table 6.9: The accuracies of relevant and nonrelevant documents under the unweighted and weighted models, with the unexpanded baseline and uniform expansion for comparison. Only runs with Wikipedia as the expansion collection are considered in this table.

though they are nonrelevant. By weighting document instances in the logistic regression cost function inversely to the overall proportion of documents with their particular relevance and score change direction, the models learn to do a better job at predicting relevant documents—at a cost to the accuracy of nonrelevant documents.

Importantly, when we compare Table 6.9 against Tables 6.4–6.7, we can see that the trade-off made by weighting—decreased accuracy on nonrelevant documents, in exchange for increased accuracy on relevant documents—generally leaves retrieval effectiveness approximately equal. However, as Tables 6.4–6.7 show, overall model accuracy across all documents is consistently and substantially lower for weighted models than for unweighted models. These findings indicate that increasing model accuracy on relevant documents is a higher priority for retrieval effectiveness than high accuracy on nonrelevant documents. This is sensible, since retrieval effectiveness scores like MAP and NDCG@20 are primarily concerned with relevant documents, and nonrelevant documents matter only so far as they “get in the way” of appropriately ranking relevant ones.
The accuracy of a uniform expansion run can also be interpreted as the proportion of documents whose scores are optimized by expansion; the accuracy of an unexpanded baseline run gives the opposite proportion. In this light, we can see that the relevant documents in AP and Robust most often benefit from expansion, whereas relevant documents in GOV2 and WT10g are usually better off unexpanded. The majority of nonrelevant documents in all four collections benefit from expansion, but GOV2 and WT10g’s nonrelevant documents are particularly prone to improvement. Given the overall harm done to relevant documents in these two collections, any retrieval effectiveness gains are likely the result of the strong overall benefit to nonrelevant documents. However, it is notable that the most consistent and strongest retrieval effectiveness improvements found in Chapter 4 were for AP and Robust, further supporting the idea that accurate expansion of relevant documents is far more important to MAP and NDCG@20 than that of nonrelevant documents.

The high accuracies achieved by uniform expansion runs demonstrates their unexpected strength as a baseline. Despite the high potential for retrieval improvement under oracle conditions, uniform expansion is simply the right choice for the large majority of documents and is therefore extremely effective overall.

<table>
<thead>
<tr>
<th>Target collection</th>
<th>Uniform MAP</th>
<th>NDCG@20</th>
<th>Rel. only MAP</th>
<th>NDCG@20</th>
<th>Nonrel. only MAP</th>
<th>NDCG@20</th>
</tr>
</thead>
<tbody>
<tr>
<td>AP</td>
<td>0.2813</td>
<td>0.4567</td>
<td>0.2696</td>
<td>0.4570</td>
<td>0.2498</td>
<td>0.4430</td>
</tr>
<tr>
<td>GOV2</td>
<td>0.2749</td>
<td>0.4170</td>
<td>0.2730</td>
<td>0.4113</td>
<td>0.2759</td>
<td>0.4299</td>
</tr>
<tr>
<td>Robust</td>
<td>0.2409</td>
<td>0.4043</td>
<td>0.2399</td>
<td>0.4053</td>
<td>0.2254</td>
<td>0.3923</td>
</tr>
<tr>
<td>WT10g</td>
<td>0.1840</td>
<td>0.3110</td>
<td>0.1753</td>
<td>0.2955</td>
<td>0.1887</td>
<td>0.3158</td>
</tr>
</tbody>
</table>

Table 6.10: Retrieval effectiveness of runs in which only the relevant or nonrelevant documents are expanded correctly, with all other documents expanded incorrectly. Uniform expansion is provided for comparison. Only runs expanding with Wikipedia are considered.

Despite evidence that relevant document accuracy should be prioritized over nonrelevant document accuracy, it remains important to achieve a reasonable degree of accuracy on both. Table 6.10 shows the retrieval effectiveness achievable with 100% accuracy on either relevant or nonrelevant documents and 0% accuracy on the other. Although very mild raw
improvement over uniform expansion is occasionally possible, the results show that in general meaningful increases to retrieval effectiveness are possible only by accurately predicting both relevant and nonrelevant documents. We therefore conclude that, although relevant document accuracy is central to retrieval effectiveness, nonrelevant document accuracy cannot be ignored.

