Exploring the benefit of contextual information for boosting TREC Genomic IR performance

Bader Aljaber , Nicola Stokes‡ , James Bailey* ‡‡ , Yi Li* ‡‡*

* Dept of Computer Science and Software Engineering, The University of Melbourne, Australia
‡School of Computer Science and Informatics, University College Dublin, Ireland
‡‡NICTA Victoria Research Laboratory, Australia

{*baljaber, jbailey, yli8*}*@csse.unimelb.edu.au, nicola.stokes@ucd.ie*

Abstract *Query Expansion is a widely used technique that augments a query with synonymous and related terms in order to address a common issue in ad hoc retrieval: the vocabulary mismatch problem, where relevant documents contain query terms that are semantically similar, but lexically distinct. Standard query expansion techniques include pseudo relevance feedback and ontology-based expansion. In this paper, we explore the use of contextual information as a means of expanding the context surrounding the unit of retrieval, rather than the query, which in this case is a document passage. The ad hoc retrieval task that we focus on in this paper was investigated at the TREC 2006 Genomic tracks, where systems were required to retrieve relevant answer passages. The most commonly reported indexing strategy was passage indexing. Although this simplifies post-retrieval processing, retrieval performance can be hurt as valuable contextual information in the containing document is lost. The focus of this paper is to investigate various contextual evidence of similarity outside of the passage such as: query/full-text similarity, query/citation sentence similarity, query/title similarity, query/abstract similarity. These similarity scores are then used to boost the rank of passages that exhibit high contextual evidence of query similarity. Our experimental results suggest that document context provides the strongest evidence of contextual information for this task.*

Keywords Passage Retrieval, Contextual Document Expansion and Ranking Strategies.

1 Introduction

Query expansion is a technique used in Information Retrieval (IR) to address the synonymy problem. More specifically, a relevant document, which contains semantically related words that are lexically dissimilar to the query, will appear less related than it actually is. This is also referred to as the*vocabulary mismatch problem* [3]. This is a very common problem, which affects IR effectiveness more than the problem of query term ambiguity [5]. An alternative to query expansion is document expansion - the process of adding related terms to the document's representation. In this paper, we explore the use of document expansion in a passage retrieval task. The TREC 2006 and 2007 (Text REtrieval Conference) Genomic track task requires the retrieval of extracted answer passages, in response to natural language questions. The most commonly used indexing strategy used by track participants for this task was passage indexing. Although this simplifies post-retrieval processing, retrieval performance can be hurt as valuable contextual information in the containing document is lost by this indexing strategy. Thus, expansion techniques are needed. Work in [7] investigated the impact of various *query* expansion term types on passage retrieval effectiveness in this Genomic IR task. The results showed that a significant improvement can be gained when ontologically related words (synonyms, hypernyms, hyponyms) are used in query expansion.

In this paper, we extend the work presented by Stokes et al. [7] by exploring different types of *contextual information* as a means of expanding the context surrounding the unit of retrieval (a passage), rather than the query. This is a type of document expansion. So, we investigate the use of various sources of contextual evidence of similarity outside of the passage such as: query/full-text similarity, query/citation sentence similarity, query/title similarity and query/abstract similarity. These similarity scores are then used to boost the rank of passages that exhibit high contextual evidence of query similarity. Our results indicate that document context is the strongest source of contextual evidence for this task.

Related Work. Similarly to query expansion, document expansion can be used to overcome the problem of synonymy. Document expansion techniques, enrich documents off-line with related terms during indexing. This type of expansion can reduce the overheads of query expansion at query time.

Billerbeck and Zobel [1] proposed two new corpus-based methods for document expansion. In the first method, each document is treated as a query and augmented by related terms. In the second method, each term in the corpus is treated as a

Proceedings of the 13th Australasian Document Computing Symposium, Hobart, Australia, 8 December 2008.

query and augmented by related terms and used to rank documents accordingly. Overall, Billerbeck and Zobel's experiments showed that compared with query expansion, document expansion methods achieved relatively poor improvements. That might be because the specific topic of the original documents is significantly changed when related terms are added.

2 TREC Evaluation Data and Metrics

In this section, we will describe the data collection, retrieval task and the evaluation metrics we use. All our experiments were conducted on the document collection used in *2006 and 2007 Genomic TREC task*[1]. The TREC collection consists of *162,259* full-text journal articles from 49 journals which are electronically published via the Highwire Press site. Besides that, 28 topics expressed as natural language questions are also provided. Participants of the task were required to implement a retrieval system and submit the first 1,000 ranked passages returned by their systems for each of the topics (Hersh et al. [4]). Passages in this task can be defined as text sequences that must occur within paragraph boundaries (delimited by HTML tags). For evaluation, human judges evaluate the relevance of passages retrieved. More precisely, passage boundaries were defined, and each relevant answer was assigned a set of topic tags (called *aspects*) from a control vocabulary of MeSH terms. MeSH stands for ***Me***dical ***S***ubject ***H***eadings[2].

