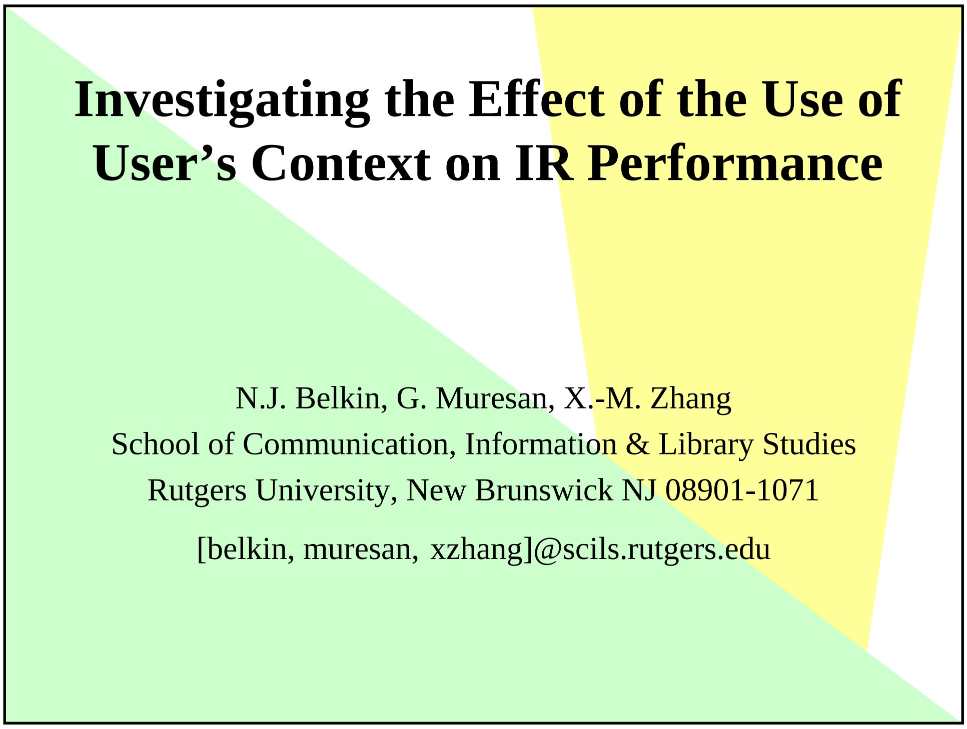
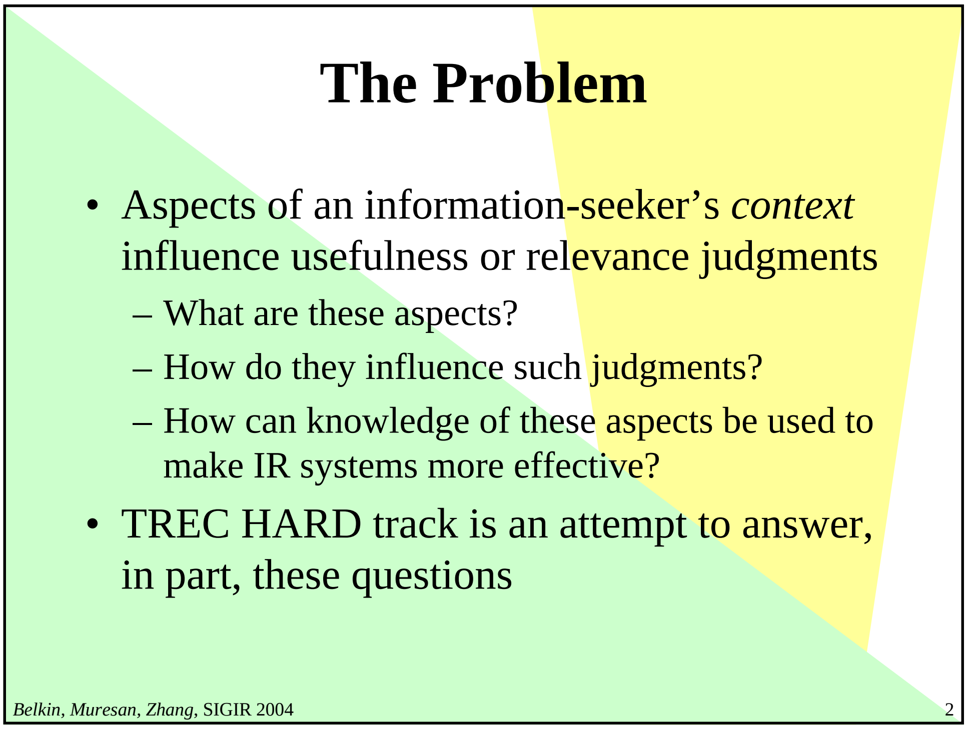


Investigating the Effect of the Use of User’s Context on IR Performance

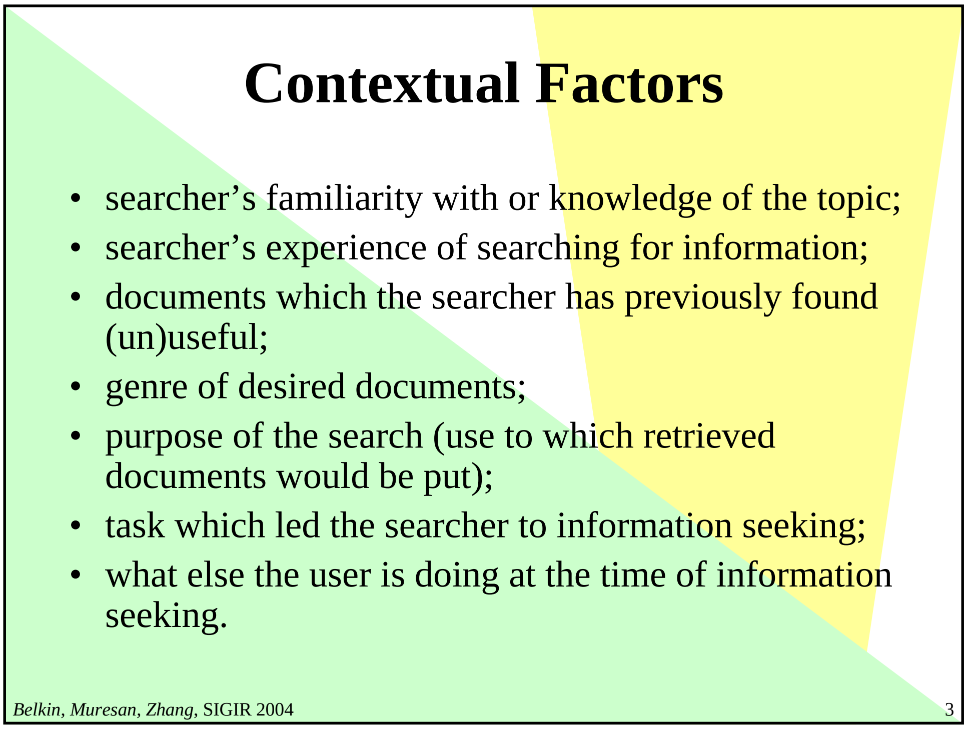
N.J. Belkin, G. Muresan, X.-M. Zhang
School of Communication, Information & Library Studies
Rutgers University, New Brunswick NJ 08901-1071

[belkin, muresan, xzhang]@scils.rutgers.edu



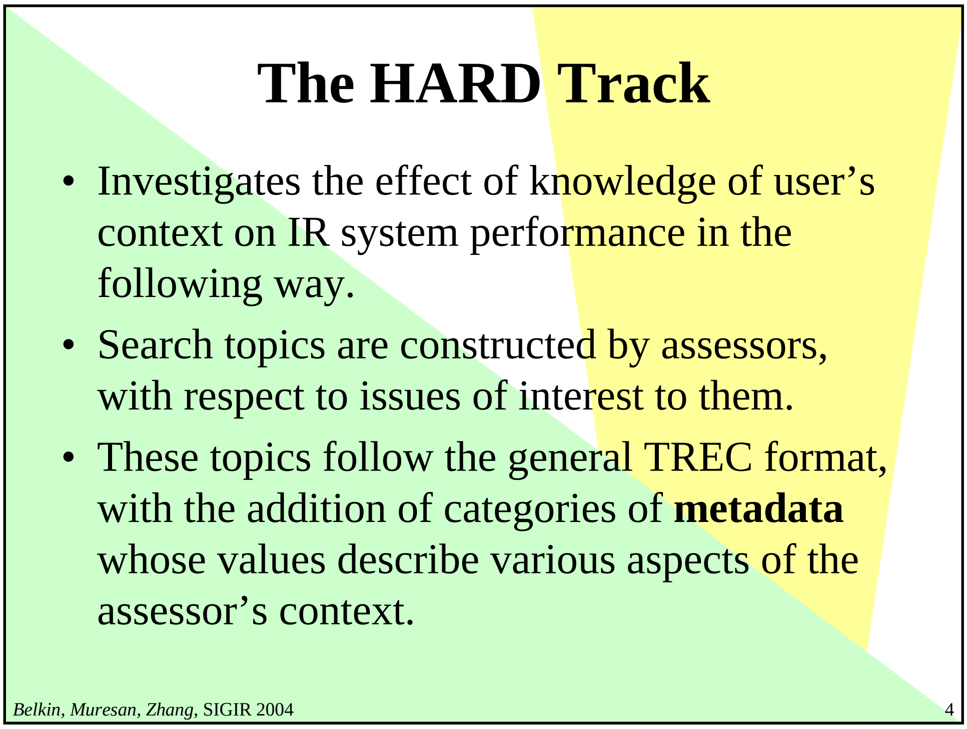
The Problem

- Aspects of an information-seeker’s *context* influence usefulness or relevance judgments
  - What are these aspects?
  - How do they influence such judgments?
  - How can knowledge of these aspects be used to make IR systems more effective?
- TREC HARD track is an attempt to answer, in part, these questions



Contextual Factors

- searcher’s familiarity with or knowledge of the topic;
- searcher’s experience of searching for information;
- documents which the searcher has previously found (un)useful;
- genre of desired documents;
- purpose of the search (use to which retrieved documents would be put);
- task which led the searcher to information seeking;
- what else the user is doing at the time of information seeking.



