Available online at www.sciencedirect.com
SCIENCE DIRECT®
Information Processing and Management 42 (2006) 1230–1247
INFORMATION PROCESSING & MANAGEMENT
www.elsevier.com/locate/infoproman

On document relevance and lexical cohesion between query terms

Olga Vechtomova [a,*], Murat Karamuftuoglu [b], Stephen E. Robertson [c]

[a] Department of Management Sciences, University of Waterloo, 200 University Avenue West, Waterloo, Ont., Canada N2L 3GE
[b] Department of Computer Engineering, Bilkent University, Bilkent, 06800 Ankara, Turkey
[c] Microsoft Research Cambridge, 7 J J Thomson Avenue, Cambridge, CB3 0FB, UK

Received 20 October 2005; received in revised form 10 January 2006; accepted 13 January 2006
Available online 15 March 2006

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. Lexical cohesion between distinct query terms in a document is estimated on the basis of the lexical-semantic relations (repetition, synonymy, hyponymy and sibling) that exist between there collocates – words that co-occur with them in the same windows of text. Experiments suggest significant differences between the lexical cohesion in relevant and non-relevant document sets exist. A document ranking method based on lexical cohesion shows some performance improvements.

Keywords: Information retrieval; Lexical cohesion; Word collocation; Document relevance

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, depend 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.

* Corresponding author. Tel.: +1 519 888 4567x2675; fax: +1 519 746 7252.
E-mail addresses: ovechtom@uwaterloo.ca (O. Vechtomova), hmk@cs.bilkent.edu.tr (M. Karamuftuoglu), ser@microsoft.com (S.E. Robertson).

0306-4573/$ - see front matter
doi:10.1016/j.ipm.2006.01.008

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 (Halliday & Hasan, 1976). 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. Lexical cohesion between distinct query terms in a document is estimated on the basis of the lexical-semantic relations (repetition, synonymy, hyponymy and sibling) that exist between their collocates, i.e., words that co-occur with them in certain spans. We also report 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" (Halliday & Hasan, 1976, 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 *cohesion*. The continuity created by cohesion consists in "expressing at each stage in the discourse the points of contact with what has gone before" (Halliday & Hasan, 1976, 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*, realised through lexis. Halliday and Hasan distinguished two broad categories of lexical cohesion: *reiteration* and *collocation*. Reiteration 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 it as a relationship between lexical items that occur in the same environment, but they fail to formulate a more precise definition.

Later, the meaning of collocation was narrowed in some works to refer only to idiomatic expressions, whose meaning cannot be completely derived from the meaning of their elements. For example Manning and Schütze (1999) 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 structure. 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 closely related, therefore they tend to occur in the same topics in texts. Beeferman, Berger, and Lafferty (1997) experimentally determined that long-span collocation effects can extend in text up to 300 words. Vechtomova, Robertson, and Jones (2003) report examples of long span collocates identified using the Z-score such as "environment–pollution", "gene–protein".

Hoey (1991) 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 lexical links between their long-span collocates formed by repetition, synonymy, hyponymy and sibling relations.

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* (Ellman & Tait, 2000; Hirst & St-Onge, 1997; Hoey, 1991; Morris & Hirst, 1991). Lexical cohesion in text is normally realised through sequences of linked words – *lexical chains*. The term "*chain*" was first introduced by Halliday and Hasan (1976) 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 (1991) define lexical chains as sequences of related words in text. One of the prerequisites for the linked words to be considered units of a chain is their co-occurrence within a certain span. Hoey (1991) 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 (Stairmand, 1997), text segmentation (Hearst, 1994) and summarisation (Manabu & Hajime, 2000).

2.2. Lexical bonds

Hoey (1991) pointed that text cohesion is formed not only by links between words, but also by 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*. A lexical bond exists between two sentences when they are connected by a certain number of lexical links. The number of lexical links the sentences must have to form a bond 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. 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 (Hoey, 1991).

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 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 for this purpose: (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.

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 Fig. 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

[1] We used the Porter stemming function (Porter, 1980).

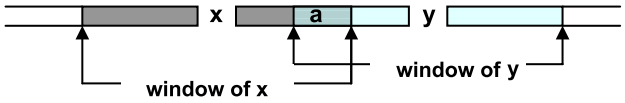

Fig. 1. Overlapping windows around query terms x and y.

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.

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' collocates, using the following two methods:

Method 1: Comparison by the number of lexical links they have.
Method 2: Comparison by the number of related types they have.

3.1.1.1. Method 1. The first method takes into account how many instances of lexically linked collocates each query term has. Fig. 2 demonstrates this method by showing links between collocates formed by simple lexical repetition. 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.

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.

We have conducted experiments with (1) using only lexical links formed by simple lexical repetition (Section 3.3.1) and (2) using lexical links formed by WordNet relations of synonymy, hyponymy and sibling in addition to lexical cohesion (Section 3.3.2).

