Cluster-Based Query Expansion for Statistical Question Answering

Lucian Vlad Lita ♠
Siemens Medical Solutions
lucian.lita@siemens.com

Jaime Carbonell
Carnegie Mellon University
jgc@cs.cmu.edu

Abstract

Document retrieval is a critical component of question answering (QA), yet little work has been done towards statistical modeling of queries and towards automatic generation of high quality query content for QA. This paper introduces a new, cluster-based query expansion method that learns queries known to be successful when applied to similar questions. We show that cluster-based expansion improves the retrieval performance of a statistical question answering system when used in addition to existing query expansion methods. This paper presents experiments with several feature selection methods used individually and in combination. We show that documents retrieved using the cluster-based approach are inherently different than documents retrieved using existing methods and provide a higher data diversity to answers extractors.

1 Introduction

Information retrieval has received sporadic examination in the context of question answering (QA). Over the past several years, research efforts have investigated retrieval quality in very controlled scenarios under the question answering task. At a first glance, document and passage retrieval is reasonable when considering the fact that its performance is often above 80% for this stage in the question answering process. However, most often, performance is measured in terms of the presence of at least one relevant document in the retrieved document set, regardless of relevant document density – where a document is relevant if it contains at least one correct answer. More specifically, the retrieval stage is considered successful even if there is a single document retrieved that mentions a correct answer, regardless of context. This performance measure is usually not realistic and revealing in question answering.

In typical scenarios, information extraction is not always able to identify correct answers in free text. When successfully found, correct answers are not always assigned sufficiently high confidence scores to ensure their high ranks in the final answer set. As a result, overall question answering scores are still suffering and considerable effort is being directed towards improving answer extraction and answer merging, yet little attention is being directed towards retrieval.

A closer look at retrieval in QA shows that the types of documents retrieved are not always conducive to correct answers given existing extraction methods. It is not sufficient to retrieve a relevant document if the answer is difficult to extract from its context. Moreover, the retrieval techniques are often very simple, consisting of extracting keywords from questions, expanding them using conventional methods such as synonym expansion and inflectional expansion, and then running the queries through a retrieval engine.

In order to improve overall question answering performance, *additional* documents and *better* documents need to be retrieved. More explicitly, information retrieval needs to: a) generate query types and query content that is designed to be successful (high precision) for individual questions and b) en-

♠ work done at Carnegie Mellon

sure that the documents retrieved by the new queries are different than the documents retrieved using conventional methods. By improving retrieval along these dimensions, we provide QA systems with additional new documents, increasing the diversity and the likelihood of extracting correct answers. In this paper, we present a cluster-based method for expanding queries with new content learned from the process of answering similar questions. The new queries are very different from existing content since they are not based on the question being answered, but on content learned from other questions.

1.1 Related Work

Experiments using the CMU Javelin (Collins-Thompson et al., 2004) and Waterloo's MultiText (Clarke et al., 2002) question answering systems corroborate the expected direct correlation between improved document retrieval performance and QA accuracy across systems. Effectiveness of the retrieval component was measured using *question coverage* – number of questions with at least one relevant document retrieved – and *mean average precision*. Results suggest that retrieval methods adapted for question answering which include question analysis performed better than ad-hoc IR methods which supports previous findings (Monz, 2003).

In question answering, queries are often ambiguous since they are directly derived from the question keywords. Such query ambiguity has been addressed in previous research (Raghavan and Allan, 2002) by extracting part of speech patterns and constructing clarification queries. Patterns are mapped into manually generated clarification questions and presented to the user. The results using the *clarity* (Croft et al., 2001) statistical measure suggest that query ambiguity is often reduced by using clarification queries which produce a focused set of documents.

Another research direction that tailors the IR component to question answering systems focuses on query formulation and query expansion (Woods et al., 2001). Taxonomic conceptual indexing system based on morphological, syntactic, and semantic features can be used to expand queries with inflected forms, hypernyms, and semantically related terms. In subsequent research (Bilotti et al., 2004), stemming is compared to query expansion using inflectional variants. On a particular question answering controlled dataset, results show that expansion using inflectional variants produces higher recall than stemming.

Recently (Riezler et al., 2007) used statistical machine translation for query expansion and took a step towards bridging the lexical gap between questions and answers. In (Terra et al., 2005) query expansion is studied using lexical affinities with different query formulation strategies for passage retrieval. When evaluated on TREC datasets, the affinity replacement method obtained significant improvements in precision, but did not outperform other methods in terms of recall.

