IJCLA VOL. 3, NO. 1, JAN-JUN 2012, PP. 107–119
RECEIVED 24/10/11 ACCEPTED 09/12/11 FINAL 13/06/12

From Co-occurrence to Lexical Cohesion for Automatic Query Expansion

HAZRA IMRAN[1] AND ADITI SHARAN[2]

[1]*Jamia Hamdard, India*
[2]*Jawaharlal Nehru University, India*

ABSTRACT

Designing an efficient Information Retrieval System (IRS) is still a big challenge and an open research problem. To overcome some of the problems of Information retrieval system, researchers have investigated query expansion (QE) techniques to help users in formulating better queries and hence improve efficiency of an Information Retrieval System. The scope of this work is limited to Pseudo Relevance Feedback based Query Expansion. Most of the work done in Pseudo Relevance Based Automatic query expansion is based on selecting the terms using co-occurrence based measures, which has some inherent limitations. Keeping in view limitations of co-occurrence based query expansion; we have tried to explore the utility of lexical based measures for expanding the query. This paper investigates the use of query expansion based on lexical links and proposes an algorithm for Lexical Cohesion Based Query Expansion (LCBQE). Based on theoretical justification and intensive experiments on TREC data set, we suggest that lexical based methods are at least as good as co-occurrence based measures and in some cases may work better than co-occurrence based measures. Depending on the nature of query, lexical based measures have great potential for improving the performance of an information retrieval system.

KEYWORDS: *Information Retrieval System, Pseudo Relevance Feedback, Automatic Query Expansion, Lexical Cohesion, Lexical Links.*

1 INTRODUCTION

An information retrieval system (IRS) is built to satisfy needs of a wide variety of users. Main objective of an IRS is to return maximum number of relevant documents corresponding to the user query, while retrieving minimum number of non relevant documents. However there are many problems in designing efficient IRS such as subjectivity, word mismatch problem and short query. Query expansion has been widely investigated as a method for improving the performance of information retrieval [8,11,12,18]. Theoretically, after query expansion, the performance of an IRS should improve. But practically, this is not always the case. Expanding a query may sometimes introduce a risk of query drift, in which the topicality of the original query may be changed, taking search into a different direction. Therefore a thorough research in Field of Query expansion is desired.

Automatic Query Expansion (AQE) refers to techniques that modify a query without user assistance. We have worked on automatic Query Expansion using Pseudo Relevance Feedback (PRF) which is similar to user relevance feedback but might be done without assistance from the user. (i.e., the approach might be fully automatic) [6]. A PRF based Automatic Query Expansion assumes that the top documents returned by the initial query are relevant ("pseudo-relevance feedback") and expansion terms are extracted from these top-ranked documents. Most of the work done in PRF based automatic query expansion is based on selecting the terms using co-occurrence based measures, which has some inherent limitations.

Keeping in view these limitations we tried to explore the use of lexical cohesion based methods for query expansion. The aim of this work was twofold: to understand the relationship between two document properties: its lexical cohesion and relevance to a query, secondly to investigate whether words that form lexical (cohesive) links between the contexts of query term instances in a document are also good query expansion (QE) terms.

The paper is organized as follows: in the next section (Section 2) basics of lexical cohesion and related terminology is introduced. In Section 3, the application of lexical cohesion for PRF based Automatic Query expansion is discusses and an algorithm for the same is provided. Section 4 presents experiments and their results. These results are analyzed thoroughly to evaluate the effectiveness of proposed method. Finally, Section 5 is dedicated for conclusion and provides suggestions for future work.