Unfortunately, the discussion so far has indicated that models have difficulty accurately predicting expansion for both relevant and nonrelevant documents at the same time. This difficulty is highlighted by the following test. Since class labels in Section 6.5 are a linear combination of relevance and score direction, providing either factor as a feature reduces the model’s task to predicting the other, which is what we examine here. We train two new types of models using the same features as before. However, one model, called “known relevance,” is additionally provided document relevance as a feature and is asked to predict the direction of score change (i.e., whether the document’s score will increase or decrease from expansion); the other model, called “known score,” is given score direction as a feature and asked to predict relevance. Training instances in each model are weighted inversely to their class proportion to counteract imbalance. Given the known information and the predicted value, we can make the expansion decision. For example, if a document is known to be relevant and predicted to increase in score, the decision is to expand the document. The results are shown in Table 6.11.

<table>
<thead>
<tr>
<th>Target collection</th>
<th>Known Relevance</th>
<th>Known Score Change</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Accuracy</td>
<td>MAP</td>
</tr>
<tr>
<td>AP</td>
<td>0.7337</td>
<td>0.3763</td>
</tr>
<tr>
<td>GOV2</td>
<td>0.6745</td>
<td>0.3242</td>
</tr>
<tr>
<td>Robust</td>
<td>0.6829</td>
<td>0.3017</td>
</tr>
<tr>
<td>WT10g</td>
<td>0.7759</td>
<td>0.2184</td>
</tr>
</tbody>
</table>

Table 6.11: Results of models that are given either relevance or score as a feature and asked to predict the other. The predicted quantity is then used in conjunction with the given quantity to formulate the expansion decision. The results show that knowing the relevance of the document was much more valuable to model performance than knowing the score.
Retrieval effectiveness of the models with known relevance tends to exceed—often very substantially—that of corresponding models with known score. In fact, most known relevance runs strongly outperform uniform expansion, whereas most known score runs achieve retrieval effectiveness scores approximately on par with, or slightly lower than, those of uniform expansion. Since relevance and score change are equal components in the class label, we may have expected these models to perform similarly. The results show, however, that when the prediction task is reduced to predicting score direction, models are much more successful than when they must predict relevance. This implies that relevance is the limiting factor to model success.

Importantly, while relevance is largely unobservable (except in limited cases, such as explicit relevance feedback), score change is entirely observable. That is, disregarding computational cost, the system could expand each document and observe the effect of expansion on the document’s score. The models with known score presented in Table 6.11 are therefore actually a realistic approach one might take to accomplish selective expansion; unfortunately, they are unsuccessful. The following section puts forward a theoretical argument to match these experimental results. It argues that, due to the inclusion of relevance in the determination of class labels, per-document selective expansion is an unrealistic and, indeed, fundamentally unsound task.

6.7 Problems with the task of selective document expansion

Our approach so far has been to train a classification algorithm to make a prediction about the effect of expansion, directly, on the score of a document but, indirectly, on the overall retrieval effectiveness metric. That is, we have modeled the decision to expand a document as a label that can be predicted: we decide to expand a document if our classifier predicts that doing so would improve retrieval effectiveness, and we leave the document unexpanded if our classifier predicts that expanding it would damage retrieval effectiveness.

Since standard retrieval effectiveness metrics, including average precision and normalized
discounted cumulative gain, are based on the ranks of relevant documents, any change in retrieval effectiveness metric is equivalent to a change in document rank. Therefore, for expansion of a relevant document to improve the overall retrieval effectiveness, it must increase the rank of that document compared to the unexpanded run; for expansion of a nonrelevant document to improve the overall retrieval effectiveness, it must decrease the rank of that document compared to the unexpanded run. These two outcomes are considered desirable, while the inverse outcomes (decreased rank of a relevant document, increased rank of a nonrelevant document) are considered undesirable. A classifier that predicts whether expansion will improve retrieval effectiveness is one that predicts whether expansion is “desirable” under this schema.

