University of Strathclyde at TREC HARD

Mark Baillie, David Elsweiler, Emma Nicol, Ian Ruthven, Simon Sweeney, Murat Yakici, Fabio Crestani, and Monica Landoni

i-lab group
Department of Computer and Information Science
University of Strathclyde
Glasgow, UK
{mb, dce, emma, ir, simon, murat, fabioc, monica}@cis.strath.ac.uk

1 Motivation

The motivation behind the University of Strathclyde's approach to this years HARD track was inspired from previous experiences by other participants, in particular research by [1], [3] and [4]. A running theme throughout these papers was the underlying hypothesis that a user's familiarity in a topic (i.e. their previous experience searching a subject), will form the basis for what type or style of document they will perceive as relevant. In other words, the user's context with regards to their previous search experience will determine what type of document(s) they wish to retrieve.

1.1 Previous Research

Belkin *et al.* stated that searchers "who are familiar with a topic will want to see documents that are detailed and terminologically specific, and people who are unfamiliar with a topic will want to see general and relatively simple documents"[1]. Documents in the corpus were assessed by how "readable" they were, using a standard measure called the Flesch readability score[2]. The Flesch score for a document is derived from the mean number of syllables per word and the number of words per sentence. For each topic, a documents Flesch score was combined with the corresponding Retrieval Status Value (RSV) estimated from the initial document ranking, also known as the baseline. It was discovered that this combination (of the normalised Flesch readability and estimated relevance scores) gave greater weight to readable documents in the ranking, aiding those user's with low topic familiarity.

In a similar theme, Harper *et al.* hypothesised that a "user's familiar with a topic will prefer documents in which highly representative terms occur, and user's with a topic will prefer documents in which highly discriminating terms will occur"[3]. In other words, by identifying terms very specific to a topic, those documents with detailed information (e.g. highly technical documents) were pushed up the original document ranking. Conversely, for those user's not familiar with the subject, expanding the original query with terms very general to the topic will boost those documents that provide an overview. Depending on the user's context, the original baseline ranking can be reordered providing more importance to those documents with a high proportion of either representative or familiar terms, respectively. Analysis of the performance found that Discriminative queries were (on average) effective at improving the original baseline document ranking, particulary when a user had previous knowledge of the topic they were searching.

Kelly *et al.* measured user familiarity gained from the meta-data of 2004s topics, against the number of times the person had searched for information about this topic in the past[4]. Expectedly, they found a degree of agreement between the number of times a user searched previously on this topic and their topic familiarity. In other words, those user's familiar with the TREC topic had on average searched more times previously on this subject than those who stated unfamiliarity. A clarification form was designed to obtain

information both on user topic familiarity and also useful information that could be utilised for improving the original baseline ranking. Questions included - what information the user previously knew about the topic, what information they wanted to know, and also a large text box was provided to allow the user to add any further keywords that they felt described the topic. This information was then utilised both to determine a user's topic familiarity, from a newly derived measure derived from the typical the length of answer as well as other responses to the clarification form, and also to expand the original queries. The original queries submitted for each the topic were expanded using different combinations of feedback from the clarification form, with varying degrees of success.

1.2 Summary

The common theme running through the above works can be summarised in the following Hypothesis:

- H1: A user's familiarity in a topic will have an impact on what type of documents they will find relevant.

A user with very little background knowledge (of a topic) will, potentially, find an overview document more helpful initially than a document with very specific (possibly technical) content. As a consequence, documents that are in someway general to the topic (e.g. little technical detail, simple introduction pieces, etc.) will more likely be judged as relevant by the user. In comparison, a user that has a high degree of familiarity or background knowledge , will (potentially) prefer documents that contain detailed comment on the topic.

To expand our motivation further, for HARD we examined another aspect of the user's context. The user's in the HARD track are not typical searchers, but TREC assessors. In some regards, these assessors do not own either the topic or the original query submitted to the Information Retrieval system. Potentially, an assessor may have both little knowledge of the topic, and importantly, little interest in researching that topic. We believe that this factor (a user's interest), will also have an impact on the type and style of document they wish to retrieve. We posit that a user with little interest and background knowledge of a subject, would prefer to read documents with little technical "jargon", accessible to read, short in length, and also stylistically pleasing (with a high number of motivating terms and phrases). Surmising our motivation, we assume that a user (in this case a TREC assessor) will often be assigned a task to search that they will have little interest in and/or previous knowledge of. Therefore we formulate the new hypothesis:

- H2: A user's interest in a topic will have an impact on what type of documents they will find relevant. A user with little interesting in reading about a topic will prefer documents with little technical material and are stylistically pleasing to read.

