A Study of Document Relevance and Lexical Cohesion between Query Terms

Olga Vechtomova
University of Waterloo
Waterloo, Canada
ovechtom@uwaterloo.ca

Murat Karamuftuoglu
Bilkent University
Ankara, Turkey
hmk@bilkent.edu.tr

Stephen Robertson
Microsoft Research Cambridge
Cambridge, UK
ser@microsoft.com

ABSTRACT
Lexical cohesion is a property of text, achieved through lexical-semantic relations between words in text. Most information retrieval systems make use of lexical relations in text only to a limited extent. In this paper we empirically investigate whether the degree of lexical cohesion between the contexts of query terms' occurrences in a document is related to its relevance to the query. Experiments suggest significant differences between the lexical cohesion in relevant and non-relevant document sets exist.

Categories and Subject Descriptors
H.3.3 [**Information Search and Retrieval**]: search process

General Terms
Experimentation, Theory.

Keywords
Lexical cohesion, Information Retrieval, Relevance, Collocation.

1. INTRODUCTION
Word instances in text depend to various degrees on each other for the realisation of their meaning. For example, closed-class words (such as pronouns or prepositions) rely entirely on their surrounding words to realise their meaning, while open-class words, having meaning of their own, rely on other open-class words in the document to realise their contextual meaning. As we read, we process the meaning of each word we see in the context of the meanings of the preceding words in text, thus relying on the lexical-semantic relations between words to understand it. Lexical-semantic relations between open-class words form the *lexical cohesion* of text, which helps us perceive text as a continuous entity, rather than as a set of unrelated sentences.

Lexical cohesion is a major characteristic of natural language texts, which is achieved through semantic connectedness between words in text, and expresses continuity between the parts of text [1]. Lexical cohesion is not the same throughout the text. Segments of text which are about the same or similar subjects (topics) have higher lexical cohesion, i.e. share a larger number of semantically-related or repeating words, than unrelated segments.

In this paper we investigate the lexical cohesion property of texts, specifically, whether there is a relationship between relevance and lexical cohesion between query terms in documents. We also report preliminary experiments to investigate whether lexical cohesion property of texts can be useful in helping IR systems to predict the likelihood of a document's relevance. From a linguistic point of view, the main problem in ad-hoc IR can be seen as matching two imperfect textual representations of meaning: a query, representing user's information need, and a document, representing author's intention. Obviously, the fact that a document and a query have matching words does not mean that they have similar meanings. For example, query terms may occur in semantically unrelated parts of text, talking about different subjects. Intuitively, it seems plausible that if we take into consideration lexical-semantic relatedness of the contexts of different query terms in a document, we may have more evidence to predict the likelihood of the document's relevance to the query. This paper sets to empirically investigate this idea.

We hypothesise that relevant documents tend to have a higher level of lexical cohesion between different query terms' contexts than non-relevant documents. This hypothesis is based on the following premise: In a relevant document, all query terms are likely to be used in related contexts, which tend to share many semantically-related words. In a non-relevant document, query terms are less likely to occur in related contexts, and hence share fewer semantically-related words.

The goal of this study is to explore whether the level of lexical cohesion between different query terms in a document can be linked to the document's relevance property, and if so, whether it can be used to predict the document's relevance to the query. Initially we formulated a hypothesis to investigate whether there is a statistically significant relation between two document properties – its relevance to a query and lexical cohesion between the contexts of different query terms occurring in it.

Hypothesis 1: There exists statistically significant association between the level of lexical cohesion of the query terms' contexts in documents and relevance.

We conducted a series of experiments to test the above hypothesis. The results of the experiments show that there is a statistically significant association between the lexical cohesion of query terms in documents and their relevance to the query. This result suggested the next step of our investigation: evaluation of the usefulness of lexical cohesion in predicting documents' relevance. We hypothesised that re-ranking document sets retrieved in response to the user's query by the documents' lexical cohesion property can yield better performance results than a term-based document ranking technique:

Hypothesis 2: Ranking of a document set by lexical cohesion scores results in significant performance improvement over term-based document ranking techniques.

The rest of the paper is organised as follows: in the next section we discuss the concept of lexical cohesion and review related work in detail; in section 3 we present the experiments comparing

the degrees of lexical cohesion between sample sets of relevant and non-relevant documents; in section 4 we describe experiments studying the use of lexical cohesion in document ranking; finally, section 5 concludes the paper and provides suggestions for future work.