To evaluate the system effectiveness, we use *M*ean *A*verage *P*recision (**MAP**), which is considered one of the most common IR evaluation metrics. In document retrieval systems, the document MAP score is calculated as follows: for a given query, the average of all the precision values at each recall point in document ranked list is first calculated. Then, the mean of all the query average precision scores is determined. The TREC Genomics Track also defines a variant of this MAP score. The Passage MAP is similar to the document MAP. However, since passage retrieval is a question answering task, a special metric which factors in the length of the passages retrieved is introduced. So, the passage MAP is calculated as the fraction of characters in the system passage overlapping with the gold standard answer, divided by the total number of characters in every passage retrieved up to that point in the ranked list. Consequently, extra characters retrieved will (negatively) affect the final MAP score.

Stokes et al. [7] defined another version of the MAP, called the *paragraph MAP* score. The paragraph MAP calculates the fraction of paragraphs retrieved that contain a correct passage, divided by the total number of paragraphs retrieved. As before, the average of these scores at each recall point is the final score for that topic. In this metric, extra characters retrieved cannot affect the final MAP score and the system will get "full-marks" if it returns the paragraph that the gold standard passage occurs in.

In our experiments, Mean Average Precision (MAP) is used to evaluate system performance at three different levels of information granularity: Passages, Documents and Paragraphs.

3 System Description

In this paper, we augment an IR query expansion system first proposed in [7]. The authors introduced a novel concept normalization ranking metric, which maximizes the impact of query expansion in the genomic domain. More specifically the system ensures that documents containing multiple unique concepts are ranked higher than those which make reference to the same concept multiple times; and expansion terms (synonyms and related terms) for the same concept are not given undue influence by the ranking metric.

Briefly, we will describe the genomic retrieval system presented in [7] with emphasis on the part that we will extend; followed by an explanation of our extension to the system. The architecture of that genomic retrieval system is shown in Figure 1. The data collection is first prepared and separately indexed on paragraph and other contextual information representations; a query is then taken and expanded with synonyms found in other external resources such as MeSH terms which stand for ***Me***dical ***S***ubject ***H***eadings. Based on that, two sets of ranked outputs are retrieved. First, a set of candidate paragraphs is retrieved based on paragraph indexing and then ranked. Second, a set of documents is retrieved based on the indexing of other contextual information such as the document contexts (that is the original document representation) and then ranked. The candidate paragraphs are reduced (that is what we will call later as a *passage reduction*) in order to extract relevant answer passages to the query (in our case, queries are natural language questions). These extracted passages are then ranked and presented to the user. The overall process is referred to as a *PASSAGE* run.

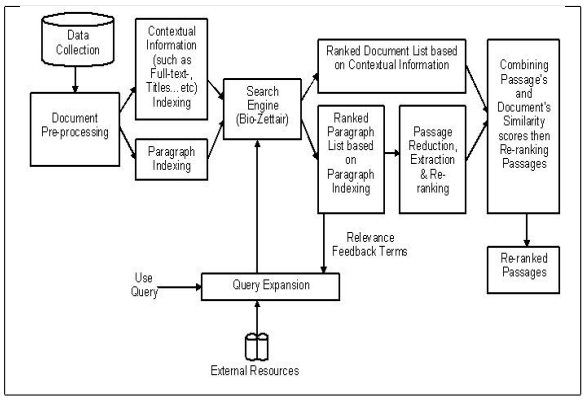

Figure 1: The architecture of the passage retrieval system.

[1] http://ir.ohsu.edu/genomics/2006protocol.html

[2] MeSH terms are managed and created by the United States National Library of Medicine (NLM), http://www.nlm.nih.gov

Stokes et al. [7] found that query expansion with synonyms from domain-specific terminology resources achieves a significant performance over a baseline system. Further, improvements in Passage MAP score were achieved when the passage reduction process was performed on retrieved paragraphs. However, although Passage MAP increased significantly (from 0.108 to 0.127), Document level MAP drop significantly (from 0.534 to 0.507).

To address this problem, Stokes et al. [7] proposed a new passage re-ranking method which considers both the relevance of the passage to the query, and incorporates the relevance score of the document containing that passage. In other words, the similarity scores of the retrieved passages are linearly combined with the similarity scores of their containing documents. This method boosts Passage and Document MAP scores and can be summarized as follows:

1. First perform query expansion, query the index, and then use passage extraction and re-ranking to find the top 1000 passages for each query. A query is defined as a set of concept and non- concept terms or phrases. For example, the query "What is the role PRNP in Mad Cow Disease?", has two concept terms 'PRNP' and 'Mad Cow Disease' and one non-concept term 'role', the rest are stopwords. By splitting the query in this manner, we can ensure that the occurrence of less informative, non-concept terms do not have an inflated weight of importance in the similarity calculation.

