Using Citations to Generate Surveys of Scientific Paradigms

Saif Mohammad[†*], Bonnie Dorr[†‡*], Melissa Egan[†‡],
Ahmed Hassan[ϕ], Pradeep Muthukrishan[ϕ],
Vahed Qazvinian[ϕ], Dragomir Radev[ϕ§], David Zajic[‡◊]

human language technology
center of excellence

Institute for Advanced Computer Studies and CLIP lab[†],
Center for Advanced Study of Language[◊],
Department of Computer Science[‡], University of Maryland.
Human Language Technology Center of Excellence[*].

Department of Electrical Engineering and Computer Science[ϕ],
School of Information[§], University of Michigan.

Rapidly growing technical information

- Number of research publications in various disciplines is growing exponentially.
- Authors of journal articles and books have to write survey articles.
- More and more of the research is cross-disciplinary.
- Inter-disciplinary review panels receive proposals on a wide range of areas.
- Need to rapidly learn about unfamiliar areas.

Our goal

- We generate surveys of multiple research publications in a given topic area by combining:
  - Bibliometric lexical link mining.
    - Exploits the structure of citations.
  - Automatic summarization techniques.
    - Exploits the content of material in both the citing and cited papers.

Data sources

- Research publications – target papers
  - E.g., from dependency parsing: Eisner (1996), Collins (1997), McDonald et al. (2006), Nivre et al. (2006),…
  - All of the content is from the author's perspective.

- Abstracts
  - Provide good summaries of individual papers
  - Also from the author's perspective.

Abstract of of Eisner (1996)

Three New Probabilistic Models For Dependency Parsing: An Exploration.

After presenting a novel O(n^3) parsing algorithm for dependency grammar, we develop three contrasting ways to stochasticize it. We propose (a) a lexical affinity model where words struggle to modify each other, (b) a sense tagging model where words fluctuate randomly in their selectional preferences, and (c) a generative model where the speaker fleshes out each word's syntactic and conceptual structure without regard to the implications for the hearer. We also give preliminary empirical results from evaluating the three models' parsing performance on annotated Wall Street Journal training text (derived from the Penn Treebank)…

Data sources *(continued)*

- Citation texts
  - Set of sentences from other papers that explicitly refer to the target.

Citation text of Eisner (1996)

-- Eisner (1996) proposed a CKY-like $O(n^3)$ algorithm.

-- Eisner (1996) introduced a data-driven dependency parser and compared several probability models on (English) Penn Treebank data.

-- Eisner (1996) gave a generative model with a cubic parsing algorithm based on an edge factorization of trees.

-- In many dependency parsing models such as (Eisner, 1996) and (MacDonald et al. , 2005), the score of a dependency tree is the sum of the scores of the dependency links.

…

Citation texts in related work

- Use in information retrieval:
  - Bradshaw (2003) used citation texts to improve the results of a search engine.
- Identifying kinds of citations:
  - Nanba et al. (2004) automatically categorize citation sentences into three groups.
- Directly summarizing citation text:
  - Qazvinian and Radev (2008) and Mei and Zhai (2008) used citation texts to create summaries of single scientific articles.

Citation texts for survey creation

- Compare and contrast multiple approaches:

  Beginning with the seminal work at IBM (Black et al, 1991; Black et al, 1992b; Black et al, 1992a), and continuing with such lexicalist approaches as (Eisner, 1996), these features have been lauded for their ability to approximate a word's semantics as a means to override syntactic preferences with semantic ones (Collins, 1999; Eisner, 2000).

  While early head-lexicalized grammars restricted the fragments to the locality of headwords (e.g. Collins 1996; Eisner 1996), later models showed the importance of including context from higher nodes in the tree (Charniak 1997; Johnson 1998).

Citation texts for survey creation

- Possess a certain amount of redundant information.
  - Multiple papers may describe the same contributions of a target paper.

    The nugget-based paradigm has been previously detailed in a number of papers (Voorhees, 2003; Hildebrandt et al, 2004; Lin and Demner-Fushman, 2005a).

    The nugget evaluation methodology (Voorhees, 2005) developed for scoring answers to complex questions is not suitable for our task.
- Indicate which contributions described in a paper were more influential over time.

This work

- Test the hypothesis:
  - An effective technical survey will have information from:
    - the perspective of its authors;
    - the perspective of others who use/commend/ discredit/add to it.
- Compare and contrast usefulness of abstracts and of citation texts in automatically generating a technical survey on a given topic.