2. LEXICAL COHESION IN TEXT

Halliday and Hasan introduced the concept of 'textual' or 'text-forming' property of the linguistic system, which they define as a "set of resources in a language whose semantic function is that of expressing relationship to the environment" [1, p.299]. They claim that it is the meaning realised through text-forming resources of the language that creates text, and distinguishes it from the unconnected sequences of sentences. They refer to text forming resources in language by the broad term of c*ohesion*. The continuity created by cohesion consists in "expressing at each stage in the discourse the points of contact with what has gone before" [1, p.299]. There are two major types of cohesion: (1) *grammatical*, realised through grammatical structures, and consisting of the cohesion categories of reference, substitution, ellipsis and conjunction; and (2) *lexical* cohesion, realised through lexis [1]. Halliday and Hasan distinguished two broad categories of lexical cohesion: *reiteration* and *collocation*. Reiteration, as defined in [1], refers to a broad range of relations between a lexical item and another word occurring before it in text, where the second lexical item can be an exact repetition of the first, a general word, its synonym or near-synonym or its superordinate. As for the second category – collocation, Halliday and Hasan understand that this is a relationship between lexical items that occur in the same environment, but they fail to formulate a more precise definition.

Later, some linguists narrowed down the meaning of collocation to refer only to restricted type of collocations, whose meaning cannot be completely derived from the meaning of their elements. For example Manning and Schütze [2] defined collocation as grammatically bound elements occurring in a certain order which are characterised by limited compositionality, i.e. the impossibility of deriving the meaning of the total from the meanings of its parts.

We recognise two major types of collocation:

1. Collocation due to lexical-grammatical or habitual restrictions. These restrictions limit the choice of words that can be used in the same grammatical structures with the word in question. Collocations of this type occur within short spans, i.e. within the bounds of a syntactic structure, such as a noun phrase, (e.g. “rancid butter”, “white coffee”, “mad cow disease”).
2. Collocation due to a typical occurrence of a word in a certain thematic environment: two words hold a certain lexical-semantic relation, i.e. their meanings are close semantically, therefore they tend to occur in the same topics in texts. Beeferman et al. experimentally determined that long-span collocation effects can extend in text up to 300 words [3]. Vechtomova et al. report examples of long span collocates identified using the Z-score such as “environment–pollution”, “gene–protein” [4].

Hoey [5] gave a different classification of lexical cohesive relationships under a broad heading of *repetition*: (1) simple lexical repetition, (2) Complex lexical repetition, (3) Simple partial paraphrase, (4) Simple mutual paraphrase, (5) Complex paraphrase, (6) Superordinate, hyponymic and co-reference repetition.

In this work we investigate the relationship between relevance and the level of lexical cohesion among query terms based on the simple lexical repetition of their long span collocates.

2.1 LEXICAL LINKS AND CHAINS

A single instance of a lexical cohesive relationship between two words is usually referred to as a *lexical link* [5, 6, 7, 8]. Lexical cohesion in text is normally realised through sequences of linked words – *lexical chains*. The term *'chain'* was first introduced by Halliday and Hasan [1] to denote a relation where an element refers to an earlier element, which in turn refers to an earlier element and so on.

Morris and Hirst [6] define lexical chains as sequences of related words, which have distance relations between them. One of the prerequisites for the linked words to be considered units of a chain is that they should co-occur within a certain span. Hoey [5] suggested using only information derivable from text to locate links in text, Morris and Hirst used Roget's thesaurus in identifying lexical chains. Morris and Hirst’s algorithm was later implemented for various tasks: IR [9], text segmentation [10] and summarisation [11].

2.2 LEXICAL BONDS

Hoey [5] pointed that text cohesion is built not only of links between words, but also of semantic relationships between sentences. He argued that if sentences are not related as whole units, even though there are some lexically linked words found in them, they are no more than a disintegrated sequence of sentences sharing a lexical context. He emphasised that it is important to interpret cohesion by taking into account the sentences where it is realised. For example, two sentences in text can enter the relation, where the second one exemplifies the statement expressed in the previous sentence. Sentences do not have to be adjacent to be related, and lexical cohesive relation can connect several sentences.

A cohesive relation between sentences was termed by Hoey as a *lexical bond*. He defines a bond between sentences as a sufficient number of lexical links between them. The number of lexical links the sentences must have to be bonded is a relative parameter, according to Hoey, depending indirectly on the relative length and the lexical density of the sentences. Hoey argues that an empirical method for estimating a minimum number of links the sentences need to have to form a bond must rely on the proportion of sentence pairs that form bonds in text. If the proportion of sentences linked by any given number of links is too high, then it is important to increase the cut-off point, until the degree of connection is not above average. In practice two or three links are considered sufficient to constitute a bond between a pair of sentences.

It is notable that in Hoey's experiments, only 20% of bonded sentences were adjacent pairs. Analysing non-adjacent sentences, Hoey made and proved two claims about the meaning of bonds. The first claim is that bonds between sentences are indicators of semantic relatedness between sentences, which is more than the sum of relations between linked words. The second claim is that a large number of bonded sentences are intelligible without

recourse to the rest of the text, as they are coherent and can be interpreted on their own [5].

