Comparing Query Formulation and Lexical Affinity Replacements in Passage Retrieval

Egidio Terra
Faculty of Informatics
PUC/RS
Porto Alegre, Brazil
egidio@inf.pucrs.br

Charles L.A. Clarke
School of Computer Science
University of Waterloo
Waterloo, Canada
claclark@plg.uwaterloo.ca

Categories and Subject Descriptors

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

General Terms

Performance, Experimentation

Keywords

Question Answering, Passage Retrieval, Query Expansion

ABSTRACT

We compare different query formulation strategies and expansion based on lexical affinities in the context of passage retrieval. Our method to expand the queries using lexical affinities replaces only the missing terms from the original query in candidate passages while scoring them. The replacement term's affinity with the missing term is used to weight the substitution, and the degree of affinity is computed using statistics generated from a terabyte corpus. The passages extracted using this replacement method and a set of passages extracted using different formulation strategies are evaluated using TREC's QA test set.

1. INTRODUCTION

In open domain question answering (QA), the process of finding answers for the questions normally takes one of these two approaches: 1) a subset of the collection is selected for further processing by the answer selection component; or 2) the whole collection is directly used by the answer selection component to find the answer. Only a certain, often small, number of documents will have one or more answers for a question, thus the first approach is often used in practice since the amount of data to be processed is reduced considerably, as it is the noise passed for posterior processing in the QA system. An important aspect in limiting the search on a sub-collection is that any imprecision in the process will prevent the system from finding the answer. The goal of sub-collection limitation is then reduce the amount of data to be further processed with the smallest error possible.

The task of sub-collection creation is accomplished by finding passages or documents that potentially contain the answer for a given question. In particular, in this work we focus on passage retrieval. In the context of QA, given a fixed retrieval model and collection, one must formulate queries that closely resemble the passages containing the answers to the questions. However, not always query terms occur in the relevant passages, either because in conjunction with other query terms it provides no or little extra information or due to the presence of an alternative term that shares a reasonably close meaning in relevant passages. This problem is normally addressed by the use of explicit query expansion or pseudo-relevance feedback. We approach this problem from a different perspective, by providing replacements for all the missing query terms using lexical affinity. The replacements can have semantic relationship with the missing terms or may be one of their morphological variants. We assume that pairs of lexical items, individual words or multiword sequences, that co-occur frequently, more often than expected by chance, have higher affinity to each other [17].

The method to use replacements by taking into account lexical affinity was introduced in Terra and Clarke [17]. We modified existing retrieval methods, one passage retrieval method and one document retrieval method, to use lexical affinity replacement. In our previous evaluation, document collection and queries were fixed in order to compare the original and the modified methods and the results showed significant improvements from the modified method to original one using the same queries. The current work extends our previous one by comparing the original passage retrieval method with expanded queries and the modified method using original queries.

We perform our experiments using the passage retrieval component of the University of Waterloo's MultiText QA system used in TREC 1999 to 2003. In particular, for the TREC 2003 QA test set [19], we generate different types of queries exploring lexical and syntactical aspects from the question, comparing results obtained using different strategies. Our results show that the replacements method based

SIGIR Workshop on Methodologies and Evaluation of Lexical Cohesion Techniques in Real-world Applications — Beyond Bag of Words '05 Salvador, Brazil

on lexical affinities outperforms in precision common explicit query expansion and formulation strategies.

The remaining of this work is organized as follows: Section 2 describes some related work on query formulation and explicit expansion. The scoring function used in our passage retrieval is presented in Section 3. The lexical affinity replacement method is presented in Section 4 and the approach to measure affinity is presented in Section 5. Some common query formulation strategies used in our comparison are presented in Section 6. The following Section 7, presents the results and discussion.

2. RELATED WORK

Radev et al. [12] examine the query formulation process for natural language questions. In order to generate the query they selectively choose which words from the original question are included by iteratively examining these words. They also explicitly expand the original question terms using WordNet and also using words with distributional similarity computed *prior* to the query processing.

