Experiments for HARD and Enterprise Tracks

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

Maheedhar Kolla
University of Waterloo,
Canada
mkolla@uwaterloo.ca

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

1. HARD track

The main theme in our participation in this year's HARD track was experimentation with the effect of lexical cohesion on document retrieval. 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 [7]. Segments of text which are about the same or similar subjects (topics) have higher lexical cohesion, i.e. share a larger number of words than unrelated segments. We have experimented with two approaches to the selection of query expansion terms based on lexical cohesion: (1) by selecting query expansion terms that form lexical links between the distinct original query terms in the document (section 1.1); and (2) by identifying lexical chains in the document and selecting query expansion terms from the strongest lexical chains (section 1.2).

1.1 Experiments with lexical cohesion between query terms

The strength of lexical cohesion between two query terms can be useful in determining whether query terms are used in related contexts. We adopt an approach proposed in [6] for estimating the strength of lexical cohesion between two distinct query terms by counting the number of lexical links that exist between their *collocates* - words that co-occur in the windows of *n* (here *n*=20) words around all instances of each query term in text. In this experiment a lexical link is defined as a relationship between two instances of the same word. For more detailed description of the algorithm see [6]. Arguably, the more lexical links exist between two distinct query terms in a document, the more evidence there is that these terms are used in the same or related contexts in the document.

1.1.1 Clarification forms CF1 and CF2

In this year's HARD track we experimented with using collocates that form links between two distinct query terms in interactive query expansion. Specifically, all collocates that form links between any two distinct query terms were extracted from the top 25 documents ranked in response to queries formed from the Title fields of the topics. The extracted collocates were then ranked by *idf*, and 50 top ranked terms were shown to the user in the clarification form. We experimented with showing terms to the users in the context of short one-line snippets (clarification form 1, CF1), and alone (clarification form 2, CF2). The hypothesis we investigate by comparing these two clarification forms is whether short contextual environments in the form of snippets around the suggested query expansion terms help users in selecting query expansion terms. In CF1 we highlighted the suggested query expansion terms shown in the context of snippets, and put a checkbox next to each snippet. By clicking on the checkbox, the user could select the highlighted query expansion term in the snippet, adding it to the expanded query. We also gave users an opportunity to cut and paste any other terms in the snippets they deem useful into a textbox provided in the form, but we have not yet experimented with adding them to the query. In CF2 we included the same terms as in CF1, but without the context. The user could click on a checkbox next to each suggested term to add it to the expanded query.

1.1.2. Query expansion runs based on CF1 and CF2

For all our runs BM25 was used [2]. UWATbaseT and UWATbaseTD are the baseline runs using terms from Title and Title+Description fields respectively. Runs UwatHARDExp1 and UwatHARDExp2 are query expansion runs, which use terms selected by the users from clarification forms 1 and 2 in addition to terms from the Title field of the topics.

1.2 Experiments with lexical chains

Lexical chains can be defined as sequences of semantically related words, spanning the document or its parts [4]. Lexical chains tend to represent the topical structure and the theme of the document.

1.2.1 Clarification form CF3

In clarification form 3 (CF3), we show phrases taken from the lexical chains extracted from the top 25 documents retrieved in response to Title+Description query terms. The phrases are selected in the following way: we first compute lexical chains for each of the top 25 documents similar to the method given in [1] and [3][1]. We then rank the chains extracted from all 25 documents based on the average *idf* of their chain members. For example, for the chain "*spacewalker* (*idf*=9.829), *astronaut* (*idf*=5.952)" the score is 7.891. A chain member is always a noun or a noun phrase, and some of them can be part of larger noun phrases. For each chain we select one member as a representative of that chain. For each such selected member, we extract the noun phrase in which it occurs. For example, for the word *astronaut*, which occurs in a number of chains, the following phrases are identified: *astronaut, the French astronaut, a NASA astronaut, a European Space Agency astronaut, one astronaut, an astronaut*. For each such set of phrases we randomly select one and present it in the clarification form. This step could be extended to select phrases based on their stability in the corpus as suggested in [5]. We include 40 phrases for each topic in CF3, along with an option for the user to delete any word in the selected phrase.