3. COMPARISON OF RELEVANT AND NON-RELEVANT SETS BY THE LEVEL OF LEXICAL COHESION

3.1 EXPERIMENTAL DESIGN

Our method of estimating the level of lexical cohesion between query terms was inspired by Hoey's method [5] of identifying lexical bonds between sentences. There is, however, a substantial difference between the aims of these two methods. Sentence bonds analysis is aimed at finding semantically related sentences. Our method is aimed at predicting whether query terms occurring in a document are semantically related, and measuring the level of such relatedness.

In both methods the similarity of local context environments is compared: in our method – fixed-size windows around query terms; in Hoey's method – sentences. Hoey's method identifies semantic relatedness between sentences in a text, whereas the objective of our method is to determine the semantic similarity of the contextual environments, i.e., collocates, of different query terms in a document.

To determine semantic similarity of the contextual environments of query terms we combine all windows for one query term, building a merged window for it. Each query term's merged window represents its contextual environment in the document. We then determine the level of lexical cohesion between the contextual environments of query terms**.** We experimented with two methods to determine the level of lexical cohesion between different query terms: (a) How many lexical links connect them, and (b) How many types they have in common. Each document is then assigned a *lexical cohesion score* (*LCS*), based on the level of lexical cohesion between different query terms' contexts.

In more detail, the algorithm for building merged windows for a query term is as follows: Fixed-size windows are identified around every instance of a query term in a document. A window is defined as *n* number of stemmed[1] non-stopwords to the left and right of the query term. We refer to all stemmed non-stopwords extracted from each window surrounding a query term as its *collocates*. In our experiments different window sizes were tested: 10, 20 and 40. These window sizes are large enough to capture collocates related topically, rather than syntactically.

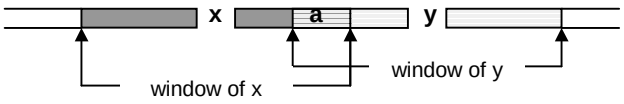

Figure 1: Overlapping windows around query terms x and y.

In this windowing technique we can encounter a situation where windows of two different query terms overlap. In such a case, we run into the following problem: let us assume that query terms *x* and *y* have overlapping windows and, hence, both are considered to collocate with term *a* (see Figure 1). We could simply add this instance of the term *a* into the merged windows of both *x* and *y*. However, when we compare these two merged windows, we would count this instance of *a* as a common term between them. This would be wrong, for we refer to the same instance of *a*, as opposed to a genuine lexical link by two different instances of *a*. Our solution to this problem is to attribute each instance of a word in an overlapping window to only one query term (node) – the nearest one.

[1] We used the weak stemming function in Okapi.

3.1.1 ESTIMATING SIMILARITY BETWEEN THE QUERY TERMS' CONTEXTS

After merged windows for all query terms in a document are built, the next step is to estimate their similarity by the collocates they have in common. We do pairwise comparisons between query terms, using the following two methods:

Method 1: Comparison by the number of lexical links they have.

Method 2: Comparison by the number of types they have in common.

3.1.1.1 METHOD 1

The first method takes into account how many instances of common collocates each query term has. In Figure 2, the first column contains collocates in the merged window of the query term *x*, the second column contains collocates in the merged window of the query term *y*. The lines between instances of the common collocates in the figure represent lexical links.

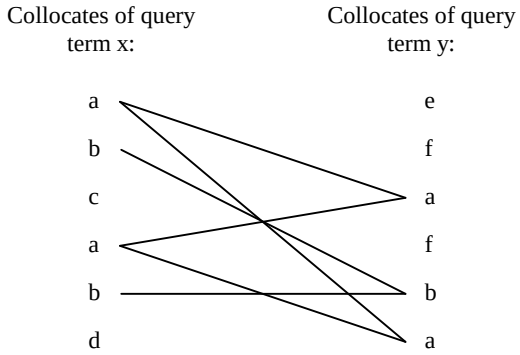

Figure 2: Links between instances of common collocates in merged windows of query terms x and y.

In this example there are altogether 6 links. If there are more than 2 query terms in a document, a comparison of each pair is done. The number of links are recorded for each pair, and summed up to find the total number of links in the document.

In our experiments, we only counted links formed by simple lexical repetition. We recognise that semantic similarity between contexts of terms might be more accurately estimated if we take into account other lexical-semantic relations between words, for example hyponymy, hypernymy, synonymy, etc. This would, require recourse to dictionaries and thesauri, such as WordNet. We plan to extend our work using such resources in the future.

A document's lexical cohesion score, calculated using method 1, will be referred to as LCS_{links}. To compare the scores across documents we need to normalise the total number of links in a document by the total size of all merged windows in a document. The normalised LCS_{links} score is:

$$LCS_{links} = \frac{L}{V} \quad (1)$$

where:

L – the total number of lexical links in a document;

V – the size (in words) of all merged windows in a document, excluding stopwords.

