A New Approach for Evaluating Query Expansion: Query-Document Term Mismatch

Tonya Custis
Thomson Corporation
610 Opperman Drive
St. Paul, MN
tonya.custis@thomson.com

Khalid Al-Kofahi
Thomson Corporation
610 Opperman Drive
St. Paul, MN
khalid.al-kofahi@thomson.com

ABSTRACT

The effectiveness of information retrieval (IR) systems is influenced by the degree of term overlap between user queries and relevant documents. Query-document term mismatch, whether partial or total, is a fact that must be dealt with by IR systems. Query Expansion (QE) is one method for dealing with term mismatch. IR systems implementing query expansion are typically evaluated by executing each query twice, with and without query expansion, and then comparing the two result sets. While this measures an overall change in performance, it does not directly measure the effectiveness of IR systems in overcoming the inherent issue of term mismatch between the query and relevant documents, nor does it provide any insight into how such systems would behave in the presence of query-document term mismatch. In this paper, we propose a new approach for evaluating query expansion techniques. The proposed approach is attractive because it provides an estimate of system performance under varying degrees of query-document term mismatch, it makes use of readily available test collections, and it does not require any additional relevance judgments or any form of manual processing.

Categories and Subject Descriptors

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

General Terms

Measurement, Experimentation

Keywords

information retrieval, evaluation, query expansion

1. INTRODUCTION

In our domain,[1] and unlike web search, it is very important for attorneys to find all documents (e.g., cases) that are relevant to an issue. Missing relevant documents may have non-trivial consequences on the outcome of a court proceeding. Attorneys are especially concerned about missing relevant documents when researching a legal topic that is new to them, as they may not be aware of all language variations in such topics. Therefore, it is important to develop information retrieval systems that are robust with respect to language variations or term mismatch between queries and relevant documents. During our work on developing such systems, we concluded that current evaluation methods are not sufficient for this purpose.

{Whooping cough, pertussis}, {heart attack, myocardial infarction}, {car wash, automobile cleaning}, {attorney, legal counsel, lawyer} are all examples of things that share the same meaning. Often, the terms chosen by users in their queries are different than those appearing in the documents relevant to their information needs. This query-document term mismatch arises from two sources: (1) the synonymy found in natural language, both at the term and the phrasal level, and (2) the degree to which the user is an expert at searching and/or has expert knowledge in the domain of the collection being searched.

IR evaluations are comparative in nature (cf. TREC). Generally, IR evaluations show how System A did in relation to System B on the same test collection based on various precision- and recall-based metrics. Similarly, IR systems with QE capabilities are typically evaluated by executing each search twice, once with and once without query expansion, and then comparing the two result sets. While this approach shows which system may have performed better overall with respect to a particular test collection, it does not directly or systematically measure the effectiveness of IR systems in overcoming query-document term mismatch.

If the goal of QE is to increase search performance by mitigating the effects of query-document term mismatch, then the degree to which a system does so should be measurable in evaluation. An effective evaluation method should measure the performance of IR systems under varying degrees of query-document term mismatch, not just in terms of overall performance on a collection relative to another system.

[1] Thomson Corporation builds information based solutions to the professional markets including legal, financial, health care, scientific, and tax and accounting.

In order to measure that a particular IR system is able to overcome query-document term mismatch by retrieving documents that are relevant to a user's query, but that do not necessarily contain the query terms themselves, we systematically introduce term mismatch into the test collection by removing query terms from known relevant documents. Because we are purposely inducing term mismatch between the queries and known relevant documents in our test collections, the proposed evaluation framework is able to measure the effectiveness of QE in a way that testing on the whole collection is not. If a QE search method finds a document that is known to be relevant but that is nonetheless missing query terms, it shows that QE technique is indeed robust with respect to query-document term mismatch.

2. RELATED WORK

Accounting for term mismatch between the terms in user queries and the documents relevant to users' information needs has been a fundamental issue in IR research for almost 40 years [38, 37, 47]. Query expansion (QE) is one technique used in IR to improve search performance by increasing the likelihood of term overlap (either explicitly or implicitly) between queries and documents that are relevant to users' information needs. Explicit query expansion occurs at run-time, based on the initial search results, as is the case with relevance feedback and pseudo relevance feedback [34, 37]. Implicit query expansion can be based on statistical properties of the document collection, or it may rely on external knowledge sources such as a thesaurus or an ontology [32, 17, 26, 50, 51, 2]. Regardless of method, QE algorithms that are capable of retrieving relevant documents despite partial or total term mismatch between queries and relevant documents should increase the recall of IR systems (by retrieving documents that would have previously been missed) as well as their precision (by retrieving more relevant documents).

In practice, QE tends to improve the average overall retrieval performance, doing so by improving performance on some queries while making it worse on others. QE techniques are judged as effective in the case that they help more than they hurt overall on a particular collection [47, 45, 41, 27]. Often, the expansion terms added to a query in the query expansion phase end up hurting the overall retrieval performance because they introduce semantic noise, causing the meaning of the query to drift. As such, much work has been done with respect to different strategies for choosing semantically relevant QE terms to include in order to avoid query drift [34, 50, 51, 18, 24, 29, 30, 32, 3, 4, 5].