WordNet relations: To identify links formed by synonymy, hyponymy and sibling relations between collocates we used WordNet (Miller, 1990). WordNet is a lexical resource, where senses of lexical units (words or phrases) are grouped into synonym sets (synsets), which are linked to other synsets via different kinds of relations, such as hyponymy and sibling. Hyponymy is a hierarchical relation between a more specific lexical unit, hyponym, and a more general unit, hypernym. An example of hyponym-hypernym relationship in WordNet is "painting – graphic art". Sibling relation occurs between lexical units which have the same hypernym, for example, "painting – print".

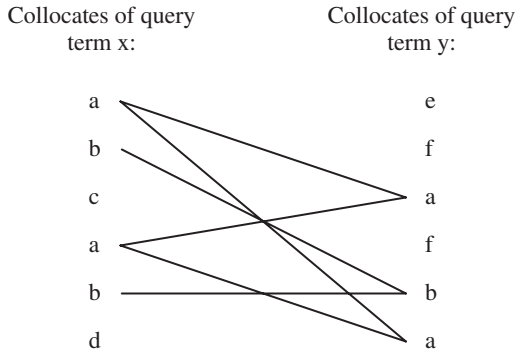

Fig. 2. Links between instances of common collocates in merged windows of query terms x and y.

The first step in the process of identifying synonymy, hyponymy and sibling relations between collocates is to map a collocate to a WordNet synset. There are several difficulties in this process: first, each lexeme may belong to several parts of speech, therefore a Part-of-Speech (POS) tagger is needed to map collocates to the correct POS forms in WordNet. Secondly, a word may have several senses in WordNet, each forming its own synset, therefore we need a method to disambiguate each collocate, and map it to the correct synset. There is a number of POS taggers (e.g., Brill, 1995), and word sense disambiguation (WSD) techniques (e.g., Gale, Church, & Yarowsky, 1992; Galley & McKeown, 2003; Yarowsky, 1995) that could be adapted for this purpose, however they are computationally expensive. An alternative approach, which we adopted in this study, is to map a collocate to the most frequent sense, which is possible as WordNet contains corpus frequencies of each word sense. A study by Mihalcea and Moldovan (2001) shows that the most frequent WordNet sense occurs with a probability of 78.52% for nouns, 61.01% for verbs, 80.98% for adjectives and 83.84% for adverbs in SemCor corpus, therefore suggesting that moderate to high levels of WSD accuracy can be achieved by mapping collocates to their most frequent WordNet sense. One other problem with using WordNet senses is that they are very fine-grained, and many of the senses are semantically close. Consider, for example, the verb *walk*, which has 10 senses in WordNet, out of which senses 1 (use one's feet to advance; advance by steps), 2 (traverse or cover by walking) and 6 (take a walk; go for a walk; walk for pleasure) are very close semantically. Arguably, applications such as Information Retrieval, do not require such fine-grained distinctions between senses, and therefore it may be advantageous to merge them, as suggested in Mihalcea and Moldovan (2001). We did not perform WordNet sense merging in this work, and its benefit for our purpose has yet to be investigated. The final difficulty in mapping collocates to WordNet synsets is that collocates in our method are always single terms, whereas WordNet synsets may contain both single terms and phrases. In the current method, if there is a phrase in a synset, we do not use it in LCS calculations. It is possible to extend our method to handle phrases in addition to words, however this remains for future work.

After collocates are mapped to WordNet synsets, we do a pairwise comparison of each collocate of query term x with each collocate of query term y as follows: first we check whether they are identical (i.e., form a link by repetition), if not we check their relationship via WordNet according to the following rules:

- if two collocates have the same synonym, they form a link by synonymy;
- if collocate a is a hyponym or hypernym of collocate b (or any of its synset members), they form a link by hyponymy;
- if two collocates have the same hypernym, they form a link as siblings.

Lexical cohesion score (links): A document's lexical cohesion score, calculated using method 1, will be referred to as $\mathrm{LCS}_{\mathrm{links}}$. 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 $\mathrm{LCS}_{\mathrm{links}}$ score is

$$\mathrm{LCS}_{\mathrm{links}} = \frac{L}{V}, \tag{1}$$

where L is the total number of lexical links in a document and V is 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 lexically related collocate *instances* each query term co-occurs with. Instead, only the number of lexically related *distinct words* (referred to as *types* throughout the rest of the paper) between each pair of merged windows is counted.

Comparison of merged windows in Fig. 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.

Synonymy, hyponymy and sibling relationships are identified in exactly the same way as in method 1, except that we count the number of related types, as opposed to tokens.

Lexical cohesion score (types): A document's lexical cohesion score estimated using this method is $\mathrm{LCS}_{\mathrm{types}}$, and is calculated by normalising the total number of common types by the total number of types in the merged windows in a document:

$$\mathrm{LCS}_{\mathrm{types}} = \frac{T}{U}, \tag{2}$$

where T is the total number of lexically related types in a document and U is 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. Out of 50 topics, only 44 had relevant documents in the Financial Times collection, therefore only these topics were used in the experiments. 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 corpora (New York Times, Associated Press Worldstream and Xinghua English, among others), and 50 topics (401–450). Five of the 50 topics had no relevant documents and were excluded from the official HARD 2004 evaluation (Allan, 2004). 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 (Spärck Jones, Walker, & Robertson, 2000). The sets of relevant and non-relevant 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.