In order to investigate both hypotheses (H1 and H2), we compared a number of approaches including Pseudo-relevance feedback, Flesch readability scores, Representative and Discriminative queries, and also a new approach that expands the original query with Motivating terms. In the following sections we introduce these techniques, and in particular the concept of expanding the original query with motivating terms. We then report the results and findings of each technique, before concluding our first attempt at both the HARD track, and TREC.

2 Methodology

In this section we introduce the different approaches that we investigated. All algorithms were implemented using the Lemur Information Retrieval framework[6]. Also, for the HARD track evaluation, the submitted runs from all groups were compared against a baseline run, declared by each participating

group. The baseline submission is the initial document retrieval that all other approaches can be compared against, measuring the improvement (or harm) in the new ranking. For example, a typical IR system will submit a query and return the same ranked list for each user, no matter what how differing their context is. However in HARD, the user's context is captured after this initial ranking, allowing for a re-ranking before the final results list is presented to the user. Depending on the user's context, different ranked lists can be presented, tailoring the system towards the user's information need. In order to evaluate how successful each personalised approach is, all techniques are compared against the original baseline ranking. For our baseline, we selected the Okapi BM25, and in particular we used the version implemented in Lemur (with the standard settings)[6]. For the baseline submission, we also used the topic titles as queries to simulate typically (poor) queries submitted by user's, which may also reflect typical behaviour of user's with little previous knowledge of the topic (or interest).

In the following sections, we outline how the user's context was captured, through the use of a clarification form, Section 2.1. We then introduce in turn each technique we investigated: a standard Pseudo-relevance feedback approach (Section 2.2), combining document Flesch readability scores (Section 2.3), query expansion using Discriminative and Representative terms (Section 2.4), and finally query expansion using Motivational terms in Section 2.5.

2.1 Capturing user context: Clarification Forms

We designed a clarification form to collect data from each assessor. Of particular interest was the user's previous topic experience or familiarity, and also their interest in finding out more about the subject. Within the clarification form, details were provided to remind the assessor about the topic, alongside a number of top ranked document summaries from the baseline ranking for the user to assess. However, we found little correlation between these summaries and the answers to the actual questions, therefore we utilised the clarification form as a way of data collection. The answers to each question were then used as a guide for setting the operation parameters in each re-ranking approach, and more importantly for grouping each user by topic familiarity and interest. Overall, we formed four groups based on the answers to the clarification forms (see Table 1).

For the final analysis, we examined the individual assessors as well. By doing so, we can examine the characteristics of each individual assessor, rather than treat the topics as independent. This was possible because for this years HARD track, the information on which assessor judged what topic was provided. There were six Assessors overall (A-F), who each judged approximately 8-10 topics.

Table 1. Group Statistics

Group	Description	Number	%
G1	Assessor familiar with topic and interested in reading about topic	33	66%
G2	Assessor not familiar with topic but interested in reading about topic	4	8%
G3	Assessor familiar with topic and has little interest in reading about topic	6	12%
G4	Assessor not familiar and has little interest in reading about topic	7	14%

2.2 Pseudo-Relevance Feedback

We compared the performance of each suggested technique against a tried and tested benchmark: Pseudo-relevance feedback. We believe this would be an interesting comparison with other approaches based on user familiarity and topic interest. This would allow us to assess whether techniques based on the inclusion

of contextual information from the user, improved over an accepted, and proven, automatic technique re-ranking. In other words, this would be our benchmark technique for comparing all other strategies.

We submitted a run re-ranking the baseline using Pseudo-relevance feedback, where the top N documents are assumed to be relevant. These documents were then used to re-rank the baseline. Therefore, the top N ranked documents were used for Pseudo-relevance feedback, re-ranking the top 1000 ranked documents from the baseline. The standard OKAPI Pseudo-relevance feedback algorithm implemented in the Lemur toolkit [6] was applied.