3.1.1.2 METHOD 2

In method 2 no account is taken of the number of common collocate *instances* each query term co-occurs with. Instead only the number of common *types* between each pair of merged windows is counted.

Comparison of merged windows in Figure 2 will return 2 types that they have in common: *a* and *b*. Again, if there are more than 2 query terms, a pairwise comparison is done. For each document we record the number of types common between each pair of merged windows, and sum them up.

A document's lexical cohesion score estimated using this method is LCS_{types}, and is calculated by normalising the total number of common types by the total number of types in the merged windows in a document:

$$LCS_{types} = \frac{T}{U} \quad (2)$$

Where:

T – the total number of common types in a document;

U – the total number of types in all merged windows in a document.

3.2 CONSTRUCTION OF SETS OF RELEVANT AND NON-RELEVANT DOCUMENTS

To test the hypothesis that lexical cohesion between query terms in a document is related to a document's property of relevance to the query, we calculated average lexical cohesion scores for sets of relevant and non-relevant documents.

We conducted our experiments on two datasets:

1) A subset of the TREC ad-hoc track dataset: FT 96[2] database, containing 210,158 Financial Times news articles from 1991 to 1994, and 50 ad-hoc topics (251 – 300) from TREC-5. We will refer to this dataset in this paper as "FT".
2) The HARD track dataset of TREC-12: 652,710 documents from 8 newswire newswire corpora (New York Times, Associated Press Worldstream and Xinghua English, among others), and 50 topics (401-450). This dataset will be referred to as "HARD".

Short queries were created from all non-stopword terms in the 'Title' fields of TREC topics. Such requests are similar to the queries that are frequently submitted by average users in practice. The queries were run in the Okapi IR system using BM25 document ranking function to retrieve top N documents for analysis. BM25 is based on the Robertson & Spärck-Jones probabilistic model of retrieval [12]. The sets of relevant and nonrelevant documents are then built using TREC relevance judgements for the top N documents retrieved.

We need to ascertain that the difference between the average lexical cohesion scores in the relevant and non-relevant document sets is not affected by the difference between the average BM25 document matching scores. To achieve this we need to build the relevant and non-relevant sets, which have similar mean and standard deviation of BM25 scores for each topic. This is achieved as follows: first all documents among the top *N* BM25-ranked documents are marked as relevant and non-relevant using TREC relevance judgements. Then each time a relevant document is found it is added to the relevant set and the nearest scoring non-relevant document is added to the non-relevant set. After the sets are composed, the mean and standard deviation of BM25 document matching scores are calculated for each topic in the relevant and non-relevant sets. If there is a significant difference between the mean and standard deviation in the two sets for a particular topic, then the sets are edited by changing some documents until the difference is minimal. We will refer to the relevant and non-relevant document sets constructed using this technique as *aligned sets*.

We created two pairs of aligned sets for FT and HARD corpora: using the top 100 BM25-ranked documents and using the top 1000 BM25-ranked documents. The sets and their sizes are presented in Table 1.

Table 1: Statistics of the aligned relevant and nonrelevant sets.

Data set	FT		HARD	
	Relevant	Non-relevant	Relevant	Non-relevant
Top100				
Number of documents	176	176	600	600
Mean BM25 document score	13.350	13.230	13.939	13.674
Stdev BM25 document score	2.200	1.905	4.254	3.864
Top1000				
Number of documents	268	268	1897	1897
Mean BM25 document score	11.515	11.472	11.306	11.219
Stdev BM25 document score	2.502	2.375	3.519	3.311

Comparison between the corresponding relevant and non-relevant sets was done by average lexical cohesion score, which was calculated as:

$$Average\ LCS = \frac{\sum_{i=1}^{S} LCS_i}{S} \quad (3)$$

where:

LCS_i – lexical cohesion score of *i*th document in the set, calculated using either formula (1), or (2) above.

S – number of documents in the set.

[2] TREC research collection, volume 4.

3.3 ANALYSIS OF RESULTS

Comparisons of pairs of relevant and non-relevant aligned sets derived from 100 and 1000 BM25-ranked documents showed large differences between the sets on some measures (Table 2). In particular, average Lexical Cohesion Scores of the relevant and non-relevant documents selected from the top 1000 BM25-ranked document sets, calculated using the Links method (LCS_{links}) have statistically significant differences (Wilcoxon[3] test at 0.05 significance level). Average LCS_{types} are also significantly different in most of the experiments.