2 Cluster-Based Retrieval for QA

In order to explore retrieval under question answering, we employ a statistical system (SQA) that achieves good factoid performance on the TREC QA task: for $\sim 50\%$ of the questions a correct answer is in the top highest confidence answer. Rather than manually defining a complete answering strategy – the type of question, the queries to be run, the answer extraction, and the answer merging methods – for each type of question, SQA learns different strategies for different types of similar questions SQA takes advantage of similarity in training data (questions and answers from past TREC evaluations), and performs question clustering. Two methods are employed constraint-based clustering and EM with similar performance. The features used by SQA clustering are surface-form n-grams as well as part of speech n-grams extracted from questions. However, any clustering method can be employed in conjunction with the methods presented in this paper.

The questions in each cluster are similar in some respect (i.e. surface form and syntax), SQA uses them to learn a complete answering strategy. For each cluster of training questions, SQA learns an answering strategy. New questions may fall in more than one cluster, so multiple answering strategies attempt simultaneously to answer it.

In this paper we do not cover a particular question answering system such as SQA and we do not examine the whole QA process. We instead focus on improving retrieval performance using a set of similar questions. The methods presented here can generalize when similar training questions are available. Since in our experiments we employ a cluster-based QA system, we use individual clusters of similar questions as local training data for learning better queries.

2.1 Expansion Using Individual Questions

Most existing question answering systems use IR in a simple, straight-forward fashion: query terms are extracted online from the test question and used to construct basic queries. These queries are then expanded from the original keyword set using statistical methods, semantic, and morphological processing. Using these enhanced queries, documents (or passages) are retrieved and the top K are further processed. This approach describes the traditional IR task and does not take advantage of specific constraints, requirements, and rich context available in the QA process. Pseudo-relevance feedback is often used in question answering in order to improve the chances of retrieving relevant documents. In web-based QA, often systems rely on retrieval engines to perform the keyword expansion. Some question answering systems associate additional predefined structure or content based on the question classification. However, there this query enhancement process is static and does not use the training data and the question answering context differently for individual questions.

Typical question answering queries used in document or passage retrieval are constructed using morphological and semantic variations of the content words in the question. However, these expanded queries do not benefit from the underlying structure of the question, nor do they benefit from available training data, which provides similar questions that we already know how to answer.

2.2 Expansion Based on Similar Questions

We introduce cluster-based query expansion (CBQE), a new task-oriented method for query expansion that is complementary to existing strategies and that leads to *different* documents which contain correct answers. Our approach goes beyond single question-based methods and takes advantage of high-level correlations that appear in the retrieval process for similar questions.

The central idea is to cluster available training questions and their known correct answers in order to exploit the commonalities in the retrieval process. From each cluster of similar questions we learn a different, *shared* query content that is used in retrieving relevant documents - documents that contain correct answers. This method leverages the fact that answers to similar questions tend to share contextual features that can be used to enhance keyword-based queries. Experiments with question answering data show that our expanded queries include a different type of content compared to and in addition to existing methods. These queries have training question clusters as a source for expansion rather than an individual test question. We show that CBQE is conducive to the retrieval of relevant documents, *different* than the documents that can be retrieved using existing methods.

We take advantage of the fact that for similar training questions, good IR queries are likely to share structure and content features. Such features can be learned from training data and can then be applied to new similar questions. Note that some of these features cannot be generated through simple query expansion, which does not takes advantage of successful queries for training questions. Features that generate the best performing queries across an entire cluster are then included in a cluster-specific feature set, which we will refer to as the *query content model*.

While pseudo-relevance feedback is performed on-line for each test question, cluster-based relevance feedback is performed across all training questions in each individual cluster. Relevance feedback is possible for training data, since correct answers are already known and therefore document relevance can be automatically and accurately assessed.

Algorithm 1 shows how to learn a query content model for each individual cluster, in particular: how to generate queries enhanced with cluster-specific content, how to select the best performing queries, and how to construct the query content model to be used on-line.