The evaluation of IR systems has received much attention in the research community, both in terms of developing test collections for the evaluation of different systems [11, 12, 13, 43] and in terms of the utility of evaluation metrics such as recall, precision, mean average precision, precision at rank, Bpref, etc. [7, 8, 44, 14]. In addition, there have been comparative evaluations of different QE techniques on various test collections [47, 45, 41].

In addition, the IR research community has given attention to differences between the performance of individual queries. Research efforts have been made to predict which queries will be improved by QE and then selectively applying it only to those queries [1, 5, 27, 29, 15, 48], to achieve optimal overall performance. In addition, related work on predicting query difficulty, or which queries are likely to perform poorly, has been done [1, 4, 5, 9]. There is general interest in the research community to improve the robustness of IR systems by improving retrieval performance on difficult queries, as is evidenced by the Robust track in the TREC competitions and new evaluation measures such as GMAP. GMAP (geometric mean average precision) gives more weight to the lower end of the average precision (as opposed to MAP), thereby emphasizing the degree to which difficult or poorly performing queries contribute to the score [33].

However, no attention is given to evaluating the robustness of IR systems implementing QE with respect to query-document term mismatch in quantifiable terms. By purposely inducing mismatch between the terms in queries and relevant documents, our evaluation framework allows us a controlled manner in which to degrade the quality of the queries with respect to their relevant documents, and then to measure the both the degree of (induced) difficulty of the query and the degree to which QE improves the retrieval performance of the degraded query.

The work most similar to our own in the literature consists of work in which document collections or queries are altered in a systematic way to measure differences query performance. [42] introduces into the document collection pseudowords that are ambiguous with respect to word sense, in order to measure the degree to which word sense disambiguation is useful in IR. [6] experiments with altering the document collection by adding semantically related expansion terms to documents at indexing time. In cross-language IR, [28] explores different query expansion techniques while purposely degrading their translation resources, in what amounts to expanding a query with only a controlled percentage of its translation terms. Although similar in introducing a controlled amount of variance into their test collections, these works differ from the work being presented in this paper in that the work being presented here explicitly and systematically measures query effectiveness in the presence of query-document term mismatch.

3. METHODOLOGY

In order to accurately measure IR system performance in the presence of query-term mismatch, we need to be able to adjust the degree of term mismatch in a test corpus in a principled manner. Our approach is to introduce query-document term mismatch into a corpus in a controlled manner and then measure the performance of IR systems as the degree of term mismatch changes. We systematically remove query terms from known relevant documents, creating alternate versions of a test collection that differ only in how many or which query terms have been removed from the documents relevant to a particular query. Introducing query-document term mismatch into the test collection in this manner allows us to manipulate the degree of term mismatch between relevant documents and queries in a controlled manner.

This removal process affects only the relevant documents in the search collection. The queries themselves remain unaltered. Query terms are removed from documents one by one, so the differences in IR system performance can be measured with respect to missing terms. In the most extreme case (i.e., when the length of the query is less than or equal to the number of query terms removed from the relevant documents), there will be no term overlap between a query and its relevant documents. Notice that, for a given query, only relevant documents are modified. Non-relevant documents are left unchanged, even in the case that they contain query terms.

Although, on the surface, we are changing the distribution of terms between the relevant and non-relevant documents sets by removing query terms from the relevant documents, doing so does not change the conceptual relevancy of these documents. Systematically removing query terms from known relevant documents introduces a controlled amount of query-document term mismatch by which we can evaluate the degree to which particular QE techniques are able to retrieve conceptually relevant documents, despite a lack of actual term overlap. Removing a query term from relevant documents simply masks the presence of that query term in those documents. It does not in any way change the conceptual relevancy of the documents.

The evaluation framework presented in this paper consists of three elements: a test collection, C; a strategy for selecting which query terms to remove from the relevant documents in that collection, S; and a metric by which to compare performance of the IR systems, m. The test collection, C, consists of a document collection, queries, and relevance judgments. The strategy, S, determines the order and manner in which query terms are removed from the relevant documents in C. This evaluation framework is not metric-specific; any metric (MAP, P@10, recall, etc.) can be used to measure IR system performance.

Although test collections are difficult to come by, it should be noted that this evaluation framework can be used on any available test collection. In fact, using this framework stretches the value of existing test collections in that one collection becomes several when query terms are removed from relevant documents, thereby increasing the amount of information that can be gained from evaluating on a particular collection.

In other evaluations of QE effectiveness, the controlled variable is simply whether or not queries have been expanded or not, compared in terms of some metric. In contrast, the controlled variable in this framework is the query term that has been removed from the documents relevant to that query, as determined by the removal strategy, S. Query terms are removed one by one, in a manner and order determined by S, so that collections differ only with respect to the one term that has been removed (or masked) in the documents relevant to that query. It is in this way that we can explicitly measure the degree to which an IR system overcomes query-document term mismatch.

