IBM in TREC 2006 Enterprise Track

Jennifer Chu-Carroll, Guillermo Averboch, Pablo Duboue, David Gondek, J William Murdock, John Prager
IBM T.J. Watson Research Center

Paul Hoffmann, Janyce Wiebe
University of Pittsburgh

November 17, 2006

© 2006 IBM Corporation – All Rights Reserved

Overview

- **Scientific Foci**
- **Discussion Task**
  - System
  - Hypotheses
  - Results
- **Expert Task**
  - System
  - Hypotheses
  - Results
- **Conclusions**

© 2006 IBM Corporation – All Rights Reserved –

Scientific Foci

- **Investigate impact of adopting multiple problem-solving strategies**
  - High-precision vs. high-recall strategies
  - Knowledge-based vs. statistical approaches
  - Search engines employing different ranking algorithms
- **Investigate combination of structured, semi-structured, and unstructured information sources**
  - High-precision extracted structured information
  - Analysis of semi-structured texts, e.g., standards documents, e-mail signature
- **Leverage NLP technologies to enhance search performance**
  - Pro/con sentiment analysis
  - Query-based multi-document summarization
  - *ExpertIn* relation detection
- **Leverage relevant external resources**
  - FOLDOC computing dictionary
  - Google Scholar

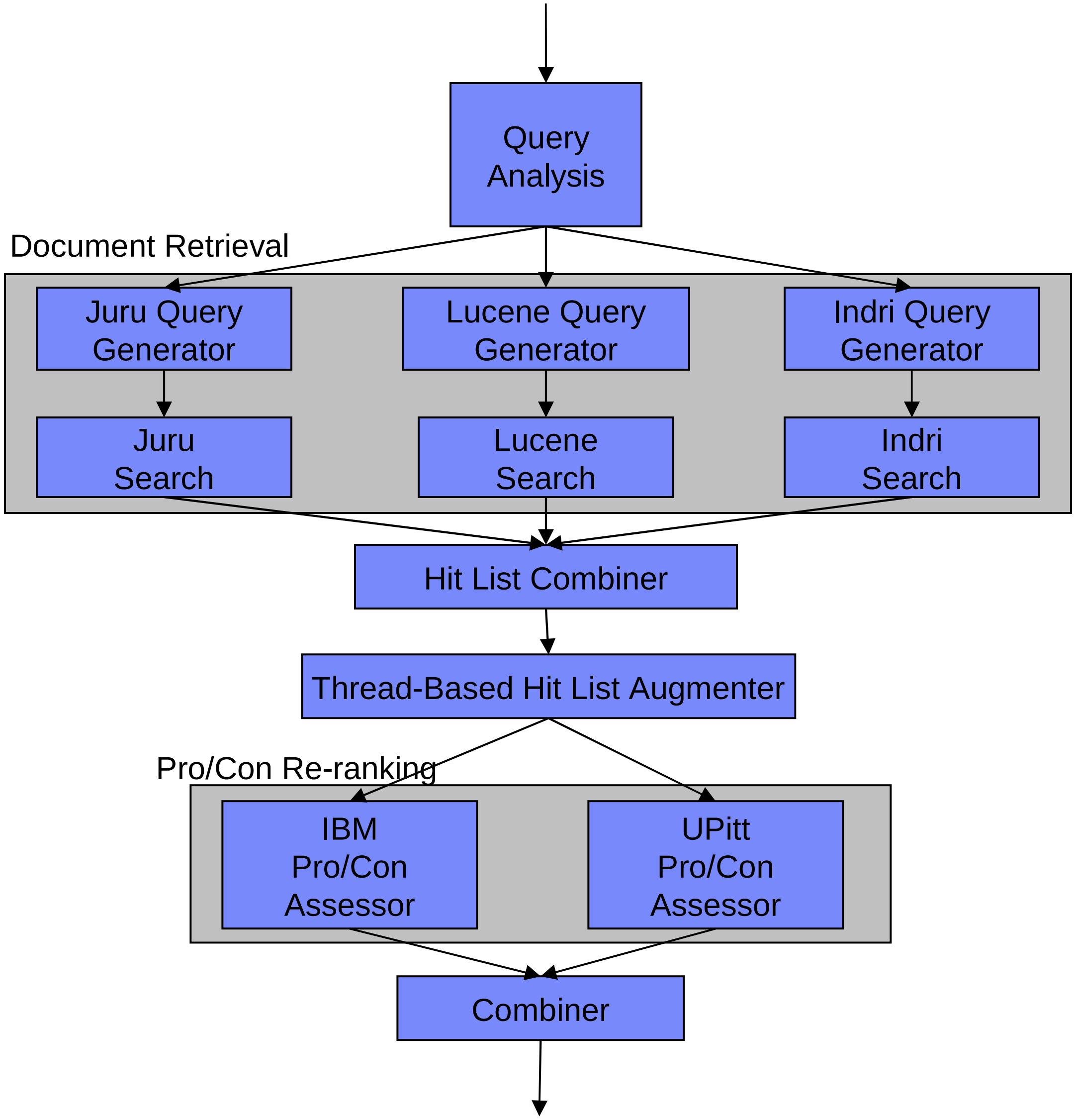
© 2006 IBM Corporation – All Rights Reserved –