The HARD Track

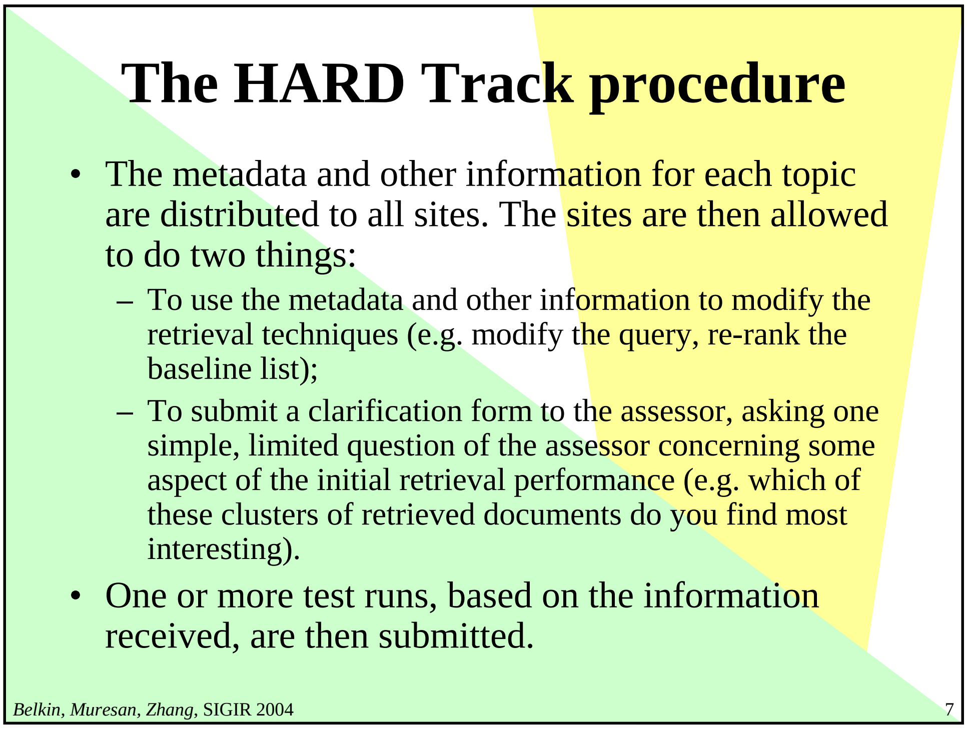
- Investigates the effect of knowledge of user’s context on IR system performance in the following way.
- Search topics are constructed by assessors, with respect to issues of interest to them.
- These topics follow the general TREC format, with the addition of categories of **metadata** whose values describe various aspects of the assessor’s context.

The HARD Track (cont’d)

- In TREC 2003, the metadata were:
  - familiarity with the topic
  - desired genre of retrieved documents
  - purpose of the search
  - specification of geographic focus of documents
- In TREC 2004, the categories of metadata were reduced to:
  - knowledge of the topic
  - desired genre of retrieved documents
  - documents should be about USA, or not USA

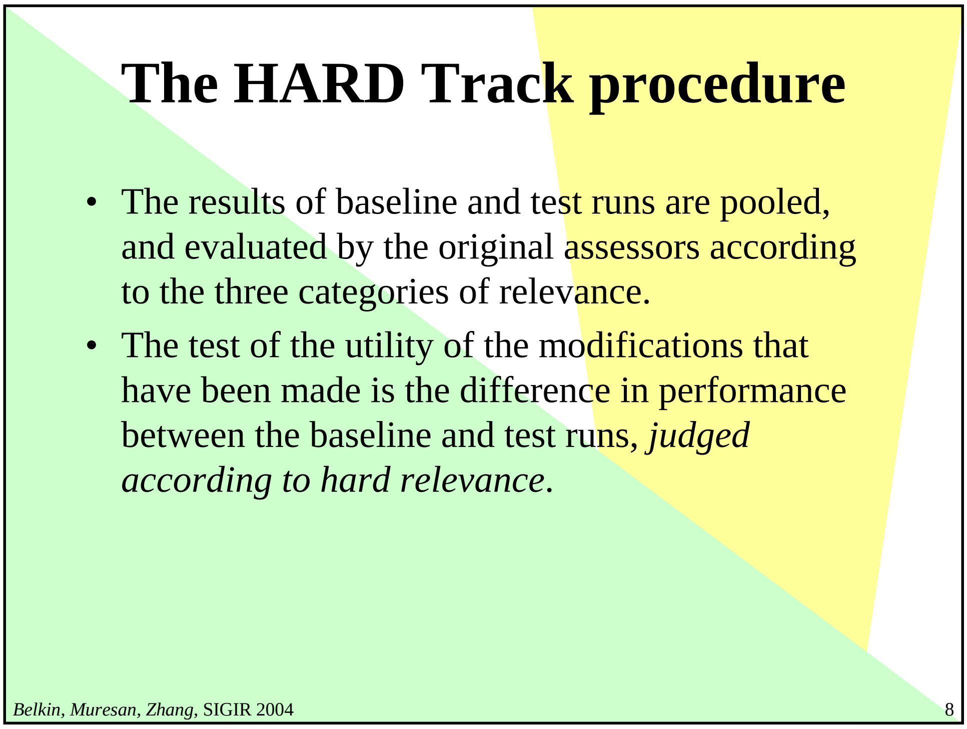
The HARD Track procedure

- Corpus and training topics are distributed to participating sites
- Topics include metadata and 100 documents which have been judged either *not* relevant, *soft* relevant (meaning on topic), and *hard* relevant (meaning on topic, *and* satisfying the metadata conditions).
- 50 test topics are distributed, *without* the contextual metadata.
- Each site constructs a query for each topic, searches the corpus and returns a ranked list of documents for each topic. This constitutes the *baseline* run.



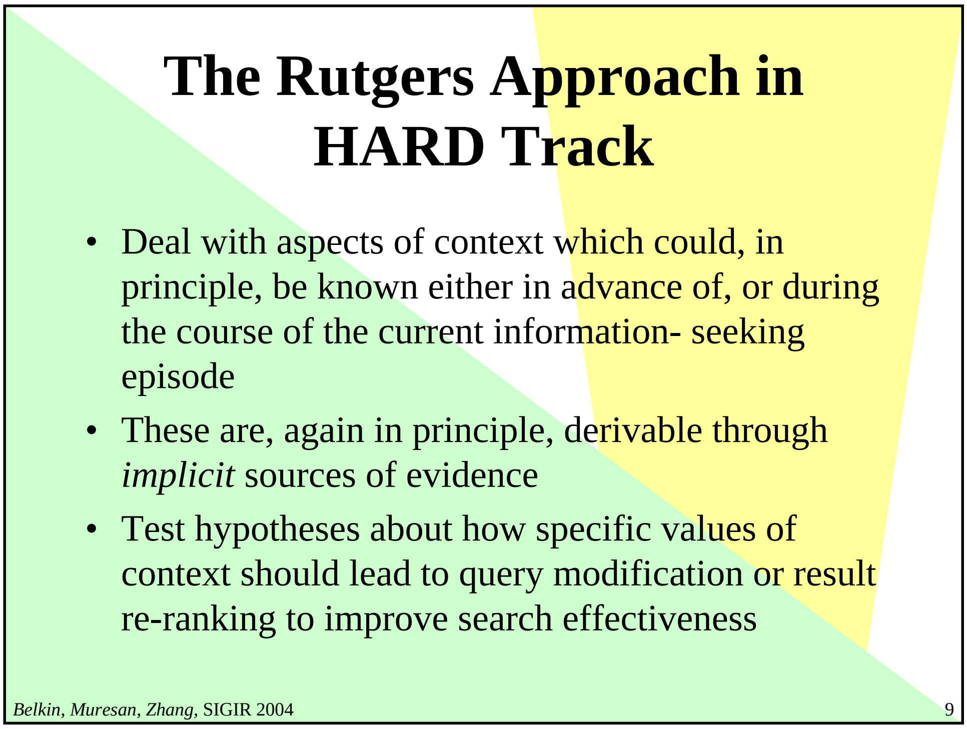
The HARD Track procedure

- The metadata and other information for each topic are distributed to all sites. The sites are then allowed to do two things:
  - To use the metadata and other information to modify the retrieval techniques (e.g. modify the query, re-rank the baseline list);
  - To submit a clarification form to the assessor, asking one simple, limited question of the assessor concerning some aspect of the initial retrieval performance (e.g. which of these clusters of retrieved documents do you find most interesting).
- One or more test runs, based on the information received, are then submitted.