The choice of a query term removal strategy is relatively flexible; the only restriction in choosing a strategy S is that query terms must be removed one at a time. Two decisions must be made when choosing a removal strategy S. The first is the *order* in which S removes terms from the relevant documents. Possible orders for removal could be based on metrics such as IDF or the global probability of a term in a document collection. Based on the purpose of the evaluation and the retrieval algorithm being used, it might make more sense to choose a removal order for S based on query term IDF or perhaps based on a measure of query term probability in the document collection.

Once an order for removal has been decided, a *manner* for term removal/masking must be decided. It must be determined if S will remove the terms individually (i.e., remove just one different term each time) or additively (i.e., remove one term first, then that term in addition to another, and so on). The incremental additive removal of query terms from relevant documents allows the evaluation to show the degree to which IR system performance degrades as more and more query terms are missing, thereby increasing the degree of query-document term mismatch. Removing terms individually allows for a clear comparison of the contribution of QE in the absence of each individual query term.

4. EXPERIMENTAL SET-UP

4.1 IR Systems

We used the proposed evaluation framework to evaluate four IR systems on two test collections. Of the four systems used in the evaluation, two implement query expansion techniques: Okapi (with pseudo-feedback for QE), and a proprietary concept search engine (we'll call it TCS, for Thomson Concept Search). TCS is a language modeling based retrieval engine that utilizes a subject-appropriate external corpus (i.e., legal or news) as a knowledge source. This external knowledge source is a corpus separate from, but thematically related to, the document collection to be searched. Translation probabilities for QE [2] are calculated from these large external corpora.

Okapi (without feedback) and a language model query likelihood (QL) model (implemented using Indri) are included as keyword-only baselines. Okapi without feedback is intended as an analogous baseline for Okapi with feedback, and the QL model is intended as an appropriate baseline for TCS, as they both implement language-modeling based retrieval algorithms. We choose these as baselines because they are dependent only on the words appearing in the queries and have no QE capabilities. As a result, we expect that when query terms are removed from relevant documents, the performance of these systems should degrade more dramatically than their counterparts that implement QE.

The Okapi and QL model results were obtained using the Lemur Toolkit.[2] Okapi was run with the parameters k1=1.2, b=0.75, and k3=7. When run with feedback, the feedback parameters used in Okapi were set at 10 documents and 25 terms. The QL model used Jelinek-Mercer smoothing, with $\lambda = 0.6$.

4.2 Test Collections

We evaluated the performance of the four IR systems outlined above on two different test collections. The two test collections used were the TREC AP89 collection (TIPSTER disk 1) and the FSupp Collection.

The FSupp Collection is a proprietary collection of 11,953 case law documents for which we have 44 queries (ranging from four to twenty-two words after stop word removal) with full relevance judgments.[3] The average length of documents in the FSupp Collection is 3444 words.

[2] www.lemurproject.org

[3] Each of the 11,953 documents was evaluated by domain experts with respect to each of the 44 queries.

The TREC AP89 test collection contains 84,678 documents, averaging 252 words in length. In our evaluation, we used both the title and the description fields of topics 151-200 as queries, so we have two sets of results for the AP89 Collection. After stop word removal, the title queries range from two to eleven words and the description queries range from four to twenty-six terms.

4.3 Query Term Removal Strategy

In our experiments, we chose to sequentially and additively remove query terms from highest-to-lowest inverse document frequency (IDF) with respect to the entire document collection. Terms with high IDF values tend to influence document ranking more than those with lower IDF values. Additionally, high IDF terms tend to be domain-specific terms that are less likely to be known to non-expert user, hence we start by removing these first.

For the FSupp Collection, queries were evaluated incrementally with one, two, three, five, and seven terms removed from their corresponding relevant documents. The longer description queries from TREC topics 151-200 were likewise evaluated on the AP89 Collection with one, two, three, five, and seven query terms removed from their relevant documents. For the shorter TREC title queries, we removed one, two, three, and five terms from the relevant documents.

4.4 Metrics

In this implementation of the evaluation framework, we chose three metrics by which to compare IR system performance: mean average precision (MAP), precision at 10 documents (P10), and recall at 1000 documents. Although these are the metrics we chose to demonstrate this framework, any appropriate IR metrics could be used within the framework.

5. RESULTS

5.1 FSupp Collection

Figures 1, 2, and 3 show the performance (in terms of MAP, P10 and Recall, respectively) for the four search engines on the FSupp Collection. As expected, the performance of the keyword-only IR systems, QL and Okapi, drops quickly as query terms are removed from the relevant documents in the collection. The performance of Okapi with feedback (Okapi FB) is somewhat surprising in that on the original collection (i.e., prior to query term removal), its performance is worse than that of Okapi without feedback on all three measures.

