Rutgers' HARD Track Experiences at TREC 2004

N.J. Belkin, I. Chaleva, M. Cole, Y.-L. Li, L. Liu, Y.-H. Liu,
G. Muresan, C. L. Smith, Y. Sun, X.-J. Yuan,X.-M. Zhang
School of Communication, Information and Library Studies, Rutgers University, New Brunswick, NJ 08901
[belkin, chaleva, mcole, lynnlee, luliu, yhliu, muresan, csmith, ysun, xjyuan, xzhang]@scils.rutgers.edu

1 Introduction

The goal of our work in the HARD track was to test techniques for using knowledge about various aspects of the information seeker's context to improve IR system performance. We were particularly concerned with such knowledge which could be gained through implicit sources of evidence, rather than explicit questioning of the information seeker. We therefore did not submit any clarification form[1], preferring to rely on the categories of supplied metadata concerning the user which we believed could, at least in principle, be inferred from user behavior, either in the past or during the current information seeking episode.

The experimental condition of the HARD track was for each site to submit at least one baseline run for the set of 50 topics, using only the title and (optionally) description fields for query construction. The results of the baseline run(s) were compared with the results from one or more experimental runs, which made use of the supplied searcher metadata, and of a clarification form submitted to the searcher, asking for whatever information each site thought would be useful in improving search results. We used only the supplied metadata, for the reasons stated above, and especially because we were interested in how to make initial queries better, rather than in how to conduct a dialogue with a searcher. There were five categories of searcher metadata for each topic (not all topics had values for all five): Genre, Familiarity, Geography, Granularity and Related text(s), which were intended to represent aspects of the searcher's context which might be useful in tailoring retrieval to the individual, and the individual situation. We made the assumption that at least some of these categories would be available to the IR system prior to (or in conjunction with) the specific search session, either through explicit or implicit evidence. Therefore, for us the HARD track experimental condition was designed to test whether knowledge of these contextual characteristics, and our specific ways of using that knowledge, would result in better retrieval performance than a good IR system without such knowledge.

We understood that there would be, in general, two ways in which to take account of the metadata. One would be to modify the initial query from the (presumed) searcher, before submitting it for search; the other would be to search with the initial query, and then to modify (i.e. re-rank) the results before showing them to the searcher. We used both, but mainly concentrated on the latter of these techniques in taking account of the different types of metadata.

2 Hypotheses for How to Take Account of Metadata Categories and Values

Our general approach was to generate hypotheses about how to take account of each of the categories of metadata information in order to improve retrieval effectiveness, to operationalize them in the TREC HARD setting, to test each hypothesis individually on the training corpus, and then to combine the best-performing ones from each category to generate a final result list for the test corpus. Below are summarized the various improvement methods used based on metadata:

Table 1. Methods for improving retrieval effectiveness based on metadata.

Genre					Familiarity			Geography
Regression model	KL distance	Query expansion	Flesch scores	Language models	Readability	Syllables per word	Abstract / concrete words	Language models

2.1 Familiarity

With respect to familiarity, we had two basic ideas. The first is that people who are familiar with a topic will want to see documents which are detailed and terminologically specific, and people who are unfamiliar with a topic will want to see general and relatively simple documents. This we operationalized in two ways. One was that readability, as measured by the Flesch Reading Ease Score[2], would approximate terminological specificity or generality, with documents with low readability being suitable for people with high familiarity with a topic, and documents with high readability being suitable for people with low familiarity with a topic. The second operationalization was to use one factor in the Flesch Reading Ease Score, the mean number of syllables per word in a document, as an indication of terminological specificity. There is good evidence that the number of

[1] See Allan, this volume, for detailed information about the goals and conditions of the HARD track.

[2] Computed using algorithms implemented in Perl by Kim Ryan (Available online: http://aspn.activestate.com/ASPN/CodeDoc/Lingua-EN-Fathom/Fathom.html#SYNOPSIS)

syllables in a word correlates negatively very highly with frequency of the word in English. We took this as a measure of terminological specificity.

The second idea concerning familiarity is based on research indicating differences in the processing of concrete and abstract words in texts (Audet & Burgess, 1999; Barsalou & Wierner-Hastings, 2004; Burgess, Livesay, & Lund, 1998; Schwanenflugel et al, 1988; Schwanenflugel, 1991). One research finding is that people are more easily able to provide distinct contexts for concrete words as compared to abstract words and that comprehension of concrete words takes place more quickly. This led us to hypothesize that people who are unfamiliar with a topic will have difficulty understanding documents that treat a topic abstractly, and will prefer documents that treat the topic with concrete terminology, hence, people with low familiarity with a topic will prefer documents which have a high proportion of concrete terms. Similarly, we hypothesize that people with high familiarity with a topic will prefer documents that have a high proportion of abstract terms.

We used Martindale's Regressive Image Dictionary (RID) for our model of concrete and abstract expression (Martindale, 1990). The RID is a taxonomy of words and word stems arranged in accordance with a psychological theory of consciousness and expression. The theory entails a definition of states of consciousness that lie along a continuum, from cognition as regressive, analogical, and concrete (Primary) to cognition as analytical, logical, and abstract (Secondary). For our purposes, we utilized only the sub-segments of the RID related to expression characterized specifically as concrete and abstract. The concrete terms are in the part of the dictionary related to "deep regression," and specifically connote spatial references, such as *at*, *where*, *over*, *out*, and *long*). Abstraction is in the part of the dictionary termed "secondary process," which is "an inverse indicator of regression"; the list includes terms such as *know*, *may*, *thought*, and *why* (Martindale, 1975).

Our application of the RID was relatively straightforward. Texts were analyzed for term frequency using the Concrete/Abstract word lists. The log of the ratio of concrete to abstract words was calculated for each text (ConAbsFreq), as was the log of the inverse ratio (AbsConFreq) and these numbers were normalized across the corpus to give "concreteness scores" and "abstractness scores".