2.3 Combing Readability Scores

The first technique we investigated, was an approach first suggested by Belkin *et al.* in the previous HARD track[1]: the Flesch Reading Ease Score [2]. A documents Flesch score is a reflection of the mean number of syllables per word, and the number of words per sentence in a document. It is assumed that the higher a documents Flesch score, the more readable a document is. We therefore assumed that a document with both a high RSV and a high Flesch score will be more appropriate for user's with low topic familiarity, and/or those user's with little interest in the topic (Groups G3 and G4).

To utilise a documents readability score, the Flesch value was combined with the RSV using a simple linear combination. To do so, the Flesch score for all documents in the corpus was first computed off-line. Then for each topic, the top 1000 ranked documents, both the RSV and Flesch score were normalised and then combined using simple weighted average (see equation 1). Other evidence combination techniques could be applied such as Dempster-Schaffer[3], but we initially wanted to test the hypothesis using a simple approach.

Hence, the normalised estimated relevance scores, $\hat{RSV}_i$, for each document i were combined with the normalised document readability score, $\hat{read}_i$ (see [2]),

$$score(i) = \hat{RSV}_i + \alpha * \hat{read}_i \tag{1}$$

where α is a weighting parameter and,

$$\hat{RSV}_i = \frac{RSV_i - \min(RSV)}{\max(RSV) - \min(RSV)} \tag{2}$$

and,

$$\hat{read}_i = \frac{read_i - \min(read)}{\max(read) - \min(read)} \tag{3}$$

Depending on a user's context, the weighting parameter α can be adjusted. For example, a user with both low topic familiarity and interest will have a higher value of α in comparison to a user with high familiarity and interest. By doing so, more weight is placed on the importance on the readability score, thus pushing such documents further up the ranking.

2.4 Query Expansion using Representative and Discriminative Terms

The second approach we investigated was the use of Representative and Discriminative terms for query expansion. For user's with little topic familiarity and / or low interest in the topic, it was believed that representative terms for the topic will be able to locate documents that are very general e.g. overview documents. In comparison, discriminative terms can be used to find detailed documents on a topic, for those user's with previous experience of the subject.

To investigate this approach, we adopted the same strategy first introduced by Harper et al [3], ranking terms in a topic model according to their contribution. A topic model being a Language Model (LM) formed from the top N documents in the original baseline ranking. The Kullback-Leibler Divergence

measure is then used to determine what each terms contribution to the topic is in the corpus vocabulary. KL is typically used for measuring the difference between two probability distributions[5]. When applied to the problem of measuring the distance between two term distributions (Language Models), KL estimates the relative entropy between the probability of a term t occurring in the actual collection Θ_a (i.e. $p(t|\Theta_a)$), and the probability of the term t occurring in the estimated Topic Language Model (LM) Θ_e (i.e. $p(t|\Theta_e)$).

KL is defined as,

$$KL(\Theta_e||\Theta_a) = \sum_{t \in V} p(t|\Theta_e) log \frac{p(t|\Theta_e)}{p(t|\Theta_a)} \tag{4}$$

where,

$$p(t|\Theta_a) = \frac{n(t, \Theta_a)}{\sum_{t \in \Theta_a} n(t, \Theta_a)} \tag{5}$$

and,

$$p(t|\Theta_e) = \frac{\sum_{d \in \Theta_e} n(t, d) + \alpha}{\sum_t (\sum_{d \in \Theta_e} n(t, d) + \alpha)} \tag{6}$$

where $n(t, d)$ is the number of times t occurs in a document d and α is a small non-zero constant (Laplace smoothing). The smaller the KL divergence the closer the topic is to the actual collection, with a zero KL score indicating two identical distributions. To account for the sparsity within the Θ_e, Laplace smoothing was applied to alleviate the zero probability problem[7].