Summarization systems

- Trimmer
  - Syntactically motivated parse-and-trim approach. (Zajic et al., 2007)
- LexRank
  - Graph-based approach that applies PageRank to identify important sentences. (Erkan and Radev, 2004)
- C-LexRank and C-RoundRobin
  - Graph clustering approaches. (Qazvinian and Radev, 2008)

Data

- ACL Anthology: a collection of about 11,000 papers from:
  - Computational Linguistics journal;
  - Proceedings of ACL conferences and workshops.
- ACL Anthology Network (AAN): a semi-automatically generated citation network in the ACL anthology. (Joseph and Radev, 2007).
  - Has 11, 773 nodes and 38,765 directed edges.

Data *(continued)*

- Topic areas: Question Answering (QA) and Dependency Parsing (DP).
- Selected 10 QA papers and 16 DP papers from AAN that each had:
  - *"question answering"* and "*dependency parsing"* in the title.
  - at least four sentences in other papers that explicitly referred to them.

Experiments

- Generated 4 x 3 x 2 automatic surveys:
  - Using 4 summarizers—Trimmer, LexRank, C-LexRank, C-RoundRobin.
  - From 3 data sources—full papers, abstracts, citation texts.
  - On two topics—QA and DP.
- Generated 3 x 2 random surveys:
  - By choosing sentences randomly from the full papers, abstracts, and citation texts of QA and DP.
- Forced a hard limit of 250 words on all surveys.

Evaluation

- Automatically generated surveys were evaluated using two separate approaches:
  - Nugget-based pyramid evaluation (Nenkova and Passonneau, 2004; Passonneau and Nenkova, 2003)
  - ROUGE evaluation (Lin, 2004)
- Reference summaries were manually created for both QA and DP:
  - Four fluent speakers of English were asked to generate 250-word surveys from the QA and DP abstracts and citation texts.

Nugget-based pyramid evaluation

- Two human judges identified 2 to 8 nuggets of information
  - from each of the QA and DP abstracts;
  - from each of the QA and DP citation texts.
- Assigned weights to the nuggets
- proportional to frequency of inclusion by the judges.

Nugget-based pyramid evaluation

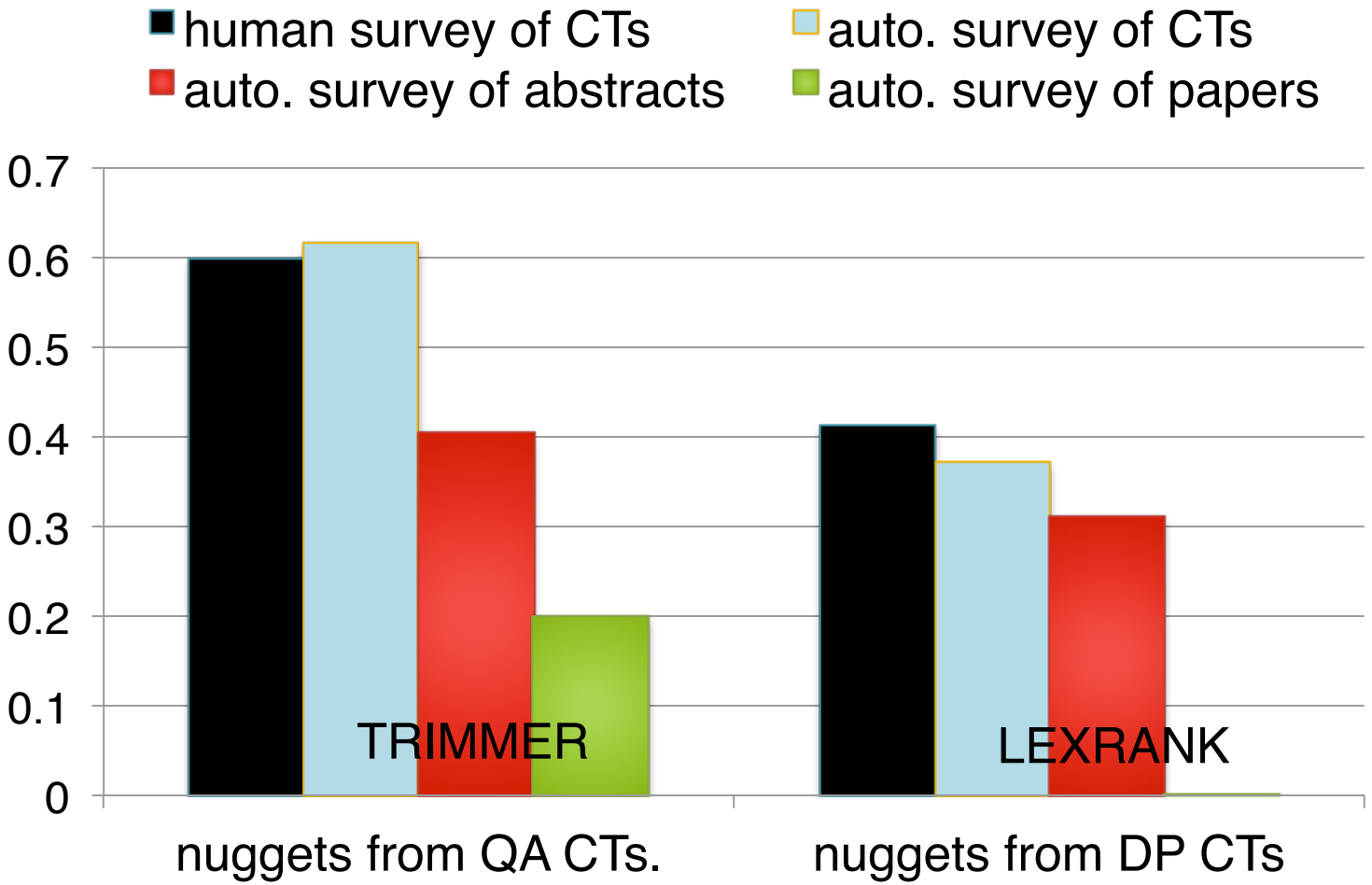
- Calculated recall and precision:

$$\text{recall} = \frac{\text{total weight of nuggets covered by a summary}}{\text{total weight of all human-identified nuggets}}$$

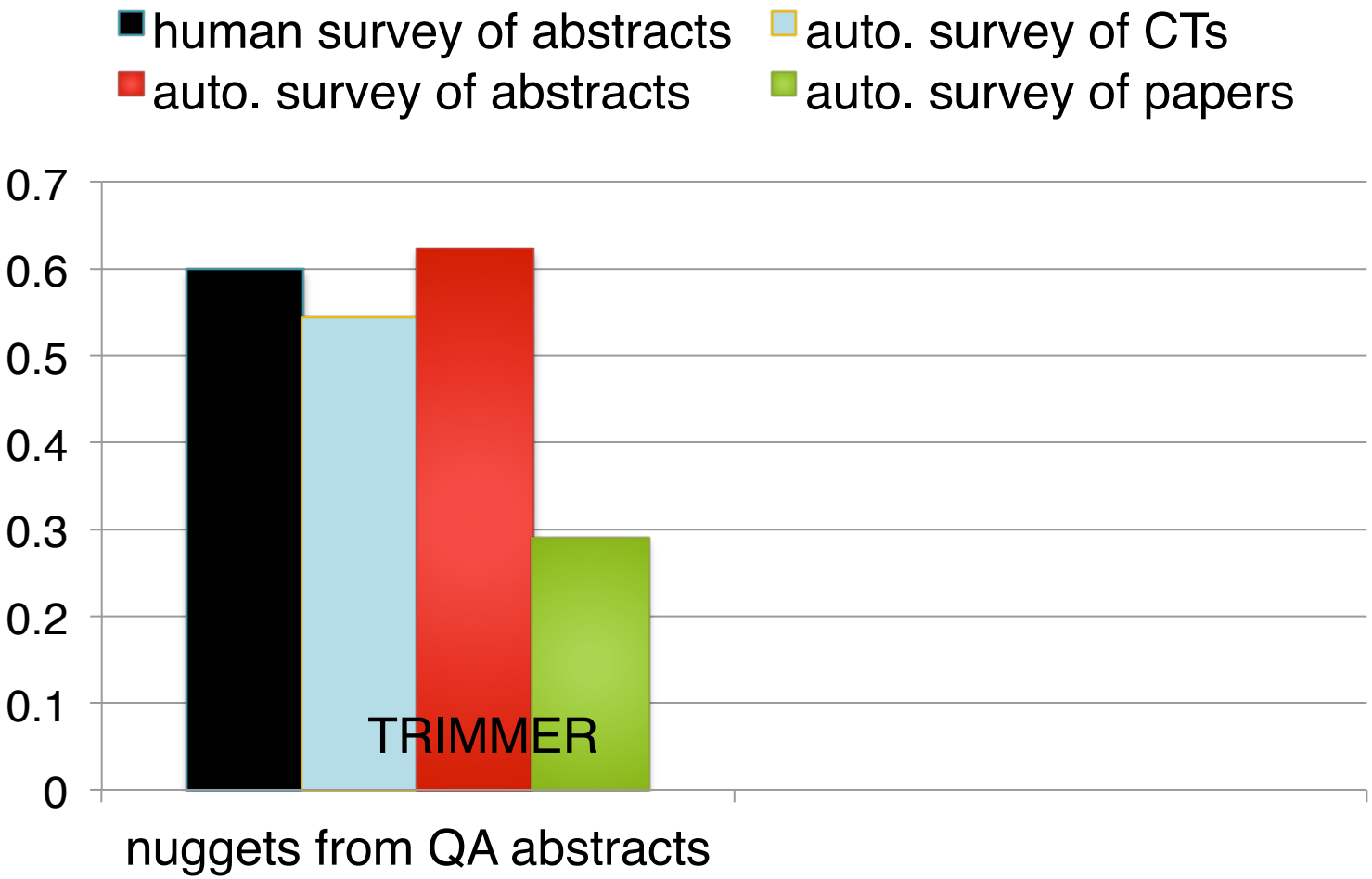
$$\text{precision} = \frac{\text{\# distinct nuggets covered in a summary}}{\text{\# sentences in the summary}}$$