Initially, simple keyword-based queries are formulated using words and phrases extracted directly from the *free* question keywords that do not appear in the cluster definition. The keyword queries are

Algorithm 1 Cluster-based relevance feedback algorithm for retrieval in question answering

```
extract keywords from training questions in a cluster and build keyword-based queries; apply traditional query expansion methods
for all keyword-based query do
    retrieve an initial set of documents
end for
classify documents into relevant and non-relevant
select top $k$ most discriminative features (e.g. n-grams, paraphrases) from retrieved documents (across all training questions).
use the top $k$ selected features to enhance keyword-based queries – adding one feature at a time ($k$ new queries)
for all enhanced queries do
    retrieve a second set of documents
end for
classify documents into relevant and non-relevant based
score enhanced queries according to relevant document density
include in the query content model the top $h$ features whose corresponding enhanced queries performed best across all training questions in the cluster – up to $20$ queries in our implementation
```

then subjected to frequently used forms of query expansion such as inflectional variant expansion and semantic expansion (table **??**). Further processing depends on the available and desired processing tools and may generate variations of the original queries: morphological analysis, part of speech tagging, syntactic parsing. Synonym and hypernym expansion and corpus-based techniques can be employed as part of the query expansion process, which has been extensively studied (Bilotti et al., 2004).

The cluster-based query expansion has the advantage of being orthogonal to traditional query expansion and can be used in addition to pseudo-relevance feedback. CBQE is based on context shared by similar training questions in each cluster, rather than on individual question keywords. Since cluster-based expansion relies on different features compared to traditional expansion, it leads to new relevant documents, different from the ones retrieved using the existing expansion techniques.

3 The Query Content Model

Simple queries are run through a retrieval engine in order to produce a set of potentially relevant documents. While this step may produce relevant documents, we would like to construct more focused queries, likely to retrieve documents with correct answers and appropriate contexts. The goal is to add query content that increases retrieval performance on training questions. Towards this end, we evaluate the discriminative power of features (n-grams and paraphrases), and select the ones positively correlated with relevant documents and negatively correlated with non-relevant documents. The goal of this approach is to retrieve documents containing simple, high precision answer extraction patterns. Features that best discriminate passages containing correct answers from those that do not, are selected as potential candidates for enhancing keyword-based queries. For each question-answer pair, we generate enhanced queries by individually adding selected features (e.g. Table 1) to simple queries. The resulting queries are subsequently run through a retrieval engine and scored using the measure of choice (e.g. average precision). The content features used to construct the top h features and corresponding enhanced queries are included in the *query content model*.

Cluster: When did **X** start working for **Y**?	
Simple Queries	*Query Content Model*
X, **Y**	"**X** joined **Y** in"
X, **Y**, start, working	"**X** started working for **Y**"
X, **Y**, "start working"	"**X** was hired by **Y**"
X, **Y**, working	"**Y** hired **X**"
…	**X**, **Y**, "job interview"
	…

Table 1: Sample cluster-based expansion features

The *query content model* is a collection of features used to enhance the content of queries which are successful across a range of similar questions (Table 1). The collection is *cluster specific* and not *question specific* - i.e. features are derived from training data and enhanced queries are scored using training question answer pairs. Building a query content model does not preclude traditional query expansion. Through the query content model we allow shared context to play a more significant role in query generation.

4 Experiments With Cluster-Based Retrieval

We tested the performance of cluster-based content enhanced queries and compared it to the performance of simple keyword-based queries and to the performance of queries expanded through synonyms and inflectional variants. We also experiment with several feature selection methods for identifying content features conducive to successful queries.

These experiments were performed with a web-based QA system which uses the Google API for document retrieval and a constraint-based approach for question clustering. Using this system we retrieved $\sim 300,000$ and built a document set of $\sim 10GB$. For each new question, we identify training questions that share a minimum surface structure (e.g. a size 3 skip-ngram in common) which we consider to be the prototype of a loose cluster. Each cluster represents a different, implicit notion of question similarity based on the set of training questions it covers. Therefore different clusters lead to different retrieval strategies. These retrieval experiments are restricted to using only clusters of size 4 or higher to ensure sufficient training data for learning queries from individual clusters. All experiments were performed using leave-one-out cross validation.

For evaluating the entire statistical question answering system, we used all questions from TREC8-12. One of the well-known problems in QA consists of questions having several unknown correct answers with multiple answer forms – different ways of expressing the same answer. Since we are limited to a set of answer keys, we avoid the this problem by using all temporal questions from this dataset for evaluating individual stages in the QA process (i.e. retrieval) and for comparing different expansion methods. These questions have the advantage of having a more restrictive set of possible answer surface forms, which lead to a more accurate measure of retrieval performance. At the same time they cover both more difficult questions such as "*When was General Manuel Noriega ousted as the leader of Panama and turned over to U.S. authorities?*" as well as simpler questions such as "*What year did Montana become a state?*". We employed this dataset for an in-depth analysis of retrieval performance.