Instead of determining the difference between two term distributions (i.e. the collection and topic LM), we are interested in the individual term contribution to the topic LM. A term contribution being the KL score for the term t. The greater the contribution to the topic model the higher the KL score. Therefore, for each term t the contribution was calculated by,

$$KL(t) = p(t|\Theta_e) log \frac{p(t|\Theta_e)}{p(t|\Theta_a)} \tag{7}$$

The top C ranked terms in a topic model are then ranked further according to each terms "representative" and "discriminative" properties. To rank a terms discriminative property (e.g. how specific a term is to the topic), the KL discriminative score for term t is calculated by,

$$KL_d(t) = log \frac{p(t|\Theta_e)}{p(t|\Theta_a)} \tag{8}$$

To calculate how general a term is to the topic LM, the KL representative score is used, calculated by,

$$KL_r(t) = p(t|\Theta_e) \tag{9}$$

For each topic, either the top K ranked terms corresponding to either the KL-representation or KL-discrimination (equations 8 and 9 respectively), are then used to expand the query. For those user's with low familiarity and/or topic interest, the top Q ranked representative terms are applied.

2.5 Query Expansion using Motivating Terms

One of the main assumptions stated earlier, is that we believe a user's interest in a topic would have a bearing on the types of documents they will find relevant. For example, a user who is searching a topic they have little interest in, would possibly find documents that are stylistically pleasing better in comparison to very verbose, technically specific documents. A particular stylistic technique for drawing a readers attention is to make use of motivating terms and phrases within the text. We therefore assume

that those documents that contain a high number of motivating terms may provide a more suitable entry point into the topic for those user's with little interest in searching on the subject.

In order to determine what motivating terms to include in the expanded query, a list of typical motivating terms and phrases were collated. A list was manually compiled of words that indicate or portray emotion. This list of motivating terms and phrases was then ranked according to each terms contribution to the topic model, which was formed from the top N ranked documents, for each query. This approach is similar to that outlined in section 2.4, however, only motivating terms are ranked according using the KL score (see equation 7). The top ranked motivating terms for each topic were then used to expand the original query. The approach we adopted is outlined below:

1. Form a topic LM with the top N ranked documents.
2. Smooth the topic LM with the reference collection using Laplace smoothing.
3. For each term $\grave{t}$ in the predefined motivating term list, we calculate the KL score (see equation 7).
4. All terms $\grave{t}$ are then ranked with respect to the KL score (highest to lowest),
5. At this stage two strategies were implemented:
   - Expand the original query with those terms with a positive contribution e.g. $KL(\grave{t}) > 0$
   - Expand the query with the top Q ranked terms for each topic.

We now illustrate some examples of an original query being expanded by motivation terms. Below is two topics in this years HARD track. For the first topic (Number 322), the original query submitted to the IR system was "International Art Crime". A topic model was formed from the top 10 ranked documents. The list of motivating terms were then ranked based on their contribution to the topic. Table 2 presents the top three ranked motivating terms for this topic. Each term could be considered related to the subject of crime. It is believed by expanding the original query with these terms we will push up those document that may be of more interest to the user, thus providing a higher likelihood being relevant. For a different topic, the title query submitted was "Black Bear Attacks". The top three ranked motivating terms , see Table 2, for topic number 336, were "wild", "stirring" and "dangerous". All three terms could be associated with the description of aggressive animal behaviour or characteristics.

Table 2. Top ranked motivating terms for TREC Topic Numbers 322 and 336 respectively

Topic Num: 322		Topic Num 336	
Term	KL score	Term	KL score
dangerous	0.00175	wild	0.0018
suspicious	0.0012	stirring	0.0014
significant	0.0004	dangerous	0.00068

3 Evaluation Results

In this Section, we discuss the results from the official runs submitted for HARD.

3.1 Submitted Runs

A summary of the runs submitted for HARD can be found in Table 3. STRA1 was our baseline submission, which was the OKAPI retrieval method[6]. All other submissions were compared against this baseline. For each submitted run, we fixed the parameters for each attempt to provide a fair comparison

```
<num> Number: 322 <Title> International Art Crime

<narr> Narrative: A relevant document is any report that
identifies an instance of fraud or embezzlement in the
international buying or selling of art objects....

<num> Number: 336 <title> Black Bear Attacks

<narr> Narrative: It has been reported that food or cosmetics
sometimes attract hungry black bears, causing them to viciously
attack humans....
```

Fig. 1. A summarised description of TREC topics 322 and 336.

across each different technique. However, after the official results were released, we re-examined a number of new runs over a wider ranger of parameters settings. The results from each run, across a wider range of varying parameters will be released as a technical report once the analysis has been completed.