We thus have three different hypotheses with respect to how to take account of a searcher's familiarity with a topic:

H1: Assessors with low familiarity with a topic are more likely to find that documents on topic that have high readability are relevant, and those with low readability are not relevant; assessors with high familiarity, vice versa.

H2: Assessors with low familiarity with a topic are more likely to find that documents on topic with a low average number of syllables per word are relevant, and those with a high average number of syllables per word are not relevant; assessors with high familiarity, vice versa.

H3: Assessors with low familiarity with a topic are more likely to find that documents on topic which have a high concreteness score are relevant, and those with a high abstractness score are not relevant; assessors with high familiarity, vice versa.

2.2 Genre

For taking account of genre, we had two basic ideas. The first was that news-report documents are objective by nature, and that opinion-editorial documents are subjective by nature. In another research project at Rutgers, HITIQA, a linear regression model had been developed, using a variety of linguistic and formal features of documents, which had proven reasonably accurate in predicting the degree of objectivity or subjectivity of a document (Ng, 2003). The features used include: average length of paragraphs in words, number of paragraphs, frequency of possessive pronouns, frequency of plural proper nouns, frequency of comparative adjectives, frequency of model auxiliary, frequency of question marks, frequency of distinct organizations, frequency of distinct person names and frequency of currency.

We therefore used this model, trained to the HARD-04 training topics and a set of 1000 documents from the New York Times (NYT) sub-collection which we manually classified according to genre, to classify the documents in the retrieved lists according to whether they were objective (for topics with news-report as genre), subjective (for topics with op-ed as genre), or both (for topics with other as genre), in this case classifying what was left as "other". Since the classification is based on a scale, we were able to assign explicit values indicating how well the particular document met each criterion. We also used a Support Vector Machine with linguistic features to accomplish the classification, on the grounds that SVMs usually outperform other learning methods. We saw this as an alternative method of classification, using features similar to those of the regression model.

Our second basic idea for genre was that different document genres can be identified by their vocabularies. This we operationalized by constructing language models for each genre based on the training topics (all the documents in the training collection which shared the same genre) and on our set of manually classified documents. We also constructed a background language model, which consisted of all the documents retrieved in the training collection plus the 1000 NYT documents that we judged ourselves. Then, for each test topic, we constructed a background model for the topic which was all of the retrieved documents for that topic. These data were used in two ways. One was to determine the Kullback-Leibler distance between the language model for each document in the retrieved list for each topic which specified a particular desired genre, and the relevant genre language model. This procedure resulted in a score for each document, indicating its closeness to that genre. The other way we used the data was to identify terms which occurred most frequently in each genre language model, and also terms which occurred significantly more frequently in the genre language model than was expected, given the background model for the training collection. These terms can then be cumulated, and a single list manually constructed of terms likely to be present in documents of the particular genre. These terms can then added to the baseline queries for the test topics to which they were relevant, and the queries run on the original baseline results lists. This will result in new scores, and a new ranking for each affected list.

Thus, our hypotheses with respect to genre are:

H4: The subjectivity or objectivity of a document will determine its membership in a genre; modifying the ranks of documents according to how well they fit to the desired genre will increase performance.

H5: Classification of documents by subjectivity or objectivity using an SVM procedure will lead to better performance than classification using a linear regression model.

H6: Documents of a single genre will have language models characteristic of that genre; the Kullback-Leibler (KL) divergence between a document's language model and the relevant genre language model indicates whether that document is a member of the genre; modifying the ranks of documents according to how well they fit the desired genre will increase performance.

H7: Documents of a single genre will include terms characteristic of that genre; re-ranking lists of documents retrieved according to a topically-based query by running a query containing both the topical terms and the relevant genre terms will increase performance.

Time constraints allowed us to complete evaluation of only **H6**, the use of language models to detect genre.

2.3 Geography

Similarly to the approach taken for genre, we hypothesized that the vocabularies of documents are specific for the geographic area that they refer to. More specifically, US documents should be distinguishable from non-US documents based on their language models (although a significant amount of noise was expected due to documents that covered US and non-US topics, such as world reactions to an event that happened in the US, or trade relations between US and other countries). Our main approach was to build language models based on the training data (positive examples of US documents and positive examples of non-US documents) and to assign "US scores" and "non-US scores" to all documents in the baseline, in order to re-rank the baseline according to how well each document fits the Geography requirement for hard relevance.

In addition, we manually extended the language models by providing geographic names taken from online atlases: for the US model we added names of US states, their capitals and other important cities, while for the non-US model we added all the names of countries and peoples in the world. Our hypotheses related to geography were:

H8: The vocabulary of documents is specific for the geographic area that they refer to. More specifically, US documents can be distinguished from non-US documents.

H9: Names of US states and important US cities are an indication of US documents; names of other countries are an indication of non-US documents.

2.4 The effect of baseline results quality

In order to check the effect of the baseline quality on the effectiveness of our methods for improving retrieval hard effectiveness, we used a range of baselines, generated with different IR systems, or with different parameters.

H10: A better baseline provides better training data and, therefore, it generates metadata scores that better predict document relevance in terms of metadata.

Time constraints did not allow us to complete the evaluation of **H10**.

3 Experimental setting

We considered two different approaches for using metadata information in order to improve baselines. One was to do query expansion based on training data, and to re-run the query. The other was to assign metadata scores for each metadata type and value ("news-report score", "opinion-editorial score", "US score", …) to each document in the baseline, indicating how well a document fits the model built for that metadata. These metadata scores can then be combined with the baseline scores, using different weights that indicate our levels of confidence in different metadata models, to generate final scores. Although both approaches have advantages and disadvantages, we concentrated on the latter due to its flexibility in changing the level of contribution for each metadata: it is possible to simply change the value of weight cells in a spreadsheet, while the former approach requires a re-run of a search. Also, in the latter approach the order in which different metadata models are considered is irrelevant. Note that some metadata models were built based on training data (hard-relevance judgments) and some, such as abstractness/concreteness or readability, were independent of training data.