Tellex et al. [15] investigate many different passage retrieval techniques using TREC2002 QA test set. They compare a probabilistic and a boolean model for initial document retrieval and find that the boolean model delivers a good performance when compared to a specific and well known probabilistic retrieval model. A boolean model is also used to retrieve an initial set of documents in other QA systems [20], Yang *et al.* acknowledge that the the number of questions covered by documents retrieved using the boolean queries is not very high, however the number of passages retrieved is also smaller. To compensate the number of questions not covered they iteratively issue boolean queries, with a "successive constraint relaxation" approach. Saggion *et al.* [14] also investigates iterative relaxation approaches for conjunctive boolean queries, including expanding individual terms, the use of some structural components in the query, such as quotes, and also deleting terms from the conjunction.

Monz [11] investigates the document retrieval component of the FlexIR system in the context of QA. He proposes the use of stemming as a way to increase effectiveness, while pseudo-relevance feedback as applied to *ad hoc* tasks yields poor performance.

Bilotti *et al.* [1] study the effect of stemming and explicit expansion using inflectional variants and found that stemming produced a lower recall while explicit expansion resulted in higher recall. Their study focused on document retrieval in the context of QA.

Roberts and Gaizauskas study some strategies to retrieve passages, either by pre or post-processing a set of retrieved documents [13]. They also formalize the concept of "coverage", used earlier by Tellex et. [15] and also used by Collins-Thompson et al. [7] and in this work.

Clarke and Terra [4] compare document and passage retrieval. Their results show that document retrieval has a slightly better coverage than passage retrieval. However, the passage retrieval provides a smaller sub-collection, thus reducing the amount of noise for down-stream components of QA systems.

3. PASSAGE RETRIEVAL

In order to investigate different expansion strategies, we use the passage retrieval component from the MultiText system. It has been used successfully in question answering [5, 3, 9] and to extract terms for pseudo-relevance feedback [21]. It can be used to retrieve passages directly from the corpus with no need for documents pre-fetching. This passage retrieval method is used for the Query formulation strategies used in this paper.

From a query $Q = \{t_1, t_2, .., t_k\}$ let $T \subseteq Q$. Given an extent of text comprising all words in the interval (u, v). The extent length is $l = v - u + 1$ and the probability of $P(t, l)$ that the extent contains one or more occurrences of t is

$$\begin{aligned} P(t,l) &= 1 - (1 - p_t)^l \\ &= 1 - (1 - lp_t + O(p_t^2)) \\ &\approx lp_t. \end{aligned}$$

The probability that an extent (u, v) contains all the terms from T is then

$$\begin{aligned} P(T,l) &= \prod_{t_i in T} P(t,l) \\ &= \prod_{t \in T} lp_t \\ &= l^{|T|} \prod_{t \in T} p_t. \end{aligned}$$

The estimation of p_t is given by the Maximum Likelihood Estimator (MLE) for t in the collection

$$p_t = f_t/N$$

where f_t is the collection frequency of t and N is the collection size in words. The score for an extent of length l containing the terms in T is the self-information of $P(T, l)$

$$\sum_{t_i \in T} \log(N/f_t) - |T| \log(l) \qquad (1)$$

The score is higher for short passages containing all terms in T and there is a trade-off on the number of terms and size of the passage.

The original passage retrieval method was presented by Clarke *et al.* [5, 6] and it provides an efficient algorithm to retrieve all passages comprising 1 to $|Q|$ query terms. The running time to extract all extents contain the terms in T is $O(|Q|\mathcal{J}_l log(N))$ where $|Q|$ is the total number of query terms, $\mathcal{J}_l$ is the number of extents containing $|T|$ query terms and N is the corpus size in words. The algorithm is based on the positions of query terms, checking for close occurrence of other query terms and skipping repetitions of the same term. This algorithm benefits from the sorted position entries in the inverted list used to index the underlying collection and quickly locate terms.