For query expansion using Motivating terms, two runs were submitted. In Section 2.1, the assessor for each topic was placed into one of four groups depending on their responses in the clarification form (see Table 1). We would posit that the performance of both runs would be better for those topics placed in groups G3-G4, who stated little interest in reading about the topic, than the other two groups. For forming the topic model prior to ranking the motivating terms, the top ten ($N = 10$) documents ranked by the baseline run were used. Then for the first submission (STRAxmta), each original query was expanded using the same number of terms (the top $Q = 6$ ranked terms). For the second run (STRAxmtg), the query was expanded with all top ranked motivating terms that recorded a KL score greater than zero i.e. $KL(\grave{t}) > 0$.

For the Pseudo-relevance feedback submission, the top N documents were also used to re-rank the first 1000 ranked documents (STRAxprfb). For consistency in our comparisons with other approaches, we again fixed N to be 10.

For comparing the Discriminative and Representative queries (STRAxqedt and STRAxqert respectively), we submitted one run each that expanded the original query with either the top sixth ranked Discriminative or Representative terms. For both runs, we would expect improved performance for groups G1 and G3 (those assessors familiar with the topic) using discriminative queries, and for groups G2 and G4, representative queries. Such a result would indicate that Representative queries rank general overview documents higher for those user's with low familiarity, while Discriminative terms would push up documents very specific to the topic. However, by submitting both sets of queries to all user's, we can also examine the effect of using discriminative queries for those user's with low familiarity, and vice versa. Again, N was set for 10 for ranking the topic terms and the original query was expanded with the top 6 Representative or Discriminative terms.

We also submitted two runs that combined the document RSV values estimated during the baseline with their Readability score. For submission STRAxreada, the same value was used for the weight α. This would help evaluate the effect of using the readability score across all groups. For the second run, STRAxreadg, we varied α, providing more to those groups with low topic familiarity and interest.

3.2 Results

Table 4 provides an overview of the performance of each of the official submissions. In the Table, we also include the proportion of topics where there was an increase over the baseline R-precision for a topic (a success), as well as the proportion of topics where an approach harmed the baseline R-precision (a fail).

Table 3. Submitted runs

Run	Description	Operation parameters
STRA1	Baseline - OKAPI	
STRAxmta	Query expansion using the top Q ranked Motivating terms	$N = 10, Q = 6$
STRAxmtg	Query expansion using the Motivating terms, with Q differing per topic	$N = 10$, all $KL(\grave{t}) > 0$
STRAxprfb	Pseudo-relevance feedback	$N = 10$
STRAxqedt	Query expansion using Discriminative terms	$N = 10, Q = 6$
STRAxqert	Query expansion using Representative terms	$N = 10, Q = 6$
STRAxreada	Combine RSV with Readability score, same weight for all groups	$\alpha = 0.1$
STRAxreadg	Combine RSV with Readability score, differing weight for all groups	$\alpha = 0.1, 0.15, 0.15, 0.2$

Examining the results irrespective of groups, it was highlighted that using Pseudo-relevance feedback was the most successful technique (R-precision = 0.263), followed by expanding the query using Representative queries (R-precision = 0.226). On average, Pseudo-relevance feedback increased R-precision by 0.048 over the baseline, while expanding all queries with Representative terms improved the baseline by 0.011. Overall, Pseudo-relevance feedback recorded more successes per topic than any other approach, improving over the baseline 66% of the time, and harming 20% of all topics. Expanding the original query with Representative terms improved 50% of topics, and harmed 38%. The other approaches marginally improved over the baseline, while expanding the original query with Motivating terms in fact harmed the original baseline ranking more often than not, worsening the original R-precision score for approximately 56% of all topics.