We used two IR tools to implement our approaches. One was the Lemur IR toolkit[3], which was used (i) to generate baseline result lists; and (ii) to build language models based on the training data for Genre and Geography and generate sets of metadata scores indicating how well each baseline document matched those language models. Based on this approach, we submitted a baseline and two official runs. The second tool was InQuery (Callan et al, 1992) Release 3.2, for which we submitted one baseline run. However, we were unable to complete InQuery test runs using the metadata in time for submission of official test results. The CMU-Cambridge statistical language model toolkit[4] was used for generating Genre language models for the documents in the Inquery baseline.

In both cases, we addressed only the document retrieval problem, and did not attempt passage-retrieval. Also, for both Lemur and InQuery runs, the initial query was constructed using both title and description fields of the topics.

4 Combination of Evidence

4.1 The approach

Figure 1 depicts, in principle, our approach to combining metadata evidence with the original baseline scores. Based on our research hypotheses, we build **metadata models**, which attempt to capture characteristics of documents that satisfy a certain metadata requirement. Some of these models are based on data external to the HARD TREC setting (e.g., readability models, objectivity/subjectivity models, concreteness/abstractness models) or at least not based on official LDC relevance judgments (e.g. Genre models based on our own judgments on a random sample of NYT documents). Other models are based on training data provided by LDC, specifically on positive and negative examples of documents that belong to certain categories of Genre and Geography, as derived from the training relevance judgments.

We then use the training judgments to evaluate these metadata models in terms of how well they support our research hypotheses. The better a model supports the corresponding hypothesis(es), the higher the **confidence level** that we attached to that model. Each metadata model contributes, according to its confidence level, to altering the baseline scores and to the generation of the final scores for the HARD TREC runs. Finally, by comparing the baseline run with the metadata-aware runs, based on LDC judgments, we can draw conclusions with regard to our hypotheses.

Note that confidence levels are relative, rather than absolute, indicating that some metadata models are somewhat more reliable than others. The researchers are expected to try various values for these confidence levels, rather than expect the model evaluation to set exact values. Moreover, using training judgments to evaluate the quality models derived from those some judgments will obviously generate over-inflated confidence levels. While not trusting these evaluation results completely, we can still observe whether the models can capture document characteristics that can predict metadata.

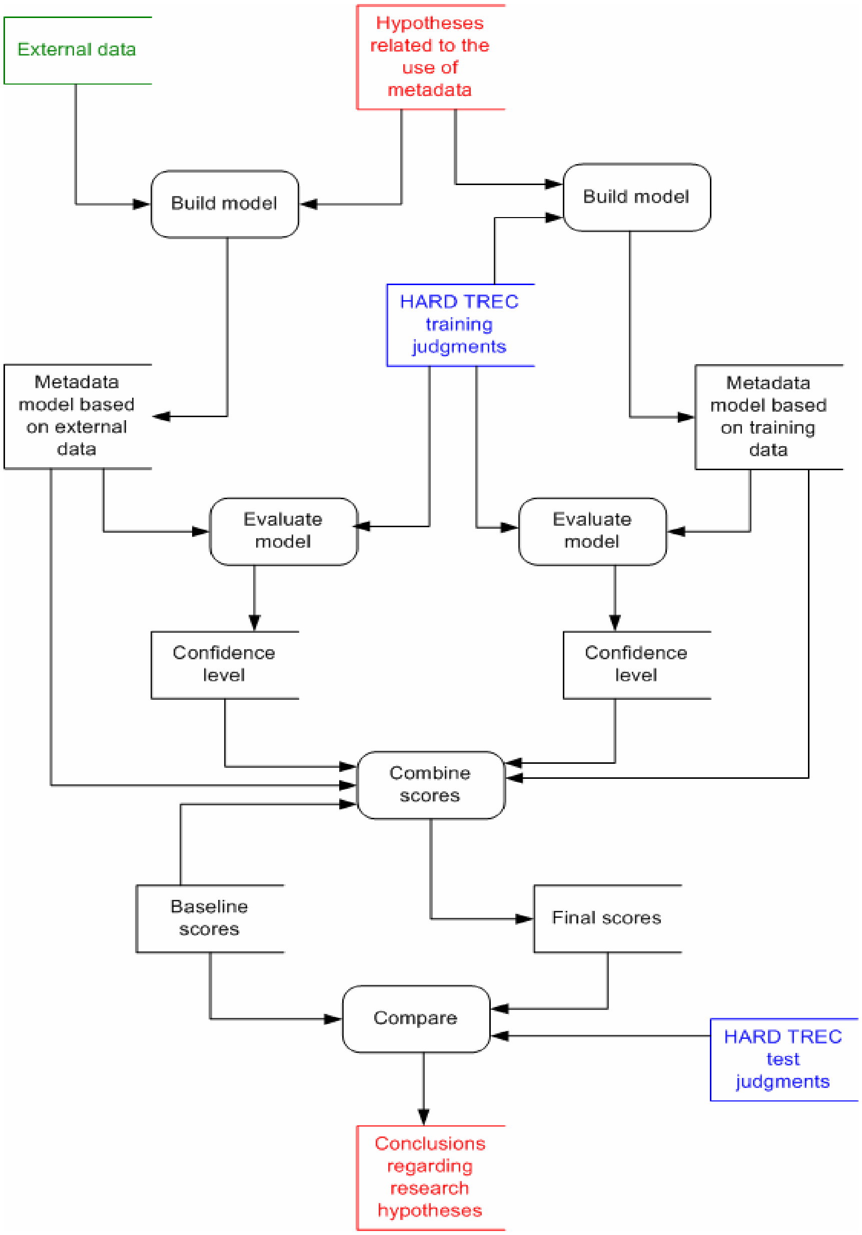
In order to evaluate a metadata model, we propose a two-stage approach:

1. Evaluate the quality of the metadata scores. This can be done by applying a t-test (or its non-parametric equivalent if the data is not normally distributed) to verify that documents that are hard-relevant tend to have better scores than documents that are non-relevant. Alternatively, the problem can be viewed as classification into a relevant and a non-relevant category, so ROC curves can be used to verify how well the metadata scores can predict the metadata category for each document.
2. Apply the metadata scores to the baseline (as described in the next section) and check whether the new scores provide better ranking. The quality of ranking can be estimated, based on LDC's test judgments, (i) by using trec_eval and looking at various measures of effectiveness, in particular R-precision; (ii) by using a measure somewhat more sensitive to effects on individual topics, such as the difference in the sum of the ranks of the relevant documents in the original list, and in the re-ranked list, or, in order to give more prominence to documents at the top of the result list, the mean reciprocal rank (MRR).

It is important to separate the two stages of evaluation, in order to distinguish between the quality of metadata scores and the formulae for combining baseline and metadata scores in order to obtain final rankings.

[3] http://www-2.cs.cmu.edu/~lemur/
[4] http://mi.eng.cam.ac.uk/~prc14/toolkit.html

Figure 1. Combining evidence


External data
Hypotheses related to the use of metadata
Build model
Build model
HARD TREC training judgments
Metadata model based on external data
Metadata model based on training data
Evaluate model
Evaluate model
Confidence level
Confidence level
Combine scores
Baseline scores
Final scores
Compare
HARD TREC test judgments
Conclusions regarding research hypotheses

4.2 The formulae for combining relevance evidence

Here we describe how we generated "new scores" for each of the baseline ranked lists by combining the evidence provided by the original or "old" score of the baseline run for each document with evidence generated by our operationalizations of the various hypotheses with respect to the metadata. We considered two general methods for combining this evidence.

1. **Simple weighted average**

The new scores are the weighted average between the old score and various sets of metadata scores

$$run\_score = baseline\_score + \sum w_i * metadata\_score_i ,$$

with baseline scores and metadata scores normalized:

$$normalized\_score = \frac{score - \min\_score}{\max\_score - \min\_score}.$$

The weights of the metadata scores reflect the confidence that we have in that metadata to improve the baseline. This is the method that we used to generate our official runs.

2. Z-scores based combination

This method offers a principled way to re-rank the baseline. The assumption is that the metadata scores are normally distributed and that values within one standard deviation from the mean should have little or no influence on the baseline. Values above one standard deviation will have an effect (positive or negative) on the baseline that increases with the value. The formula applied is:

$$new\_score = old\_score \pm d * z * k,$$

where:

- d is the average score difference between adjacent documents in the baseline:

  $d = \frac{\max\_score - \min\_score}{n-1}$, n being the number of documents in the baseline;

- z is the Z-score of the metadata:

  $z = \frac{metadata\_score - \mu}{\sigma}$, where the mean is $\mu = \frac{\sum metadata\_score}{n}$,

  and the standard deviation is $\sigma = \sqrt{\frac{\sum (metadata\_score - \mu)^2}{n}}$,

  or, in computational form, $\sigma = \sqrt{\frac{\sum metadata\_score^2 - (\sum metadata\_score)^2}{n}}$.

- k is a coefficient that indicates how confident we are in the metadata and implicitly what influence it should have on the baseline; one can start with value 1, and then increase it; k-values corresponding to different metadata can be chosen to reflect the confidence level attached to the corresponding model;
- + or – is determined according to specific characteristics of the metadata values and the desired re-ranking effect

5 Results of Official Runs

While successfully preparing the groundwork for such a high number of hypotheses, we were unable to complete the experimental process described in the previous section before the deadline for submitting official runs. In the rest of this paper we concentrate on describing and analyzing our official runs, and describe plans for future work in view of generating and analyzing unofficial runs, which will potentially shed light on the accuracy of our hypotheses.

We submitted two official runs obtained by combining the Lemur baseline scores with Genre and Geography scores obtained by building language models based on the LDC training judgments, and assigning each document sets of metadata scores indicating how well the document fit the model. We had not managed to compute confidence levels or to implement the planned combination of metadata scores, so we simply applied a weighted average of scores, with a weight of 1.0 assigned to the baseline score (indicating topicality) and a weight of 0.1 for the first run, respectively 0.2 for the second run, to all the metadata scores.

Table 2. Effectiveness of baseline and test runs(* p< .05, ** p < .01)

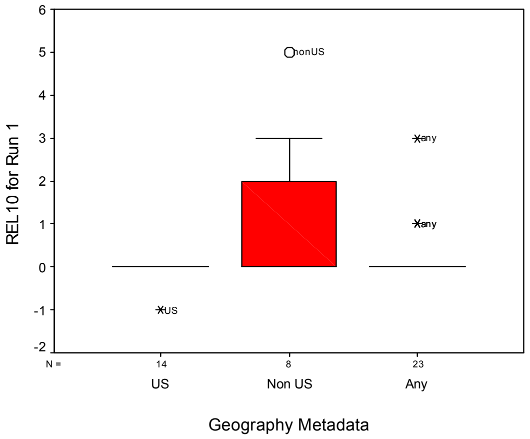
	HARD			SOFT		
	Avg. Prec.	Rel10	R-Prec.	Avg. Prec.	Rel10	R-Prec.
Baseline	0.1697	2.356	0.1957	0.1690	2.867	0.1957
(SD)	*(0.1999)*	*(2.789)*	*(0.2133)*	*(0.1841)*	*(3.020)*	*(0.1809)*
Run 1	0.1845**	2.644*	0.1867	0.1788**	3.133*	0.2103**
(SD)	*(0.2065)*	*(3.053)*	*(0.2196)*	*(0.1861)*	*(3.209)*	*(0.1888)*
Run 2	0.1624	2.333	0.1672	0.1608	2.911	0.1911
(SD)	*(0.1983)*	*(2.852)*	*(0.2080)*	*(0.1746)*	*(3.029)*	*(0.1666)*

Table 2 depicts our results. As expected, the three performance measures (average precision (AvgPrec), relevance at top 10 documents (Rel10), and R-precision (RPrec)) for the baseline and two experimental runs were not normally distributed. Tests for the significance of mean differences were therefore done using non-parametric methods, specifically Wilcoxon and Kruskall-Wallis.