Table 2: Difference between the aligned relevant and non-relevant sets (FT dataset)

Method	Window	Rel	Nonrel	Difference (%)	Wilcoxon P(2-tail)	Significant
FT, Top 1000						
Links	10	0.097	0.076	28.795	0.025	Y
Links	20	0.151	0.119	26.727	0.002	Y
Links	40	0.197	0.165	19.868	0.008	Y
Types	10	0.056	0.043	30.454	0.009	Y
Types	20	0.071	0.057	24.733	0.001	Y
Types	40	0.082	0.071	14.333	0.031	Y
FT, Top 100						
Links	10	0.091	0.069	31.562	0.061	N
Links	20	0.144	0.109	32.703	0.001	Y
Links	40	0.187	0.146	28.016	0.001	Y
Types	10	0.048	0.036	33.920	0.024	Y
Types	20	0.063	0.047	32.928	0.001	Y
Types	40	0.074	0.061	21.010	0.005	Y
HARD, Top 1000						
Links	10	0.090	0.074	21.39	0.000	Y
Links	20	0.145	0.122	15.76	0.000	Y
Links	40	0.195	0.166	17.49	0.000	Y
Types	10	0.053	0.050	7.17	0.003	Y
Types	20	0.071	0.069	2.65	0.167	N
Types	40	0.086	0.084	1.36	0.387	N
HARD, Top 100						
Links	10	0.102	0.089	15.66	0.032	Y
Links	20	0.167	0.143	16.68	0.003	Y
Links	40	0.218	0.188	16.24	0.000	Y
Types	10	0.059	0.054	9.01	0.087	N
Types	20	0.080	0.075	5.91	0.175	N
Types	40	0.095	0.091	4.32	0.105	N

The first method of comparison by counting the number of links between merged windows appears to be somewhat better than the second method of comparison by types. This suggests that the density of repetition of common collocates in the contextual environments of query terms offers some extra relevance discriminating information.

To investigate other possible differences between the documents in the relevant and non-relevant sets, we have calculated various document statistics (Table 3). In both FT and HARD document collections the relevant documents, on average are longer, have more query term occurrences, and consequently have more collocates per query term. The latter finding is interesting, given that we selected relevant and non-relevant document pairs with the similar BM25 scores. However, BM25 scores do not depend on query term occurrences only. A number of other factors affect BM25 score: a) document length; b) *idf* weights of the query terms; c) non-linear within-document term frequency function which progressively reduces the contribution made by the repeating occurrences of a query term to the document score, on the assumption of verbosity[4].

Table 3: Averaged document characteristics (FT and HARD document sets created from top1000 documents)

	Rel	Nonrel	Difference (%)	t-test P
FT (Top 1000)				
Ave. number of collocate tokens per query term	95.900	71.331	34.444	0.000
Ave. query term instances	11.704	8.719	34.230	0.000
Ave. document length	332.012	224.658	47.786	0.000
Ave. distance between query terms	19.444	14.976	29.832	0.027
Ave shortest distance between query terms	6.533	4.617	41.498	0.085
HARD (Top 1000)				
Ave. number of collocate tokens per query term	86.848	66.561	30.479	0.000
Ave. query term instances	11.297	8.693	29.962	0.000
Ave. document length	282.740	220.419	28.274	0.000
Ave. distance between query terms	18.077	17.705	2.099	0.633
Ave shortest distance between query terms	6.164	7.113	15.389	0.091

An interesting, though somewhat counter-intuitive, finding is the *average distance* between query term instances, which is significantly longer in relevant documents. To calculate the average distance between query terms, we take all possible pairs of different query term instances, and for each pair find the shortest matching strings, using the cgrep program [13]. The shortest matching string is a stretch of text between two different query terms (say, *x* and *y*) that do not contain any other query term instance of the same type as either of the query terms (i.e., *x* or *y*). Once the shortest matching strings are extracted for each pair of query terms, the distances between them are calculated (as the number of non-stopwords) and averaged over the total number of pairs. The closer the query terms occur to each other, the more their windows overlap, and hence the fewer collocates they have. In the nonrelevant documents query terms occur on average closer to each other (Table 3), which may contribute to the fact that they have fewer collocates. Longer distances between query terms in the relevant documents may be explained by the higher document length values in the relevant set, compared to the nonrelevant set.

Another statistic, *average shortest distance* between query terms, is calculated by finding the shortest matching string for each distinct query term combination. In this case, only one value, the shortest distance between each distinct pair, is returned. The shortest distances of all distinct pairs are then summed and

[3] The distribution of the data is non-Gaussian.

[4] The term frequency effect can be adjusted in BM25 by means of the tuning constant k_1. In our experiments we used k_1=1.2, which showed optimal performance on TREC data [12]. This chosen value means that repeating occurrences of query terms contribute progressively less to the document score.

averaged. As Table 3 shows, this value is larger in the relevant documents than in the nonrelevant in the FT corpus, and smaller in the HARD corpus. The differences are not statistically significant, though.