Table 4. Results for each submitted run - R-precision

Run	R-precision	Ave-precision	MAP@10	% Success	% Fail
STRA1	0.215	0.1598	0.338		
STRAxmta	0.174	0.132	0.3	22%	58%
STRAxmtg	0.176	0.132	0.298	22%	56%
STRAxprfb	**0.263**	**0.209**	**0.402**	66%	20%
STRAxqedt	0.219	0.171	0.364	40%	16%
STRAxqert	0.226	0.185	0.376	50%	38%
STRAxreada	0.216	0.160	0.339	28%	16%
STRAxreadg	0.216	0.160	0.339	28%	18%

For HARD, the performance of each technique across the four predefined groups was of more interest (see Table 1). Table 5 provides an overview of the results across these groups, with both the R-precision (R-prec) and success rate per topic (suc) presented. Analysing the performance of each group, we discovered that Pseudo-relevance feedback again performs best for three out of the four groups (G1, G3 and G4). For group G1, (user's with a high familiarity and interest), there was some evidence that using Representative terms improved the original baseline ranking. Also, for group G4 (user's with both low familiarity and little interest), there was some evidence that using Representative terms to expand the query improved the baseline ranking (3.4%), with a success rate of 57% per topic.

No approaches expect Pseudo-relevance feedback improved the baseline for group G3. Also, for group G2, those user's with low topic familiarity but a willingness to read more about the topic, all approaches improved only marginally over the baseline, with the Representative terms performing best. Although all techniques improved on the baseline, the actual baseline R-precision value was very low, at 0.0385. All other runs appeared to be affected with this poor initial baseline ranking.

Table 5. R-precision and number of success across Groups

	Groups							
	G1		G2		G3		G4	
Stra1	0.244		**0.0385**		0.143		0.245	
Run	R-prec	% suc	R-prec	% suc	R-prec	% suc	R-prec	% suc
STRAxmta	0.1968	0.242	0.044	0.5	0.1389	0.1667	0.171	0
STRAxmtg	0.201	0.242	0.044	0.5	0.1379	0.1667	0.171	0
STRAxprfb	**0.300**	**0.727**	0.041	0.25	**0.1658**	**0.5**	**0.3016**	**0.7143**
STRAxqedt	0.2487	0.454	0.044	0.5	0.133	0.1667	0.254	0.285
STRAxqert	**0.2515**	**0.515**	0.096	0.5	0.1256	0.333	**0.2697**	**0.571**
STRAxreadA	0.2457	0.303	0.041	0.25	0.141	0.1667	0.242	0.2857
STRAxreadg	0.2454	0.303	0.041	0.25	0.141	0.167	0.242	0.286

We also examined the performance of each approach across the different TREC assessors, with the results of this analysis presented in Table 6. Again, both the R-precision and success rate for each run is presented along with the percentage of topics that provided an improvement over the baseline ranking. There were 6 TREC assessors overall, who on average judged 8 topics each, with assessor F judging 10.

From this analysis, we discovered that yet again Pseudo-relevance feedback worked better for five out of the six assessors. However, for assessor D, no approach improved over the baseline. In fact, all approaches harmed the initial baseline ranking for this assessor. For assessors C, E and F, there was some evidence that Representative queries performed well, and for assessor A, Discriminative queries showed improved performance over the baseline.

Table 6. R-precision and number of success across Assessors

	Assessors											
	A		B		C		D		E		F	
Stra1	0.208		0.197		0.207		**0.190**		0.240		0.243	
Run	R-prec	# suc	R-prec	# suc	R-prec	# suc	R-prec	# suc	R-prec	# suc	R-prec	# suc
STRAxmta	0.180	(1/8)	0.145	(1/8)	0.196	(3/8)	0.133	(1/8)	0.160	(2/8)	0.219	(3/10)
STRAxmtg	0.184	(1/8)	0.147	(1/8)	0.196	(3/8)	0.135	(1/8)	0.153	(2/8)	0.230	(3/10)
STRAxprfb	**0.250**	(**6/8**)	**0.236**	(**4/8**)	**0.313**	(**7/8**)	0.185	(4/8)	**0.281**	(**5/8**)	**0.304**	(**7/10**)
STRAxqedt	**0.236**	(3/8)	0.187	(2/8)	0.228	(3/8)	0.178	(4/8)	0.220	(3/8)	0.257	(5/10)
STRAxqert	0.183	(2/8)	0.210	(5/8)	**0.269**	(**5/7**)	0.144	(3/8)	0.274	(4/8)	0.269	(5/10)
STRAxreadA	0.208	(1/8)	0.201	(4/8)	0.209	(2/8)	0.188	(3/8)	0.238	(3/8)	0.246	(4/10)