In the experimental work presented in this chapter, we have substituted document retrieval score in place of document rank. This is because the score of a given document is largely independent of the scores of all other documents, particularly in comparison to rank. While this makes score a more consistent basis for prediction (since the model need not consider interdependencies among documents in making a prediction), it is fundamentally a substitute for rank in much the same way that retrieval score is a substitute for rank—theoretically, a retrieval function could directly assign ranks, rather than scores, to documents it retrieves; however, to do so would require simultaneous consideration of all documents with respect to the query, whereas assigning scores can be done individually. Nevertheless, it is ranks, and not scores, that both retrieval models and selective expansion models intend to optimize.

Although our experiments have been unsuccessful, imagine that we had trained a classification function to predict the desirability of expansion with a reasonably high rate of success. Say that we observe a document at rank five in the baseline run, and our classifier predicts that expansion of this document is desirable, so we proceed with expanding the document. We observe that, as a result of expansion, the document is now at rank three.

From observation, we know that the document’s rank has increased. From our classifier,
we know (with high probability) that the effect of expansion was desirable. Since we know
that increased rank is only desirable when the document is relevant, we can infer that the
document must be relevant (again, with high probability). This conclusion holds true for
score-based classifiers, too: if a highly accurate classifier predicts that expansion is desirable,
and we observe that the document’s score increases as a result of expansion, we can infer
that the document must be relevant.

The predicted variable, desirability, is comprised of two factors: relevance and rank
change (or score). Since we are able to observe one of these two factors (rank change, or
score), we can deduce the other (relevance). Put another way, this problem could be re-
formulated, as in the previous section, to include rank change as a feature in the classifier,
which essentially leaves relevance as the label to predict. Predicting a property correlated
with desirability of expansion would not alleviate the problem, either, since accurate predic-
tion of the correlated property would be equivalent to predicting desirability and therefore
suffers from the problem of having to predict relevance.

While predicting relevance is a laudable goal—it is in some sense the goal of IR generally—
it is a much broader and more difficult task than we intended to accomplish with the selective
document expansion component of our model. It also reverses the proper order of the
document expansion process: the aim was to predict whether the expanded or unexpanded
document language model would be more accurate for the retrieval model to use in assessing
relevance, but we now see that deciding between the expanded and unexpanded language
models requires us to first determine the relevance.

6.7.1 Generalizing from document expansion

Given a query \( Q \) and a collection of documents \( C \), we can compute for each document
\( D \in C \) a retrieval score \( s(D, Q) \). Ranking documents by \( s(D, Q) \), we get the results list
\( R_Q = (D_1, D_2, ..., D_n) \), which we can use to compute an evaluation metric \( M(R_Q, Q) \).

We can optionally substitute a given document \( D \) with some modified form of the doc-
ument, $D'$, yielding results list $R^{(D')}_Q$. We express the relevance of $D$ with respect to $Q$ as $\text{rel}(D, Q)$, assuming a $\text{rel}$ of zero for nonrelevant documents. Importantly, $\text{rel}(D', Q) = \text{rel}(D, Q)$ but $s(D', Q)$ does not necessarily equal $s(D, Q)$. In the case of document expansion, $D'$ is the expanded form of the document, but any substitution of the document is valid. To optimize the overall retrieval effectiveness, we should substitute $D'$ for $D$ whenever $\text{sign}[M(R^{(D')}_Q, Q) - M(R_Q, Q)] = 1$ (assuming larger $M$ is optimal; this discussion applies equally to cases where smaller $M$ is preferred, but we will assume that greater $M$ is better). This is the notion of “desirability” described above, referring to the relative improvement of the results list.

Robertson & Zaragoza show that the retrieval metric $M$ may respond to any individual document rank “flip” in which a document changes place with its neighbor [76]. Though they are concerned with finding optimal parameter settings, their discussion of $M$’s reaction to an individual flip holds in the case described here: $M$ may respond to an individual flip, and we know that in our case the flip must have been caused not by movement through parameter space but by the substitution of $D$, since it is the only difference between $R^{(D')}_Q$ and $R_Q$.