TCS outperforms the QL keyword baseline on every measure except for MAP on the original collection (i.e., prior to removing any query terms). Because TCS employs implicit query expansion using an external domain specific knowledge base, it is less sensitive to term removal (i.e., mismatch) than the Okapi FB, which relies on terms from the top-ranked documents retrieved by an initial keyword-only search. Because overall search engine performance is frequently measured in terms of MAP, and because other evaluations of QE often only consider performance on the entire collection (i.e., they do not consider term mismatch), the QE implemented in TCS would be considered (in another evaluation) to hurt performance on the FSupp Collection. However, when we look at the comparison of TCS to QL when query terms are removed from the relevant documents, we can see that the QE in TCS is indeed contributing positively to the search.

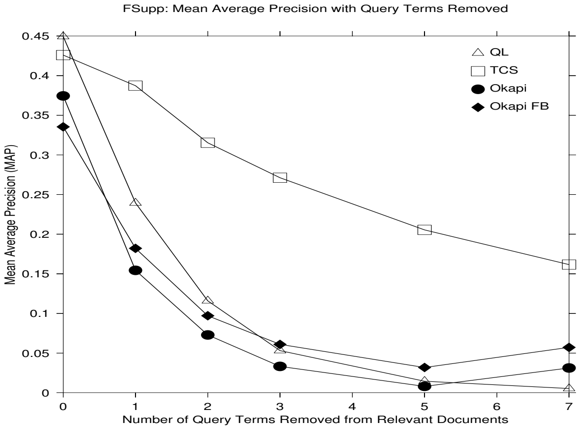

Figure 1: The performance of the four retrieval systems on the FSupp collection in terms of Mean Average Precision (MAP) and as a function of the number of query terms removed (the horizontal axis).

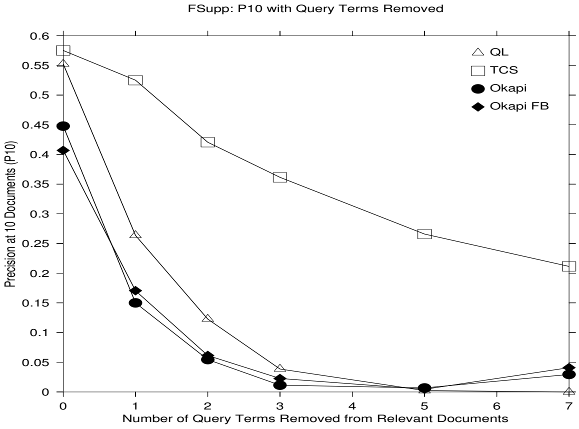

Figure 2: The performance of the four retrieval systems on the FSupp collection in terms of Precision at 10 and as a function of the number of query terms removed (the horizontal axis).

5.2 The AP89 Collection: using the description queries

Figures 4, 5, and 6 show the performance of the four IR systems on the AP89 Collection, using the TREC topic descriptions as queries. The most interesting difference between the performance on the FSupp Collection and the AP89 collection is the reversal of Okapi FB and TCS. On FSupp, TCS outperformed the other engines consistently (see Figures 1, 2, and 3); on the AP89 Collection, Okapi FB is clearly the best performer (see Figures 4, 5, and 6). This is all the more interesting, based on the fact that QE in Okapi FB takes place after the first search iteration, which

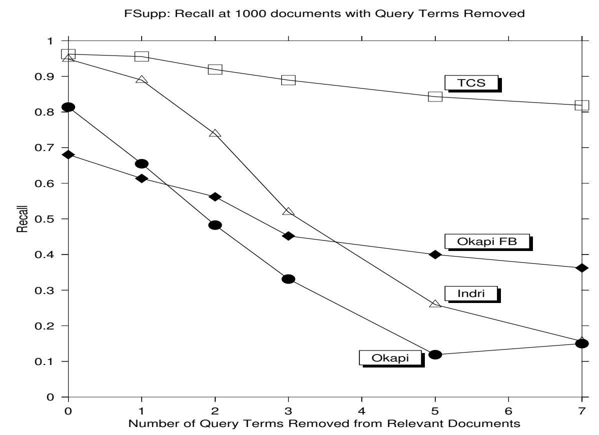

Figure 3: The Recall (at 1000) of the four retrieval systems on the FSupp collection as a function of the number of query terms removed (the horizontal axis).

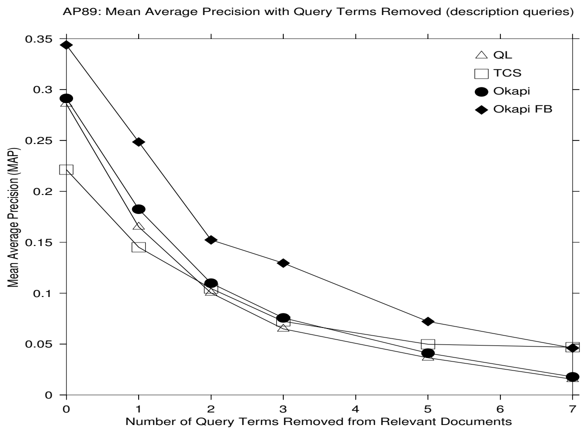

Figure 4: MAP of the four IR systems on the AP89 Collection, using TREC description queries. MAP is measured as a function of the number of query terms removed.