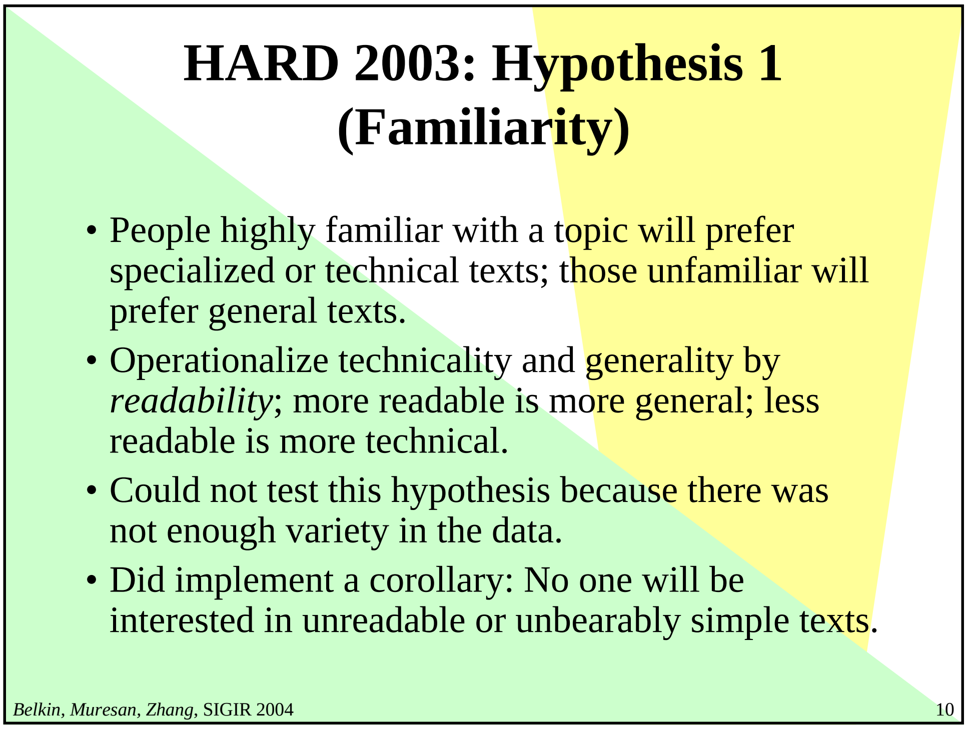
The HARD Track procedure

- The results of baseline and test runs are pooled, and evaluated by the original assessors according to the three categories of relevance.
- The test of the utility of the modifications that have been made is the difference in performance between the baseline and test runs, *judged according to hard relevance*.



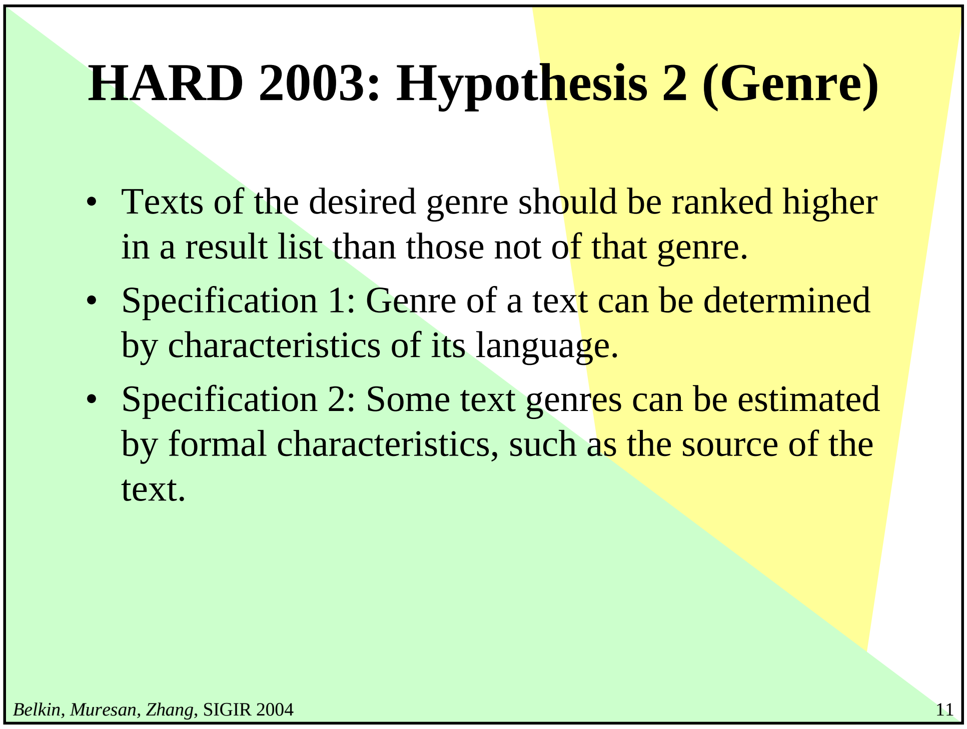
The Rutgers Approach in HARD Track

- Deal with aspects of context which could, in principle, be known either in advance of, or during the course of the current information- seeking episode
- These are, again in principle, derivable through *implicit* sources of evidence
- Test hypotheses about how specific values of context should lead to query modification or result re-ranking to improve search effectiveness



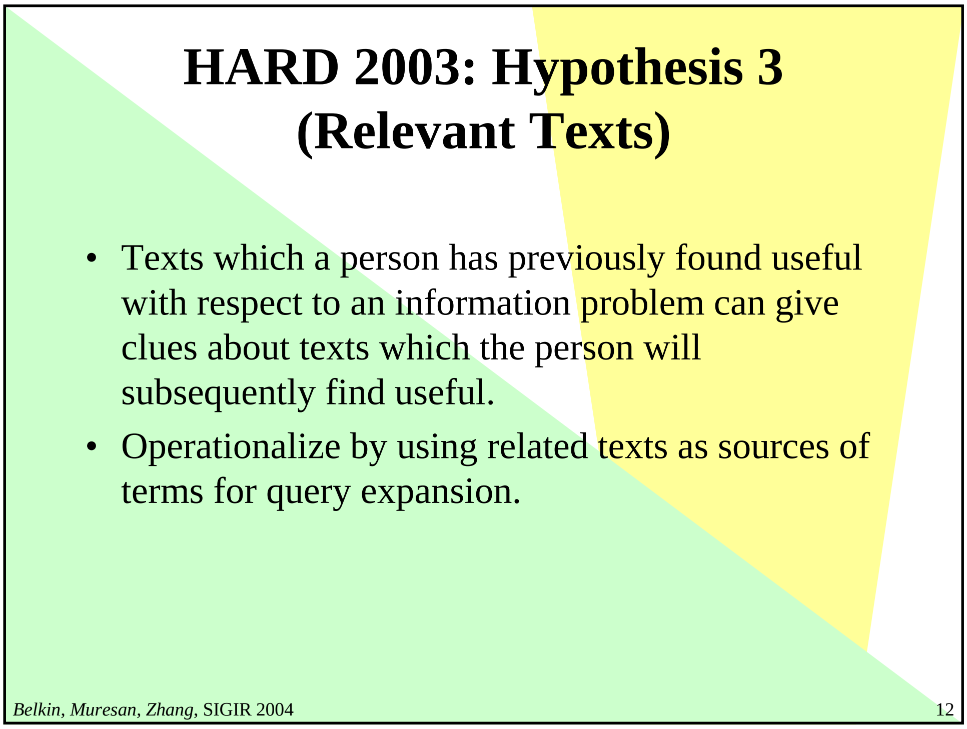
HARD 2003: Hypothesis 1 (Familiarity)

- People highly familiar with a topic will prefer specialized or technical texts; those unfamiliar will prefer general texts.
- Operationalize technicality and generality by *readability*; more readable is more general; less readable is more technical.
- Could not test this hypothesis because there was not enough variety in the data.
- Did implement a corollary: No one will be interested in unreadable or unbearably simple texts.



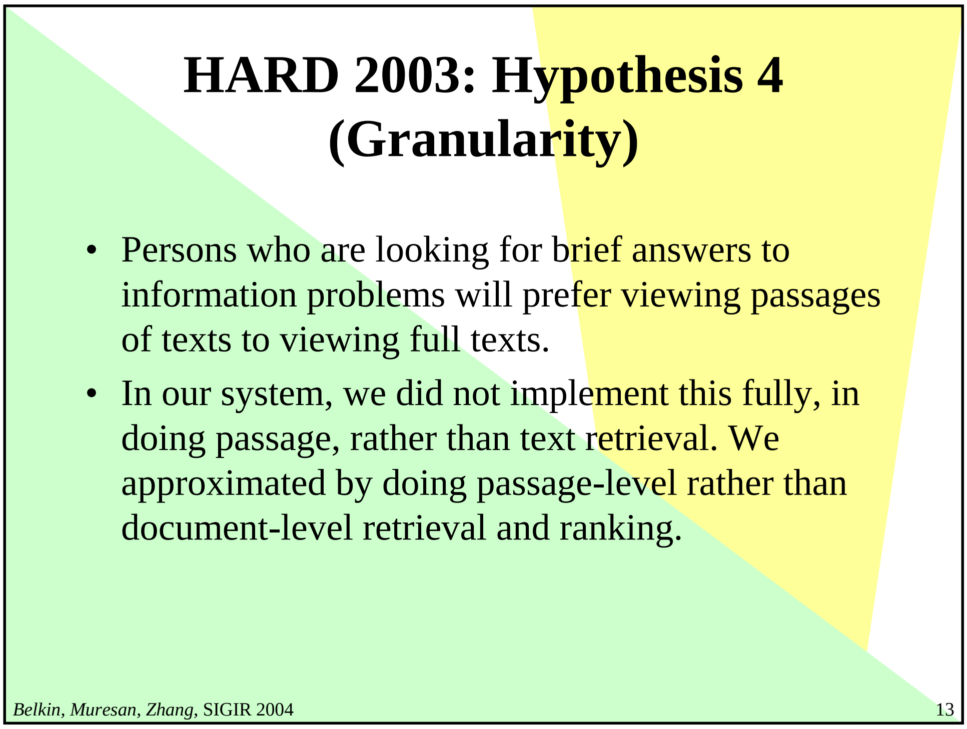
HARD 2003: Hypothesis 2 (Genre)

- Texts of the desired genre should be ranked higher in a result list than those not of that genre.
- Specification 1: Genre of a text can be determined by characteristics of its language.
- Specification 2: Some text genres can be estimated by formal characteristics, such as the source of the text.



