University of Chicago at TREC 2004: HARD Track

Gina-Anne Levow
University of Chicago
levow@cs.uchicago.edu

Abstract

The University of Chicago participated in the Text Retrieval Conference 2004 (TREC 2004) HARD track. HARD track experiments focused on passage retrieval and exploitation of metadata information to increase accuracy under HARD-relevance criteria. Passage retrieval employed query-based repeated merger of 2-3 sentence pseudo-documents to extract tightly focused relevant passages. Lexical cues and statistical language modeling were applied to identify documents consistent with specified metadata criteria. Retrieval lists were reranked based on the quality of metadata match.

1 Introduction

The University of Chicago participated in the Text Retrieval Conference 2004 (TREC 2004) HARD track. The HARD track emphasizes high accuracy retrieval constrained additional metadata annotations, restricting otherwise on-topic documents to conform to additional criteria including genre, geography, familiarity, and retrieval unit size in order to be assessed as relevant. HARD track experiments focused on passage retrieval and exploitation of metadata information to increase accuracy under HARD-relevance criteria. The submitted runs contrast the use of pure lexical cues to metadata-based classification with integration of evidence from statistical language models.

Section 2 presents the general retrieval architecture under which the HARD retrieval runs were performed. Section 2.2 describes the methodology employed for passage extraction. Section 3 describes the lexical and language modeling approaches to identifying consistency with metadata specifications and explains the reranking procedure through which this classification influenced the retrieval results. Section 4 briefly describes the overall experimental results.

2 Retrieval Architecture

The information retrieval framework used for the 2004 HARD track experiments at the University of Chicago employed the INQUERY information retrieval system version 3.1p1 licensed from the University of Massachusetts(Callan et al., 1992), based on belief networks.

2.1 Query Formulation

The University of Chicago submitted official runs in the title, description, and combined title and description conditions. Queries were formed by concatenating the appropriate sections of the topic specification. Stopwords were removed based on INQUERY's default stopword list, and stemming was performed with the integrated *kstem* stemmer. No addition stop structure removal was done.

Queries were enriched through blind-pseudo-relevance feedback. We employed the INQUERY API to identify enriching terms based on the top 10 ranked retrieved documents and integrated these terms with the original query forms. The terms were added by concatenation with the original query with no additional re-weighting. Our hope was that this enrichment process would capture additional on-topic terminology absent from the relatively short query forms.

2.2 Passage Construction

We viewed the passage extraction process as analogous to the story segmentation task in spoken document retrieval with unknown story boundaries. Using an approach inspired by (Abberley et al., 1999), we performed passage extraction as follows. First we created 2 sentence segments with a single sentence overlap from the body of the original document. These units were then indexed using the INQUERY retrieval system version 3.1p1 with both stemming and standard stopword removal.

2.2.1 Retrieval Segment Construction

To produce suitable retrieval segments, we merged the fine-grained segments returned by the base retrieval process on a per-query basis. For each query, we retrieved 5000 fine-grained segment windows. We then stepped through the ranked retrieval list merging overlapping segments, assigning the rank of the higher ranked segment to the newly merged segment. We cycled through the ranked list until convergence. The top ranked 1000 resulting merged passages formed the final ranked retrieval results submitted for evaluation.

2.3 Document Retrieval

For retrieval of full documents, indexing, with standard stopword removal and stemming, was performed on all text fields of the document. The top 1000 documents were returned.

3 Metadata-based Document Selection

For the HARD 2004 experiments, we chose to focus on exploiting topic metadata for HARD relevance. Due to a paucity of training data for FAMILIARITY metadata, we chose to focus on GENRE and GEOGRAPHY to constrain and re-evaluate baseline retrieval results. Two approaches were explored for assessing possible relevance relevant to the metadata annotations: direct lexical cues and statistical language model perplexity.

Direct Lexical Cues We performed a series of simple classification experiments to distinguish US/NOTUS geographic references and OP-ED/NEWS genres. Based on the 2004 HARD training data, we employed Boostexter (Schapire and Singer, 2000) to train appropriate genre and geographic classifiers. The full text of articles that were annotated on-topic but not relevant due to metadata were used as exemplars of the contrasting metadata value; articles annotated as HARD-relevant were used as positive exemplars of their corresponding metadata values. Boostexter was set to learn n-gram features of varying lengths to perform the classification. While the classification accuracy was too low to use the classification directly, the features selected by the classifiers provided evidence for the type of information that could be useful in these decisions. Specifically, geographic references - by region, country, or city - in header and byline information yielded useful cues, while genres were often distinguished by the presence, in editorial pieces, of first and second

Query Form	Baseline	Lexical Only	Lexical & LM
Doc Title+Descr	0.2683	0.2645	
Pass Title+Descr	0.2514	0.2550	0.2398

Table 1: Document-based R-Precision for Title+Descr Queries

Query Form	Baseline	Lexical Only	Lexical & LM
Pass Title	0.198	0.197	0.1473
Pass Title+Descr	0.156	0.155	0.134

Table 2: Passage-based bpref@12000 for Title+Descr Queries

person pronouns and their absence in news articles. The geographic locations identified by Boostexter were augmented with city, country, and region name information from on-line lists, divided into US and non-US groupings. Genre information in the non-body fields of the article was also employed.

Statistical Language Modeling We also chose to employ a smoother statistically-based measure of geographic or genre consistency. Again using the HARD 2004 training data labeled for geographic and genre category as above, we employed the CMU-Cambridge language modeling toolkit(Clarkson and Rosenfeld, 1997) to build trigram language models of each class. We determine metadata consistency by pairwise comparison of the perplexity of the text of the current article relative to each class for each metadata tag.

Metadata-based Reranking The lexical and language model information about consistency with metadata annotation was used to rerank the documents or passages in the baseline ranked list. Each explicit metadata value match promoted the corresponding article, while each explicit mismatch demoted the corresponding article. Rank was increased or decreased a fixed step size. When both evidence from both lexical cues and language models was available concurrently, additional consistent evidence increased step size, while conflicting evidence decreased step size.

4 Results and Discussion

Runs for the best query formulation - generally combined topic and description with relevance feedback - yielded middle-of-the-road effectiveness. Both document and passage retrieval based measures matched the median closely. Title only and description only queries performed less well on document based measures. In passage-based evaluation, though, title-only queries often performed competitively. Lexically based metadata-based reranking yielded no significant change over baseline retrieval effectiveness overall. Incorporation of the language modeling based measures appeared to somewhat decrease effectiveness. Inspection of the training data indicated some apparently noisy annotations that could possibly have limited the accuracy of trained statistically based measures.

Future experiments will include modifications to the passage retrieval strategy to integrate a deeper model of topical cohesion and the use of clarification forms to provide a richer source of relevance information.

References

Abberley, D., Renals, S., Cook, G., and Robinson, T. (1999). Retrieval of broadcast news documents with the thisl system. In Voorhees, E. and Harman, D., editors, *Proceedings of the Seventh Text REtrieval Conference (TREC-7)*, pages 181–190. NIST Special Publication 500-242.

Callan, J. P., Croft, W. B., and Harding, S. M. (1992). The INQUERY retrieval system. In *Proceedings of the Third International Conference on Database and Expert Systems Applications*, pages 78–83. Springer-Verlag.

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

Schapire, R. E. and Singer, Y. (2000). Boostexter: A boosting-based system for text categorization. *Machine Learning*, 39(2–3):135–168.