Approaches to Using Word Collocation in Information Retrieval

Olga Vechtomova

Thesis submitted for the degree of Doctor of Philosophy

Department of Information Science
City University
London

November 2001

Contents

Acknowledgements viii
Declaration of copyright ix
Abstract x

1. Introduction 1

2. Linguistic perspective of text and lexis 4
2.1 Natural language text 4
2.1.1 Text organisation 4
2.1.2 Text elements 6
2.1.3 Cohesion in text 9
2.2 Lexis 10
2.2.1 Lexical unit and lexical meaning 10
2.2.2 Syntagmatic and paradigmatic relations 12
2.3 Lexical cohesion 13
2.3.1 Types of lexical cohesion 14
2.3.2 Lexical links and lexical chains 17
2.3.3 Bonds between sentences and nets of bonds 20
2.4 Collocation 23
2.4.1 Definitions of collocation 23
2.4.2 Patterns of collocation in text 26

3. Natural language text in IR context 30
3.1 Document and query representation 30
3.1.1 Controlled language indexing 31
3.1.2 Natural language indexing 33
3.2 Statistical approach vs. natural language processing (NLP) approach to IR . 34
3.2.1 Limitations and benefits of statistical methods and NLP in IR 34
3.2.2 Phrases. Statistical and linguistically-motivated approaches 38
3.3 Approaches to collocation in IR 40
3.3.1 Document-wide co-occurrences 42
3.3.2 Long-span collocations 44
3.4 Applications of lexical cohesion in IR 49
3.5 Document passages 51

4. Probabilistic information retrieval 55
4.1 Robertson & Sparck Jones model of probabilistic IR 55
4.1.1 Term relevance weighting 56
4.1.2 Relevance feedback 59
4.2 Term independence assumption 60

5. Collocation analysis in this project 62
5.1 Research question 62
5.2 Overview of the experiments 63
5.2.1 Global collocation analysis 64
5.2.2 Local collocation analysis 65
5.2.3 Lexical cohesion analysis using local collocations 67

5.3 Windowing technique 67
5.4 Choice of statistical measures of collocation significance 70
5.4.1 Mutual information (MI) 71
5.4.2 Z score 73
5.5 Document collection and topics 74
5.5.1 Choice of the database 74
5.5.2 Indexing the database 75
5.5.3 Topics 77

6. Global collocation analysis experiments 79

6.1 Introduction 79
6.2 Construction of the database of global collocates 79
6.3 Lexical-semantic analysis of collocations 82
6.3.1 Differences between collocates selected with MI and Z statistics 82
6.3.2 Collocates of polysemantic words 84
6.3.3 Comparison of collocates with terms from engineered term networks .85
6.4 Retrieval experiments 87
6.4.1 Experimental design 87
6.4.2 Analysis of results 88
6.5 Statistical analysis 89
6.5.1 Presence of global collocates in relevant documents 89
6.5.2 Regression analysis experiments 90
6.6 Concluding remarks 92

7. Local collocation analysis experiments 94

7.1 Introduction 94
7.2 Query expansion with local collocates ranked by Z score 94
7.2.1 Experimental design 95
7.2.2 Analysis of results 97
7.3 Collocation relevance weighting 104
7.3.1 Experimental design 104
7.3.2 Analysis of results 106
7.4 Evaluation of performance by categories of terms in the expanded query . . 106
7.5 Concluding remarks 113

8. Lexical cohesion analysis using local collocations 116

8.1 Introduction 116
8.2 Comparison of relevant and non-relevant sets by the level of lexical cohesion 116
8.2.1 Experimental design 116
8.2.2 Analysis of results 121
8.3 Re-ranking of document sets by lexical cohesion scores 122
8.3.1 Experimental design 122
8.3.2 Analysis of results 123
8.4 Concluding remarks 124

9. Conclusions and recommendations 126

Appendices **136**

A Global collocation analysis **137**
A.1 Selected lists of top collocates ranked by MI and Z 137
A.2 Trec_eval performance results of global runs 153
A.3 Regression analysis results 154