Query type	Coverage C@100	Questions Covered	Documents Correct	# Documents	Precision P@100	Precision P@20
Okapi BM25 + AQUAINT	0.903	327	5,368	36,200	0.1483	0.2381
Okapi BM25 + Terabyte	0.887	319	9,146	34,738	0.2633	0.3229

Table 1: Effectiveness of the document retrieval in the initial set

4. LEXICAL AFFINITIES REPLACEMENT METHOD

In explicit query expansion, new terms are added to the original query in order to prevent the loss of the related concept to missing original query term. For instance, in stemming, an occurrence of inflected form of a query term is to be accepted as its own. The replacement method presented here was introduced by Terra and Clarke [17]. It does not use any additional term if the original query terms are in the passage. If an original query term is missing then a new term is used. All the terms in the passage have their lexical affinity with the missing term computed and the term with the highest affinity is chosen to replace the missing term.

This modified passage retrieval only considers the whole query Q since every extent has a representative for missing query terms and uses the degree of affinity between the missing query term and its representative to adjust the scoring function of the original method. We make a simplifying assumption that a sequence of terms containing the term t also have a co-occurrence of t and itself, i.e. $p_{t,t} = p_t$. If the term t is not in the document a replacement term r will be used. The weight of the replacement is the conditional probability $p_{t|r}$, which is calculated by estimating the maximum likelihood for p_r from the corpus and estimating the joint probability by

$$p_{t,r} = f_{r,t}/N'$$

where $f_{r,t}$ is the joint frequency and N' is the total number of pairs considered for the joint frequency in the corpus.

We take a winner-takes-it-all approach and choose the best r in the extent,

$$\arg\max_{r \in (u,v)} \; p_{t|r}$$

Finally, the modified version of equation 1 using replacements is given by

$$\sum_{t_i \in Q} \log(N/f_{ti}) \cdot p_{ti|r} - |Q| \log(l) \qquad (2)$$

We should note that since every extent has a representative for a query term, we can make arbitrary decisions on the extent size. This creates a trade-off between extent size and replacement quality. On the other hand, the fact that any extent can have a representative does not allow us to use the efficient algorithm existing for the original method. Instead of selecting the extent in sub-linear time complexity (log of the corpus size) as the original method, our approximation extracts the passages in linear time.

This method can be considered a query expansion technique but not a traditional explicit expansion where new terms are added prior to the query execution. It is neither a pseudo-relevance feedback since there is not an initial retrieval stage from which new terms are added to the query. The replacement of missing terms is done while executing the query, by finding replacements when scoring each passage.

5. COMPUTING LEXICAL AFFINITIES

To compute lexical affinity, we use the approach used by Terra and Clarke [17]. For lexical affinity the pointwise mutual information (PMI) is used to score relatedness between pairs of terms.

$$PMI(w_1, w_2) = log \frac{P_{w1,w2}}{P_{w1} P_{w2}} \qquad (3)$$

The reason for choosing PMI is twofold [17]. First, it was demonstrated to be effective for language phenomena [18]. Second, it has a relationship with the inverse collection frequency —*icf* (or *idf* if document frequencies are used) . This relationship comes from the assumption that $P_{w,w} = P_w$, thus

$$\begin{aligned} PMI(w,w) &= log \; \tfrac{P_{w,w}}{P_w \cdot P_w} \\ &= log \tfrac{P_w}{P_w \cdot P_w} \\ &= -log \; P_w \\ &= icf_w \end{aligned} \qquad (4)$$

In the case of the pair of words w_1 and w_2, the maximum value for the pointwise mutual information is bounded by $PMI(w_1, w_2) \leq icf_{w1}$ and $PMI(w_1, w_2) \leq icf_{w2}$. This can be easily verified since the PMI formula has maximum value when the joint probability is equal to the smallest marginal (if marginals are different). Therefore, we can use *icf* to normalize the PMI for a given word we want to replace

$$CondPMI(w_1, w_2) \; = \; \frac{log \; (P_{w1,w2})/(P_{w1} \cdot P_{w2})}{log \; (1)/(P_{w1})} \qquad (5)$$

which is monotonic to

$$\frac{(P_{w1,w2})/(P_{w1} \cdot P_{w2})}{1/P_{w1}} \; = \; P_{w1|w2}$$

Thus, if we fix one word, in this case the missing query term, we can rank the affinity of remaining words of the passage. Since the goal is to find a replacement for one query term at each time, the denominator of the equation 5 is fixed for every replacement. We should note that there is a problem with the normalization in the conditional PMI. The problem occurs when PMI is negative, in which case we just set it to zero. Setting the negative value to zero could be avoided if we offset both *icf* and PMI by the minimal PMI value. We ignore negative PMI and set its value to zero, thus we use a self-regulated cut-off for the minimal value for a conditional PMI. We assume that any word in the document with a negative PMI with respect to the missing query term is not a good candidate for replacement.