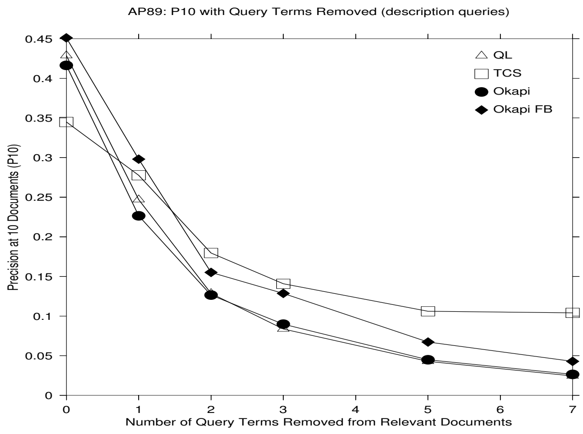

Figure 5: Precision at 10 of the four IR systems on the AP89 Collection, using TREC description queries. P at 10 is measured as a function of the number of query terms removed.

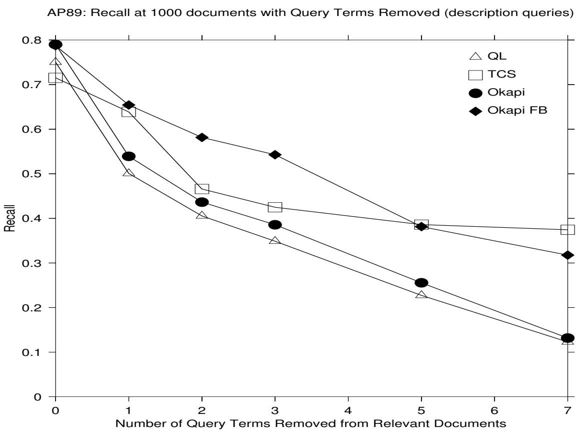

Figure 6: Recall (at 1000) of the four IR systems on the AP89 Collection, using TREC description queries, and as a function of the number of query terms removed.

we would expect to be handicapped when query terms are removed.

Looking at P10 in Figure 5, we can see that TCS and Okapi FB score similarly on P10, starting at the point where one query term is removed from relevant documents. At two query terms removed, TCS starts outperforming Okapi FB. If modeling this in terms of expert versus non-expert users, we could conclude that TCS might be a better search engine for non-experts to use on the AP89 Collection, while Okapi FB would be best for an expert searcher.

It is interesting to note that on each metric for the AP89 description queries, TCS performs more poorly than all the other systems on the original collection, but quickly surpasses the baseline systems and approaches Okapi FB's performance as terms are removed. This is again a case where the performance of a system on the entire collection is not necessarily indicative of how it handles query-document term mismatch.

5.3 The AP89 Collection: using the title queries

Figures 7, 8, and 9 show the performance of the four IR systems on the AP89 Collection, using the TREC topic titles as queries. As with the AP89 description queries, Okapi FB is again the best performer of the four systems in the evaluation. As before, the performance of the Okapi and QL systems, the non-QE baseline systems, sharply degrades as query terms are removed. On the shorter queries, TCS seems to have a harder time catching up to the performance of Okapi FB as terms are removed.

Perhaps the most interesting result from our evaluation is that although the keyword-only baselines performed consistently and as expected on both collections with respect to query term removal from relevant documents, the performances of the engines implementing QE techniques differed dramatically between collections.

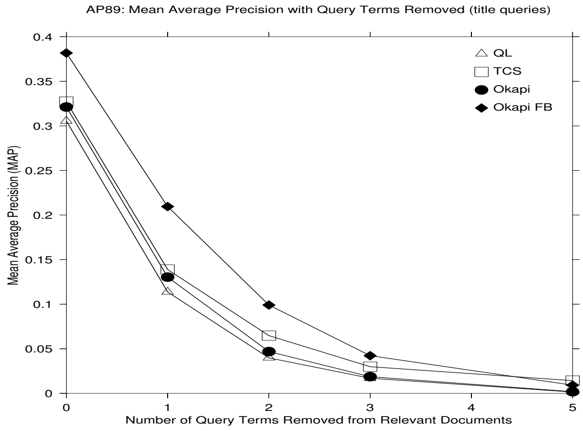

Figure 7: MAP of the four IR systems on the AP89 Collection, using TREC title queries and as a function of the number of query terms removed.

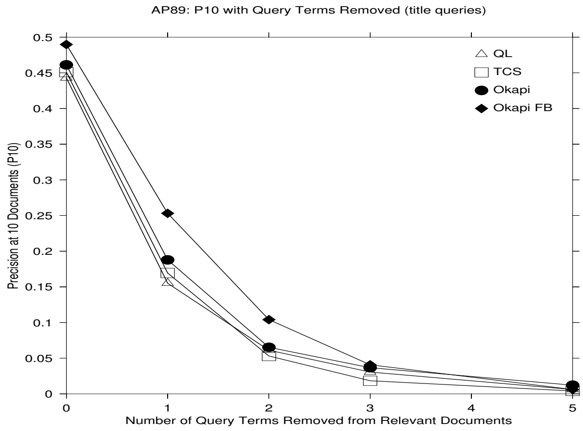

Figure 8: Precision at 10 of the four IR systems on the AP89 Collection, using TREC title queries, and as a function of the number of query terms removed.