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

$$\text{Average LCS} = \frac{\sum_{i=1}^{S} \mathrm{LCS}_i}{S}, \tag{3}$$

where LCS_i is the lexical cohesion score of ith document in the set, calculated using either formula (1), or (2) above; and S is the number of documents in the set.

[2] TREC research collection, volume 4.

Table 1
Statistics of the aligned relevant and non-relevant sets

Data set	FT		HARD	
	Relevant	Non-relevant	Relevant	Non-relevant
Top 100				
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
Top 1000				
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

In the next subsection we analyse the results of comparison between relevant and non-relevant documents. We compare average lexical cohesion scores calculated by using simple lexical repetition in Section 3.3.1, and by using repetition, synonymy, hyponymy and sibling relations in Section 3.3.2.

3.3. Analysis of results

3.3.1. Links formed by simple lexical repetition

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

Table 2
Difference between the aligned relevant and non-relevant sets

Method	Window	Relevant	Non-relevant	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

Table 3
Averaged document characteristics (FT and HARD document sets created from top 1000 documents)

	Relevant	Non-relevant	Difference (%)	*t*-Test *P*
FT, top 1000				
Average number of collocate tokens per query term	95.900	71.331	34.444	0.000
Average query term instances	11.704	8.719	34.230	0.000
Average document length	332.012	224.658	47.786	0.000
Average 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				
Average number of collocate tokens per query term	86.848	66.561	30.479	0.000
Average query term instances	11.297	8.693	29.962	0.000
Average document length	282.740	220.419	28.274	0.000
Average 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

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.[3] Average LCS_{types} are also significantly different in most of the experiments.

The first method of comparison by counting the number of links between merged windows appears to be 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]

An interesting, though somewhat counter-intuitive, finding is the *average distance* between query term instances, which is 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 (Clarke & Cormack, 1995). 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 non-relevant 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 non-relevant 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

[3] We used Wilcoxon test as 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 (Spärck Jones et al., 2000). This chosen value means that repeating occurrences of query terms contribute progressively less to the document score.

each distinct pair, is returned. The shortest distances of all distinct pairs are then summed and averaged. As Table 3 shows, this value is larger in the relevant documents than in the non-relevant 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 when their windows overlap, 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 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 non-relevant, and this difference is not due to the larger number of query term instances.

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 collocates in the windows of all query term instances. We refer to the normalised link-forming collocate count (excluding query terms) per document as *link_cols*. The data (Table 4) shows that there exist large differences in *link_cols* between the relevant and non-relevant 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.

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 non-relevant 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.

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	Relevant	Non-relevant	Difference (%)	Wilcoxon *P* (2-tail)	Significant
FT, top 1000					
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, top 100					
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	Y
HARD, top 1000					
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, top 100					
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

Table 5
Difference between the aligned relevant and non-relevant sets in average LCS calculated using WordNet relations (HARD 2004 corpus, top 1000)

Method	Window	Relevant	Non-relevant	Difference (%)	Wilcoxon P (2-tail)	Significant
HARD, top 1000						
Links	10	0.107	0.089	19.280	0.000	Y
Links	20	0.172	0.146	18.019	0.000	Y
Links	40	0.234	0.199	17.695	0.000	Y
Types	10	0.057	0.053	5.994	0.002	Y
Types	20	0.077	0.074	4.049	0.037	Y
Types	40	0.093	0.089	4.390	0.039	Y

3.3.2. Links formed by repetition, synonymy, hyponymy and sibling relations

We compared the average lexical cohesion scores between the aligned relevant and non-relevant sets, derived from top 1000 documents of the HARD corpus, where LCS were calculated using WordNet relations of synonymy, hyponymy and sibling in addition to simple lexical repetition. The results of the comparison are presented in Table 5.

As seen from the table, WordNet relations overall do not contribute much to differentiating between relevant and non-relevant sets, compared to the use of only simple lexical repetition (cf. data under the heading "HARD, top 1000" in Table 2). Experiments with various parameters, such as excluding the sibling relations, and assigning different weights to relations as proposed in Galley and McKeown (2003), led to similar results.

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. For this purpose, we conducted 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 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 (Spärck Jones et al., 2000).

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 regulate the effect of LCS:

$$\text{COMB-LCS} = \text{MS} + x * \text{LCS}. \tag{4}$$

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:

$\text{LCS}_{\text{links}}$ – calculated using method 1 of comparing query terms' collocation environments by the number of links they have;
$\text{LCS}_{\text{types}}$ – calculated using method 2 of comparing query terms' collocation environments by the number of related types they have.

The window sizes tested were 10, 20 and 40.

4.2. Analysis of results

4.2.1. Links formed by simple lexical repetition

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 6 (HARD corpus) and Table 7 (FT corpus).