1.2.2 Query expansion runs based on CF3

Two runs were conducted following user relevance feedback to CF3. (1) user-selected phrases were added to the original query (UwatHARDExp3); and (2) user-selected phrases plus members of all lexical chains containing the head nouns of these phrases were added to the query (UWAThardLC1). For example, if the user selected the phrase *a NASA astronaut*, we also add to the query other members of the same chain. In this example *spacewalker* is added, which is the member of the chain (*spacewalker, astronaut*).

1.3 Results

Due to a mistake in the indexing process, only part of the AQUAINT corpus was indexed, which negatively affected all our submitted runs and led to the incorrect official results. Table 1 contains the official incorrect results and the correct ones, which were obtained after TREC experiments when the problem was detected and corrected.

[1] we did not consider the sibling relation

Run name	AveP			P@10			R-Prec		
	Incorrect	**Correct**	Improvement over baseline	Incorrect	**Correct**	Improvement over baseline	Incorrect	**Correct**	Improvement over baseline
UWATbaseT	0.1235	**0.1766**		0.3460	**0.3520**		0.2002	**0.2358**	
UwatHARDExp1	0.1653	**0.2323**	+31.5% *	0.4640	**0.4780**	+35.8% *	0.2375	**0.2834**	+20.2% *
UwatHARDExp2	0.1666	**0.2372**	+34.3% *	0.4120	**0.4220**	+19.9% *	0.2342	**0.2806**	+19% *
UWATbaseTD	0.1408	**0.1972**		0.4020	**0.4140**		0.2267	**0.2592**	
UwatHARDExp3	0.1525	**0.2432**	+23.3% *	0.4580	**0.4820**	+16.43 *	0.2250	**0.2949**	+13.8% *
UWAThardLC1	0.1403	**0.2276**	+15.4% *	0.4100	**0.4480**	+8.2%	0.2097	**0.2802**	+8.1%

Table 1: Evaluation results ("incorrect" are the official results which were obtained with an incomplete corpus due to a mistake). Improvements marked with * are statistically significant (t-test at .05 significance level)

Use of the terms selected from CF1 (run UwatHARDExp1) leads to a 20.2% improvement in R-Precision compared to the corresponding baseline run (UWATbaseT), while terms selected from CF2 (run UwatHARDExp2) result in 19% improvement in R-Precision. There is a substantial difference of 13% between the P@10 of UwatHARDExp1 and UwatHARDExp2. This suggests that showing terms to the user in the context of snippets (CF1) leads to the selection of better query expansion terms, than showing them without a context (CF2).

Query expansion with the user-selected phrases from CF3 results in a 13.8% improvement in R-Precision over the baseline UWATbaseTD. Expansion with both user-selected phrases, and other members of the same lexical chains leads to an improvement of 8.1%.

2 Enterprise Track

In this year's Enterprise Search track, we took part in the following tasks:

- Discussion Search;
- Known-Item Search.

2.1 Discussion Search

In discussion search, the scenario is that the user is looking for messages contributing towards the discussion about some topic. In particular, the user is interested to find messages that express a pro or con view about the topic. Our main focus in the discussion search experiments is to determine whether thread properties would enable us to identify the discussions. This is a step towards condensing the information from various messages to answer a query. This year, we carried out experiments to answer the following research questions:

1. Whether grouping messages into one large document, based on their subject title, helps us to identify the discussion?
2. Whether removal of redundant text quoted from previous emails results in better performance than retaining it.