Discussion Search Task

- **Task: given a topic, return ranked list of e-mail messages that discuss pro/con aspects of the topic**
- **Basic approach**
  - Search for topic-relevant documents
  - Analyze documents for presence of pro/con sentiments
- **Experimental foci**
  - Investigate impact of adopting multiple problem-solving strategies
    - Adopted multiple search engines for document retrieval
    - Developed and leveraged multiple pro/con sentiment analysis engines
  - Leverage NLP technologies to enhance search performance
    - Developed a rule-based sentiment analyzer based on syntactic parses
    - Developed a statistical sentiment analyzer based on POS-driven bag of words and extraction patterns
  - Leverage relevant external resources
    - Processed FOLDOC to extract acronym/expansion pairs and phrases highly associated with each term for query expansion

© 2006 IBM Corporation – All Rights Reserved –

Discussion Search System Architecture



- Utilizes “query” and “description” from topic
- Performs query expansion
- Produces one or more abstract query representations

- Leverages multiple search engines with different query languages and ranking algorithms

- Augment hitlist with documents in the same e-mail thread as retrieved e-mails using Webber’s threading information

- Leverages multiple sentiment analyzers
- IBM Pro/Con assessor: rule-based sentence-level analyzer based on syntactic parses
- UPitt Pro/Con assessor: statistical document-level analyzer based on words and extraction patterns

© 2006 IBM Corporation – All Rights Reserved –

Discussion Search Results

	MAP		bpref		p@10	
	topic	pro/con	topic	pro/con	topic	pro/con
JQ	0.2745	0.1654	0.3218	0.2082	0.4950	0.2800
IBM06JAQ	0.3146	0.2030	0.3572	0.2337	0.5440	0.3391
JILQ	0.3017	0.1762	0.3472	0.2083	0.5360	0.2978
JILQD	0.3095	0.1835	0.3559	0.2174	0.5360	0.3065
IBM06JILAPQD	0.3310	0.2021	0.3709	02323	0.5640	0.3391

- Document search only
- Three document search engines
- Query and description

- **Summary of results**
  - Multiple problem-solving strategies
    - Employing multiple document retrieval engines improved MAP by 9.9%
    - Multiple pro/con analyzers yielded marginal improvement
  - Leverage NLP technologies
    - Single pro/con analyzer improved pro/con MAP score by 22.7%
    - IBM06JAQ: one of three runs with greater rank increase from topic MAP to pro/con MAP
  - External resources
    - Query expansion using description field (with FOLDOC) yielded marginal improvement