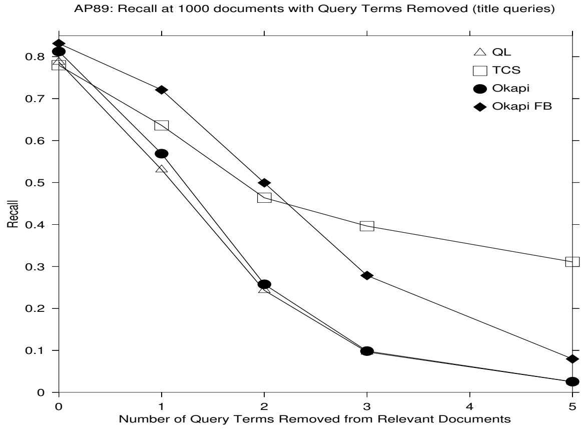

Figure 9: Recall (at 1000) of the four IR systems on the AP89 Collection, using TREC title queries and as a function of the number of query terms removed.

6. DISCUSSION

The intuition behind this evaluation framework is to measure the degree to which various QE techniques overcome term mismatch between queries and relevant documents. In general, it is easy to evaluate the overall performance of different techniques for QE in comparison to each other or against a non-QE variant on any complete test collection. Such an approach does tell us which systems perform better on a complete test collection, but it does not measure the ability of a particular QE technique to retrieve relevant documents despite partial or complete term mismatch between queries and relevant documents.

A systematic evaluation of IR systems as outlined in this paper is useful not only with respect to measuring the general success or failure of particular QE techniques in the presence of query-document term mismatch, but it also provides insight into how a particular IR system will perform when used by expert versus non-expert users on a particular collection. The less a user knows about the domain of the document collection on which they are searching, the more prevalent query-document term mismatch is likely to be. This distinction is especially relevant in the case that the test collection is domain-specific (i.e., medical or legal, as opposed to a more general domain, such as news), where the distinction between experts and non-experts may be more marked. For example, a non-expert in the medical domain might search for "whooping cough", but relevant documents might instead contain the medical term "pertussis".

Since query terms are masked only the in relevant documents, this evaluation framework is actually biased against retrieving relevant documents. This is because non-relevant documents may also contain query terms, which can cause a retrieval system to rank such documents higher than it would have before terms were masked in relevant documents. Still, we think this is a more realistic scenario than removing terms from all documents regardless of relevance.

The degree to which a QE technique is well-suited to a particular collection can be evaluated in terms of its ability to still find the relevant documents, even when they are missing query terms, despite the bias of this approach against relevant documents. However, given that Okapi FB and TCS outperformed each other on two different collection sets, further investigation into the degree of compatibility between QE expansion approach and target collection is probably warranted. Furthermore, the investigation of other term removal strategies could provide insight into the behavior of different QE techniques and their overall impact on the user experience.

As mentioned earlier, our choice of the term removal strategy was motivated by (1) our desire to see the highest impact on system performance as terms are removed and (2) because high IDF terms, in our domain context, are more likely to be domain specific, which allows us to better understand the performance of an IR system as experienced by expert and non-expert users.

Although not attempted in our experiments, another application of this evaluation framework would be to remove query terms individually, rather than incrementally, to analyze which terms (or possibly which types of terms) are being helped most by a QE technique on a particular test collection. This could lead to insight as to when QE should and should not be applied.

This evaluation framework allows us to see how IR systems perform in the presence of query-document term mismatch. In other evaluations, the performance of a system is measured only on the entire collection, in which the degree of query-term document mismatch is not known. By systematically introducing this mismatch, we can see that even if an IR system is not the best performer on the entire collection, its performance may nonetheless be more robust to query-document term mismatch than other systems. Such robustness makes a system more user-friendly, especially to non-expert users.

This paper presents a novel framework for IR system evaluation, the applications of which are numerous. The results presented in this paper are not by any means meant to be exhaustive or entirely representative of the ways in which this evaluation could be applied. To be sure, there is much future work that could be done using this framework.

In addition to looking at average performance of IR systems, the results of individual queries could be examined and compared more closely, perhaps giving more insight into the classification and prediction of difficult queries, or perhaps showing which QE techniques improve (or degrade) individual query performance under differing degrees of query-document term mismatch. Indeed, this framework would also benefit from further testing on a larger collection.