2 Lexical Cohesion and Related Terminology

Lexical Cohesion is the cohesion that arises from semantic relationships between words. Segments of text, which are about the same or similar subjects, have higher lexical cohesion, i.e., share a larger number of semantically related or repeating words, than unrelated segments. The strength of lexical cohesion between two words can be useful in determining whether words are used in related contexts. Hence lexical cohesion has been used in identifying collocates. Sinclair and Jones [14] were the first to attempt corpus-based analysis of collocations based on lexical links. The major notions of collocation analysis were introduced in Sinclair [13] and systemized further in terms of 'node', 'collocate', ‘window’ and 'span'. A 'node' is defined as "an item whose total pattern of co-occurrence with other words is under examination". While a ‘collocate’ is “any item which appears with the node within a specified window.” The term ‘span’ is used to refer to the stretch of text (with a predefined window size) around the node within which words are considered to be its collocates.

The identified collocates, as discussed above can be used to find the links between the words. Hoey used the term 'link' to denote an instance of repetition. A single instance of a lexical cohesive relationship between two words is usually referred to as a lexical link [1,5,9]. A lexical link is a relationship between two instances of the same lexeme (simple lexical repetition), its morphological derivatives (complex lexical repetition) or semantically related words (such as hyponyms, synonyms, meronyms, etc).

3 Pseudo Relevance Feedback Based Query Expansion based on Co-occurrence and Lexical Cohesion

3.1 *Co-Occurrence based Query Expansion*

In the majority of works on pseudo-relevance feedback-based automatic query expansion, co-occurrence based approach has been used for selecting query expansion terms. These are the terms that are most frequently co-occurring with the query. Co-occurrence aspects can be captured in different ways. Two methods for extracting terms are used in this paper: one is based on Jacquard coefficient of co-occurring terms and another based on frequency of co-occurring terms [10].

The in depth analysis of co-occurrence based query expansion shows mix chances of success or failure. Thus major drawbacks of co-occurrence based automatic query expansion were investigated. One of the important limitations of co-occurrence based query expansion is that it selects frequently used terms for expansion. A frequently occurring term generally does not allow discriminating between relevant and irrelevant documents, so it is not good for expansion. If the co-occurring terms are selected from top ranked documents, discrimination does occur to a certain extent. However still there are chances that a term that is frequent in top n relevant documents is also frequent in entire collection.

Another limitation of co-occurrence based measure is that terms co-occurring with individual query terms are selected first. These results are then combined to select final expansion terms. In such cases if some query term dominates, most of the co-occurring terms may come from subset of these terms. In most of the cases it may be desirable to select those terms which are co-occurring with most of the query terms and at different instances of occurrence of query terms. This aspect can be captured by lexical links. Therefore, motivation of this work was to explore the use of lexical links for automatic query expansion.

3.2 *Query Expansion based on Lexical Links*

Lexical links [1,5,9] have been found useful in finding words that are related in certain context. However, most of the IRSs make use of lexical relations to a limited extent. The basic research question of this work is whether the use of lexical cohesion and lexical links for query expansion can improve performance of information retrieval. The use of lexical cohesion for query expansion is based on following intuition. A user query is likely to describe a relevant topic. Therefore in a relevant document terms correspond to same topic and they tend to cohere with each other and have similar collocation environment. Whereas in a non-relevant document, the occurrence of these query terms is not motivated by the presence of a relevant topic, but due to other factors, therefore they are less likely to occur in the same semantic context.

The query expansion methods presented in this paper rely on the method of calculating lexical cohesion between query terms' contexts in a document introduced by Vechtomova [16, 17]. In Vechtomova [16] it was suggested that simple lexical repetition alone performed as well as the use of repetition plus semantically related words (determined using Word Net).Their observation was that much difference was not found in the above two cases. Therefore in this work, simple lexical repetition is used to identify the lexical links.

Our main interest was in finding those terms from the top N relevant documents which are lexically cohesive to the context of query terms. Lexical cohesion between query terms' contexts is calculated by counting the number of lexical links between them. The context of a query term in a document is defined as a set of stemmed non-stop terms extracted from fixed-sized windows around each occurrence of the query term in the document. All the context terms of a query term are combined to form collocates for that query term. Collocation environments of the query terms were then compared to find terms having lexical links. The terms having lexical links were termed as link terms. These link terms were then re-ranked based on number of lexical links (Equation 1) and top N link terms were used to reformulate the initial query.