The above analysis clearly shows that relevant documents are longer and have more query term occurrences. So, could any of these factors possibly be the reason for the higher average Lexical Cohesion Scores in relevant documents? As instances of the original query terms can be collocates of each other, and form links between the collocational contexts of each other or other query terms, we need to find out what is the number of link-forming collocates (i.e. those which form links with collocates of other query terms), which are not query terms themselves. The following hypothesis was formulated to investigate this possibility:

Hypothesis 1.1: Collocational environments of different query terms are more cohesive in the relevant documents than in the nonrelevant, and this difference is not due to the larger number of query terms.

To investigate the above hypothesis, we counted in each document the total number of link-forming collocate instances (excluding the query terms) and normalised this count by the total number of all collocates in the windows of all query term instances. We refer to the normalised link-forming collocate count per document as *link_cols*, and the total number of collocates of query terms in the document as *total_cols*. The data (Table 4) shows that there exist large differences between the relevant and nonrelevant sets. Seven out of twelve experiments demonstrate statistically significant differences. This indicates that the contexts of different query terms in the relevant documents on average are more cohesive than in the non-relevant documents, and that this difference is not due to the higher number of query term instances. The fact that we normalise the count by the total number of collocates of query terms in the document eliminates the possibility of larger collocate numbers affecting this difference.

Table 4: Average number of link-forming collocates (excluding original query terms), normalised by the total number of collocates of query terms in the document

Window	Rel	Nonrel	Difference (%)	Wilcoxon P(2-tail)	Significant?
FT, Top1000					
10	0.071	0.065	9.607	0.000	Y
20	0.100	0.095	5.849	0.002	Y
40	0.123	0.118	4.636	0.010	Y
FT, Top100					
10	0.070	0.065	7.630	0.067	N
20	0.101	0.096	5.019	0.300	N
40	0.123	0.115	6.963	0.045	N
HARD, Top1000					
10	0.063	0.055	14.408	0.066	N
20	0.085	0.071	19.567	0.009	Y
40	0.103	0.090	14.465	0.013	Y
HARD, Top100					
10	0.063	0.053	18.441	0.083	N
20	0.086	0.067	27.904	0.004	Y
40	0.105	0.086	21.992	0.002	Y

To find out whether the normalised link-forming collocate count can be statistically predicted by the number of query term instances we conducted linear regression analysis on the data of one of the experiments (HARD, top 1000 document dataset, window size 10), with the normalised link-forming collocate count per document (*link_cols*) as the dependent variable, and the number of query term instances in the document (*qterms*) as the independent variable. The R Square for the relevant document set was found to be 0.182, and for the nonrelevant document set, R Square was 0.122. Rather low R Square values support the Hypothesis 3 stated above. The result of the analysis indicates that the linear model using *qterms* can predict only about 18% of the *link_cols* values.

4. RE-RANKING OF DOCUMENT SETS BY LEXICAL COHESION SCORES

4.1 EXPERIMENTAL DESIGN

Statistically significant differences in the average lexical cohesion scores between relevant and non-relevant sets, discovered in the previous experiments, prompted us to evaluate LCS as a document ranking function.

It was decided to conduct experiments on re-ranking the set of top 1000 BM25-ranked documents by their LCS scores. Document sets were formed by using weighted search with the queries for 45[5] topics of the HARD corpus. The queries were created from all non-stopword terms in the 'Title' fields of the TREC topics. Okapi IR system with the search function set to BM25 (without relevance information) was used for searching. Tuning constant k_1 (controlling the effect of within-document term frequency) was set to 1.2 and b (controlling document length normalisation) was set to 0.75 [12].

BM25 function outputs each document in the ranked set with its document matching score (MS). We decided to test re-ranking with a simple linear combination function (*COMB-LCS*) of MS and LCS. Tuning constant x was introduced into the function to

$$COMB\text{-}LCS = MS + x * LCS \qquad (4)$$

regulate the effect of LCS:

The following values of x were tried: 0.25, 0.5, 0.75, 1, 1.5, 3, 4, 5, 6, 7, 8, 10 and 30.

We conducted experiments with both types of lexical cohesion scores:

LCS_{links} – calculated using method 1 of comparing query terms' collocation environments by the number of links they have;

LCS_{types} – calculated using method 2 of comparing query terms' collocation environments by the number of types they have in common.

The window sizes tested were 40, 20 and 10.

[5] Five of the 50 topics had no relevant documents and were excluded from the official HARD 2004 evaluation [15].

4.2 ANALYSIS OF RESULTS

Precision results of re-ranking with the combined linear function of MS and LCS with different values for the tuning constant *x* are presented in Table 5.