4.2.1.1. HARD corpus. The results show that there is a significant increase in precision at the cut-off point of 10 documents (P@10) when LCS_{links} scores are combined with the MS as given in Eq. (4) above, with $x = 8$ and window size of 40. The precision at 10 for BM25 and LCS_{links} 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 Eq. (4). The highest gain in average precision (5.7%) is achieved when x is 5 and window size is 20, and the highest gain in R-Precision (5.8%) is achieved when x is 5 or 6 and window size is 20. The last two gains are not, however, statistically significant.

The analysis of results shows that 65.39% of documents have LCS score of zero. This is mainly because a large proportion of documents (52.64%) only have one distinct query term, making the scope for improvement rather limited. Five of the 45 topics contain only one query term in the title. In the remaining 40 topics, 49.7% of all retrieved documents have only one distinct query term. It is also important to note that the retrieved documents with one distinct query term constitute 19% of all relevant documents for these topics,

Table 6
Results of re-ranking BM25 document sets by COMB-LCS (HARD corpus; LCS is calculated using simple lexical repetition only)

Runs with different x values	Window size 40			Window size 20			Window size 10		
	AveP	P@10	R-Prec	AveP	P@10	R-Prec	AveP	P@10	R-Prec
BM25	0.2196	0.3089	0.2499						
Method 1 (links)									
0.25	0.2201	0.3156	0.2506	0.2199	0.3178	0.2502	0.2198	0.3156	0.2504
0.5	0.2208	0.3200	0.2507	0.2207	0.3200	0.2507	0.2200	0.3178	0.2506
0.75	0.2213	0.3222	0.2514	0.2217	0.3156	0.2512	0.2202	0.3178	0.2507
1	0.2213	0.3200	0.2531	0.2217	0.3133	0.2523	0.2209	0.3156	0.2509
1.5	0.2217	0.3244	0.2530	0.2223	0.3156	0.2519	0.2214	0.3200	0.2512
3	0.2242	0.3267	0.2505	0.2241	0.3200	0.2511	0.2230	0.3222	0.2551
4	0.2240	0.3311	0.2536	0.2268	0.3222	0.2623	0.2230	0.3133	0.2535
5	0.2205	0.3400	0.2464	**0.2322**	0.3333	**0.2644**	0.2231	0.3244	0.2519
6	0.2227	0.3444	0.2586	0.2316	0.3378	**0.2644**	0.2230	0.3267	0.2526
7	0.2227	0.3489	0.2574	0.2314	0.3356	0.2637	0.2258	0.3289	0.2636
8	0.2265	**0.3556**	0.2602	0.2311	0.3422	0.2635	0.2258	0.3356	0.2628
10	0.2217	**0.3556**	0.2584	0.2303	0.3356	0.2634	0.2254	0.3333	0.2597
30	0.1964	0.3200	0.2349	0.2097	0.3244	0.2430	0.2179	0.3156	0.2464
Method 2 (types)									
0.25	0.2196	0.3089	0.2496	0.2196	0.3067	0.2497	0.2196	0.3111	0.2495
0.5	0.2197	0.3133	0.2497	0.2197	0.3111	0.2499	0.2196	0.3133	0.2496
0.75	0.2199	0.3133	0.2498	0.2197	0.3111	0.2499	0.2197	0.3111	0.2495
1	0.2200	0.3133	0.2503	0.2198	0.3156	0.2500	0.2197	0.3133	0.2497
1.5	0.2201	0.3133	0.2513	0.2200	0.3178	0.2508	0.2199	0.3178	0.2518
3	0.2200	0.3044	0.2503	0.2203	0.3156	0.2514	0.2209	0.3200	0.2540
4	0.2199	0.3044	0.2476	0.2203	0.3156	0.2504	0.2210	0.3200	0.2545
5	0.2200	0.2978	0.2468	0.2205	0.3133	0.2503	0.2216	0.3244	0.2540
6	0.2199	0.3022	0.2464	0.2203	0.3133	0.2498	0.2216	0.3200	0.2524
7	0.2172	0.3022	0.2388	0.2207	0.3133	0.2481	0.2216	0.3222	0.2500
8	0.2168	0.3022	0.2402	0.2217	0.3111	0.2480	0.2213	0.3244	0.2495
10	0.2161	0.3044	0.2397	0.2215	0.3111	0.2469	0.2211	0.3244	0.2481
30	0.2030	0.3178	0.2343	0.2133	0.3200	0.2457	0.2142	0.3089	0.2426

Table 7
Results of re-ranking BM25 document sets by COMB-LCS (FT corpus; LCS is calculated using simple lexical repetition only)