Following from [76], the effect of a flip between documents $D$ and $\bar{D}$ can be calculated as follows:

$$g(D, \bar{D}) = \text{sign}[s(D, Q) - s(\bar{D}, Q)] \times \text{sign}[\text{rel}(D, Q) - \text{rel}(\bar{D}, Q)] \quad (6.1)$$

where the flip is, in Robertson & Zaragoza’s terms, “good” (i.e., improves $M$) when $g(D, \bar{D}) = 1$, “bad” (i.e., penalizes $M$) when $g(D, \bar{D}) = -1$, and “neutral” (i.e., does not affect $M$) when $g(D, \bar{D}) = 0$. For clarity, “improves $M$” refers to the case where $\text{sign}[M(R^{(2)}_Q, Q) - M(R^{(1)}_Q, Q)] = 1$, where $R^{(1)}_Q$ is the initial results list and $R^{(2)}_Q$ is the results list after the flip. For a single flip, therefore, $\text{sign}[M(R^{(2)}_Q, Q) - M(R^{(1)}_Q, Q)] = g(D, \bar{D})$.

Say that our substitution of $D'$ for $D$ induces a single flip between $D'$ and some other
document $\bar{D}$. Our document substitution decision is therefore reduced to predicting $g(D', \bar{D})$. Since $s(D', Q)$ and $s(\bar{D}, Q)$ are known, Eq. 6.1 becomes:

$$
\begin{align*}
g(D', \bar{D}) &= c \times \text{sign}[\text{rel}(D', Q) - \text{rel}(\bar{D}, Q)] \\
&= c \times \text{sign}\left[\sum_{D \in \bar{D}} \text{rel}(D, Q) - \sum_{D \in \bar{D}} \text{rel}(\bar{D}, Q)\right] \\
\end{align*}
$$

(6.2)

where $c \in \{-1, 0, 1\}$ is a known constant equaling $\text{sign}[s(D', Q) - s(\bar{D}, Q)]$. Predicting $g(D', \bar{D})$ is therefore equivalent to predicting $\text{sign}[\text{rel}(D', Q) - \text{rel}(\bar{D}, Q)]$, the relative relevance of $D'$ and the document it trades places with, $\bar{D}$.

While the preceding discussion is concerned with a flip of neighboring documents, document substitution can cause a document to “leap” more than one rank. When this occurs, the sublist of documents $\mathcal{D}$ from the former rank of $D$ to the rank of $D'$ will shift by one rank in the opposite direction of $D$’s leap. This constitutes a sort of flip between $D$ and $\bar{D}$ that can complicate Eq 6.2. For example, when $M$ is average precision, $\text{rel}(\bar{D}, Q) = \text{sign}\left[\sum_{D \in \bar{D}} \text{rel}(D, Q)\right]$. This quantity is more complex than that of a single flip of neighboring documents. However, our conclusion remains the same: $\text{sign}[s(D', Q) - s(\bar{D}, Q)]$ is known (since $s(D', Q)$ is either greater than or less than all $s(\bar{D}, Q)$ for $\bar{D} \in \mathcal{D}$), so predicting $g(D', \bar{D})$ is equivalent to predicting $D$’s relevance relative to $\bar{D}$.

Note that the sole result of substituting $D'$ for $D$ is the potential change in the value of scoring function $s$—all other consequences follow from this change. It is therefore the case that, for a constant $D$, any substitution of the scoring function $s$ with some more optimal function $s'$ also requires us to predict the relative relevance of the documents that flip as a result of the substitution. Indeed, viewing document substitution as a form of

*Average precision calculates precision values at each level of recall, which are then averaged. If a flip or “leap” results in higher (or lower) precision at the maximum rank affected, it is an indication that average precision has increased (or decreased) overall. The definition of $\text{rel}(\bar{D}, Q)$ above defines the precision of $\bar{D}$, against which we can compare the relevance of $D'$. When $D'$ is relevant and $\text{Prec}(\bar{D}) < 1$ (i.e., precision can be increased), average precision is affected by the rank change of $D'$ (it increases if $s(D', Q) > s(\bar{D}, Q)$ and decreases otherwise). The inverse is true when $D'$ is nonrelevant and $\text{Prec}(\bar{D}) > 0$. This can be proven by comparing the precision at any recall in $\bar{D}$ before and after the rank change of $D'$. For example, if $D'$ is relevant and its rank increases, the precision at each level of recall is always at least as large as prior to the rank change: if precision at a given relevant document $D_{Rel}$ is $\text{Prec}(D_{Rel}) = x/y$ prior to the rank change, it becomes $\text{Prec}'(D_{Rel}) = (1 + x)/(1 + y)$ following $D'$’s promotion (this promotion is the cause of the additional count in both the numerator and denominator). Since $y \geq x$, $\text{Prec}' \geq \text{Prec}$ is guaranteed.
scoring function substitution helps reveal its unsound premise, since documents should not reasonably be ranked against one another on the basis of differing scoring functions (e.g., it would not make sense to rank documents scored with query likelihood against those scored with BM25 unless some sort of rank fusion is used [99, 11, 46, 43, 9]).