For our first experimental run (Run 1), soft relevance performance for was superior for all three. For Run 1, HARD relevance performance was also greater for AvgPrec and Rel10 measures. Our second experimental run resulted in no significant difference in performance relative to baseline, but the performance was consistently below the baseline.

We analyzed Run 1 HARD relevance performance, looking for differences related to Genre, Geography, and Familiarity metadata. In a comparison of the four types of genre preference, performance did not vary significantly within the metadata group. No significant performance differences were found in a comparison of the two familiarity levels. Geography was the only metadata type for which a significant performance difference was found between specified values (*US*, *non-US*, and *any*) ($\chi^2 = 8.263$, $df = 2$, $p < .05$). Visual inspection of box-plots for geography revealed that topics specifying non-US geography appear to have been the major factor differentiating performance among the three geography preferences.

Fig. 2 Performance and geography metadata



6 Analysis and Interpretation

6.1. *Overall results*

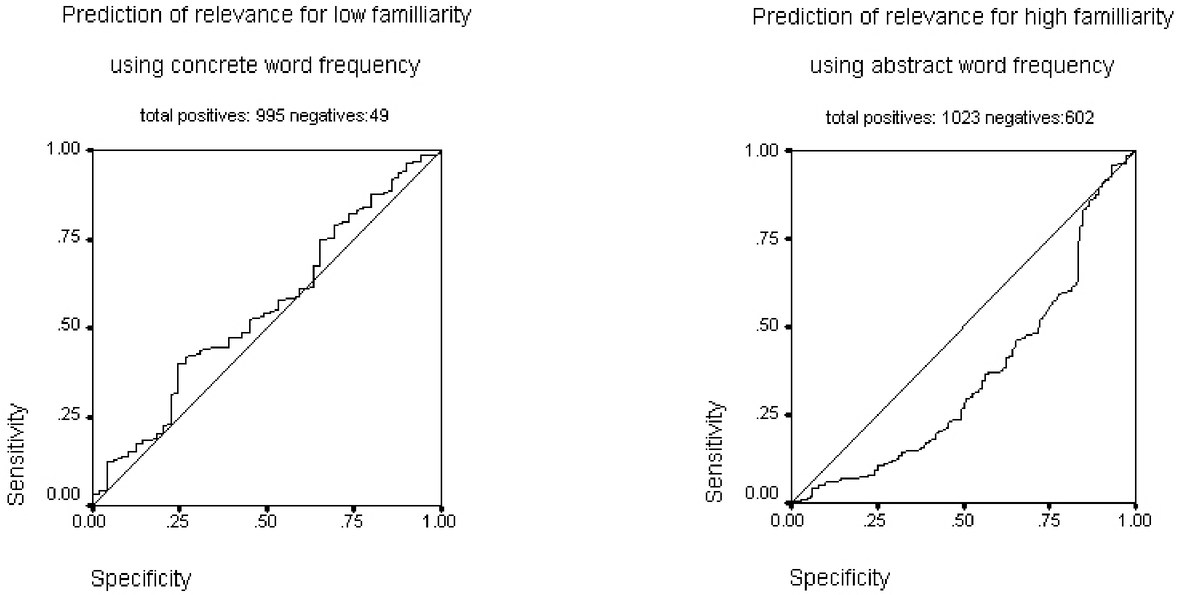
Our official runs reflect a very narrow part of our investigation, namely that related to Genre hypothesis **H6** and Geography hypothesis **H8**. The results indicate support for H8 in Run1, i.e. taking Geography into account can potentially improve effectiveness. However, when too much importance is given to metadata compared to topicality, as in Run2, performance actually decreases.

The purpose of the HARD TREC experiment is to try and improve retrieval effectiveness, as measured by hard relevance, with the expectation that soft (topical) effectiveness will be sacrificed. It is surprising, therefore, to observe that our test run RUN1 shows improvement over the baseline in terms of both soft and hard effectiveness. This could be attributed to poor relevance judgments, but it is more likely due to a correlation between Genre and topicality; for example, some topics may be likely to appear in US documents, while other topics may be more typical for non-US documents.

6.2 *Evaluation of metadata scores*

In order to have a better understanding of these results, we also analyzed the quality of metadata scores, as reflected by evaluation judgments from LDC. The ROC curves below depict the capacity of metadata scores to distinguish between documents judged relevant or non-relevant in terms of their metadata (Genre, Geography, Familiarity).

Fig. 3 Abstract and concrete word frequencies are poor predictors of documents selected for familiarity.



For some of the metadata, the training and testing sets had large imbalances of positive and negative judgments, so the interpretation of the ROC curves for general comment is difficult. For example, in the case of the opinion-editorial genre there were 123 positive assessments and 1185 negative assessments. In general though, if the scores curve is under the LDC judgment line, it means the document scoring has performed poorly and high document scores are not well correlated with the metadata judgments. Such a result should raise serious doubts as to whether that metadata can be reliably detected and so whether reference to that metadata should be included in a new relevance prediction system. However, if a significant portion of the curve is better than the LDC judgment prior, it may indicate that a range of scores is a good signal for the metadata.