Table 5: Results of re-ranking BM25 document sets by *COMB-LCS* (HARD corpus)

Runs with different *x* values	Window size 40		Window size 20		Window size 10	
	AveP	P@10	AveP	P@10	AveP	P@10
BM25	0.2196	0.3089				
Method 1 (links)						
0.25	0.2201	0.3156	0.2199	0.3178	0.2198	0.3156
0.5	0.2208	0.3200	0.2207	0.3200	0.2200	0.3178
0.75	0.2213	0.3222	0.2217	0.3156	0.2202	0.3178
1	0.2213	0.3200	0.2217	0.3133	0.2209	0.3156
1.5	0.2217	0.3244	0.2223	0.3156	0.2214	0.3200
3	0.2242	0.3267	0.2241	0.3200	0.2230	0.3222
4	0.2240	0.3311	0.2268	0.3222	0.2230	0.3133
5	0.2205	0.3400	0.2322	0.3333	0.2231	0.3244
6	0.2227	0.3444	0.2316	0.3378	0.2230	0.3267
7	0.2227	0.3489	0.2314	0.3356	0.2258	0.3289
8	0.2265	0.3556	0.2311	0.3422	0.2258	0.3356
10	0.2217	0.3556	0.2303	0.3356	0.2254	0.3333
30	0.1964	0.3200	0.2097	0.3244	0.2179	0.3156
Method 2 (types)						
0.25	0.2196	0.3089	0.2196	0.3067	0.2196	0.3111
0.5	0.2197	0.3133	0.2197	0.3111	0.2196	0.3133
0.75	0.2199	0.3133	0.2197	0.3111	0.2197	0.3111
1	0.2200	0.3133	0.2198	0.3156	0.2197	0.3133
1.5	0.2201	0.3133	0.2200	0.3178	0.2199	0.3178
3	0.2200	0.3044	0.2203	0.3156	0.2209	0.3200
4	0.2199	0.3044	0.2203	0.3156	0.2210	0.3200
5	0.2200	0.2978	0.2205	0.3133	0.2216	0.3244
6	0.2199	0.3022	0.2203	0.3133	0.2216	0.3200
7	0.2172	0.3022	0.2207	0.3133	0.2216	0.3222
8	0.2168	0.3022	0.2217	0.3111	0.2213	0.3244
10	0.2161	0.3044	0.2215	0.3111	0.2211	0.3244
30	0.2030	0.3178	0.2133	0.3200	0.2142	0.3089

The results show that there is a significant increase in precision at the cut-off point of 10 documents (P@10) when LCS scores are combined with the MS as given by equation 4 above, with x=8 and window size of 40. The precision @10 for BM25 and LCS scores are 0.3089 and 0.3556, respectively. The 15% increase is statistically significant (Wilcoxon test at P=0.001). Thirteen topics have higher precision and none – lower. Average precision (AveP) also increases, although by a smaller amount when documents are re-ranked with equation 4. The highest gain in average precision (5.7%) is achieved when x is 5 and window size is 20. This result is not, however, statistically significant. It is also worth mentioning that 5 out of 45 topics used in evaluation have only one query term in the topic title, and our method can only be applied to queries with two or more query terms.

A number of factors need to be considered in the context of the re-ranking experiments: 65.39% of documents have LCS score of zero. This is mainly because a very large proportion of documents (52.64%) only have one distinct query term. Also, we only compared lexical environments of query terms through the repetition of their collocates. It is likely that only a certain proportion of lexical links is determined in this way. For a fuller analysis, other types of lexical-semantic relations should be investigated. The above factors may have a significant impact on the results of re-ranking, and we expect to have better results if the above points are successfully addressed in future studies. It should also be noted that the combined function used in re-ranking is rather simple and alternatives (e.g. non-linear functions) are worth investigating.

5. CONCLUSIONS

In this study we explored the property of lexical cohesion between query terms in documents: whether it is related to relevance, and whether it can be used to predict relevance in document ranking. Two hypotheses were put forward. The first hypothesis we studied was:

Hypothesis 1: There exists statistically significant association between the level of lexical cohesion of the query terms in documents and relevance.

We conducted experiments by building sets of relevant and non-relevant documents, calculating their lexical cohesion scores and comparing the averages of these scores. The experiments showed that there exists a statistically significant difference between the average lexical cohesion scores of relevant and non-relevant documents extracted from the top 100 and top 1000 BM25-ranked sets. We also proved that this difference is genuine, and is not affected by differences in BM25 scores or other document characteristics.

The experimental results provided support for Hypothesis 1, demonstrating that there exists a statistically significant relation between relevance and the level of lexical cohesion between query terms.

Having discovered that on the whole relevant documents have more instances of query terms than non-relevant documents, we explored another hypothesis:

Hypothesis 1.1: Collocational environments of different query terms are more cohesive in the relevant documents than in the nonrelevant, and this difference is not due to the larger number of query terms.

Our experiments supported the above hypothesis and showed that on average relevant documents have larger numbers of link-forming collocates, which are not original query terms, compared to nonrelevant documents. Following these experiments, we explored another hypothesis:

Hypothesis 2: Ranking of a document set by lexical cohesion scores results in significant performance improvement over term-based document ranking techniques.