2. The top 1000 passages are then divided into different concept level groups. That is, we group documents in the ranked list based on the number of query concept terms they contain, where documents with all query concepts reside at the top of the ranked list.

3. Within each group, passages are re-ranked by combining their similarity scores with their containing document's similarity score.

So, for a passage i which has a similarity score *Pi* and whose containing document has a similarity score *Di*, the final combination score *Si* is calculated as:

$$Si = Pi \times \frac{Di}{Dmax} \times Pmax \qquad (1)$$

where *Pmax* and *Dmax* are the maximum similarity scores of all the 1000 passages and their containing documents.

Our Contribution: Our extension to the work in [7] is based on the investigation of additional types of contextual information for re-ranking the retrieved passages, in order to attain better IR performance for this task. In other words, as well as using the document context's similarity scores in re-ranking the retrieved passages, we also examine and evaluate the effect of using similarity scores based on different types of contextual information, which are outlined in the next section.

4 Experimental Results

It has been shown in [7] that when the document context's similarity scores are used to re-rank retrieved passages, MAP score increases for both Passage level MAPs (that is, from 0.108 to 0.137) and document level MAPs (that is, from 0.534 to 0.543) are observed. Hence, as already mentioned, use this result to motivate our investigation of exploring the benefit of using other contextual information for boosting TREC Genomic IR performance such as Citation Contexts, Titles, Abstracts and MeSH terms.

A citation context is essentially the text surrounding the reference markers (e.g the 'cite' command in LaTeX) used to refer to other scientific works. These citation contexts are essentially descriptive fragments and are likely to contain synonymous or related terms to the document being cited. Consequently, they can be used as an alternative representation of the contents of a document. Citation contexts have been used by a number of techniques in information retrieval [6].

From the entire collection, we were able to extract the citation contexts for *3475* documents. More specifically, we have omitted documents which have been rarely cited by other documents in our collection, as no meaningful citation representations can be used for these documents. Also, we have used a fixed citation window size of 50 terms before and after the citation marker as suggested by Bradshaw [2].

Experimental results shown in this paper present the MAP scores of the system at the Passage level, Document level and Paragraph level (that are paslev, doclev and parlev respectively) for the following system runs:

- Baseline: the best expansion run presented in [7].
- PASSAGE: the baseline system when paragraphs are reduced to answer passages (called passage reduction in the previous section).
- PASSAGE + Doc: the PASSAGE run where the retrieved passages are re-ranked using the *Document* context's similarity scores.
- PASSAGE + Cit: the PASSAGE run where retrieved passages were re-ranked using the *Citation* context's similarity scores.
- PASSAGE + Title: the PASSAGE run where the retrieved passages were re-ranked using the *Title* context's similarity scores.
- PASSAGE + Abstract: the PASSAGE run where the retrieved passages were re-ranked using the *Abstract* context's similarity scores
- PASSAGE + MeSH: the PASSAGE run where the retrieved passages were re-ranked using the MeSH context's similarity scores.

Looking at Table 1, we can see that at the passage (paslev) and paragraph (parlev) MAP scores, show performance improvements over the baseline run. Not only that, but at the passage level MeSH contexts (PASSAGES +MeSH) can marginally outperforms document contexts (PASSAGES + Doc). While, at the document evaluation level (doclev), no representation can obtain better than the PASSAGE + Doc run score.

	paslev MAP		doclev MAP		parlev MAP	
Baseline	0.108		0.534		0.356	
PASSAGES	0.127	17.84%	0.507	-5.05%	0.362	1.78%
PASSAGES + Doc	**0.137**	27.13%	**0.543**	1.67%	**0.384**	8.01%
PASSAGES + Cit	0.119	10.33%	0.500	-6.42%	0.353	-0.83%
PASSAGES + Title	0.126	16.94%	0.508	-4.88%	0.363	2.15%
PASSAGES + Abstract	0.123	14.42%	0.525	-1.71%	0.376	5.75%
PASSAGES + MeSH	**0.138**	28.21%	0.519	-2.74%	0.365	2.71%

Table 1: Table showing the effectiveness of re-ranking passage retrieval results (that is PASSAGES) with other contextual evidence of query/passage similarity.