We generated four sets of queries and we tested their performance. We are interested in observing to what extent different methods produce additional relevant documents. The initial set of queries are constructed by simply using a bag-of-words approach on the question keywords. These queries are run through the retrieval engine, each generating 100 documents. The second set of queries builds on the first set, expanding them using synonyms. Each word and potential phrase is expanded using synonyms extracted from WordNet synsets. For each enhanced query generated, 100 documents are retrieved. To construct the third set of queries, we expand the queries in the first two sets using inflectional variants of all the content words (e.g. verb conjugations and noun pluralization (Bilotti et al., 2004)). For each of these queries we also retrieve 100 documents.

When text corpora are indexed without using stemming, simple queries are expanded to include morphological variations of keywords to improve retrieval and extraction performance. Inflectional variants include different pluralizations for nouns (e.g. *report, reports*) and different conjugations for verbs (e.g. *imagine, imagines, imagined, imagining*). Under local corpus retrieval inflectional expansion bypasses the unrelated term conflation problem that stemmers tend to have, but at the same time, recall might be lowered if not all related words with the same root are considered. For a web-based question answering system, the type of retrieval depends on the search-engine assumptions, permissible query structure, query size limitation, and search engine bandwidth (allowable volume of queries per time). By using inflectional expansion with queries that target web search engines, the redundancy for supporting different word variants is higher, and has the potential to increase answer extraction performance. Finally, in addition to the previous expansion methods, we employ our cluster-based query expansion method. These queries incorporate the top most discriminative ngrams and paraphrases (section 4.1) learned from the training questions covered by the same cluster. Instead of further building an expansion using the original question keywords, we expand using contextual features that co-occur with answers in free text. For all the training questions in a cluster, we gather statistics about the co-occurrence of answers and potentially beneficial features. These statistics are then used to select the best features and apply them to new questions whose answers are unknown. Figure 1 shows that approx-

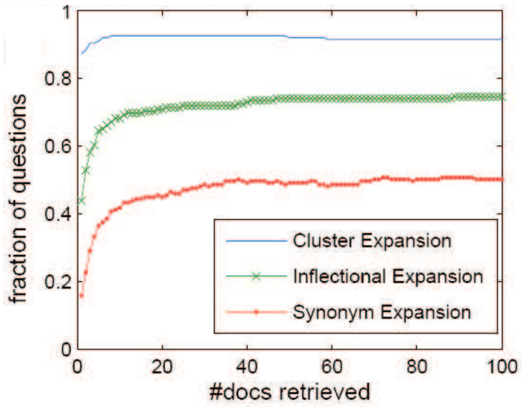

Figure 1: Cumulative effect of expansion methods

imately 90% of the questions **consistently** benefit from cluster-based query expansion when compared to approximately 75% of the questions when employing the other methods combined. Each question can be found in multiple clusters of different resolution. Since different clusters may lead to different selected features, questions benefit from multiple strategies and even though one cluster-specific strategy cannot produce relevant documents, other cluster-specific strategies may be able to.

The cluster-based expansion method can generate a large number of contextual features. When comparing feature selection methods, we only select the top 10 features from each method and use them to enhance existing question-based queries. Furthermore, in order to retrieve, process, extract, and score a manageable number of documents, we limited the retrieval to 10 documents for each query. In Figure 1 we observe that even as the other methods retrieve more documents, $\sim$ 90% of the questions still benefit from the cluster-based method. In other words, the cluster-based method generates queries using a different type of content and in turn, these queries retrieve a different set documents than the other methods. This observation is true even if we continue to retrieve up to 100 documents for simple queries, synonym-expanded queries, and inflectional variants-expanded queries.

This result is very encouraging since it suggests that the answer extraction components of question answering systems are exposed to a different type of relevant documents, previously inaccessible to them. Through these new relevant documents, cluster-based query expansion has the potential to provide answer extraction with richer and more varied sources of correct answers for 90% of the questions.

	new relevant documents	
simple	4.43	100%
synonyms	1.48	33.4%
inflect	2.37	53.43%
cluster	1.05	23.65%
all	9.33	210.45%
all - synonyms	7.88	177.69%
all - inflect	6.99	157.69%
all - cluster	8.28	186.80%

Table 2: Keyword-based ('simple'), synonym, inflectional variant, and cluster-based expansion. Average number of new relevant documents across instances at 20 documents retrieved.