7. CONCLUSION

The proposed evaluation framework allows us to measure the degree to which different IR systems overcome (or don't overcome) term mismatch between queries and relevant documents. Evaluations of IR systems employing QE performed only on the entire collection do not take into account that the purpose of QE is to mitigate the effects of term mismatch in retrieval. By systematically removing query terms from relevant documents, we can measure the degree to which QE contributes to a search by showing the difference between the performances of a QE system and its keyword-only baseline when query terms have been removed from known relevant documents. Further, we can model the behavior of expert versus non-expert users by manipulating the amount of query-document term mismatch introduced into the collection.

The evaluation framework proposed in this paper is attractive for several reasons. Most importantly, it provides a controlled manner in which to measure the performance of QE with respect to query-document term mismatch. In addition, this framework takes advantage and stretches the amount of information we can get from existing test collections. Further, this evaluation framework is not metric-specific: information in terms of any metric (MAP, P@10, etc.) can be gained from evaluating an IR system this way.

It should also be noted that this framework is generalizable to any IR system, in that it evaluates how well IR systems evaluate users' information needs as represented by their queries. An IR system that is easy to use should be good at retrieving documents that are relevant to users' information needs, even if the queries provided by the users do not contain the same keywords as the relevant documents.

8. REFERENCES

[1] Amati, G., C. Carpineto, and G. Romano. Query difficulty, robustness and selective application of query expansion. In *Proceedings of the 25th European Conference on Information Retrieval (ECIR 2004)*, pp. 127-137.

[2] Berger, A. and J.D. Lafferty. 1999. Information retrieval as statistical translation. In *Research and Development in Information Retrieval*, pages 222–229.

[3] Billerbeck, B., F. Scholer, H. E. Williams, and J. Zobel. 2003. Query expansion using associated queries. In *Proceedings of CIKM 2003*, pp. 2-9.

[4] Billerbeck, B., and J. Zobel. 2003. When Query Expansion Fails. In *Proceedings of SIGIR 2003*, pp. 387-388.

[5] Billerbeck, B. and J. Zobel. 2004. Questioning Query Expansion: An Examination of Behaviour and Parameters. In *Proceedings of the 15th Australasian Database Conference (ADC2004)*, pp. 69-76.

[6] Billerbeck, B. and J. Zobel. 2005. Document Expansion versus Query Expansion for Ad-hoc Retrieval. In Proceedings of the 10th Australasian Document Computing Symposium.

[7] Buckley, C. and E.M. Voorhees. 2000. Evaluating Evaluation Measure Stability. In *Proceedings of SIGIR 2000*, pp. 33-40.

[8] Buckley, C. and E.M. Voorhees. 2004. Retrieval evaluation with incomplete information. In *Proceedings of SIGIR 2004*, pp. 25-32.

[9] Carmel, D., E. Yom-Tov, A. Darlow, D. Pelleg. 2006. What Makes A Query Difficult? In *Proceedings of SIGIR 2006*, pp. 390-397.

[10] Carpineto, C., R. Mori and G. Romano. 1998. Informative Term Selection for Automatic Query Expansion. In *The 7th Text REtrieval Conference*, pp.363:369.

[11] Carterette, B. and J. Allan. 2005. Incremental Test Collections. In *Proceedings of CIKM 2005*, pp. 680-687.

[12] Carterette, B., J. Allan, and R. Sitaraman. 2006. Minimal Test Collections for Retrieval Evaluation. In *Proceedings of SIGIR 2006*, pp. 268-275.

[13] Cormack, G.V., C. R. Palmer, and C.L. Clarke. 1998. Efficient Construction of Large Test Collections. In *Proceedings of SIGIR 1998*, pp. 282-289.

[14] Cormack, G. and T.R. Lynam. 2006. Statistical Precision of Information Retrieval Evaluation. In *Proceedings of SIGIR 2006*, pp. 533-540.

[15] Cronen-Townsend, S., Y. Zhou, and W.B. Croft. 2004. A Language Modeling Framework for Selective Query Expansion, CIIR Technical Report.

[16] Efthimiadis, E.N. Query Expansion. 1996. In Martha E. Williams (ed.), *Annual Review of Information Systems and Technology (ARIST)*, v31, pp 121- 187.

[17] Evans, D.A. and Lefferts, R.G. 1995. CLARIT-TREC Experiments. *Information Processing & Management.* 31(3): 385-295.

[18] Fang, H. and C.X. Zhai. 2006. Semantic Term Matching in Axiomatic Approaches to Information Retrieval. In *Proceedings of SIGIR 2006*, pp. 115-122.

[19] Gao, J., J. Nie, G. Wu and G. Cao. 2004. Dependence

language model for information retrieval. In Proceedings of SIGIR 2004, pp. 170-177.

[20] Harman, D.K. 1992. Relevance feedback revisited. In *Proceedings of ACM SIGIR 1992*, pp. 1-10.

[21] Harman, D.K., ed. 1993. *The First Text REtrieval Conference (TREC-1): 1992.*

[22] Harman, D.K., ed. 1994. *The Second Text REtrieval Conference (TREC-2): 1993.*

[23] Harman, D.K., ed. 1995. *The Third Text REtrieval Conference (TREC-3): 1994.*

[24] Harman, D.K., 1998. Towards Interactive Query Expansion. In *Proceedings of SIGIR 1998*, pp. 321-331.