B Local collocation analysis **172**
B.1 Trec_eval performance results of retrospective MI/Z collocate runs 172
B.2 Trec_eval performance results of predictive MI/Z collocate runs 177
B.3 Trec_eval performance results of CRW runs 183
B.4 Selected expanded queries from the run 'PRED 8 Z COL + 20 OK (100 window size)' 184
B.5 Evaluation of performance by categories of terms in the expanded queries of the run 'PRED 16 MI COL + 20 OK (100 window size)' 195

C Lexical cohesion analysis using local collocations **201**
C.1 Mean and standard deviation of Okapi scores in the aligned sets of relevant and non-relevant documents 201
C.2 Distribution of lexical cohesion scores in the relevant and non-relevant sets 206
C.3 Trec_eval performance results of re-ranking Okapi sets by *COMB-LCS* . . 215

D Programs **218**
D.1 Selected Perl scripts for global collocation analysis 218
D.2 Selected Perl scripts for local collocation analysis 223
D.3 GSL codes 234

References **235**

List of Figures

2.1 Stratal organisation of language elements 6
2.2 Semantic structure types by [Skorochod'ko72, p. 1180] 22
2.3 Distance distribution of token types in the windows of 50 words either sides of the nodes 27
5.1 Window around node *x*, defined as spans of *S* words to the left/right of *x* . . 68
5.2 Window truncated by hitting the document boundary 68
5.3 Right-hand half of the window truncated by hitting another occurrence of x after the node 69
5.4 Left-hand half of the window ignored when another occurrence of *x* is found before the node 69
6.1 Query expansion with global collocates 88
7.1 Overlapping windows of two different query terms *x* and *y* 105
7.2 Influence of categories of terms in the expanded queries 'PRED 16 Z COL + 20 OK (200 window size)' on average precision 109
7.3 Distribution of precision differences for category 1. *Collocate* 110
7.4 Distribution of precision differences for Category 2: *Collocate of 2 or more query terms* 110
7.5 Distribution of precision differences for Category 3: *Collocate of 1 query term* 111
7.6 Distribution of precision differences for Category 4: *Okapi RF term* 111
7.7 Distribution of precision differences for Category 5: *Original query term* . 112
7.8 Distribution of precision differences for Category 9: *Collocate and an original query term* 112
7.9 Distribution of precision differences for Category 10: *Okapi term and an original query term* 113
8.1 Links between instances of common collocates in merged windows of query terms **x** and **y** 118

List of Tables

6.1 Top 20 collocates of the query term *cigarette* (topic 257) ranked by Z score 81
6.2 Top 20 collocates of the query term *alien* (topic 252) ranked by Z score . . . 81
6.3 Lists of top collocates for the term *acquire* sorted by MI and Z statistics . . . 82
6.4 Lists of top collocates for the synonym group @0104 (*environment, environmental*) sorted by MI and Z statistics 83
6.5 Lists of top collocates for the term *education* sorted by MI and Z statistics . 83
6.6 Lists of top collocates for the term *nitrogen* sorted by MI and Z statistics . . 84
6.7 Lists of top collocates for the term *gene* sorted by MI and Z statistics 84
6.8 Lists of top collocates for the term *pyramid* sorted by MI and Z statistics . . 85
6.9 Lists of top collocates for the term *pressure* sorted by MI and Z statistics . . 86
6.10 Terms related to the above two senses of the term *pressure* in WordNet . . . 86
6.11 Top terms in Z-ranked collocation list matching INSPEC and WordNet terms related to the term *fuel* 87
6.12 Summary of retrieval results for query expansion with global collocates . . . 89
6.13 Presence of global collocates in relevant documents 90
6.14 Summary for regression using MI, Z, JF and NOPOS to predict OW 91
6.15 Summary of retrieval results for PREDICTED OW run 92
7.1 Retrospective performance results of Okapi runs 97
7.2 Retrospective performance results of query expansion runs with local collocates ranked by global Z 98
7.3 Retrospective performance results (in average precision) of query expansion runs with local collocates ranked by local Z 99
7.4 Retrospective performance results of query expansion runs with local collocates ranked by global MI 100
7.5 Retrospective performance results (in average precision) of query expansion runs with local collocates ranked by local MI 100
7.6 Predictive performance results of Okapi runs 101
7.7 Predictive performance results of query expansion runs with local collocates ranked by global Z 101
7.8 Predictive performance results (in average precision) of query expansion runs with local collocates ranked by local Z 102
7.9 Predictive performance results of query expansion runs with local collocates ranked by global MI 103
7.10 Predictive performance results (in average precision) of query expansion runs with local collocates ranked by local MI 103
7.11 Retrospective performance results (in average precision) of query expansion runs with local collocates ranked by *CRW * c* 106
7.12 Predictive performance results (in average precision) of query expansion runs with local collocates ranked by *CRW * c* 106
7.13 Influence of categories of terms in the expanded queries 'PRED 8 Z COL + 20 OK (100 window size)' on average precision 108
7.14 Influence of categories of terms in the expanded queries 'PRED 16 Z COL + 20 OK (200 window size)' on average precision (in percentage) 108
8.1 Difference between the non-aligned relevant and non-relevant sets (documents taken from the top 100 Okapi documents) 121