For estimation of $P_{w1,w2}$ use the maximum likelihood :

$$P_{w1,w2} = f_{w1,w2}/N' \tag{6}$$

where the joint-frequency $f(w_1, w_2)$ is the number of the co-occurrences of w_1 and w_2 at distances ranging from four to 40 words apart. The lower cut-off prevents phrasal relationships (e.g. if the term "New" is a query term but "York" is not, then the latter is probably not a good replacement for the first). As most of the co-occurrences of "New" and "York" happen at distance one, then this cut-off will avoid this bias for pairs in the same phrase. Terra and Clarke [18] showed that windows of 32 words are a good setting for an upper bound on the distance. Our upper cut-off was arbitrarily set close to it (40). The value of N' is the size of the window times the corpus size $(36 \cdot N)$.

6. QUERY FORMULATION STRATEGIES

To compare the lexical affinities replacement method with query formulation we created a series of queries using some common strategies. For all of them we perform stop word exclusion.

- Bag-of-Words

  This is one of the most common forms to specify a query. In particular the vector space, probabilistic and many language models employ this strategy to generate queries. In the scope of QA, the query is comprised of the question terms and the order in which terms are specified is not important. The query terms are considered to be independent from each other.

- Stemming (Bag+Stem)

  A common strategy in information retrieval is to apply a stemmer to the query terms. The intuition is that using the word stem, and not the original form from the question, will help overcome mismatching vocabulary problems. These queries are comprised of the question terms with stemming. The collection index contains both the stem and original forms.

- Boolean conjunction (Bool)

  To ensure that all questions terms are present in the query, some QA systems use the boolean expressions [1, 14, 15, 20]. Our boolean queries are a conjunction of the question terms after stop words exclusion.

- Quotes

  In these queries we keep the original quotes when supplied in the question, e.g., What country is know as the "Land of the Rising Sun?". For the purpose of retrieval, these quotations are treated as phrases and their constituent words are not used in query other than in the phrasal component. The remaining of the question words (not stop words) are used as in the bag-of-words approach.

- Quotes plus Noun Phrases (Phrases)

  To further investigate phrases in our passage retrieval method we explore noun phrases in the questions that are not part of quotes. The words in the questions were tagged using a standard POS tagger and adjacent pairs were concatenated if the sequence matches one of the following : 1) adjective followed by noun; 2) a non-proper noun followed by any noun; 3) foreign word followed by any noun; 4) any noun followed by a foreign word; 5) proper-noun followed by proper noun; and 5) numeral followed by any noun. Quotations were kept from the question. We must note that the POS tagger sometimes fails but that does not happen often, e.g. How/WRB did/VBD Jerry/NNP Garcia/NNP Die/NNP ? where the main verb is tagged as a proper noun.

- Verb expansion (VE)

  The Waterloo's MultiText QA system of TREC expanded verbs as a way to improve effectiveness [6]. We use a probabilistic version of Earley's parser and a grammar created to handle QA questions. Each regular verb is stemmed and all irregular verbs are expanded. Bilotti *et al.* [1] expanded verbs, along with other expansions, in a "conjunction of disjunction" boolean queries (query terms are ANDed and expansion for individual terms are ORed).

- Verb expansion plus Quotes (VE+Quotes)

  These queries have both expanded verbs and quotes from the original questions. These components, along with some heuristics expansions such as expanding "U.S" to ("U.S" or "United States"), form queries used in our last TREC-QA participations. The heuristics expansions were removed in the experiments reported here, leaving only verb expansion and quotes as phrases.