[25] Hofmann, T. 1999. Probabilistic latent semantic indexing. In *Proceedings of SIGIR 1999*, pp 50-57.

[26] Jing, Y. and W.B. Croft. 1994. The Association Thesaurus for Information Retrieval. In *Proceedings of RIAO 1994*, pp. 146-160

[27] Lu, X.A. and R.B. Keefer. Query expansion/reduction and its impact on retrieval effectiveness. In: D.K. Harman, ed. *The Third Text REtrieval Conference (TREC-3).* Gaithersburg, MD: National Institute of Standards and Technology, 1995,231-239.

[28] McNamee, P. and J. Mayfield. 2002. Comparing Cross-Language Query Expansion Techniques by Degrading Translation Resources. In *Proceedings of SIGIR 2002*, pp. 159-166.

[29] Mitra, M., A. Singhal, and C. Buckley. 1998. Improving Automatic Query Expansion. In *Proceedings of SIGIR 1998*, pp. 206-214.

[30] Peat, H. J. and P. Willett. 1991. The limitations of term co-occurrence data for query expansion in document retrieval systems. *Journal of the American Society for Information Science*, 42(5): 378-383.

[31] Ponte, J.M. and W.B. Croft. 1998. A language modeling approach to information retrieval. In *Proceedings of SIGIR 1998*, pp.275-281.

[32] Qiu Y., and Frei H. 1993. Concept based query expansion. In *Proceedings of SIGIR 1993*, pp. 160-169.

[33] Robertson, S. 2006. On GMAP - and other transformations. In *Proceedings of CIKM 2006*, pp. 78-83.

[34] Robertson, S.E. and K. Sparck Jones. 1976. Relevance Weighting of Search Terms. *Journal of the American Society for Information Science*, 27(3): 129-146.

[35] Robertson, S.E., S. Walker, S. Jones, M.M. Hancock-Beaulieu, and M. Gatford. 1994. Okapi at TREC-2. In D.K. Harman (ed). 1994. *The Second Text REtrieval Conference (TREC-2): 1993*, pp. 21-34.

[36] Robertson, S.E., S. Walker, S. Jones, M.M. Hancock-Beaulieu, and M. Gatford. 1995. Okapi at TREC-3. In D.K. Harman (ed). 1995. *The Third Text REtrieval Conference (TREC-2): 1993*, pp. 109-126

[37] Rocchio, J.J. 1971. Relevance feedback in information retrieval. In G. Salton (Ed.), *The SMART Retrieval System.* Prentice-Hall, Inc., Englewood Cliffs, NJ, pp. 313-323.

[38] Salton, G. 1968. *Automatic Information Organization and Retrieval.* McGraw-Hill.

[39] Salton, G. 1971. *The SMART Retrieval System: Experiments in Automatic Document Processing.* Englewood Cliffs NJ; Prentice-Hall.

[40] Salton,G. 1980. Automatic term class construction using relevance-a summary of work in automatic pseudoclassification. *Information Processing & Management.* 16(1): 1-15.

[41] Salton, G., and C. Buckley. 1988. On the Use of Spreading Activation Methods in Automatic Information Retrieval. In *Proceedings of SIGIR 1998*, pp. 147-160.

[42] Sanderson, M. 1994. Word sense disambiguation and information retrieval. In *Proceedings of SIGIR 1994*, pp. 161-175.

[43] Sanderson, M. and H. Joho. 2004. Forming test collections with no system pooling. In *Proceedings of SIGIR 2004*, pp. 186-193.

[44] Sanderson, M. and Zobel, J. 2005. Information Retrieval System Evaluation: Effort, Sensitivity, and Reliability. In *Proceedings of SIGIR 2005*, pp. 162-169.

[45] Smeaton, A.F. and C.J. Van Rijsbergen. 1983. The Retrieval Effects of Query Expansion on a Feedback Document Retrieval System. *Computer Journal.* 26(3):239-246.

[46] Song, F. and W.B. Croft. 1999. A general language model for information retrieval. In Proceedings of the Eighth International Conference on Information and Knowledge Management, pages 316-321.

[47] Sparck Jones, K. 1971. *Automatic Keyword Classification for Information Retrieval.* London: Butterworths.

[48] Terra, E. and C. L. Clarke. 2004. Scoring missing terms in information retrieval tasks. In *Proceedings of CIKM 2004*, pp. 50-58.

[49] Turtle, Howard. 1994. Natural Language vs. Boolean Query Evaluation: A Comparison of Retrieval Performance. In *Proceedings of SIGIR 1994*, pp. 212-220.

[50] Voorhees, E.M. 1994a. On Expanding Query Vectors with Lexically Related Words. In Harman, D. K., ed. *Text REtrieval Conference (TREC-1): 1992.*

[51] Voorhees, E.M. 1994b. Query Expansion Using Lexical-Semantic Relations. In *Proceedings of SIGIR 1994*, pp. 61-69.