The weighting criteria used in this paper is also innovative in a sense that in most of the work studied by us the weights are given to expanded query terms based on their weight in document and not on the importance of the expanded query term with respect to original query term. In this work weights were assigned to expanded query terms basing on rank of the expanded query term, giving more importance to those terms that are more meaningful. For the expanded terms the following criteria was used for weighting k^{th} expansion term basing on number of lexical links:

$$weight(t_k) = \frac{number\_of\_lexical\_links(t_k)}{\max\limits_{1 \le i \le m}(number\_of\_lexical\_links(t_i))} \qquad (1)$$

Once weights are given to expanded query a new vector $\vec{eq}$ containing expansion terms can be constructed. Now the new query can be represented as

$$\vec{q} = \vec{q} + \vec{eq} \qquad (2)$$

Note that $\vec{eq}$ does not contain any of the original query terms.

On the basis of above motivation following idea we propose following algorithm for Lexical Cohesion based query expansion (LCBQE). The input to the algorithm is: documents and query; output is the expanded query. The algorithm is as follows:

- Represent documents and query in vector space model.
- Using tf-idf weighting scheme, give weights to the document and query terms.
- Match the query terms with the documents using okapi similarity measure.
- Retrieve the top *N* documents corresponding to the initial query.
- For each top *N* documents do:

  For each query term q_i do:
  - Identify a snippet around each instance that contains 3 non-stop words before and after it in document.
  - Add the snippet to the collocates of the query term.
- Find all link-terms from the collocates of all individual query terms.
- Rank the link terms by *idf.*
- Extract the top *N* link terms.
- Assign weight to the link term based on the number of links.
- Expand the initial query with these extracted link terms.
- Return the expanded query.

4 Experiments and Results

Experiments were conducted on TIPSTER document collection,(TREC data set) a standard test collection in the IR community. Volume 1 is a 1.2 GByte collection of full-text articles and abstracts. The documents came from the following sources:

- WSJ – Wall Street Journal (1986, 1987, 1988, 1989, 1990, 1991, and 1992),
- AP – AP Newswire (1988,1989 and 1990),
- ZIFF – information from Computer Select disks (Ziff-Davis, Publishing),
- FR – Federal Register (1988),
- DOE – short abstracts from Department of Energy.

For queries: 50 queries were used that were formed through 50 TREC topic sets. Average word length for query was 2.3 words.

4.1 *Experiments using Co-occurrence and Lexical Cohesion Based Query Expansion*

The purpose of experiments was to find out the usefulness of the lexical information for automatic query expansion and compare it with baseline and co-occurrence based query expansion. For Co-occurrence based query expansion Jaccard coefficient based and frequency based measures were used for selecting the expansion terms. The terms were then re-ranked with entropy based measure. For lexical based query expansion local collocates were extracted from the windows around every occurrence of the query terms in top *N* documents as discussed in above algorithm. Further link terms were extracted from local collocates. Top n link terms were used expanding the query. The weights of candidate expansion terms were assigned based on the number of lexical links.

Parameters used in the experiments were: number of top *N* documents, window size (for lexical cohesion based expansion), and number of top ranked terms used for expanding the query. Values of parameters were decided empirically after intensive experimentation. In our experiments we used top *N* documents = 100, window size as 3 non-stop words on either side of query term, and number of top ranked terms were 10. The results are shown through Table 1 and Figure 1.

Table 1. Results Without Query Expansion, Co-occurrence and Lexical Based Query Expansion

	Without query expansion	Co-occurrence Method		Lexical Link Method
		Jaccard	Freq	
Number of Queries (num_q)	50	50	50	50
Number of Retrieved Documents (num_ret)	5000	5000	5000	5000
Number of relevant documents (num_rel)	16386	16386	16386	16386
Number of relevant retrieved documents (num_rel_ret)	1156	1237	1220	**1256**
Map	0.0443	0.0602	.0592	**.0611**
gm_map	0.0079	0.0102	.010	0.01
Rprec	0.0973	0.1032	.1032	**0.1043**

Table 1 shows the result of experiments for the baseline (without query expansion), co-occurrence, and lexical cohesion based query expansion respectively. The best results are shown in boldface. The MAP of lexical is 0.0611, which is better than co-occurrence and baseline. Similarly, Rprec also shows better result when compared with other two methods.