7. RESULTS AND DISCUSSION

We evaluate the performance of the different query formulation strategies in passage retrieval using TREC 2003 QA test set. We focus on the 413 factoid questions from which we use the 362 that have available regular expression patterns, created from all submissions after human judgments were done. These patterns can be used in a script to perform automatic assessments, called lenient in TREC, as opposed to human judgments, called strict in TREC. Two target corpora were used. The official TREC corpus — AQUAINT — and a terabyte collection [4, 17, 3, 18].

The replacement weights were all extracted from the terabyte corpus, using MLE, as discussed in Section 5. However, it is interesting to note that using a window of larger size increases the chance of observing co-occurrence and, along with proper normalization, this can be viewed as a sort of smoothing [16].

To measure effectiveness of the passage retrieval with the different strategies we calculated the *coverage*, the percentage of the 362 questions where at least one retrieved passage containing the answer at 20 documents (C@20), the same metric used by Tellex et al. [15] and Roberts and

Query type	Coverage C@20	Questions Covered	Passages Correct	# Passages	Precision P@20
Bag of words	0.738	267	1269	7240	0.1753
Bag+stem	0.710	257	1251	7240	0.1728
Boolean (bool)	0.483	175	669	3787	0.1767
Quotes	0.735	266	1261	7240	0.1742
Quotes+phrases	0.669	242	1076	7032	0.1530
Verb expansion (VE)	0.746	270	1223	7240	0.1689
VE +quotes	0.749	271	1226	7240	0.1693
Replacement	0.749	271	1412	7240	0.1950

Table 2: Passage Retrieval from top 100 okapi documents in the AQUAINT Corpus

Query type	Coverage C@20	Questions Covered	Passages Correct	# Passages	Precision P@20
Bag of words	0.751	272	1894	7240	0.2616
Bag+stem	0.735	266	1835	7240	0.2535
Boolean (bool)	0.702	254	1474	5640	0.2613
Quotes	0.754	273	1891	7240	0.2612
Quotes+phrases	0.718	260	1681	7090	0.2371
Verb expansion (VE)	0.785	284	1877	7240	0.2593
VE +quotes	0.785	284	1899	7240	0.2623
Replacements	0.757	274	2033	7240	0.2808

Table 3: Passage Retrieval from top 100 okapi documents in the Terabyte Corpus

Gaizauskas [13]. We also used *precision* at 20 documents (P@20).

To restrict the passage selection we first retrieve a set of documents, using Okapi BM25, to which we apply all the query formulations and the replacement method. As such, the effectiveness of the passage selection is bounded by the original effectiveness of the document retrieval on the two collections, as presented in Table 1. The number of questions covered using the different collections is similar, a little bit higher in the AQUAINT collection but the precision is higher in the terabyte collection, as reported at 100 documents used in the initial retrieval. The same pattern occur at 20 documents.

The different strategies for explicit expansion and the lexical affinity replacement method were then applied to the 100 documents in each question to select the best 20 passages. For each query formulation a single passage is extracted from each document using equation 1. The same procedure is executed for the replacement method: one passage per document, passages scored by equation 2. While the matching span of a query is variable in the passage retrieval method used here, we extend all the returned passages to be 170 words long, roughly 1000 bytes-long and one quarter of the document average size.

The results of the passage selection in the AQUAINT corpus are shown in Table 2. Both verb expansion and the replacement methods cover the highest number of questions. In precision at 20 passages the replacement method is better: the difference with any other query formulation is statistically significant at 99% significance level using Wilcoxon signed rank test, as shown in Table 4. For the Terabyte corpus the results are shown in Table 3. Once again, the verb expansion strategies yield better coverage. The replacement method is worse than verb expansion in coverage but it is again the best in precision, with the differences between the replacement and other methods, with exception of the boolean queries, being statistically significant at 99% using Wilcoxon signed rank test. It is interesting to note that the trends in coverage are maintained across the two collections: poor performance of boolean queries and phrases. Also, stemming seems to harm more than help in precision, contradicting Monz [11] and in accordance with Bilotti *et al.* [1], although the metrics used here are different.