8.2 Difference between the aligned relevant and nonrelevant sets (documents selected from the top 100 Okapi documents) 121
8.3 Difference between the aligned relevant and non-relevant sets (documents selected from the top 1000 Okapi documents) 121
8.4 Results (in average precision) of re-ranking Okapi document sets by COMB-LCS 123

Acknowledgements

I am thankful to everyone who contributed to this project and helped to make it possible. In particular I would like to thank:

Steve Robertson and Susan Jones for their inspiring ideas, wise supervision, support and guidance.

Steve Walker for his advice and help with Okapi.

My parents and friends for their support and encouragement.

Declaration of Copyright

I grant powers of discretion to the University Librarian to allow this thesis to be copied in whole or in part without further reference to me. This permission covers only single copies made for study purposes, subject to normal conditions of acknowledgement.

Abstract

The thesis explores long-span collocation and its application in information retrieval. The basic research question of the thesis is whether the use of long-span collocates can improve performance of a probabilistic model of IR. The model used in the project is the Robertson & Sparck Jones probabilistic model.

The basic research question was explored by investigating three different ways of integrating collocation information with the probabilistic model:

1. Global collocation analysis. The method consists in expanding the original query with long-span global collocates of query terms. Global collocates of a query term are selected from large fixed-size windows around all occurrences of a term in the corpus and ranked by statistical measures of Mutual Information (MI) and Z score. A fixed number of top-ranked collocates is used in query expansion.

   Query expansion with global collocates did not show to be superior to the original queries, the possible reason being the fact that query terms often have a fairly broad meaning and, hence, a rather semantically heterogeneous pattern of occurrence.

2. Local collocation analysis. This method is a form of iterative query expansion following relevance or pseudo-relevance (blind) feedback. The original query is expanded with the query terms' collocates which are extracted from the long-span windows around all occurrences of query terms in the known relevant documents, and selected using statistical measures of MI and Z. Some parameters whose effect was systematically studied in this experiment set are: window size, measure of collocation significance for collocate ranking, number of query expansion collocates and categories of terms in the expanded queries.

   Some results showed a tendency towards performance gain over relevance feedback in the probabilistic model, however it was not significant enough to conclude that this method is superior to the existing relevance feedback used in the model.

3. Lexical cohesion analysis using local collocations. This experiment set aimed to explore whether the level of lexical cohesion between query terms in a document can be linked to the document's relevance property, and if so, whether it can be used to predict documents' relevance to the query. Lexical cohesion between different query terms is estimated from the number of collocates they have in common.

   The experiments proved that there exists a statistically significant association between the level of lexical cohesion of the query terms in documents and relevance. Another set of experiments, aimed at using lexical cohesion to improve probabilistic document ranking, showed that sets re-ranked by their lexical cohesion scores have similar performance as the original ranking.