We therefore see that attempts to selectively optimize the representation of $D$ or the scoring function $s$ applied to $D$ require us to directly predict the relative relevance of $D$ and any document with which it changes rank. Given this conclusion, the negative experimental results reported in Section 6.5 are unsurprising.

6.8 Conclusions

Although selective document expansion has proven unsuccessful, the work presented in this chapter has revealed useful insights. Specifically, this work has contributed:

- Specific results for the novel task of selective document expansion
- Detailed insights into the success of the uniform document expansion runs first presented in Chapter 4 relating to the benefits of expansion for relevant and nonrelevant documents
- Generalized theoretical arguments explaining the difficulty of selective document expansion despite its evident potential for benefit

We hope that these insights prove useful in future work.
Chapter 7
Conclusions

7.1 Revisiting research questions

This thesis was motivated by the intuition that document expansion may be a solution to
the problem of sparse data. We sought to explore the utility and function of language model-
based document expansion through experimentation and analysis. The results of this work
have provided insights into the expansion process, its effects, and its potential for future use,
answering the research questions posed at the start of this thesis:

1. How can document expansion that exclusively employs language modeling techniques be used to improve retrieval effectiveness, and can query expansion and/or use of external document collections further improve effectiveness when paired with document expansion?

We have shown that document expansion using consistent language modeling techniques is able to improve retrieval effectiveness over baselines without expansion. By building relevance models on the basis of expanded documents, query expansion may be paired with document expansion to achieve some improvement over standard relevance models, and has therefore proven worthy of consideration. While model success is not dependent on the choice of expansion collection, use of external collections generally yields higher retrieval effectiveness than use of the target collection. Overall, we have found that a consistent language model-based approach to document expansion is capable of success.

2. What is the relationship between document language models and document expansion, and how does document expansion improve retrieval effectiveness?
Document language models generally improve in their representation of topical content as a result of document expansion. Analysis has shown that this improvement is tied to the quality of document pseudo-queries and, thereby, of expansion documents. Specifically, document expansion improves retrieval effectiveness by incorporating appropriate expansion documents into the target document. It identifies appropriate expansion documents by means of pseudo-queries that are both topical and discriminating. Analysis has shown, however, that our document expansion model is unlikely to ameliorate some retrieval difficulties, pointing to differences between the topical focus of the document expansion model and the definition of relevance used in some TREC relevance judgments. Overall, our analysis showed that successful document language model re-estimation by way of document expansion is highly dependent on the quality of document pseudo-queries.

3. **Can learning algorithms be used to identify and optimize opportunities for document expansion?**

   Experimental results and theoretical analysis have indicated that learning techniques for predicting opportunities for document expansion are limited primarily by the need to predict relevance. Although we were unsuccessful at optimizing document expansion through predictive modeling, the attempt revealed important insights about the strength of uniform expansion baselines and the difficulty of any per-document optimization efforts in IR, which we hope may prove beneficial to other pursuits.

### 7.2 Limitations

As with any research, the work presented in this thesis is necessarily limited by the decisions we have made and the scenarios we have considered. In this section, we discuss only a few such limitations. In the following section, we discuss possible avenues for future work based in part on addressing these limitations.

Many of the decisions made in the course of this thesis may have had a significant
impact on our results. For example, we chose to employ relevance models in exploring the
use of query expansion methods alongside our document expansion retrieval model. While
relevance models are a reasonable choice due to their excellent performance on related tasks,
other query expansion methods exist, and some may have outperformed relevance models
in our particular context. Similar choices, such as the document collections studied, the
classification methods employed, and the evaluation metrics utilized all lead us to draw
specific, and therefore limited, conclusions in response to the research questions motivating
this thesis.