From all the strategies, the use of phrases has the worst outcome. Phrases can be decomposable or undecomposable [8]. The decomposable ones can be rewritten in different forms and, as consequence, be absent from some relevant passages, which we can call "mismatching decomposable phrases problem". This outcome can also be explained by fact that the scoring function used is designed to handle individual terms in order to address the bag-of-word approach and by assuming independence among query terms. The same is not observed when using quotes, since quotes are important as specified and tend not be rewritten, i.e. they are undecomposable phrases. Verb expansion consistently improves coverage but results in precision at 20 are mixed, mostly not statistically significant when compared to other query formulation strategies but the lexical affinity replacement method.

Boolean queries in conjunctive form are more restrictive: fewer passages are retrieved when these queries are used. This reduction helps final precision since every correct passage will have a greater impact. The coverage of boolean queries is smaller, a result of the reduced number of passage (i.e. less chance to cover questions). These findings suggest an explanation for the successful adoption of boolean

p-values	Bag+stem	Boolean	Quotes	Phrases	VE	VE+Quotes	Replacement
Bag of words	0.2096	0.1287	0.1003	2.76E-005	0.1632	0.1634	0.0003
Bag+stem	-	0.3234	0.3864	0.0148	0.9305	0.9258	1.29E-005
Boolean (bool)		-	0.1568	0.8451	0.4246	0.4067	0.0014
Quotes			-	8.90E-005	0.3033	0.3109	7.51E-005
Quotes+phrases				-	0.0086	0.0104	2.78E-009
Verb expansion (VE)					-	1.0000	1.69E-005
VE +quotes						-	1.91E-005

Table 4: p-values for p@20 pairwise Wilcoxon signed rank test - AQUAINT

p-values	Bag+stem	Boolean	Quotes	Phrases	VE	VE+Quotes	Replacement
Bag of words	0.0283	0.6338	0.5062	0.0001	0.8277	0.9937	0.0005
Bag+stem	-	0.1358	0.0354	0.1455	0.1474	0.1078	1.43E-006
Boolean (bool)		-	0.6657	0.0060	0.4872	0.5710	0.0540
Quotes			-	2.96E-005	0.8786	0.9336	0.0010
Quotes+phrases				-	0.0037	0.0007	1.80E-008
Verb expansion(VE)					-	0.2839	0.0013
VE +quotes						-	0.0031

Table 5: p-values for p@20 pairwise Wilcoxon signed rank test - Terabyte

queries, used in multiple iterations, in some QA systems [1, 10, 14, 15, 20]. Nonetheless, it is arguable that a QA system that can take advantage of the redundancy [5, 2] to find answers to questions would benefit from a large number of passages, in particular if the precision is at the same level or higher, as delivered by the lexical affinities replacement method. In fact, since the replacement method guarantees that exactly one representative for each query term is always present, we can view this method as performing a boolean conjunctive query of the original terms, either by themselves or through a proxy.

8. CONCLUSIONS

We compare different query formulation strategies and a lexical affinity replacement method in passage retrieval in the context of QA. We used lexical affinities to identify replacements for missing query terms while scoring passages. The replacement term weight is adjusted by its affinity with the missing one. This term replacement method produced consistent and significant improvements in precision in comparison with other query strategies. In terms of coverage, the replacement method has not outperformed query formulations, in particular verb expansion, which may suggest that a combination of verb expansion and the replacement method may produce both better coverage and precision. Our findings on phrases mirror the results of other works [8]. In particular, quotations are undecomposable phrases and must be used as they appear in the questions. Further investigation on decomposable phrases as suggested by Spark-Jones [8], with different scoring for phrases and individual terms may also improve effectiveness. In particular, these phrases can have their degree of lexical affinity taken into account in a modified scoring function.

9. REFERENCES

[1] M. W. Bilotti, B. Katz, and J. Lin. What works better for question answering: Stemming or morphological query expansion. In *SIGIR 2004 IR4QA: Information Retrieval for Question Answering Workshop*, 2004.

[2] E. Brill, J. Lin, M. Banko, S. Dumais, and A. Ng. Data-intensive Question Answering. In *Proceedings of 2001 Text REtrieval Conference*, Gaithersburg, MD, 2001.