4 Discussion

Both the use of Readability scores and Motivating terms resulted in poor performance, when applied in isolation. Expanding the original query with Motivating terms, more often than not resulted in harming the original baseline ranking. The probable reason for this result, was that many of the relevant documents will either share few Motivating terms, or contain none at all. Although, we were encouraged by the ranking of Motivating terms with respect to each TREC topic. Those Motivating terms related to the subject were often ranked highly. This was indication that there was some potential in investigating this approach further. We now plan to investigate the use of Motivating terms further both re-examining dif-

ferent weighting schemes for query expansion, and also new approaches that could utilise these terms effectively.

There was evidence (although marginal) that using both Representative and Discriminative terms for expanding the original query for user's with low and high topic familiarity, respectively, did work. For group G1, Discriminative queries were successful in over 50% of the topics, while for group G4, Representative queries were successful in 57% of topics. Again, further examination of how to successfully use these terms is warranted.

For G2, the baseline was very poor, which had a negative effect on all other runs. The reason for this could be explained by a number of factors such as poor initial queries, difficult topics, or also the number of relevant topics for the topic (on average 69 partially relevant and 14 highly relevant documents per topic). This group contained four topics (344, 345, 397, 401), each judged by a different assessor. Compared to other topics, there was a similar number of judged relevant documents. Also, comparing the typical performance for these topics for all participants of HARD, the median R-precision for the baseline and final submissions was found to be 0.0873 and 0.078 respectively. This would indicate that the four grouped topics are difficult in someway. Whether this is due to the context of the assessors of some other factor would require further analysis. However, the result does indicate the instability of all submissions evaluated in this work. The majority of techniques are reliant on the original baseline ranking to contain a number of relevant documents in the top N. If this is not the case, then each approach performs poorly as a result.

Overall, Pseudo-relevance feedback was found to be the best approach both on average, and across the each group and assessor. Only for group G2, and also for assessor D, did Pseudo-relevance feedback not perform well. The other suggested techniques based on a user's context did not record similar improvement over the baseline R-precision in comparison with this approach. It would be of interest in future research to identify new approaches, either based on combining methods such as Discriminative / Representative queries with a tried and tested strategy approach such as Pseudo-relevance feedback. Also, further research into new techniques that are not so reliant on the initial baseline ranking would be beneficial for difficult topics.

5 Acknowledgements

We would like to thank Ellen Voorhees, James Allan and Ian Soboroff for solving the mystery as to why our second clarification form failed. A lesson learned for next year.

References

1. N.J. Belkin, I. Chaleva, M. Cole, L. Liu Y.-L. L and, Y.-H. Liu, G. Muresan, C.L. Smith, Y. Sun, X.-J. Yuan, and X.-M. Zhang. Rutgers' hard track experiences at trec 2004. In *Proceedings of the 13th Text REtrieval Conference (TREC 2004)*. NIST, 2004.
2. D. Foulger. A simplified flesch reading ease formula. Web, 1997. http://www.foulger.info/davis/papers/SimplifiedFleschReadingEaseFormula.htm (Last visited August 2005).
3. D.J. Harper, G. Muresan, B. Liu, I. Koychev, D. Wettschereck, and N. Wiratunga. The robert gordon university's hard track experiments at trec 2004. In *Proceedings of the 13th Text REtrieval Conference (TREC 2004)*. NIST, 2004.
4. D. Kelly, V. Deepak Dollu, and X. Fu. University of north carolina's hard track experiments at trec 2004. In *Proceedings of the 13th Text REtrieval Conference (TREC 2004)*. NIST, 2004.
5. S. Kullback. Information theoery and statistics. *Wiley, New York*, 1959.
6. Lemur Language Modeling Toolkit. http://www.lemurproject.org/. Web, 2005. Last Visited October 2005.
7. J. Xu and W. B. Croft. Cluster-based language models for distributed retrieval. In *SIGIR '99: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 254–261, New York, NY, USA, 1999. ACM Press.