Although expansion methods generate additional relevant documents that simpler methods cannot obtain, an important metric to consider is the density of these new relevant documents. We are interested in the number/percentage of new relevant documents that expansion methods contribute with. Table 2 shows at retrieval level of twenty documents how different query generation methods perform. We consider keyword based methods to be the baseline and add synonym expanded queries ('synonym'), inflectional variants expanded queries ('inflect') which build upon the previous two types of queries, and finally the cluster enhanced queries ('cluster') which contain features learned from training data. We see that inflectional variants have the most impact on the number of new documents added, although synonym expansion and cluster-based expansion also contribute significantly.

4.1 Feature Selection for CBQE

Content features are learned from the training data based on observing their co-occurrences with correct answers. In order to find the most appropriate content features to enhance our cluster-specific queries, we have experimented with several feature selection methods (Yang and Pederson, 1997): information gain, chi-square, and scaled chi-square (phi). Information gain (IG) measures the reduction in entropy for the pre presence/absence of an answer in relevant passages, given an n-gram feature. Chi-square (χ^2) is a non-parametric measure of association that quantifies the passage-level association between n-gram features and correct answers.

Given any of the above methods, individual n-gram scores are combined at the cluster level by averaging over individual questions in the cluster. In figure 2 we compare these feature selection methods on our dataset. The selected features are used to enhance queries and retrieve additional documents. We measure the fraction of question instances for which enhanced queries obtain at least one new relevant document. The comparison is made with the document set generated by keyword-based queries, synonym expansion, and inflectional variant expansion. We also include in our comparison the combination of all feature selection methods ('All'). In this experiment, average precision on training data proves to be the best predictor of additional relevant documents: $\sim$71% of the test questions benefit from queries based on average precision feature selection. However, the other feature selection methods also obtain a high performance, benefiting $\sim$68% of the test question instances.

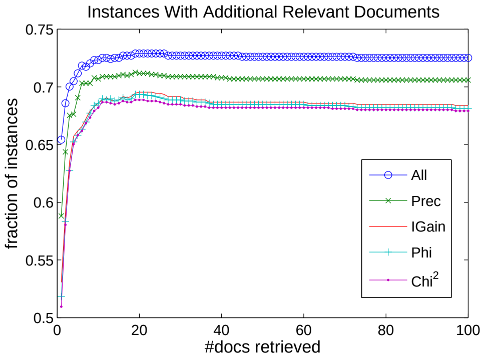

Figure 2: Selection methods for cluster-based expansion

Since these feature selection methods have different biases, we expect to observe a boost in performance (73%) from merging their feature sets (Figure 2). In this case there is a trade-off between a 2% boost in performance and an almost double set of features and enhanced queries. This translates into more queries and more documents to be processed. Although it is not the focus of this research, we note that a clever implementation could incrementally add features from the next best selection method only after the existing queries and documents have been processed. This approach lends itself to be a good basis for utility-based models and planning (Hiyakumoto et al., 2005). We investigate to what extent the scores of the selected features are meaningful and correlate with actual retrieval performance on test data by measuring the average precision of these queries at different number of documents retrieved. Figure 3 shows precision at one, five, and ten documents retrieved. We observe that feature scores correlate well with actual retrieval performance, a result confirmed by all three retrieval levels, suggesting that useful features learned. The average precision also increases with more documents retrieved, which is a desirable quality in question answering.

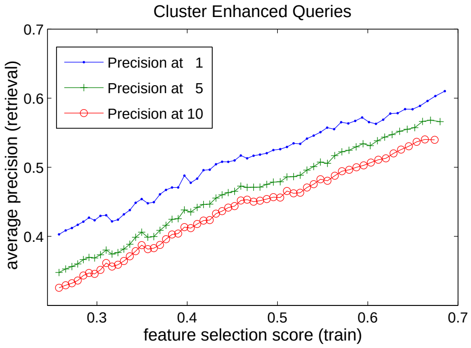

Figure 3: Average precision of cluster enhanced queries

4.2 Qualitative Results

The cluster-based relevance feedback process can be used to discover several artifacts useful in question answering. For several of the clusters, we observe that the feature selection process consistently and with high confidence selected features such as "*noun NP1 has one meaning*" where $NP1$ is the first noun phrase in the question. The goal is to add such features to the keyword-based queries to retrieve high precision documents. Note that our example, *NP1* would be different for different test questions.