[3] C. Clarke, G. Cormack, M. Laszlo, T. Lynam, and E. Terra. The impact of corpus size on question answering performance. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 369–370, Tampere, Finland, 2002.

[4] C. Clarke and E. L. Terra. Passage retrieval vs. document retrieval for factoid question answering. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 427–428, Toronto, Canada, 2003.

[5] C. L. A. Clarke, G. V. Cormack, and T. R. Lynam. Exploiting redundancy in question answering. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 358–365. ACM Press, 2001.

[6] C. L. A. Clarke, G. V. Cormack, T. R. Lynam, and E. Terra. *Advances in Open Domain Question Answering*, chapter Question answering by passage selection. Kluwer Academic Publishers. To appear, 2004.

[7] K. Collins-Thompson, E. Terra, J. Callan, and C. L. A. Clarke. The effect of document retrieval quality on factoid question answering. In *ACM SIGIR Conference on Research and development in Information Retrieval*, 2004.

[8] K. S. Jones, S. Walker, and S. E. Robertson. A probabilistic model of information retrieval: development and comparative experiments - part 2. *Information Processing and Management*, 36(6):809–840, 2000.

[9] J. Lin, A. Fernandes, B. Katz, G. Marton, and S. Tellex. Extracting answers from the web using data annotation and knowledge mining techniques. In *The Eleventh Text REtrieval Conference (TREC 2002)*, Gaithersburg, MD, 2002.

[10] D. Moldovan, M. Paca, S. Harabagiu, and M. Surdeanu. Performance issues and error analysis in an open-domain question answering system. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 33–40, 2002.

[11] C. Monz. Document retrieval in the context of question answering. In F. Sebastiani, editor, *Proceedings of the 25th European Conference on Information Retrieval Research (ECIR-03), Lecture Notes in Computer Science 2633*, pages 571–579. Springer-Verlag Heidelberg, 2003.

[12] D. R. Radev, H. Qi, Z. Zheng, S. Blair-Goldensohn, Z. Zhang, W. Fan, and J. Prager. Mining the web for answers to natural language questions. In *Proceedings of the tenth international conference on Information and knowledge management*, pages 143–150. ACM Press, 2001.

[13] I. Roberts and R. Gaizauskas. Evaluating passage retrieval approaches for question answering. In *Proceedins of the 26th European Conference on Information Retrieval Research (ECIR 2004), Lecture Notes in Computer Science 2997*, pages 72–84. Springer-Verlag Heidelberg, 2004.

[14] H. Saggion, R. Gaizauskas, M. Hepple, I. Robrts, and M. Greenwood. Exploring the performance of boolean retrieval strategies for open domain question answering. In *SIGIR 2004 IR4QA: Information Retrieval for Question Answering Workshop*, 2004.

[15] S. Tellex, B. Katz, J. Lin, A. Fernandes, and G. Marton. Quantitative evaluation of passage retrieval algorithms for question answering. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 41–47, 2003.

[16] E. Terra. *Lexical Affinities and Natural Language Applications*. Ph.D. thesis, School of Computer Science, University of Waterloo, October 2004.

[17] E. Terra and C. L. Clarke. Scoring missing terms in information retrieval tasks. In *13th ACM Conference on Information and Knowledge Management(CIKM)*, 2004.

[18] E. Terra and C. L. A. Clarke. Frequency estimates for statistical word similarity measures. In *Proceedings of Human Language Technology conference / North American chapter of the Association for Computational Linguistics annual meeting*, pages 244–251, Edmonton, Alberta, 2003.

[19] E. M. Voorhees. Overview of the TREC 2003 Question Answering track. In *In proceedings of 2003 Text REtrieval Conference*, pages 14–27, Gaithersburg, MD, 2003.

[20] H. Yang, T.-S. Chua, S. Wang, and C.-K. Koh. Structured use of external knowledge for event-based open domain question answering. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 33–40, 2003.

[21] D. L. Yeung, C. L. A. Clarke, G. V. Cormack, T. R. Lynam, , and E. Terra. Task-specific query expansion (multitext experiments for trec 2003). In *2002 Text REtrieval Conference*, Gaithersburg, MD, 2003.