In order to test the first question, we group all messages in the archive into threads and index each thread as single document. This is done as follows: first we group the messages based on their subject title. When grouping the messages, we do not consider the prefixes such as "Re:" or "fwd:". The messages in each thread are then sorted based on their timestamp value. Each message is then processed using OAK system[2] to divide it into sentences. We then eliminate the redundant or quoted text, i.e. text already seen in a previous mail in the thread. We then concatenate the text from the messages in a given thread into one document, and use it for the run UwatEntDSth. In the run UwatEntDS quoted text is removed from messages, and each message is indexed separately. Finally, for the run UwatEntDSq, we index each message separately, but without removing the quoted text.

2.2 Experimental runs

For runs UwatEntDS and UwatEntDSq, we retrieve the top 1000 messages using BM25 function [2], and the title field of the query topic. For run UwatEntDSth, we retrieved the messages as follows: we first retrieve the top 1000 threads using BM25 and the Title field as query. Messages in each thread are re-ranked using the index created for the run UwatEntDS. Once the messages within the threads are re-ranked, we submit the top 1000 messages.

2.3 Results

Messages are judged on a three point scale: irrelevant, partially relevant and relevant. The evaluation results for the 59[3] topics are as shown in the table 2.

Run ID	Mean Average Precision (MAP)	R-Prec	bpref	Reciprocal rank	P@10
UwatEntDSth	0.2762	0.3149	0.2977	0.5315	0.4153
UwatEntDSq	0.3187	0.3514	0.3185	0.6860	0.4831
UwatEntDS	0.2705	0.3117	0.2782	0.6273	0.4508

Table 2: Discussion search task evaluation results

As shown in Table 2, the Mean Average Precision and R-Precision values of UwatEntDSth are similar to those of UwatEntDS. UwatEntDSq has better performance than both other runs.

2.4 Known-Item Search

In the Known-Item search, the scenario is that the user is looking for a message containing the information that s/he knows to exist, for example the contact information of some company. The goal of the task is to find the message which contains this information. In UWATEntKI we used the BM25 function [2] to retrieve the top 100 messages for a given query. The evaluation results for the 125 topics of the known-item search task are shown in Table 3.

Run	Reciprocal rank	Target found in Top 10	No target found
UWATEntKI	0.519	89(71.2%)	14(11.2%)

Table 3: Known item search task evaluation results

Of the all 125 topics, 89 topics had the target message in the top 10 retrieved documents and no target was found in the top 100 for 14 topics.

[2] http://nlp.cs.nyu.edu/oak/

[3] one topic did not have any relevant messages in the judgement pool

References

[1] M. Galley and K. McKeown. (2003) Improving word sense disambiguation in lexical chaining. In Proceedings of 18th International Joint Conference on Artificial Intelligence (IJCAI'03), pages 1486-1488, Acapulco, Mexico.

[2] K. Sparck-Jones, S. Walker, and S. E. Robertson. (2000) A probabilistic model of information retrieval: development and comparative experiments - part 2. Information Processing and Management, 36(6):809-840.

[3] M. Kolla. Automatic text summarization using lexical chains: Algorithms and experiments. Master's thesis, Department of Mathematics and Computer Science, University of Lethbridge, (2005).

[4] J. Morris and G. Hirst. (1991) Lexical cohesion computed by thesaural relations as an indicator of the structure of text. Computational Linguistics, 17(1):21-48.

[5] O. Vechtomova and M. Karamuftuoglu (2004) Approaches to high accuracy retrieval:. phrase-based search experiments in the hard track. In Proceedings of the 13th Text Retrieval Conference, Gaithersburg, MD, November 16-19.

[6] O. Vechtomova, M. Karamuftuoglu and S.E. Robertson (2005) A Study of Document Relevance and Lexical Cohesion between Query Terms. In Proceedings of the Workshop on Methodologies and Evaluation of Lexical Cohesion Techniques in Real-World Applications (ELECTRA 2005), the 28th Annual International ACM SIGIR Conference, August 19, 2005, Salvador, Brazil, pp. 18-25

[7] M.A.K. Halliday and R. Hasan (1976) Cohesion in English. Longman.