Figure 1 shows the comparison of recall precision graph between unexpanded query and query expansion based on lexical links. From Figure 1 it is observed that expansion of initial query with statistically significant local collocates following pseudo relevance feedback results in significant performance improvement over unexpanded under the same conditions.

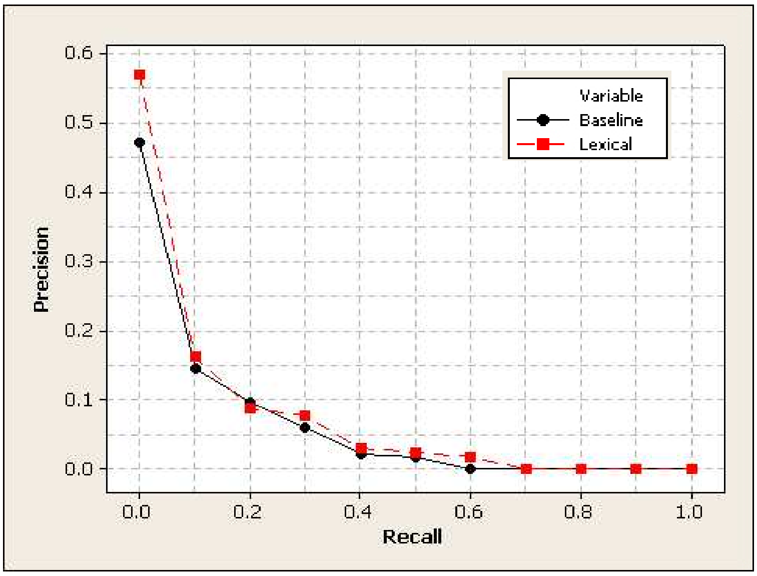

Figure. 1. Recall Precision Curve showing result of experiment for QE based on Base line and Lexical Cohesion

From paired t-test it can be concluded that the means differ at the 0.05 level of significance. The mean of lexical is significantly different from the baseline. (p<.005) The true difference is between 0.0078235 and 0.025844.

Apart from overall analysis of result an in depth analysis of result was done for individual queries, specially focusing on queries for which there was no significant improvement in result. Figure 2 shows the differences in average precision of individual queries expanded using LCBQE method from the unexpanded query. Here an attempt is to analyze the performance of individual queries from baseline to lexical method is done.

The results given in Figure 2 show that in a considerable number of cases lexical method yielded the most gain in performance (74% queries shows improvement). This suggests that there is a room for better improvement in the retrieval if lexical terms are used for expansion.

There result was analyzed for the queries for which result does not show significant improvement. Here, it was found that the lexical terms are not adding more meaning to the original query. Example in query 88 (*Crude oil Price Trends*) the lexical terms found were: *hishan, countries, megaprojects, 5084, 5209, 5273*. These terms are not adding significant meaning to the original query and hence the performance degrades.