Runs with different x values	Window size 40			Window size 20			Window size 10		
	AveP	P@10	R-Prec	AveP	P@10	R-Prec	AveP	P@10	R-Prec
BM25	0.1274	0.1523	0.1383						
Method 1 (links)									
0.25	0.1278	0.1568	0.1382	0.1278	0.1568	0.1391	0.1276	0.1568	0.1391
0.5	0.1286	0.1568	0.1435	0.1279	0.1568	0.1386	0.1275	0.1545	0.1389
0.75	0.1283	0.1636	0.1436	0.1275	0.1659	0.1384	0.1271	0.1545	0.1460
1	0.1286	0.1614	0.1438	0.1276	0.1659	0.1441	0.1264	0.1568	0.1457
1.5	0.1282	0.1659	0.1439	0.1269	0.1659	0.1444	0.1262	0.1591	0.1449
3	0.1270	0.1636	0.1411	0.1255	0.1636	0.1422	0.1258	0.1591	0.1363
4	0.1256	0.1636	0.1370	0.1254	0.1636	0.1411	0.1244	0.1659	0.1358
5	0.1252	0.1659	0.1364	0.1251	0.1636	0.1408	0.1235	0.1636	0.1378
6	**0.1297**	0.1636	0.1370	0.1247	0.1636	0.1401	0.1236	0.1614	0.1435
7	0.1284	0.1682	0.1371	0.1241	0.1682	0.1394	0.1235	0.1614	0.1412
8	0.1237	0.1682	0.1362	0.1235	0.1682	0.1342	0.1230	0.1568	0.1399
10	0.1138	0.1659	0.1273	0.1220	**0.1727**	0.1335	0.1218	0.1545	0.1404
30	0.0891	0.1318	0.0945	0.0981	0.1591	0.1007	0.1051	0.1477	0.1112
Method 2 (types)									
0.25	0.1279	0.1568	0.1383	0.1278	0.1568	0.1384	0.1276	0.1568	0.1384
0.5	0.1276	0.1545	0.1397	0.1277	0.1568	0.1384	0.1277	0.1545	0.1384
0.75	0.1279	0.1568	0.1384	0.1276	0.1568	0.1395	0.1278	0.1568	0.1395
1	0.1283	0.1568	0.1429	0.1280	0.1591	0.1395	0.1279	0.1568	0.1393
1.5	0.1286	0.1614	0.1429	0.1287	0.1614	0.1453	0.1277	0.1591	0.1456
3	0.1292	0.1636	0.1442	0.1274	0.1636	0.1446	0.1271	0.1614	0.1458
4	0.1290	0.1682	0.1407	0.1275	0.1636	0.1451	0.1273	0.1636	0.1444
5	0.1276	0.1705	0.1407	0.1273	0.1636	0.141	0.1269	0.1591	0.1436
6	0.1274	0.1705	0.1408	0.1277	0.1682	0.1408	0.1269	0.1614	**0.1480**
7	0.1267	0.1705	0.1406	0.1265	0.1682	0.1400	0.1268	0.1591	0.1478
8	0.1296	0.1705	0.1397	0.1262	0.1682	0.1394	0.1266	0.1591	0.1478
10	0.1292	0.1636	0.1365	0.1261	0.1682	0.1447	0.1273	0.1591	0.1477
30	0.1136	0.1455	0.1017	0.1111	0.1455	0.1076	0.1179	0.1409	0.1373

all of which were either demoted in the ranked list or retained their original rank following the LCS-based re-ranking. Relevant documents containing only one distinct query term may contain some other semantically related word(s) instead of the user's original query term. For example, there is a document judged relevant for the topic "Identity Theft", which contains only one query term "identity". The document, however, contains the term "fraud", which is close in meaning to "theft" and could be used as its replacement in calculating the document's lexical cohesion score. A method that attempts to find a replacement for a missing query term may be useful for identifying lexical cohesion between query concepts in a document. One such approach, proposed by Terra and Clarke (2005), relies on corpus statistics to identify a replacement word for a missing query term in each document. The method was evaluated in the passage retrieval task, and showed statistically significant improvements in P@20 over the baseline Multitext passage retrieval function.

4.2.1.2. FT corpus. There is a maximum increase of 13.4% in P@10 with $x = 10$ and window size 20 when LCS_{links} is combined with the BM25 document matching score (P@10 for BM25 and LCS scores are 0.1523 and 0.1727, respectively). Nine out of 44 topics have higher P@10 and three – lower. Increase in the average precision is low: 1.8% (LCS_{links}, $x = 6$; window size = 40), while the highest increase in R-Precision (7%) is achieved with LCS_{types}, $x = 6$ and window size of 10. The LCS_{links} run with $x = 8$ and window size of 40, which showed the best performance in P@10 in the HARD corpus, has P@10 of 0.1682, and an increase of 10% over the baseline. None of the above improvements are statistically significant, but there is a statistically significant improvement of 11% in P@10 for the run LCS_{types} ($x = 8$; window size = 40).

4.2.2. Links formed by repetition, synonymy, hyponymy and sibling relations

We conducted document re-ranking experiments with the HARD corpus using WordNet relations in calculating lexical cohesion scores. The use of WordNet relations in addition to simple lexical repetition in calculating LCS, does not change notably the performance of the methods using simple lexical repetition alone (Table 8).

We analysed the distribution of different types of WordNet relations that form lexical links to see whether lack of improvement is due to small numbers of the WordNet relations. The number of links formed between collocates (window size 20) by means of different relations is shown in Table 9.