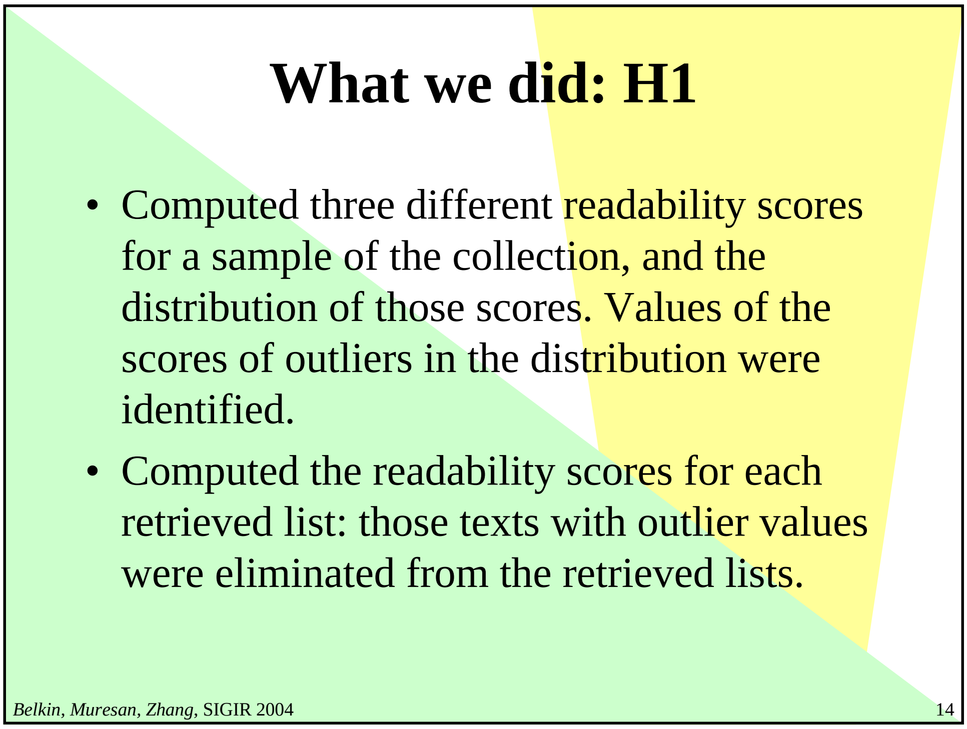
HARD 2003: Hypothesis 3 (Relevant Texts)

- Texts which a person has previously found useful with respect to an information problem can give clues about texts which the person will subsequently find useful.
- Operationalize by using related texts as sources of terms for query expansion.



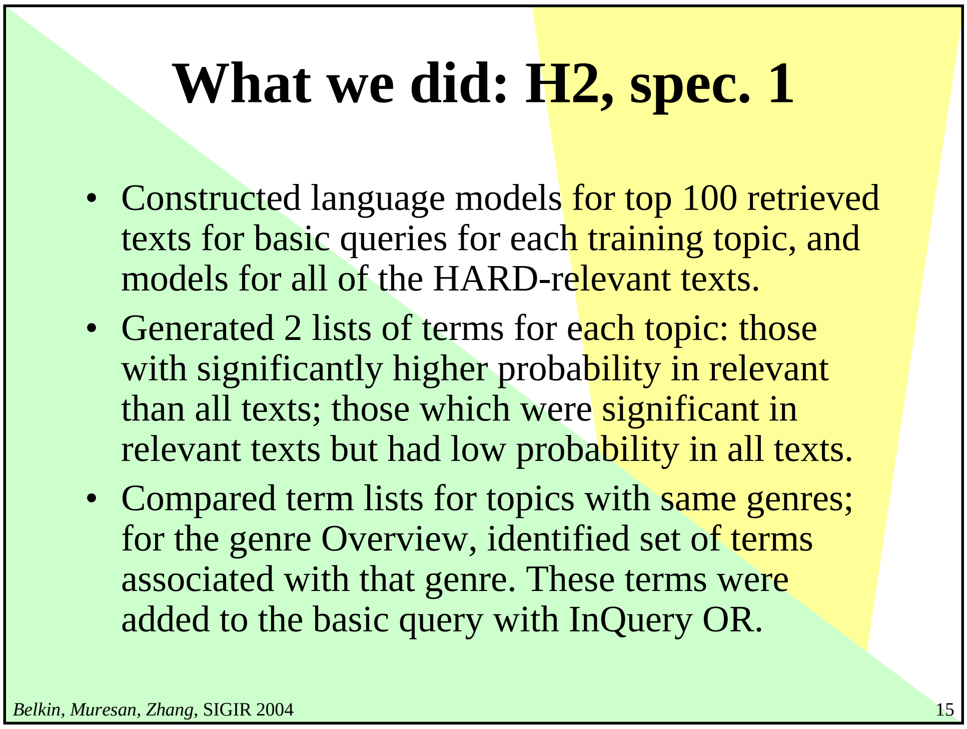
HARD 2003: Hypothesis 4 (Granularity)

- Persons who are looking for brief answers to information problems will prefer viewing passages of texts to viewing full texts.
- In our system, we did not implement this fully, in doing passage, rather than text retrieval. We approximated by doing passage-level rather than document-level retrieval and ranking.



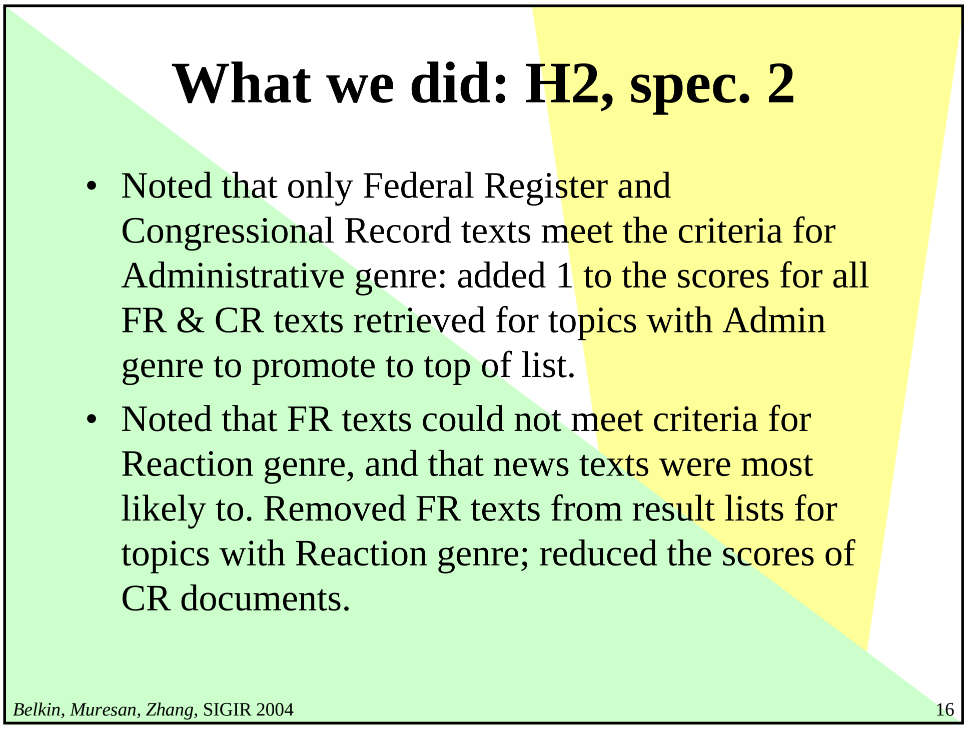
What we did: H1

- Computed three different readability scores for a sample of the collection, and the distribution of those scores. Values of the scores of outliers in the distribution were identified.
- Computed the readability scores for each retrieved list: those texts with outlier values were eliminated from the retrieved lists.