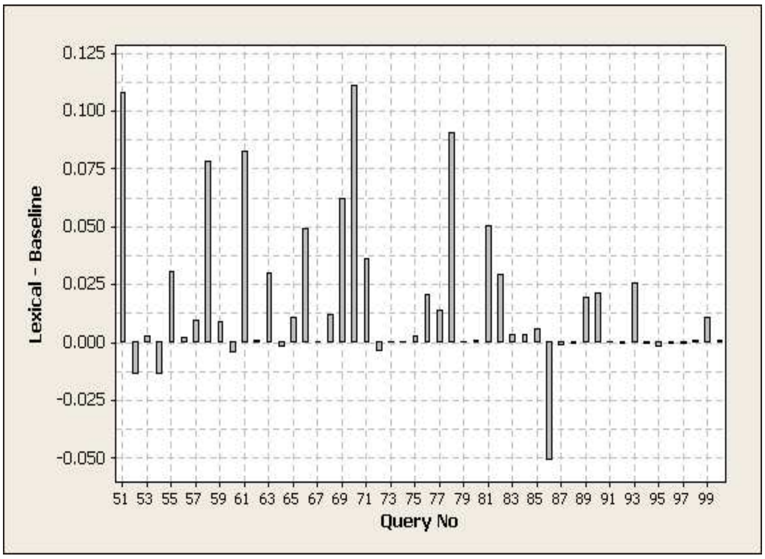

Figure. 2. Differences in average precision of individual queries after query expansion from the unexpanded query

It is observed that results obtained by lexical cohesion based methods are almost similar to those obtained by co-occurrence based automatic query expansion methods. As lexical cohesion based methods have not been explored much for automatic query expansion, the results of this work motivates us to perform in-depth analysis of these methods. An investigation into individual query performance presents another significant aspect of lexical based automatic query expansion method. This aspect as percentage of queries for which the result improved is the greatest in this method (74%). Therefore, it can be said that the chances of failure of this method in expanding the query are comparatively less than other methods.

4.2 *Analysis of Relation between Lexical Cohesion and the Relevance of Documents*

In order to gain some qualitative understanding of the relation between lexical cohesion and the relevance property of documents, the top N documents retrieved in initial run and retrieved after automatic query expansion (based on lexical links) were compared. It was examined on a small sample from the top *N* (100) documents. Although, it is not possible to generalize, experiments suggest that certain patterns of lexical links could be identified in the documents promoted and demoted after initial query expansion with lexically cohesive terms. In the set examined, it was noticed that documents that are promoted, contain most of the query terms, and there are several instances of each of them. It also appears that the instances of query terms are spread throughout the documents, i.e., they are not concentrated in isolated sections of the documents. As expected, the instances of query terms are well connected by lexical links in the promoted documents.

An example of this type of document is SJMN91-06252124, retrieved in response to query “Natural Language Processing” (query 66). There are many instances of the query terms (11 instances of “natural,” 5 instances of “language” and 3 instances of Processing) in the text and they are extensively connected with each other by lexical links.

In the demoted documents, it was seen three different patterns. Some of the demoted documents are made up of disjoint pieces of text that cover separate and unrelated information. In second type of demoted document, query terms (such as “Natural Language Processing”) occur but not all in the same context. In the third type of demoted documents, the query topic is treated marginally.

4.3 *Analysis of Terms Obtained for Co-Occurrence and Lexical Cohesion based AQE*

The quality of terms extracted with both the methods:-co-occurrence and lexical cohesion was analyzed. For example for Query 87 (*Criminal Actions Against Officers of Failed Financial Institutions*) the Co-occurring terms (*failed actions, enforcement, justice, charges, prosecutors, civil, convicted, investigation, alleged, attorney, jury*) are more relevant than the lexical terms (*officers, criminal, failed, actions, mijalis, hallada, khoo, neidorf, esm, lytl, zodiac, dmv, disheartening, mndrg*). However, this trend is only seen in 32% of queries. In rest 68% queries result is improving with lexical method. For example in query 67(Politically Motivated Civil Disturbances) the lexical terms (*pungently, hurbon, meacher, fitzhugh, corporacion, cispes, insinuated, mouawad, deceitful, amnesties*) appears to be more relevant than co-occurring terms (*liberties, unrest, violence, racial, criminal, rights, racially, blacks, riots, rioting*).

5 CONCLUSION

Most of the work done for PRF based query expansion till now has been based on using co-occurrence based measures for selecting expansion terms. Co-occurrence based query expansions has certain limitations. The research question investigated in the paper was whether the use of lexical link information can improve performance of PRF based Automatic Query Expansion and whether it can overcome some of the limitations of co-occurrence based query expansion.

