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
Overview

- Scientific Foci
- Discussion Task
  - System
  - Hypotheses
  - Results
- Expert Task
  - System
  - Hypotheses
  - Results
- Conclusions
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
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
Discussion Search System Architecture

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

- 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

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

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

Diagram:
- Query Analysis
  - Juru Query Generator
  - Lucene Query Generator
  - Indri Query Generator

- Hit List Combiner
- Thread-Based Hit List Augmenter

- Pro/Con Re-ranking
  - IBM Pro/Con Assessor
  - UPitt Pro/Con Assessor

- Combiner
### Discussion Search Results

<table>
<thead>
<tr>
<th></th>
<th>MAP</th>
<th>bpref</th>
<th>p@10</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>topic</td>
<td>pro/con</td>
<td>topic</td>
</tr>
<tr>
<td>JQ</td>
<td>0.2745</td>
<td>0.1654</td>
<td>0.3218</td>
</tr>
<tr>
<td>IBM06JAQ</td>
<td>0.3146</td>
<td>0.2030</td>
<td>0.3572</td>
</tr>
<tr>
<td>JILQ</td>
<td>0.3017</td>
<td>0.1762</td>
<td>0.3472</td>
</tr>
<tr>
<td>JILQD</td>
<td>0.3095</td>
<td>0.1835</td>
<td>0.3559</td>
</tr>
<tr>
<td>IBM06JILAPQD</td>
<td>0.3310</td>
<td>0.2021</td>
<td>0.3709</td>
</tr>
</tbody>
</table>

#### 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

- **Document search only**
- **Three document search engines**
- **Query and description**
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
Expert Search System Architecture

- **Query Analyzer**

  - **Pseudo Doc Agt**
    - Lucene
    - Vector
    - Indri

  - Helix Document Filter
    - 5w
    - 50w
    - 20t
    - 100t
    - summ
    - full

- **Hit List Combiner**

  - Heuristic-Based Expert Post-Processor
  - Heuristic-Based Support Document Post-Processor

- **Heuristic-Based Expert Post-Processor**

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

- **Heuristic-Based Support Document Post-Processor**

  - 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
Expert Search Results

<table>
<thead>
<tr>
<th></th>
<th># ques</th>
<th>MAP</th>
<th>bpref</th>
<th>p@5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>answered</td>
<td>expert</td>
<td>support</td>
<td>expert</td>
</tr>
<tr>
<td>pseudo lucene</td>
<td>49</td>
<td>0.3970</td>
<td>0.2490</td>
<td>0.4039</td>
</tr>
<tr>
<td>pseudo vector</td>
<td>49</td>
<td>0.4122</td>
<td>0.2558</td>
<td>0.4144</td>
</tr>
<tr>
<td>pseudo indri</td>
<td>49</td>
<td>0.3997</td>
<td>0.2267</td>
<td>0.4118</td>
</tr>
<tr>
<td>metadata</td>
<td>19</td>
<td>0.2026</td>
<td>0.1107</td>
<td>0.2013</td>
</tr>
<tr>
<td>ekdb</td>
<td>28</td>
<td>0.0735</td>
<td>0.0105</td>
<td>0.0793</td>
</tr>
<tr>
<td>google</td>
<td>27</td>
<td>0.0500</td>
<td>--</td>
<td>0.0622</td>
</tr>
<tr>
<td>IBM06QO</td>
<td>49</td>
<td>0.4536</td>
<td>0.2863</td>
<td>0.4402</td>
</tr>
</tbody>
</table>

### 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
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**