Figure 4 The highest readability scores have some correlation with documents relevant for readers with low familiarity to the topic and are neutral for readers expressing high familiarity.

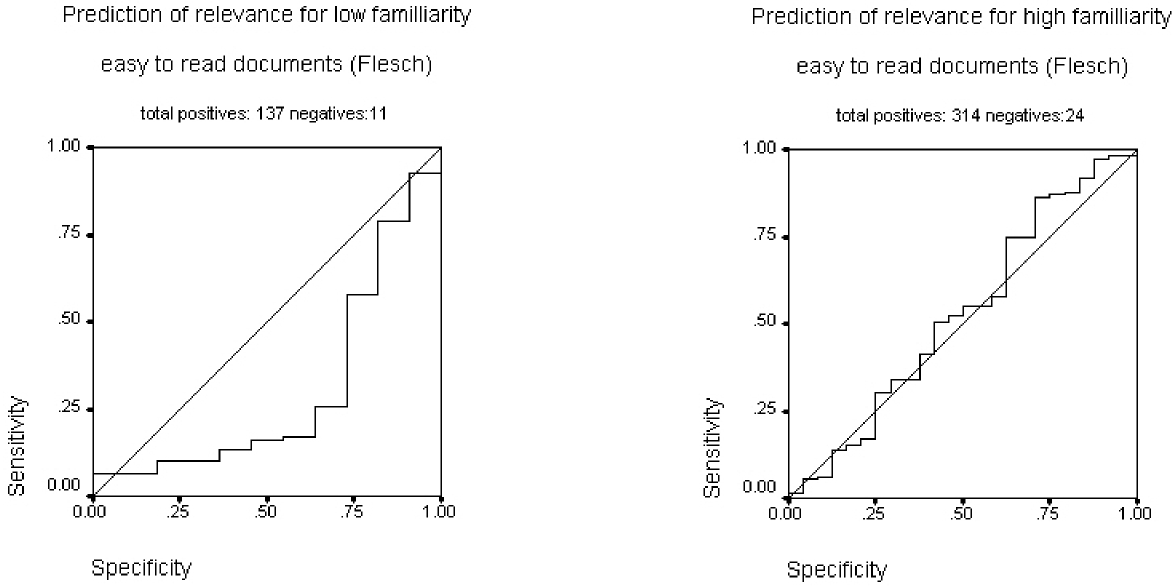


Figure 5 Language model prediction of geography may be promising

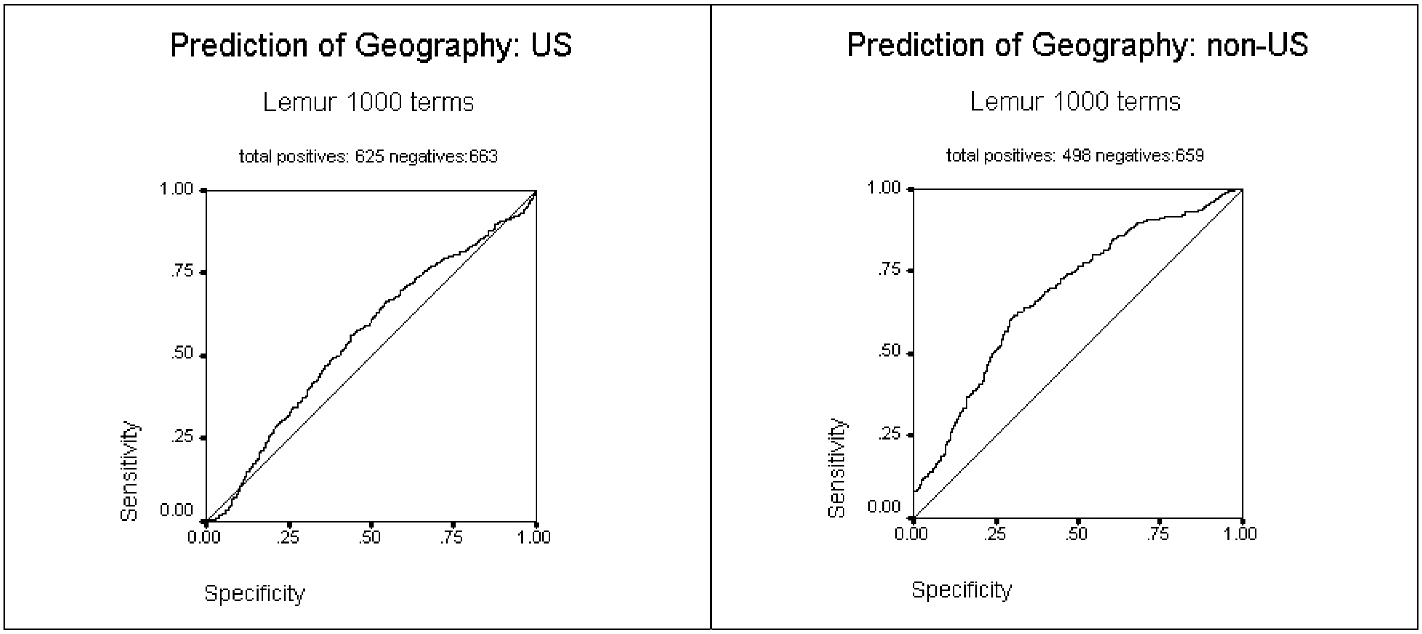
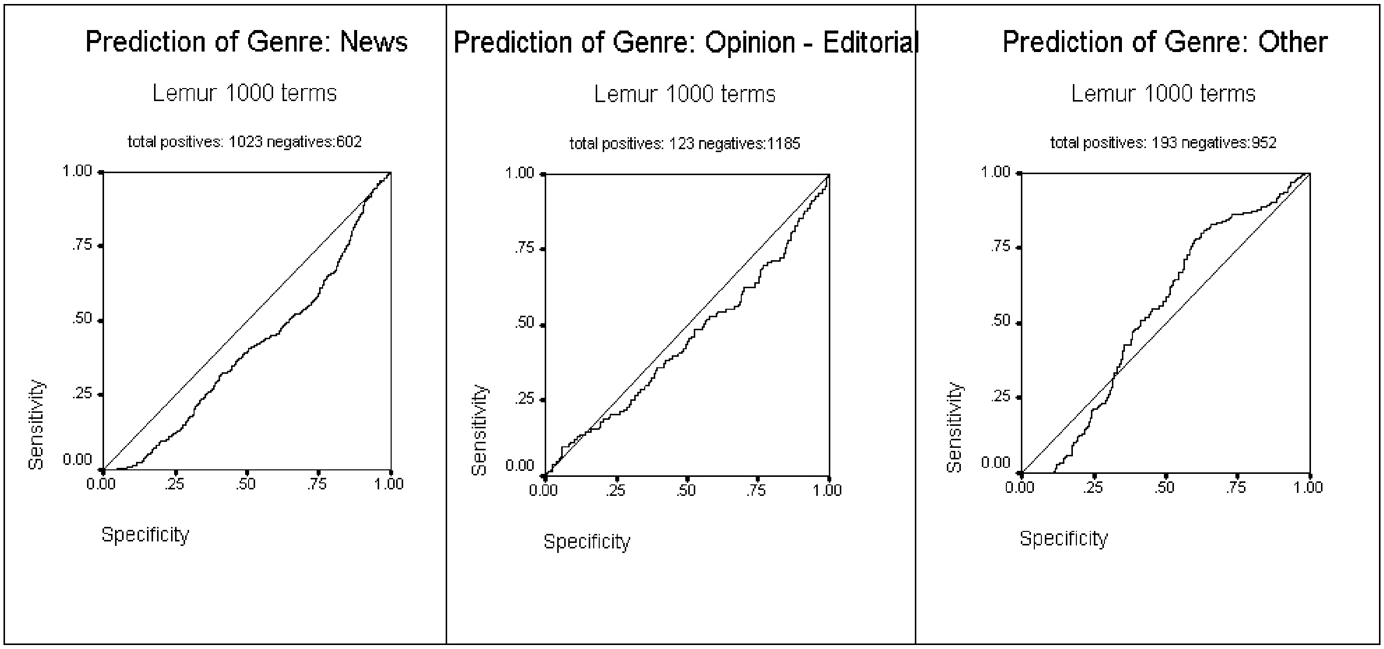