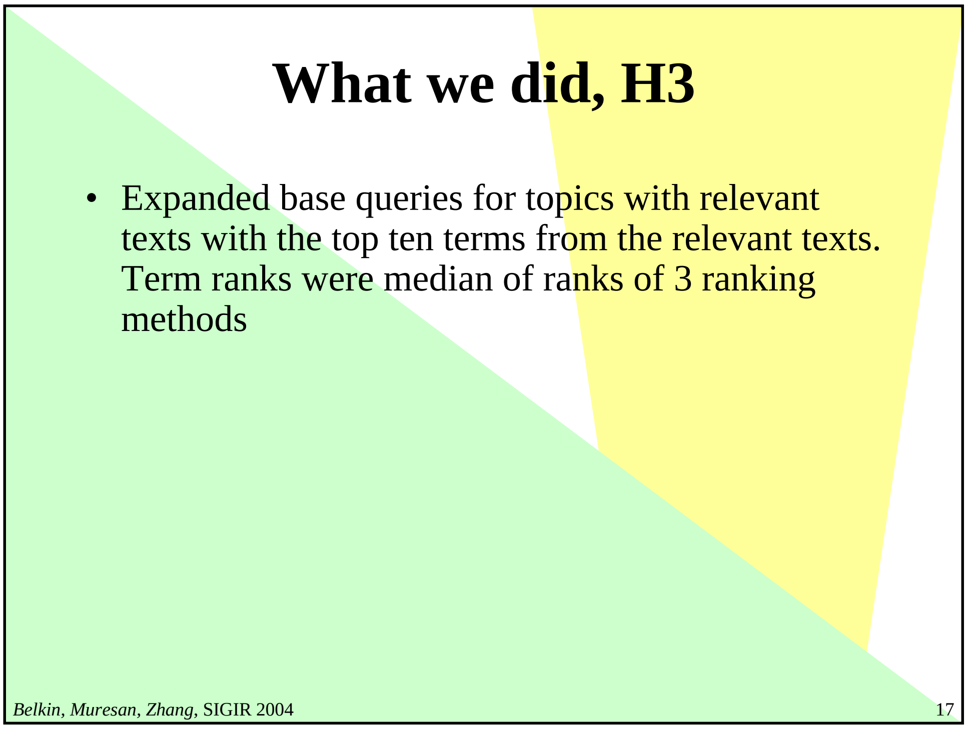
What we did: H2, spec. 1

- Constructed language models for top 100 retrieved texts for basic queries for each training topic, and models for all of the HARD-relevant texts.
- Generated 2 lists of terms for each topic: those with significantly higher probability in relevant than all texts; those which were significant in relevant texts but had low probability in all texts.
- Compared term lists for topics with same genres; for the genre Overview, identified set of terms associated with that genre. These terms were added to the basic query with InQuery OR.



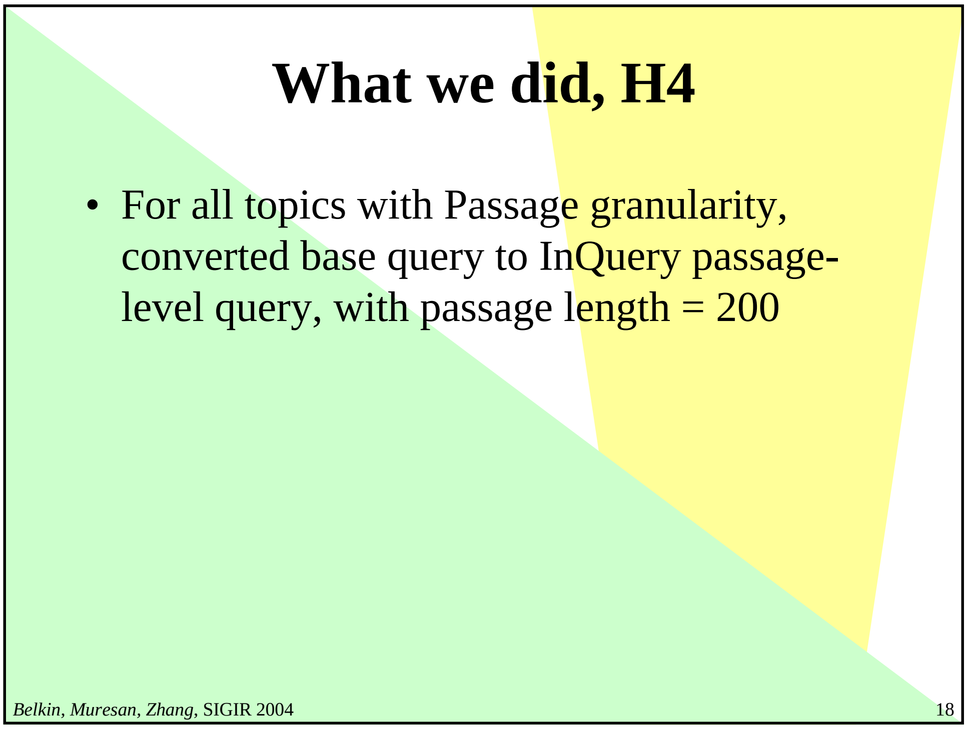
What we did: H2, spec. 2

- Noted that only Federal Register and Congressional Record texts meet the criteria for Administrative genre: added 1 to the scores for all FR & CR texts retrieved for topics with Admin genre to promote to top of list.
- Noted that FR texts could not meet criteria for Reaction genre, and that news texts were most likely to. Removed FR texts from result lists for topics with Reaction genre; reduced the scores of CR documents.



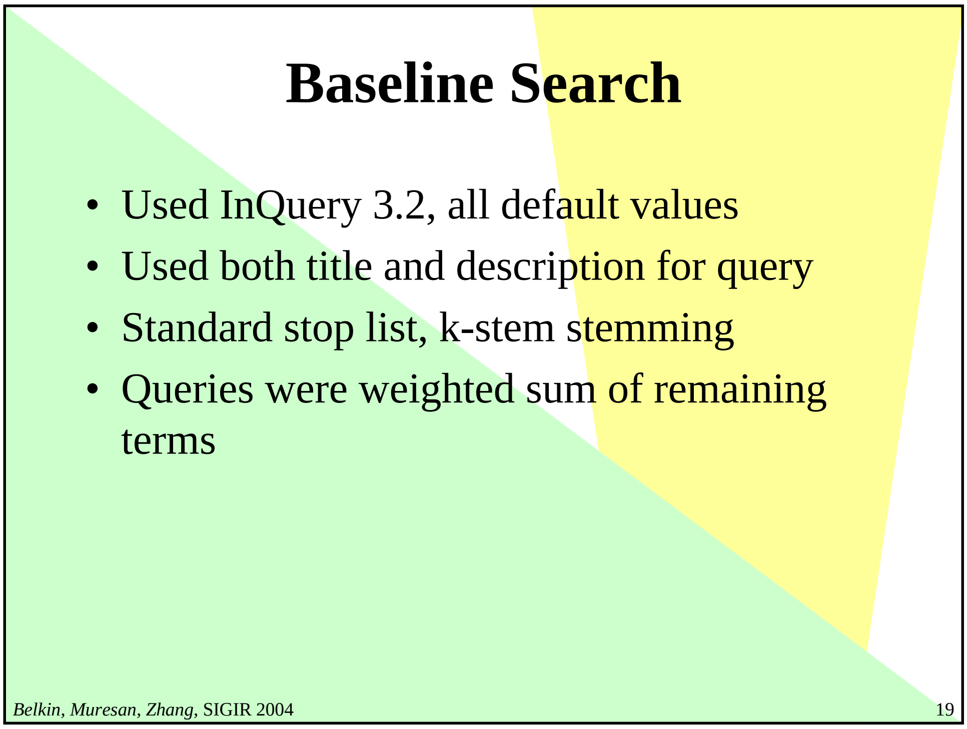
What we did, H3

- Expanded base queries for topics with relevant texts with the top ten terms from the relevant texts. Term ranks were median of ranks of 3 ranking methods



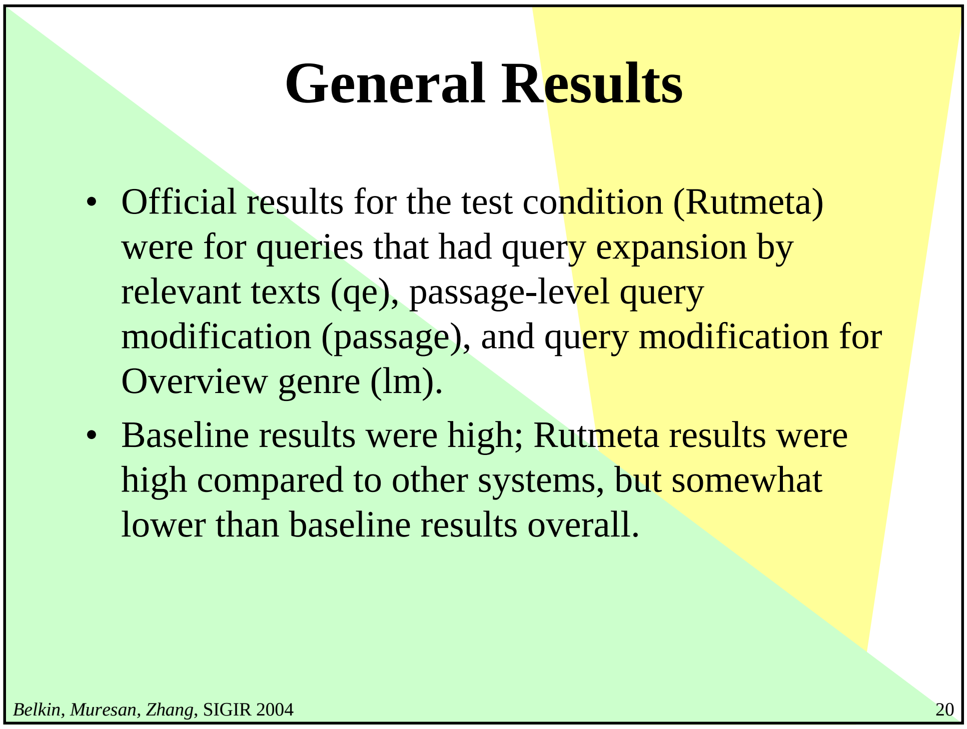
What we did, H4

- For all topics with Passage granularity, converted base query to InQuery passage-level query, with passage length = 200



Baseline Search

- Used InQuery 3.2, all default values
- Used both title and description for query
- Standard stop list, k-stem stemming
- Queries were weighted sum of remaining terms



General Results

- Official results for the test condition (Rutmeta) were for queries that had query expansion by relevant texts (qe), passage-level query modification (passage), and query modification for Overview genre (lm).
- Baseline results were high; Rutmeta results were high compared to other systems, but somewhat lower than baseline results overall.