The most frequent relationship is simple lexical repetition (83.4%), followed by sibling and hyponymy relationships. Only a very small percentage of links (1.8%) is formed by means of synonymy. An earlier analysis of lexical link distribution by Ellman and Tait (2000) also showed that the most common link type is repetition of the same word. However, according to their results, repetition was closely followed by the relationship between words of the same category in Roget thesaurus, which was in turn followed by links between words belonging to the same group of categories in Roget and, finally, links between words connected by one level of internal thesaurus pointers. In their study, Ellman and Tait used the lexical chaining algorithm by Morris and Hirst (1991) to identify lexical links between words, and a small corpus of long texts of different genres. In our experiments, small numbers of synonymy relations between collocates could be due to, firstly, rather fine-grained partitioning of words into senses in WordNet, as a result of which many synsets consist of very few or only one word. Secondly, compound synset members are not used in our method of lexical link construction (see Section 3.1.1).

Table 8
Results of re-ranking BM25 document sets by COMB-LCS (HARD corpus; LCS is calculated using simple lexical repetition and WordNet relations)

Runs with different x values	Window size 40			Window size 20			Window size 10		
	AveP	P@10	R-Prec	AveP	P@10	R-Prec	AveP	P@10	R-Prec
BM25	0.2196	0.3089	0.2499						
Method 1 (links)									
0.25	0.2202	0.3133	0.2507	0.2200	0.3178	0.2501	0.2198	0.3156	0.2501
0.5	0.2210	0.3200	0.2511	0.2207	0.3200	0.2505	0.2202	0.3178	0.2503
0.75	0.2215	0.3244	0.2513	0.2212	0.3178	0.2517	0.2205	0.3178	0.2504
1	0.2216	0.3222	0.2512	0.2222	0.3133	0.2524	0.2210	0.3156	0.2503
1.5	0.2220	0.3289	0.2508	0.2227	0.3111	0.2515	0.2218	0.3178	0.2544
3	0.2229	0.3311	0.2517	0.2245	0.3267	0.2499	0.2236	0.3200	0.2547
4	0.2205	0.3356	0.2458	0.2272	0.3333	0.2638	0.2235	0.3289	0.2522
5	0.2185	0.3444	0.2498	**0.2321**	0.3356	**0.2641**	0.2263	0.3311	0.2639
6	0.2187	0.3511	0.2506	0.2319	0.3378	0.2637	0.2262	0.3333	0.2627
7	0.2145	**0.3533**	0.2511	0.2292	0.3400	0.2560	0.2315	0.3311	0.2628
8	0.2137	**0.3533**	0.2489	0.2271	0.3400	0.2556	0.2312	0.3311	0.2621
10	0.2100	**0.3533**	0.2406	0.2258	0.3422	0.2568	0.2283	0.3244	0.2522
30	0.1943	0.3178	0.2309	0.2062	0.3289	0.2420	0.2127	0.3178	0.2441
Method 2 (types)									
0.25	0.2197	0.3089	0.2499	0.2196	0.3067	0.2499	0.2196	0.3089	0.2496
0.5	0.2199	0.3111	0.2497	0.2198	0.3089	0.2498	0.2198	0.3133	0.2493
0.75	0.2202	0.3156	0.2505	0.2198	0.3111	0.2506	0.2201	0.3178	0.2511
1	0.2202	0.3133	0.2503	0.2199	0.3111	0.2506	0.2201	0.3156	0.2513
1.5	0.2204	0.3156	0.2511	0.2203	0.3133	0.2504	0.2211	0.3200	0.2526
3	0.2206	0.3111	0.2496	0.2204	0.3178	0.2529	0.2176	0.3156	0.2517
4	0.2205	0.3067	0.2483	0.2166	0.3111	0.2512	0.2179	0.3178	0.2517
5	0.2205	0.3067	0.2479	0.2169	0.3133	0.2505	0.2178	0.3178	0.2528
6	0.2206	0.3067	0.2478	0.2170	0.3089	0.2500	0.2181	0.3200	0.2534
7	0.2204	0.3111	0.2498	0.2171	0.3067	0.2499	0.2178	0.3178	0.2533
8	0.2179	0.3089	0.2443	0.2171	0.3044	0.2489	0.2153	0.3222	0.2454
10	0.2162	0.3133	0.2423	0.2150	0.3133	0.2435	0.2150	0.3178	0.2415
30	0.2093	0.3267	0.2413	0.2108	0.3267	0.2446	0.2165	0.2933	0.2428

Table 9
The number of lexical links formed by different relations

Relationship	Number of collocates	Percent of collocates (%)
Simple lexical repetition	708,050	83.4
Synonymy	14,864	1.8
Hyponymy	28,117	3.3
Sibling	97,856	11.5

4.3. Qualitative analysis

In order to gain some qualitative understanding of the relation between lexical cohesion and the relevance property of texts, we have examined a few documents from the collection used in the experiments described above. Although, it is not possible to generalise, the small sample of documents examined suggests that certain patterns of lexical links could be identified in the documents promoted and demoted after LCS re-ranking. In the set examined, we 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 text, i.e., they are not concentrated in isolated sections of the text. As expected, the instances of query terms are well connected by lexical links in the LCS promoted documents. An example of this type of document is NYT20030711.0053, retrieved in response to query "AIDS in Africa" (topic no: HARD-409). There are many instances of the both query terms (10 instances of "AIDS" and 9 instances of "Africa") in the text and they are extensively connected with each other by lexical links.