Figure 6 Language model prediction of Genre ("news-report", "opinion-editorial", "other") performs poorly.



It seems that the geography metadata, especially non-US, were well-detected in the Lemur testing runs. The genre metadata is another matter. The "news-report" and "other" scores are anti-correlated at the highest scoring levels, while "opinion-editorial" does not seem to have any effect. A possible explanation is that, as the language models were built based on the training judgments, they captured the 23 training topics rather than the genre of the documents (i.e. the style of the vocabulary). As the test topics are distinct from training topics, not only cannot we expect an improvement in performance, but a decrease in performance will result, as documents with the wrong topic are pushed towards the top of the ranked result list.

In order to test if our explanation is reasonable, we will build Genre models independent of the training topics and judgments; for example, we can use the random sample of 1000 Documents that whose genre we judged. As those documents were picked at random, a diversity of topics can be expected, so the model should not capture some training topics.

6.3 Using KL divergence to detect genre and re-rank documents

To test hypothesis **H6**, we used the CMU-Cambridge statistical LM toolkit (Clarkson.and Rosenfeld, 1997). to construct a language model for each genre based on training topics. We also constructed a background language model for all the retrieved documents. Based on these language models, we can use KL divergence to evaluate the difference between them. KL divergence measures the distance between two probability distributions. Smaller KL divergence values indicate greater similarity between the two distributions. For two identical distributions the KL divergence is equal to 0.

Given two distributions *q* and *r*, the KL distribution is calculated (Lee, 2001).

$$D(q,r) = \sum_y q(y)(\log q(y) - \log r(y)),$$

where y ranges over all words in the vocabulary

In our case, q(y) is the distribution for all the retrieved documents and r(y) is the distribution for the relevant genre language model.

This basic KL divergence calculation cannot be applied directly to our problem because a given word from the retrieved documents may not exist in the genre model vocabulary. In such a case log r(y) is infinite and the formula breaks down. A skewed KL divergence formula (Lee, 2001) can be used to avoid this problem. It replaces *r* with an average of *r* and *q* in the original KL divergence formula:

$$D(q, \alpha * q + (1-\alpha) * r) = \sum_y q(y) * (\log q(y) - \log(\alpha * q(y) + (1-\alpha) * r(y))$$

where y ranges over all words in the vocabulary and α is a constant.

Using the skewed KL divergence formula, we calculated a new score for the LDC training and re-ranked the documents according to the new score. The effectiveness of the treatment is evaluated as:

Effectiveness = (sum of the ranks of hard relevant documents with old scores) –

(sum of the ranks of hard relevant documents with new scores) ,

so a positive result indicates an improvement in document rankings.

To our surprise, we found the value of the effectiveness measure to be negative. The re-ranking treatment pushed most hard relevant documents to higher ranks, towards the bottom of the ranked list. These results go against our intuition, so we plan more experiments to understand what happened.