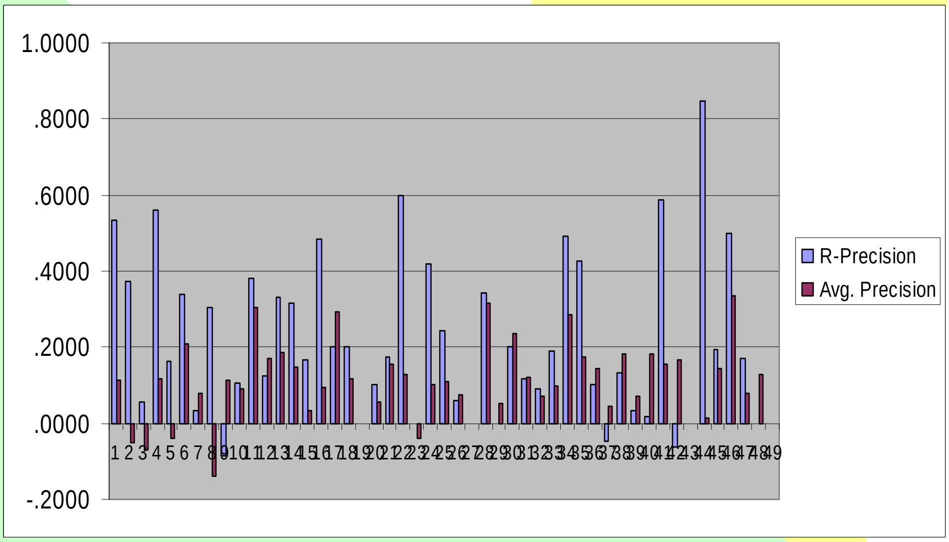


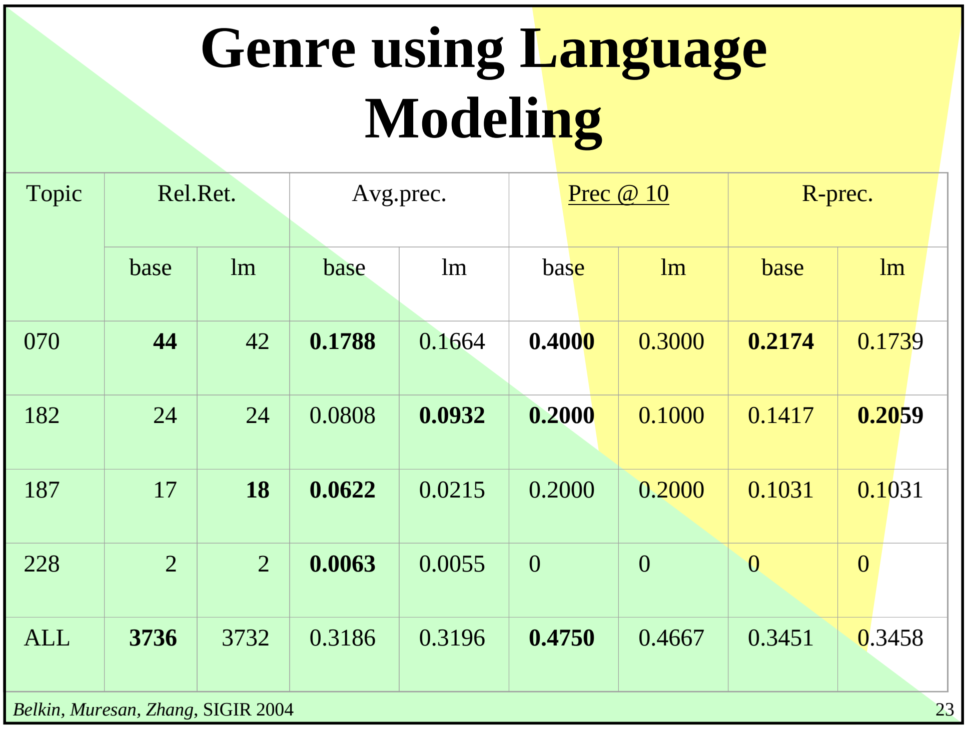
Figure 1. Difference between median values and Rutgers experimental run results for R-precision and Average Precision, by topic

Run	**Precision @ 10**	**R-precision**	**Avg. Precision**	**Rel. Ret.**
base	0.4750	0.3451	0.3186	3736
Rutmeta	0.4750	0.3308	0.3019	3728

Mean values of performance measures for baseline and test Rutgers runs.

	Rel. Ret. @ 10		**R-Precision**		**Avg. Precision**		**Rel. Ret**	
	Rutmeta	base	Rutmeta	base	Rutmeta	base	Rutmeta	base
Better	11	15	16	19	26	17	12	17

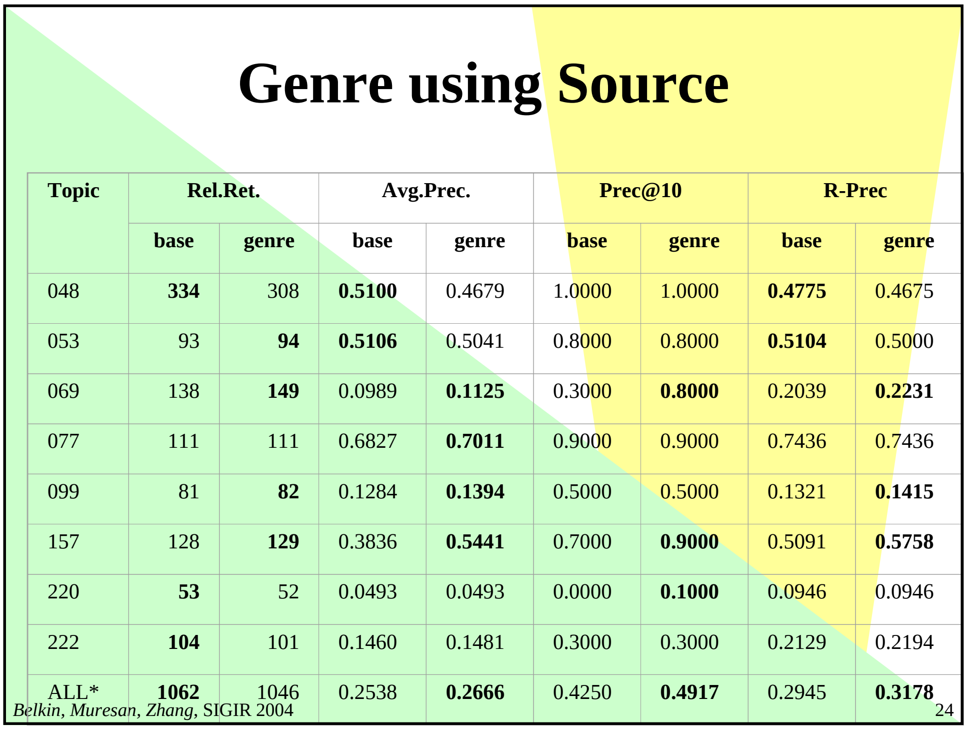
Topic-by-topic comparison of performance between baseline and experimental runs.



Genre using Language Modeling

Topic	Rel.Ret.		Avg.prec.		Prec @ 10		R-prec.	
	base	lm	base	lm	base	lm	base	lm
070	**44**	42	**0.1788**	0.1664	**0.4000**	0.3000	**0.2174**	0.1739
182	24	24	0.0808	**0.0932**	**0.2000**	0.1000	0.1417	**0.2059**
187	17	**18**	**0.0622**	0.0215	0.2000	0.2000	0.1031	0.1031
228	2	2	**0.0063**	0.0055	0	0	0	0
ALL	**3736**	3732	0.3186	0.3196	**0.4750**	0.4667	0.3451	0.3458