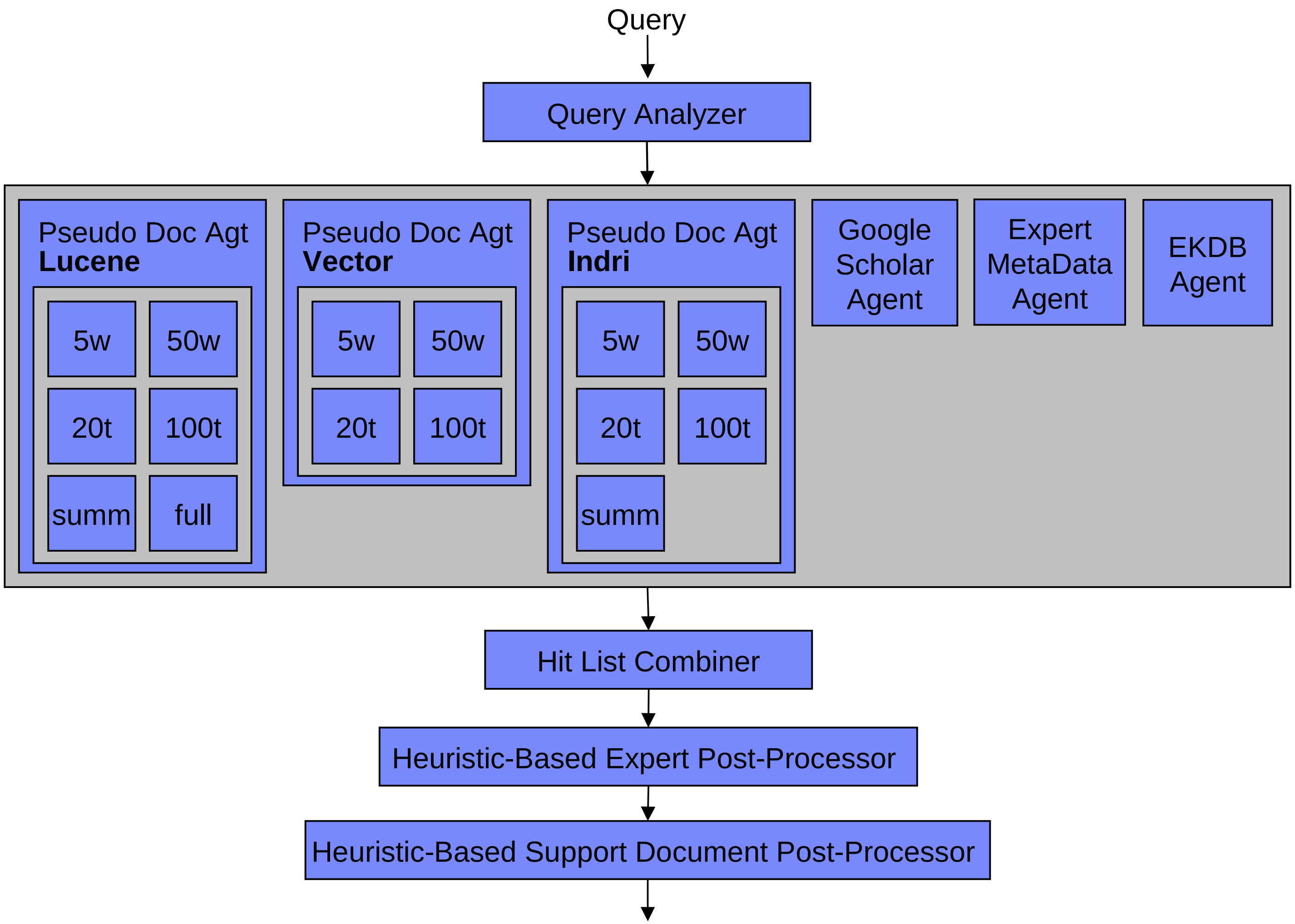
© 2006 IBM Corporation – All Rights Reserved –