7 Discussion and Conclusions

With regard to running the HARD TREC experiment, we have learnt that planning and prioritizing are essential. We had an extensive set of research hypotheses based on an extensive literature survey, had an elaborate plan of work for systematically exploring our hypotheses, spent time creating our own Genre annotations for a random sample of 1000 documents, and even used two different IR systems in order to make our investigation more complete. Progressing in parallel on all the fronts proved to be the wrong approach: before the submission deadline we were way beyond the plan and had to submit runs based on guesswork rather than on the planned intermediary results. More specifically, we had no estimation of confidence levels for different metadata scores and simply assigned all metadata the same importance. In retrospect, an iterative approach could have worked better, with the most important hypotheses tested first; if time allows it, the experiment can be extended to test other hypotheses.

With regard to building models for personalization, we can conclude that, once a searcher's preferences are known based on some implicit sources of evidence (implicit relevance feedback), models built independently of the learning sample seem more reliable than those based on exemplar documents. For example, instead of using language models that may capture the topics of the sample documents offered as exemplars of relevant documents, one should seek document features that can capture the style, rather than the topic of such documents: readability, measures of concreteness/abstractness, etc.

Using language models to capture Genre preference was a complete failure, presumably because the language models captured the topics of the training documents. Somewhat surprising was the success of language models for Geography metadata; we need to investigate further, in order to understand the difference between Genre and Geography models.

Flesch readability scores were successful in detecting the simplest documents for people with low familiarity to the topic. We were less successful in providing the right documents for people with high familiarity to the topic. This may be due to the existence of a large number of "bad" documents (in a foreign language, or containing just list of articles, or lists of phone numbers) in the corpus, and our model may not have distinguished between them and valid but complex documents. More work on examining the results and filtering out bad documents may help.

Some future work that we plan to do in order to extend this work and better understand our results is describe below:

- When indexing the document collection, we applied stopword removal. It later occurred to us that the distribution of function words in documents may be an indication of the documents' genre. Therefore, we plan to re-run the part of the experiment that employs language models for generating genre scores, this time without removing stopwords.
- The hypotheses that document vocabulary properties will determine membership in a certain genre seem to be promising and we plan to continue our work to determine if genre-specific terms exist and whether genre can be detected by measures of the subjectivity or objectivity of the genre documents. We also plan to run experiments using classifiers other than the simple linear regression model, for example SVMs. Application of language models to characterize terms specific to a genre will also be pursued, with more attention given to the experimental design, so that the training captures the genres rather than the topics of the exemplary documents.
- When using Lemur to assign Genre and Geography scores to each document in a baseline, we used smoothing based on the baseline. We plan to repeat that part of the experiment, but to use smoothing based on the entire HARD-04 document collection; this way we hope that the language models will better capture the genre of the documents used as positive example, rather than their topics.
- On a related issue we plan to try and build Genre language models based on the random sample of 1000 NYT documents that we judged ourselves, instead of the training documents judged and provided by LDC. The LDC documents were all results of searches based on 23 training queries; using for training a random sample of documents will increase the topical diversity of the sample and will increase the chances that the language models capture genre rather than topicality.

Finally, we suggest that the quality of the test collection be improved for future similar experiments. The collection contained a high number of bad documents, which should have been filtered out, the quality of the training judgments was rather poor (only two documents were rejected based on familiarity mismatch, only two positive examples of "opinion-editorial" documents were offered, etc.), and some of the relevance judgments used in evaluation were clearly questionable (some documents were rejected on grounds of genre mismatch, even if the request was for "any").

8 Acknowledgments

Jussi Karlgren, of the Swedish Institute of Computer Science, collaborated with us in investigating determination of genre. David Harper, of the Robert Gordon University, Aberdeen, Scotland, helped us with useful suggestions, Perl scripts, and many hours of fruitful discussions.

The HITIQA project is supported by the Advanced Research and Development Activity (ARDA)'s Advanced Question Answering for Intelligence (AQUAINT) Program under contract number 2002-H790400-000 to SUNY Albany, with Rutgers as a subcontractor.

9 References

Audet, C and Burgess, C. (1999) *Using a high-dimensional memory model to evaluate the properties of abstract and concrete words*, in Proceedings of the 20th Annual Conference of the Cognitive Science Society, Vancouver, BC, Canada, pp 37-42.

Barsalou,L., and Wierner-Hastings, K. (2004). *Situating abstract concepts*. To appear in D. Pecher and R. Zwaan (Eds.), Grounding cognition: The role of perception and action in memory, language, and thought. New York: Cambridge University Press.

Burgess, C., Livesay, K, Lund, K. (1998). *Explorations in Context Space: Words, sentences, discourse*. Discourse Processes, 25, 211-257.

Callan, James C. and Croft, W. Bruce and Harding, Stephen (1992) *The INQUERY Retrieval System*, in Proceedings of the 3rd International Conference on Database and Expert Systems.

Clarkson, P.R.and Rosenfeld. R. (1997). Statistical language modeling using the CMU-Cambridge toolkit. Proceedings ESCA Eurospeech 1997.

Lee, L. (2001).*On the effectiveness of the skew divergence for statistical language analysis*. Artificial Intelligence and Statistics 2001 pp 65-72.

Martindale, C. (1975). *Romantic progression: The psychology of literary history*. New York: John Wiley & Sons.

Martindale, C. (1990). *The clockwork muse: The predictability of artistic change*. Basic Books.

Ng, K.B., Kantor, P., Tang, R, Rittman, R., Small, S., Song, P., Strzalkowski, T. Sun, Y. and Wacholder, N. (2003) *Identification of effective predictive variables for document qualities*. In Proceedings of 2003 Annual Meeting of American Society for Information Science and Technology. Medford, NJ: Information Today, Inc.

Schwanenflugel, P. (1991). *Why are abstract concepts hard to understand?* In. P.J. Schwanenflugel (Ed.), The psychology of word meaning (pp. 223-250), Mahwah, NJ: Erlbaum.

Schwanenflugel, P., Harnishfeger, K., & Stow, R. (1988). *Context availability and lexical decisions for abstract and concrete words*, Journal of Memory & Language, 27, 499-520.