In the demoted documents, we noticed three different patterns. Some of the demoted documents are made up of disjoint pieces of text that cover separate and unrelated news stories. The query terms appear in different parts (different stories) in these documents and therefore have no or few lexical links between them. An example of this type of document is NYT20030616.0015 which is retrieved in response to query "Chimpanzee Language Ability" (topic no: HARD-407). The document consists of several short (a paragraph-long) science-stories. The term "chimpanzee" appears in a story about a conservation project, "language" appears in a story about a study of neurological activities in humans and refers to "differential equations and other mathematical *language* beloved by economists", and "ability" appears in a short story on "Child-rearing problems" and "anxiety among parents already uncertain about their *ability* to be parents". Another type of demoted document exhibits a different structure. In the document AFE20031110.0344 (topic no: HARD-411) three of the four query terms ("Natural Disasters and Global Warming") occur but not all in the same context. The term "disaster" appears several times in this document and refers to an ecological disaster caused by oil leakage from a tanker. However, this is not a natural disaster as required in the topic description and narrative. The term "natural" appears only once and in the context of "The sinking dealt a huge blow to a region of outstanding *natural* beauty". Similarly, the term "global" appears once and out of context: "disaster was the worst of its kind in Europe and second only on a *global* scale to the 1989 slick from the Exxon Valdez catastrophe". Clearly the term does not here refer to "global warming" which was the context specified in the topic description and narrative. In the third type of demoted documents, the query topic is

Table 10
Summary of the document qualitative analysis

Topic	Promoted	Demoted	Reason	LCS
409	NYT20030711.0053		Extensive lexical links	0.348
426	XIE20031021.0398		Extensive lexical links	0.244
407		NYT20030616.0015	Disjoint stories	0.017
405		AFE20031210.0021	Disjoint stories	0
411		AFE20031110.0344	Unrelated contexts	0.037
407		AFE20030523.0366	Unrelated contexts	0
407		NYT20030804.0022	Unrelated contexts	0
419		NYT20030909.0070	Unrelated contexts	0.018
415		AFE20031015.0922	Marginal	0.040

treated marginally. In the document AFE20031015.0922, which is retrieved in response to "Life on Mars" (topic no: HARD-415), there are several instances of the term "Mars" but only one instance of the term "life". The document is mainly about space missions to Mars, and although the query terms occur in related contexts, the subject of "life on Mars" was mentioned in passing in the document, and therefore, there are only few links between these two query terms. Table 10 summarises the findings of this small-scale qualitative analysis.

5. Conclusions and future work

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. The first hypothesis and the related sub-hypothesis were:

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

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

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, number of instances of query terms or other document characteristics. 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 demonstrate that with certain x values significant improvements over BM25 document ranking function can be achieved, thus providing support for Hypothesis 2. The experiments reported suggest that there is strong association between lexical cohesion and document relevance. To achieve further benefit from lexical cohesion in document ranking, more experimentation is needed. In particular, the issues discussed below need further investigation.

The use of lexical relationships between words obtained by using WordNet did not lead to any noticeable differences in performance. The reason is that links formed by means of WordNet relations only constitute 16.6% of all links, the rest being due to simple lexical repetition. One possible cause is rather fine-grained partitioning of senses in WordNet. It may be useful to use more coarse-grained partitioning of senses, as suggested in Mihalcea and Moldovan (2001). The second possible cause is that we did not use compound synset members in LCS calculation. It may be useful to develop methods to include compound terms in lexical link calculation.

In the method reported in this paper, documents which contain query terms close or adjacent to each other do not receive any special treatment compared to documents where query terms are separated by longer distances. Intuitively, query terms located in close proximity are more likely to be related topically. We experimented with attributing collocates in the overlapping windows of two distinct query terms to both of them, which led to the formation of more links between the collocates of closely located query terms, and consequently higher LCS. But, the results were inferior to those of the reported method. Interestingly, our study also shows that there is no significant difference between the average shortest distances between distinct query terms in the relevant and non-relevant sets in two TREC collections. However, it has been demonstrated in some studies that term proximity can be useful for document retrieval tasks (e.g., Clarke & Cormack, 2000), therefore possible combination of the two approaches to document ranking needs to be investigated further. In particular, queries which consist of a stable multi-word unit (e.g., "United Nations") may benefit more from proximity search, whereas queries consisting of a set of separate words (e.g., "China trade") or

"loose" phrases, whose components can occur separately in text (e.g., "AIDS in Africa"), may benefit more from lexical cohesion-based methods.