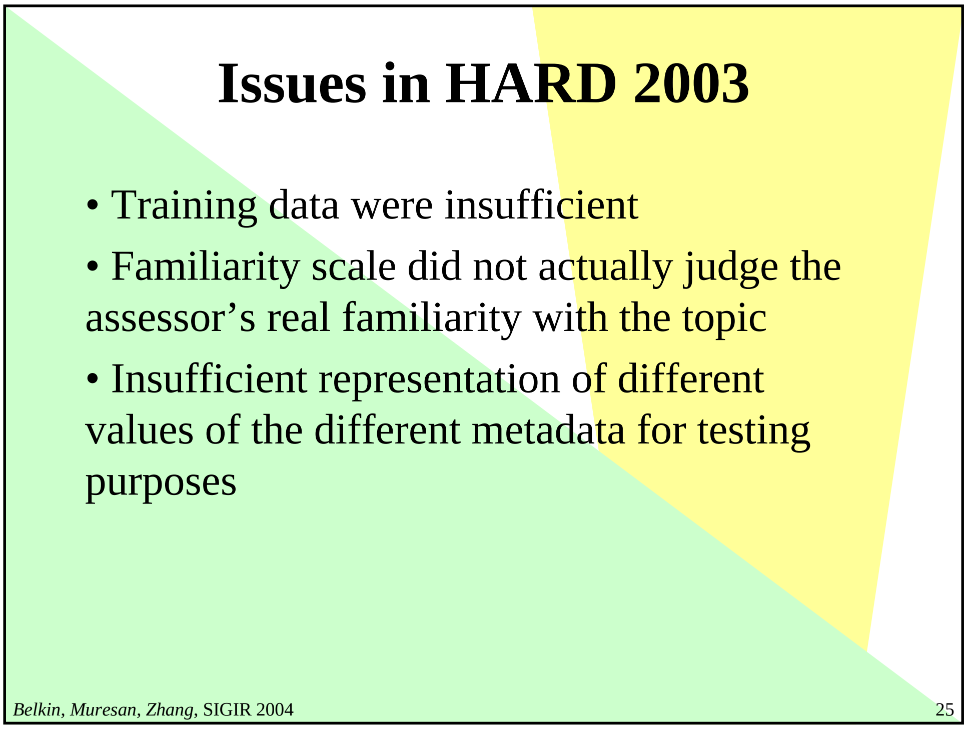
Belkin, Muresan, Zhang, SIGIR 2004 23



Genre using Source

Topic	**Rel.Ret.**		**Avg.Prec.**		**Prec@10**		**R-Prec**	
	base	**genre**	**base**	**genre**	**base**	**genre**	**base**	**genre**
048	**334**	308	**0.5100**	0.4679	1.0000	1.0000	**0.4775**	0.4675
053	93	**94**	**0.5106**	0.5041	0.8000	0.8000	**0.5104**	0.5000
069	138	**149**	0.0989	**0.1125**	0.3000	**0.8000**	0.2039	**0.2231**
077	111	111	0.6827	**0.7011**	0.9000	0.9000	0.7436	0.7436
099	81	**82**	0.1284	**0.1394**	0.5000	0.5000	0.1321	**0.1415**
157	128	**129**	0.3836	**0.5441**	0.7000	**0.9000**	0.5091	**0.5758**
220	**53**	52	0.0493	0.0493	0.0000	**0.1000**	0.0946	0.0946
222	**104**	101	0.1460	0.1481	0.3000	0.3000	0.2129	0.2194
ALL*	**1062**	1046	0.2538	**0.2666**	0.4250	**0.4917**	0.2945	**0.3178**

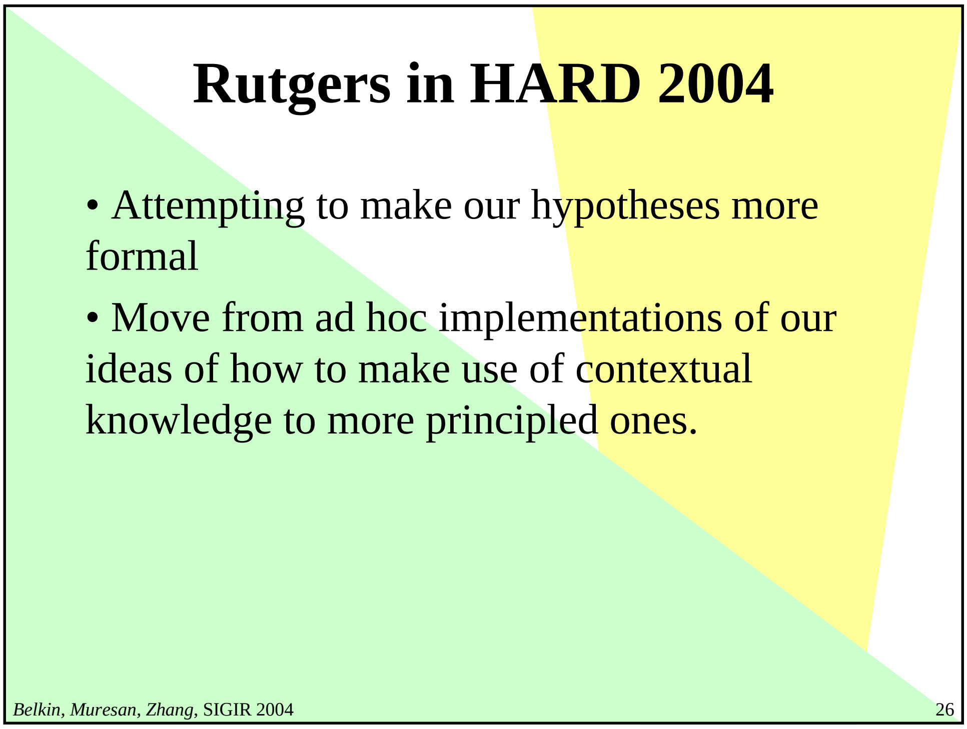
Belkin, Muresan, Zhang, SIGIR 2004 24



Issues in HARD 2003

- Training data were insufficient
- Familiarity scale did not actually judge the assessor’s real familiarity with the topic
- Insufficient representation of different values of the different metadata for testing purposes

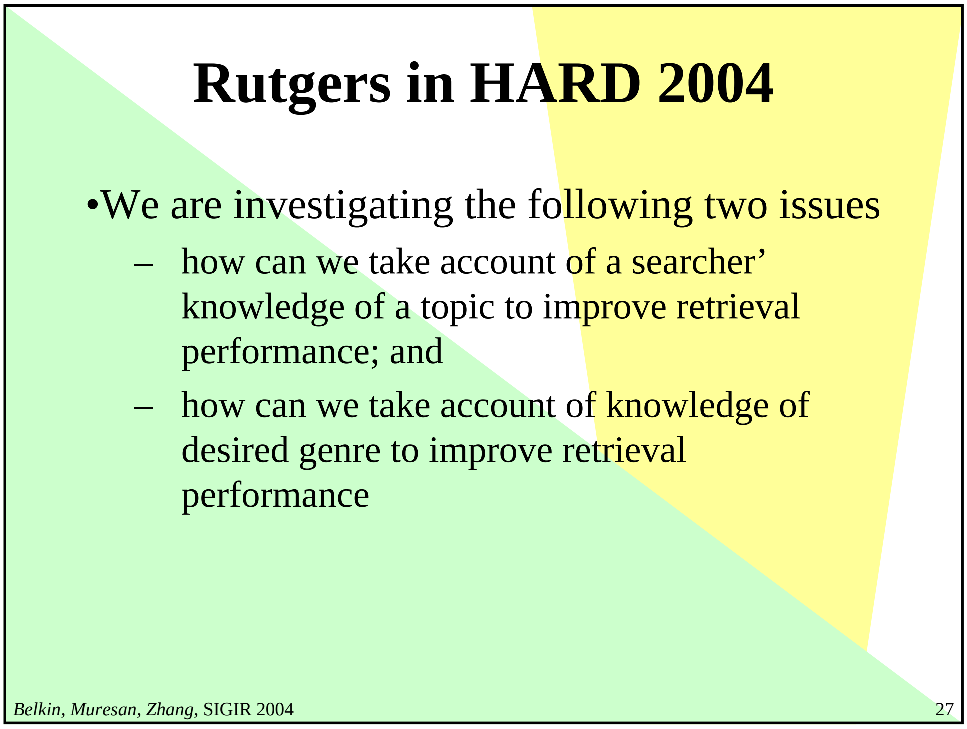
Belkin, Muresan, Zhang, SIGIR 2004 25



Rutgers in HARD 2004

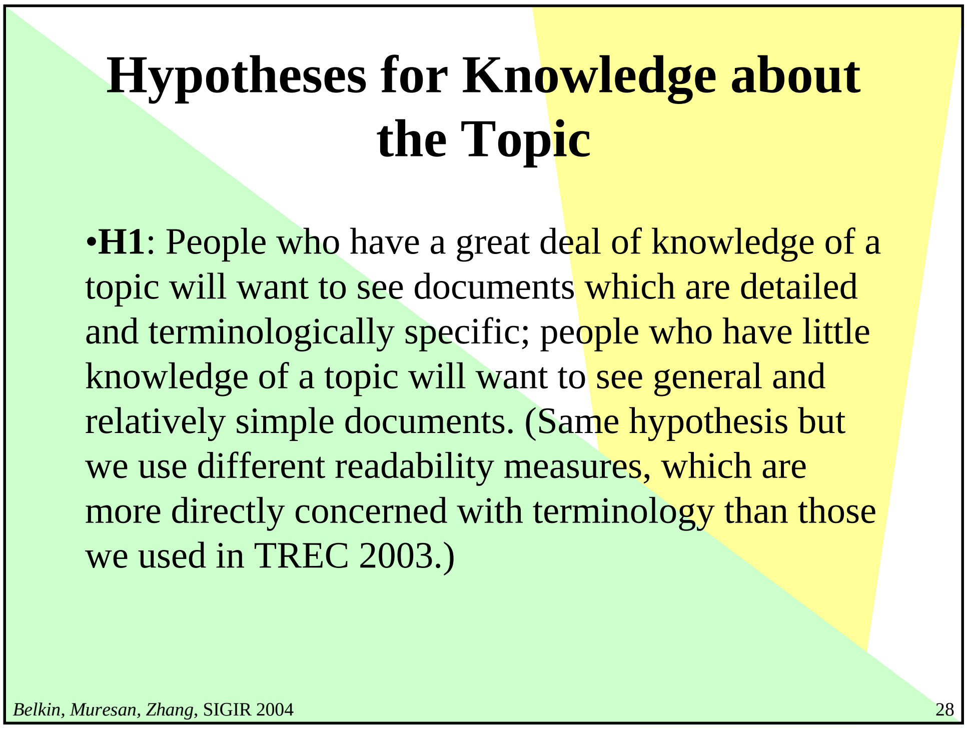
- Attempting to make our hypotheses more formal
- Move from ad hoc implementations of our ideas of how to make use of contextual knowledge to more principled ones.