Importantly, this work lacks analysis of the computational costs associated with document
expansion. Though our model is feasible with the TREC collections selected, it may be too
costly to use in some real world scenarios, such as when computational resources are lacking
or when collection sizes are prohibitively large. Deeper understanding of the costliness of
our method and the impact of that costliness on its applications may yield fruitful avenues
for research helping to reduce computational expense.

Because our approach to IR deals solely with language models, this thesis is also limited
in its conception of expansion methods and expansion data. Retrieval models that harness
more than just unigram features—such as hyperlinks, user click data, or inferred data like
part of speech tags—may also benefit from document expansion, and may be able to employ
different techniques for identifying useful expansion data sources. Further, they may be
capable of incorporating expansion data from sources other than documents, broadening the
potential impact of document expansion for retrieval.

Even without extension to other data types, our model may be reformulated in various
ways. For example, like most modern language modeling approaches to IR, we have assumed
a multinomial distribution over terms, but this is not required, and other distributional
assumptions may yield better retrieval effectiveness. Similarly, although we have chosen to
select the top \( n \) most similar expansion documents, alternative approaches are possible, such
as setting a threshold for inclusion.
### 7.3 Future work

The findings presented in this thesis, and the limitations discussed above, point to several opportunities for future work that may prove beneficial to the field of IR. We discuss only a few possible avenues of investigation.

Results from Chapter 4 showed that document pseudo-queries are a more effective means of identifying appropriate expansion documents than word embedding techniques. Paired with Chapter 5’s findings that pseudo-query quality is associated with language model improvement, it seems clear that improved methods for identifying expansion documents are likely to yield further improvements to retrieval effectiveness. This may be a more fruitful task than selective expansion for the application of learning approaches in the vein of [12].

As mentioned in Section 7.2, our document expansion approach may be compatible with non-language features. For example, hyperlinks from one web document to another imply the relevance of the linked document to at least some portion of the linking document. This structure may be exploited as a means to identify expansion documents, e.g. following [29]. In documents that lack hyperlinks, entity links—links between textual mentions of some entity and a “canonical” reference in Wikipedia or another database—may be utilized in a similar fashion.

One approach to reformulating our model considered but not explored in this thesis was that of query-biased document expansion, similar to the query-specific clustering techniques discussed in Section 2.3. We might model a preference for expansion documents that are relevant to the query. Drawbacks to this approach deterred us from investigation here. For example, query biasing may have adverse effects on the estimation of nonrelevant documents, since it risks artificially and incorrectly emphasizing query terms in documents that are by definition not relevant to the query. Further, document expansion, which is a time-and resource-intensive process, would have to occur at query time, damaging user experiences. However, query biasing may also prove particularly useful in the context of full length document retrieval. For example, documents relevant to the query but which additionally
contain information about other topics are likely to benefit from query-biased expansion.

A related avenue for future exploration is what might be called passage expansion: expanding documents by expanding their passages in an individual, piecemeal fashion. By treating portions of documents independently from one another, a passage expansion retrieval model may be able to more fully reflect the topical makeup of documents than more traditional document expansion methods (like the one proposed in this thesis), in which less prominent topics are likely to be drowned out by more prevalent ones. Passage expansion methods may also aid with passage retrieval tasks [55].

In Section 5.4.1, we found that the relationship between pseudo-query quality and query likelihood change was difficult to detect, although analysis suggests that the relationship does exist. Further work refining the metrics quantifying both pseudo-query quality and query likelihood change may help elucidate the nature of this relationship.

We are also hopeful that the document topicality data collected for Chapter 5 may prove useful to future research, whether or not it relates to document expansion. The centrality of document topicality to IR suggests that access to such data may prove beneficial in the development of future retrieval models and in the analysis of the workings of existing ones.

It is also worth remembering that because our document expansion model simply enriches document representations without substantially modifying them, it is complementary to many, if not most, LM-based retrieval models, and can therefore be used as part of future IR research and practice not specifically concerned with document expansion per se. Doing so may also help reveal situations in which document expansion is particularly helpful—or unhelpful—to the retrieval process, which may add to our understanding of document expansion, but which may also add to our understanding of the retrieval models being tested as well: their robustness to data sparsity, their ability to handle noise, etc.
References