This paper proposed a new method for query expansion: Lexical Cohesion Based Query Expansion (LCQBE) and provided an algorithm for the same. The experiments were done to compare the results with co-occurrence based query expansion. Experiments were performed on standard TREC data set. It was found that result of lexical based query expansion is generally as good as query expansion using co-occurrence methods, in some cases it is even better. As lexical cohesion based query expansion is a comparatively unexplored topic, there is an immense potential for exploring the usefulness of these methods in order to improve information retrieval efficiency.

An in depth analysis of the results obtained in this paper motivates us to work in new direction. In this work we may be able to find out different types of queries (categorized on some basis) for which one can guess that whether retrieval efficiency can be improved by co-occurrence based query expansion, lexical based query expansion, by none of them or both of them. An important contribution of this work can be that we can use a switch to decide a priori that which of the query expansion (co-occurrence or lexical) can be more useful for a specific query or the query should not be extended at all.

REFERENCES

1. Ellman, J., and Tait, J. (1998). Meta searching the web using exemplar texts: Initial results. Proceedings of the 20th BCSIRSG.
2. Halliday, M. A., and Hasan, R. (1976). Cohesion in English. Longman.
3. Hearst, M. (1994). Multi-paragraph segmentation of expository text. Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics.
4. Hirst, G., and St Onge, D. (1997). Lexical chains as representation of context for the detection and correction of malapropisms. MIT Press.
5. Hoey, M. (1991). Patterns of Lexis in Text. Oxford University Press.
6. Imran, H. and Sharan, A. (2010). A Framework for Automatic Query Expansion, Lecture Notes in Computer Science (LNCS), Volume 6318, 2010, Springer, 386–393.
7. Manabu, O., and Hajime, M. (2000). Query-biased summarization based on lexical chaining. Computational Intelligence, 578–585.
8. Molto, M., and Svenonious, E. (1991). Automatic Recognition of title page names. Information Processing and Management, 83–95.
9. Morris, J., and Hirst, G. (1991). Lexical cohesion computed by thesaural relations as an indicator of the structure of text. Computational Linguistics, 21–48.
10. Rijsenberg, C. J. (1979). Information Retrieval (Second, Ed.) Butterworth Heinemann.
11. Sakai, T., and Robertson, S. E. (2001). Flexible pseudo-relevance feedback using optimization tables, Louisiana, 396–397.
12. Salton, G. (1998). Automatic text processing: the transformation, analysis, and retrieval of information by computer. Addison-Wesley.
13. Sinclair, J. M. (1991). Corpus, concordance, collocation. Oxford University Press.
14. Sinclair, J. M., and Jones, S. (1996). Lexis and Lexicography. Singapore: UniPress.
15. Stairmand, M. A. (1997). Textual context analysis for information retrieval. Procedings of the ACM SIGIR , 140–147.

16. Vechtomova, O. (2006). Noun phrases in interactive query expansion and document ranking. Information Retrieval, 399–420.
17. Vechtomova, O., Karamuftuoglu, M., and Robertson, S. E. (2006). On document relevance and lexical cohesion between query terms. Information Processing and Management, 1230–1247.
18. Witten, I., Moffat, A., and Bell, T. (1999). Managing Gigabytes: Compressing and Indexing Documents and Images. Morgan Kaufmann.

HAZRA IMRAN
DEPARTMENT OF COMPUTER SCIENCE,
JAMIA HAMDARD ,
NEWDELHI, INDIA
E-MAIL: <HIMRAN@JAMIAHAMDARD.AC.IN>

ADITI SHARAN
SCHOOL OF COMPUTERS AND SYSTEM SCIENCES,
JAWAHARLAL NEHRU UNIVERSITY,
NEW DELHI, INDIA
E-MAIL: <ADITISHARAN@MAIL.JNU.AC.IN>