One of the characteristics of the reported method for estimating lexical cohesion between query terms is that it requires presence of at least two query terms in a document. Documents containing one query term constitute 19.5% of all relevant documents in the experiments reported. A method proposed by Terra and Clarke (2005), which attempts to find replacement terms for missing query terms, may prove useful in capturing more of the lexical cohesive relations between the concepts underlying query terms.

In our approach, 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.

The corpora used in this study consist of news articles which are usually relatively short documents (see Table 2 earlier in the paper). Arguably, lexical cohesion could be more useful for ranking longer documents, as there is more scope for topic diversity in longer documents compared to short ones. In the future we plan to conduct more experimentation on corpora containing longer documents.

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.

References

Allan, J. (2004). HARD track overview in TREC 2004 (Notebook). High accuracy retrieval from documents. In E. Voorhees, & L. Buckland (Eds.), *TREC 2004 conference notebook proceedings*. Gaithesburg, MD.

Beeferman, D., Berger, A., & Lafferty, J. (1997). A model of lexical attraction and repulsion. In *Proceedings of ACL-EACL joint conference*. Madrid, Spain.

Brill, E. (1995). Transformation-based error-driven learning and natural language processing: A case study in part of speech tagging. *Computational Linguistics, 21*(4), 543–565.

Clarke, C. L. A., & Cormack, G. V. (1995). 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.

Clarke, C. L. A., & Cormack, G. V. (2000). Shortest-substring retrieval and ranking. *ACM Transactions on Information Systems, 18*(1), 44–78.

Ellman, J., & Tait, J. (2000). On the generality of thesaurally derived lexical links. In *Proceedings of 5th JADT* (pp. 147–154).

Gale, W., Church, K., & Yarowsky, D. (1992). One sense per discourse. In *Proceedings of the DARPA speech and natural language workshop*. Harriman, New York.

Galley, M., & McKeown, K. (2003). Improving word sense disambiguation in lexical chaining. In *Proceedings of 18th international joint conference on artificial intelligence (IJCAI'03)*.

Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English*. London: Longman.

Hearst, M., 1994. Multi-paragraph segmentation of expository text. In *Proceedings of the 32nd annual meeting of the association for computational linguistics*.

Hirst, G., & St-Onge, D. (1997). Lexical chains as representation of context for the detection and correction of malapropisms. In C. Fellbaum (Ed.), *WordNet. An electronic lexical database* (pp. 305–332). Cambridge, MA: MIT Press.

Hoey, M. (1991). *Patterns of lexis in text*. Oxford: Oxford University Press.

Manabu, O., & Hajime, M. (2000). Query-biased summarization based on lexical chaining. *Computational Intelligence, 16*(4), 578–585.

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, MA: The MIT Press.

Mihalcea, R., & Moldovan, D., 2001. Automatic generation of a coarse grained WordNet. In *Proceedings of NAACL workshop on WordNet and other lexical resources* (pp. 35–41). Pittsburgh, PA.

Miller, G. A. (1990). WordNet: An on-line lexical database. *International Journal of Lexicography, 3*(4), 235–312.

Morris, J., & Hirst, G. (1991). Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics, 17*, 21–48.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program, 14*(3), 130–137.

Spärck Jones, K., Walker, S., & 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).

Stairmand, M. A. (1997). Textual context analysis for information retrieval. In *Proceedings of the 20th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 140–147). Philadelphia, PA.

Terra, E., & Clarke, C. L. A. (2005). Comparing query formulations and lexical affinity replacements in passage retrieval. In *Proceedings of the ACM-SIGIR workshop on methodologies and evaluation of lexical cohesion techniques in real-world applications (ELECTRA 2005)* (pp. 11–17). Salvador, Brazil.

Vechtomova, O., Robertson, S., & Jones, S. (2003). Query expansion with long-span collocates. *Information Retrieval, 6*(2), 251–273.

Yarowsky, D. (1995). Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd annual meeting of the association of computational linguistics (ACL'95)*. Cambridge, MA.

Olga Vechtomova is an Assistant Professor in the Department of Management Sciences, University of Waterloo, Canada. She received Ph.D. in Information Science from City University, London in 2001. Her research interests are in Information Retrieval, Natural Language Processing applications for IR and user interaction with IR systems. She is a member of ACM SIGIR and the Association for Computational Linguistics. Her works are published in Information Retrieval, Information Processing and Management and Journal of Information Science (http://ovecht2.uwaterloo.ca).

Murat Karamuftuoglu is an Assistant Professor in the Computer Engineering Department, Bilkent University, Turkey. Dr. Karamuftuoglu received Ph.D. in Information Science from City University, London in 1998. His research interests are in Interactive Information Retrieval, Knowledge Management and Computer Mediated Communication. His works are published in Journal of the American Society for Information Science and Technology, Journal of Information Science, Information Processing and Management.

Stephen E. Robertson is a researcher at Microsoft Research Cambridge and a professor at City University, London. He has published extensively in information retrieval over a period of almost forty years, specifically on probabilistic models, ranking algorithms and evaluation methods. He received the Salton Award of the ACM SIGIR in 2000.