Belkin, Muresan, Zhang, SIGIR 2004 26



Rutgers in HARD 2004

- We are investigating the following two issues
  - how can we take account of a searcher’ knowledge of a topic to improve retrieval performance; and
  - how can we take account of knowledge of desired genre to improve retrieval performance



Hypotheses for Knowledge about the Topic

- **H1**: People who have a great deal of knowledge of a topic will want to see documents which are detailed and terminologically specific; people who have little knowledge of a topic will want to see general and relatively simple documents. (Same hypothesis but we use different readability measures, which are more directly concerned with terminology than those we used in TREC 2003.)

Hypotheses for Knowledge about the Topic

•**H2**: People who have little knowledge of the topic will prefer documents with a low ratio of abstract words to total words, and a high ratio of concrete words to total words. People who have good knowledge of a topic will prefer documents which have a high ratio of abstract words to total words, and abstract words to concrete words. This hypothesis leads to a re-ranking strategy.

Hypotheses for Knowledge about the Topic

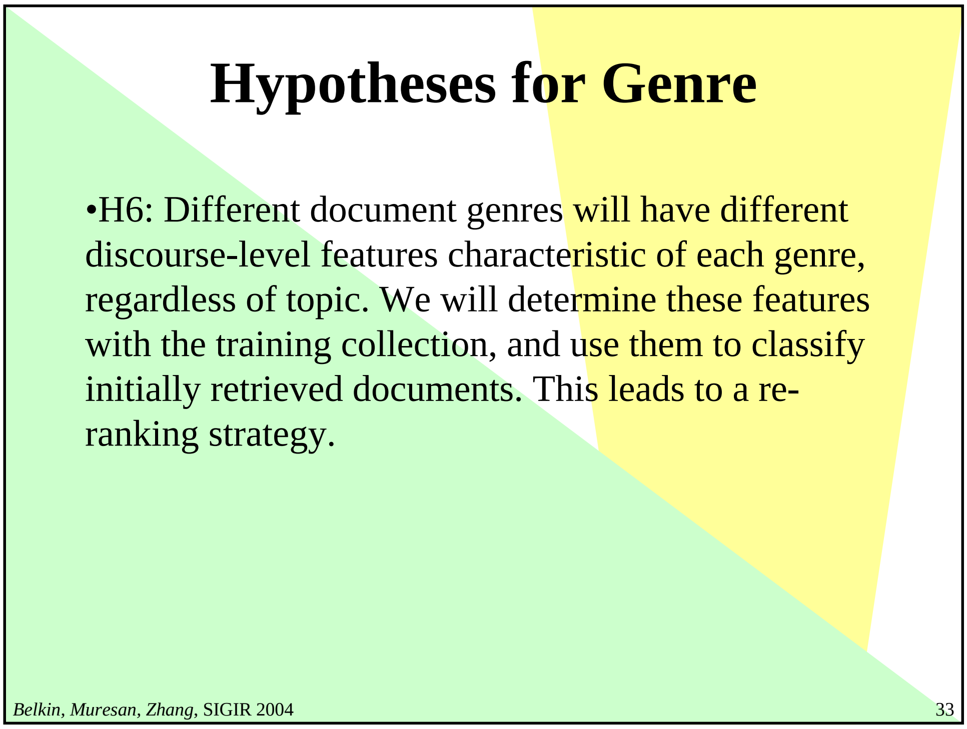
•**H3**: Adding concrete terms to the initial query will lead to more effective results for people with little knowledge of the topic; adding abstract terms from the topic domain will lead to more effective results for people with a great deal of knowledge of the topic. This is a query modification strategy.

Hypotheses for Genre

- H4: The differences between the genres of news-report and opinion can be identified according to the degree of subjectivity or objectivity of a document, as determined by various linguistic features of the documents (cf. Rittman, 2004). This leads to a classification and re-ranking strategy.

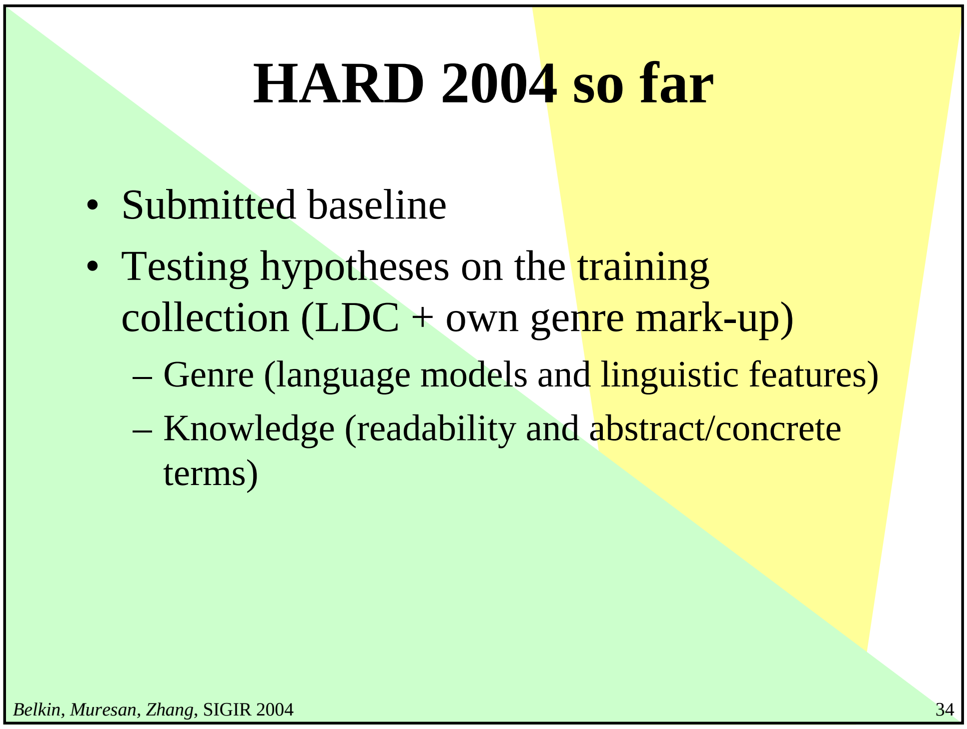
Hypotheses for Genre

- H5: Different document genres will have different characteristic vocabularies, regardless of topic. This is essentially the same hypothesis that we had last year, and we investigate it by again developing language models for the topic in general (i.e. soft relevant), and those for the different genres within each topic. Words which occur with greater than expected frequency with respect to the topic models for a particular genre, across all topics, will be indicative of the genre’s vocabulary. This technique can be used both to identify words which can be added to a query (query modification strategy), and to classify documents which belong to a specific genre (re-ranking strategy).



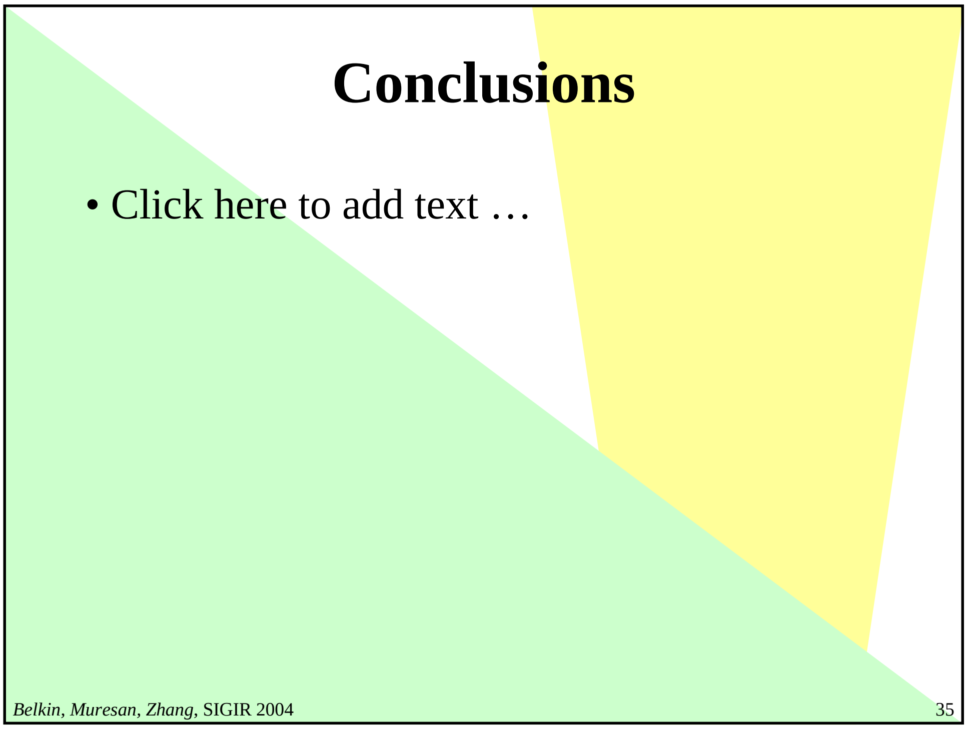
Hypotheses for Genre

•H6: Different document genres will have different discourse-level features characteristic of each genre, regardless of topic. We will determine these features with the training collection, and use them to classify initially retrieved documents. This leads to a re-ranking strategy.



HARD 2004 so far

- Submitted baseline
- Testing hypotheses on the training collection (LDC + own genre mark-up)
  - Genre (language models and linguistic features)
  - Knowledge (readability and abstract/concrete terms)



Conclusions

- Click here to add text …