Table 2 presents a second set of context experiments, where in this case every containing document of a passage is now represented by a combined representation, which combines either its title, abstract or MeSH terms with its citation contexts (if any), that is *PASSAGE + (Title+Cit)*, *PASSAGE + (Abstract+Cit)* and *PASSAGE + (MeSH+Cit)*). For the addition of citation context terms, despite the fact that some minor increases at particular MAP levels can be seen, overall the results are inconsistent. This may be explained by the fact that in many cases we do not have sufficient citation sentences to make up a citation representation for a combining document. A final combination run is also included in this table which combines all contextual information (PASSAGES + (Abs+Title+MeSH+Cit)). However, the PASSAGE+DOC run still performs this. Using a paired Wilcoxon signed-rank test, this PASSAGE+DOC run was found to be statistically significant better (at all MAP levels) when compared with the baseline and PASSAGE runs at the 0.05 confidence interval.

	paslev MAP		doclev MAP		parlev MAP	
Baseline	0.108		0.534		0.356	
PASSAGES	**0.127**	17.84%	0.507	-5.05%	**0.362**	1.78%
PASSAGES + Doc	**0.137**	27.13%	**0.543**	1.67%	**0.384**	8.01%
PASSAGES + (Title+Cit)	**0.124**	15.17%	0.508	-4.88%	**0.363**	2.22%
PASSAGES + (Abstract+Cit)	**0.123**	14.32%	0.526	-1.58%	**0.376**	5.77%
PASSAGES + (MeSH+Cit)	**0.135**	25.43%	0.518	-2.90%	**0.362**	1.86%
PASSAGES + (Abs+Title+MeSH+Cit)	**0.127**	17.87%	0.528	-1.16%	**0.380**	6.79%

Table 2: Table showing additional combinations of context information, which are used to re-rank passages returned by the PASSAGE run.

The results in Table 3 show MAP scores for the top performing systems on the TREC 2006 Genomic Track tasks. TREC_MEDIAN refers to the median values of each MAP score for the official TREC results. Okapi BM25 is the baseline used in the task. Some of the other top performing runs have a detailed description given in [7]. We can see that the majority of MAP scores achieved by our context re-ranking runs outperform the scores of these system, with the exception of UIC_SIGIR and UIC_SIGIR for document level MAP score.

Discussion. In summary, a number of conclusions can be drawn from our experiments:

Run	paslev MAP	doclev MAP	parlev MAP
TREC_MEDIAN	0.037	0.308	0.124
UIC_GenRun3	0.123	0.532	0.342
THU2	0.099	0.434	0.265
NLMinter	0.084	0.473	0.272
UIC_SIGIR	NA	0.539	NA
Okapi	0.048	0.336	0.137

Table 3: Table showing performance of the top performing TREC systems on the Genomics Track.

- The use of the document context brings the best IR performance at passage, document and paragraph level MAPs
- The use of other contextual information, especially abstracts and MeSH terms, can also boost IR performance compared with baseline system, particularly for passage and paragraph level MAPs
- In the absence of the availability of the document context (since the source of many documents is not easily/freely available), the use of other, publically available information contexts (such as MeSH terms, Abstracts and Titles), is a useful way to improve IR performance.
- Use of citation contexts appears promising for improving IR performance. However, it is difficult to obtain citation contexts for most documents. In our collection, only 2.14% of documents could be sufficiently described using citation contexts.

5 Conclusions

A successful implementation of the retrieval ranking method using different types of contextual information can deliver further improvements. In particular, for this passage retrieval task we found that MeSH terms and Document similarity contexts, further boost the performance of an already competent query expansion information retrieval system.

References

[1] B. Billerbeck and J. Zobel. Document expansion versus query expansion for ad-hoc retrieval. In *the 10th ADCS*, pages 34–41, 2005.

[2] S. Bradshaw. Reference directed indexing: Redeeming relevance for subject search in citation indexes. In *the 7th ECDL*, pages 499–510, 2003.

[3] G. Furnas, T. Landauer, L. Gomez and S. Dumais. The vocabulary problem in human-system communication. *Communications of the ACM*, Volume 30, Number 11, pages 964–971, 1987.

[4] W. Hersh, A. Cohen, P. Roberts and H. Rekapalli. Trec 2006 genomics track overview. In *The 15th TREC*, November 2006.

[5] R. Krovetz and W. Croft. Lexical ambiguity and information retrieval. *ACM Trans. Inf. Syst.*, Volume 10, Number 2, pages 115–141, 1992.

[6] A. Ritchie, S. Teufel and S. Robertson. Using terms from citations for ir: Some first results. In *the 30th ECIR*, pages 211–221, 2008.

[7] N. Stokes, Y. Li, L. Cavedon and J. Zobel. Exploring criteria for successful query expansion in the genomic domain. *Information Retrieval*, 2008.