Expert Search Task

- **Task: given a topic, return a ranked list of experts on that topic**
- **Basic approach**
  - Adopt multiple expert finding strategies and combine results
  - Re-rank/Filter experts/support documents
- **Experimental foci**
  - Investigate impact of adopting multiple problem-solving strategies
    - Adopted multiple agents for expert finding
  - Investigate combination of structured, semi-structured, and unstructured information sources
    - Utilized unstructured information for pseudo-document generation
    - Analyzed semi-structured standards documents for expert identification
    - Extracted high-precision structured information using relation recognizers
  - Leverage NLP technologies to enhance search performance
    - Utilized MEAD [Radev et al., 2003], a query-based multi-document summarization system for pseudo-document generation
    - Developed *ExpertIn* relation recognizer for identifying expert-topic associations
  - Leverage relevant external resources
    - Queried Google Scholar for authors of scholarly publications on topic

© 2006 IBM Corporation – All Rights Reserved –

Expert Search System Architecture



- Employs multiple expert finding strategies
- Some targets high precision and others high recall

- Affinity-based expert reranker

- Acknowledgements document filter
- Duplicate document filter
- EKDB document reranker

© 2006 IBM Corporation – All Rights Reserved –

Expert Search Agent Details

- **Pseudo-document agents: generate one pseudo-document per expert to capture their expertise [Fu et al, 2006]**
  - Windowing approach: n sentences before/after each mention of a candidate expert
  - Top sentence approach: first n sentences in documents where candidate appears
  - Whole document approach: all documents in which a candidate appears
  - Summarization approach: summarization generated for each candidate by MEAD
- **Expert MetaData agent**
  - Identifies standards documents and associates authors/editors with topic
- **EKDB agent**
  - Determines expertise from extracted structured data based on *ExpertIn* relation and e-mail author/subject pairs
- **Google Scholar agent**
  - Extracts authors of papers on given topic, and filter for experts on candidate list

© 2006 IBM Corporation – All Rights Reserved –

Expert Search Results

	# ques answered	MAP expert	MAP support	bpref expert	bpref support	p@5 expert	p@5 support
pseudo lucene	**49**	**0.3970**	**0.2490**	**0.4039**	**0.5431**	**0.4980**	**0.3796**
pseudo vector	**49**	**0.4122**	**0.2558**	**0.4144**	**0.5545**	**0.5**	**0.3918**
pseudo indri	**49**	**0.3997**	**0.2267**	**0.4118**	**0.4695**	**0.5469**	**0.3796**
metadata	**19**	**0.2026**	**0.1107**	**0.2013**	**0.1170**	**0.7263**	**0.4211**
ekdb	**28**	**0.0735**	**0.0105**	**0.0793**	**0.0150**	**0.3357**	**0.0714**
google	**27**	**0.0500**	**--**	**0.0622**	**--**	**0.2444**	**--**
IBM06QO	**49**	**0.4536**	**0.2863**	**0.4402**	**0.3711**	**0.6653**	**0.4857**

- **Summary of results**
  - Effective combination of multiple strategies leveraging structured, semi-structured, and unstructured information yielded 11.9% improvement in support MAP
  - NLP technologies
    - Current use of summarization system did not yield improvement over other approaches
    - *ExpertIn* relation detection was key contributor in EKDB agent performance
  - External resource Google Scholar resulted in minimal improvement

© 2006 IBM Corporation – All Rights Reserved –

Conclusions

- **Our adoption of multiple strategies for problem-solving was highly effective**
  - 9.9% MAP improvement in discussion task with three search engines vs. one
  - 11.9% MAP improvement in expert task with six agents vs. best performing agent
  - Multiple pseudo-document generation strategies also improved upon a single-strategy approach
- **Select NLP technologies had high impact**
  - Pro/Con sentiment analyzers increased pro/con MAP score by 22.7%
  - *ExpertIn* relation detector enabled of extraction of high quality data for EKDB agent
  - Summarization as currently used did not result in performance improvement
- **External resources utilized in our experiments yielded minimal improvement**

© 2006 IBM Corporation – All Rights Reserved –