The indirect reason for selecting such features is in fact the discovery of *authorities*: websites that follow a particular format and which have a particular type of information, relevant to a cluster. In the example above, the websites *answers.com* and *wordnet.princeton.edu* consistently included answers to clusters relevant to a person's biography. Similarly, *wikipedia.org* often provides answers to definitional questions (e.g. "*what is uzo?*"). By including non-intuitive phrases, the expansion ensures that the query will retrieve documents from a particular authoritative source – during feature selection, these authorities supplied high precision documents for all training questions in a particular cluster, hence features specific to these sources were identified.

Q: When did Bob Marley die? [A: answers.com]
The noun Bob Marley has one meaning:
Jamaican singer who popularized reggae (1945-81)

Born:	*6 February 1945*
Birthplace:	*St. Ann's Parish, Jamaica*
Died:	*11 May 1981 (cancer)*
Songs:	*Get Up, Stand Up, Redemption Song* …

In this example, profiles for many entities mentioned in a question cluster were found on several *authority* websites. Due to unlikely expansions such as "*noun Bob Marley has one meaning*" the entity "Bob Marley", the answer to the question "*When did Bob Marley die?*" can easily be found. In fact, this observation has the potential to lead to a cluster-based authority discovery method, in which certain sources are given more credibility and are used more frequently than others. For example, by observing that for most questions in a cluster, the *wikipedia* site covers at least one correct answer (ideally that can actually be extracted), then it should be considered (accessed) for test questions before other sources of documents. Through this process, given a set of questions processed using the IBQA approach, a set of authority answer sources can be identified.

5 Conclusions & Future Work

We presented a new, cluster-based query expansion method that learns query content which is successfully used in answering other similar questions. Traditional QA query expansion is based only on the individual keywords in a question. In contrast, the cluster-based expansion learns features from context shared by similar training questions from a cluster.

Since the features of cluster-based expansion are different from the features used in traditional query expansion, they lead to new relevant documents that are different from documents retrieved using existing expansion techniques. Our experiments show that more than 90% of the questions benefit from our cluster-based method when used in addition to traditional expansion methods.

Retrieval in local corpora offers more flexibility in terms of query structure and expressivity. The cluster-based method can be extended to take advantage of structure in addition to content. More specifically, different query structures could benefit different types of questions. However, learning structure might require more training questions for each cluster. Further research can also be done to improve the methods of combining learned content into more robust and generalizable queries. Finally we are interested modifying our cluster-based expansion for the purpose of automatically identifying authority sources for different types of questions.

References

M. W. Bilotti, B. Katz, and J. Lin. 2004. What works better for question answering: Stemming or morphological query expansion? In *IR4QA, SIGIR Workshop*.

C. Clarke, G. Cormack, G. Kemkes, M. Laszlo, T. Lynam, E. Terra, and P. Tilker. 2002. Statistical selection of exact answers.

K. Collins-Thompson, E. Terra, J. Callan, and C. Clarke. 2004. The effect of document retrieval quality on factoid question-answering performance.

W.B. Croft, S. Cronen-Townsend, and V. Lavrenko. 2001. Relevance feedback and personalization: A language modeling perspective. In *DELOS-NSF Workshop on Personalization and Recommender Systems in Digital Libraries*.

L. Hiyakumoto, L.V. Lita, and E. Nyberg. 2005. Multi-strategy information extraction for question answering.

C. Monz. 2003. From document retrieval to question answering. In *Ph. D. Dissertation, Universiteit Van Amsterdam*.

H. Raghavan and J. Allan. 2002. Using part-of-speech patterns to reduce query ambiguity.

S. Riezler, A. Vasserman, I. Tsochantaridis, V. Mittal, and Y. Liu. 2007. Statistical machine translation for query expansion in answer retrieval. In *ACL*.

E. Terra, C.L., and A. Clarke. 2005. Comparing query formulation and lexical affinity replacements in passage retrieval. In *ELECTRA, SIGIR Workshop*.

W.A. Woods, S.J. Green, P. Martin, and A. Houston. 2001. Aggressive morphology and lexical relations for query expansion.

Y. Yang and J. Pederson. 1997. Feature selection in statistical learning of text categorizatio n.