We conducted experiments on re-ranking BM25-ranked document sets with a simple linear combination function of BM25 document

matching score and the lexical cohesion score. Different values of a tuning constant *x*, regulating the effect of LCS were tried. The results suggested that there are some significant improvements over BM25 document ranking function, thus providing support for Hypothesis 2. We are aware of the fact that the function used in re-ranking the documents is simple and more elaborate methods need to be investigated.

Results achieved in the first half of this study – i.e., difference between relevant and non-relevant documents by their average lexical cohesion scores are promising. Our approach to using LCS in document ranking in the second half of the study also proved to be useful. The experiments reported suggest that the concept of lexical cohesion has strong association with document relevance, and therefore is worth further investigation. To achieve further benefit from lexical cohesion in document ranking, more experimentation is needed. In particular, problems of documents with zero LCS score and better ways of combining LCS with BM25 scores need to be investigated.

Lexical cohesion, as a text property, is formed not only through word repetition, but other more complex lexical relations. So far we looked into lexical cohesion between query terms achieved only through repetition of their collocates. Other lexical cohesion forming phenomena, such as synonymy, antonymy, hyponymy and meronymy [14] could also be taken into account in identifying lexical cohesive links between the environments of query terms. A more complete analysis of lexical environments of query terms can be expected to provide more support to the ideas behind this study. It is noteworthy to mention that an earlier analysis of lexical link distribution by Ellman [8] showed that the most common link type, repetition of the same word, is closely followed by the type of links, formed by words belonging to the same thesaurus category. A possible future development of our method could, thus, consist of defining links on the basis of repeated words and words related through either manually or automatically constructed lexical resources and thesauri.

In the reported work, all links formed by repetition are treated equally. Arguably, links formed by collocates with high inverse document frequency (*idf*) are more indicative of a strong lexical cohesion between the contexts of query terms, than links formed by words with low *idf*. For example, some collocates could be discourse-forming or topic-neutral words (e.g., "say", "report", "argue"), which tend to have low *idf*. One possible future extension of this work is to weight links using *idf* weights of the terms forming them.

Apart from being a potential aid as a ranking function, the proposed method of estimating the degree of lexical cohesion between query terms could be useful in other tasks such as query expansion and summarisation. It is likely that query terms with a strong lexical cohesion belong to the same topic, therefore they are more likely to collocate with relevant query expansion terms, than query terms with weak lexical cohesion.

6. REFERENCES

[1] Halliday, M.A.K. and Hasan, R. *Cohesion in English.* Longman, 1976.

[2] Manning C.D., Schütze H. *Foundations of Statistical Natural Language Processing*, The MIT Press, Cambridge, Massachusetts, 1999.

[3] Beeferman D., Berger A., Lafferty J. A model of lexical attraction and repulsion. *In Proc. ACL-EACL Joint Conference*, Madrid, Spain, 1997.

[4] Vechtomova O., Robertson S., Jones S. Query expansion with long-span collocates. Information Retrieval, 6(2), pp. 251-273, 2003.

[5] Hoey, M. *Patterns of Lexis in Text.* Oxford University Press; 1991.

[6] Morris, J. and Hirst, G. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 17, 1991, pp. 21-48.

[7] Hirst, G. and St-Onge, D. Lexical chains as representation of context for the detection and correction of malapropisms. *Wordnet. An Electronic Lexical Database.* C.Fellbaum (ed.), MIT Press, 1997, pp.305-332.

[8] Ellman, J. and Tait, J. On the generality of thesaurally derived lexical links. *In the Proceedings of 5th JADT*, 2000, pp.147-154.

[9] Stairmand M.A. Textual context analysis for information retrieval. *In Proc. ACM-SIGIR*, 1997, pp. 140-147.

[10] Hearst, M. Multi-paragraph segmentation of expository text. In the Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics. 1994.

[11] Manabu O., Hajime M. Query-biased summarization based on lexical chaining. *In Computational Intelligence*, Vol. 16, N 4, 2000, pp. 578-585.

[12] Spärck Jones, K., Walker, S. and Robertson, S.E. (2000). A probabilistic model of information retrieval: development and comparative experiments. *Information Processing and Management*, 36(6), 779-808 (Part 1); 809-840 (Part 2).

[13] Clarke, C.L.A. and Cormack, G.V. On the use of regular expressions for searching text. University of Waterloo Computer Science Department Technical Report number CS-95-07, University of Waterloo, Canada, February 1995.

[14] Hasan, R. Coherence and cohesive harmony. *In Flood, J. (ed.) Understanding Reading Comprehension*. 1984. pp.181-219. Delaware: International Reading Association.

[15] Allan, J. HARD Track overview in TREC 2004 (Notebook). High Accuracy Retrieval From Documents. In Voorhees, E. and Buckland, L. (Eds.) TREC 2004 Conference Notebook Proceedings, November 2004.