- Calculated F-score with $\beta = 3$.
  - Weighted recall higher to reward surveys that included higher weighted nuggets.

F-scores of human and best automatic surveys: using CT nuggets



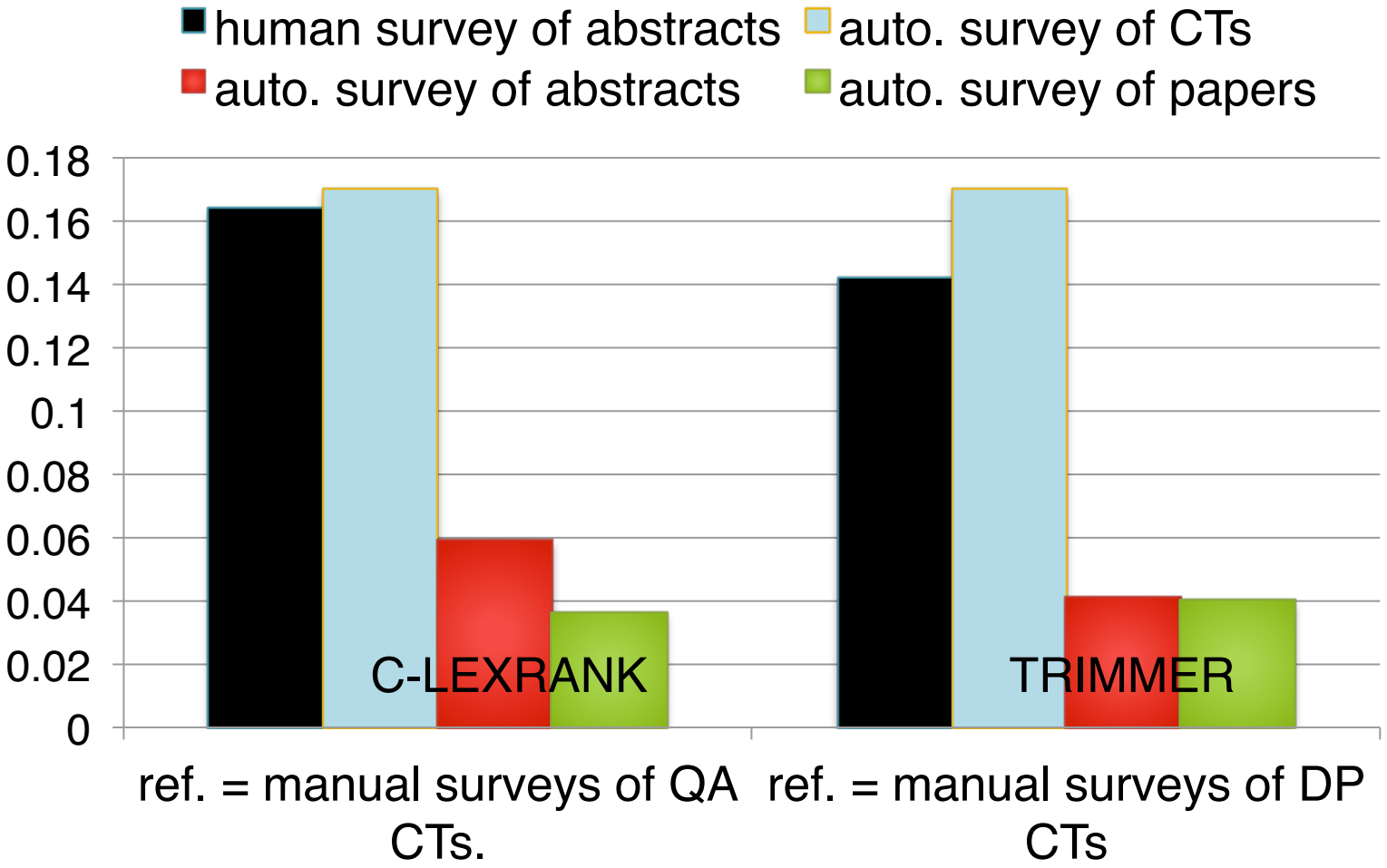
F-scores of human and best automatic surveys: using abstract nuggets



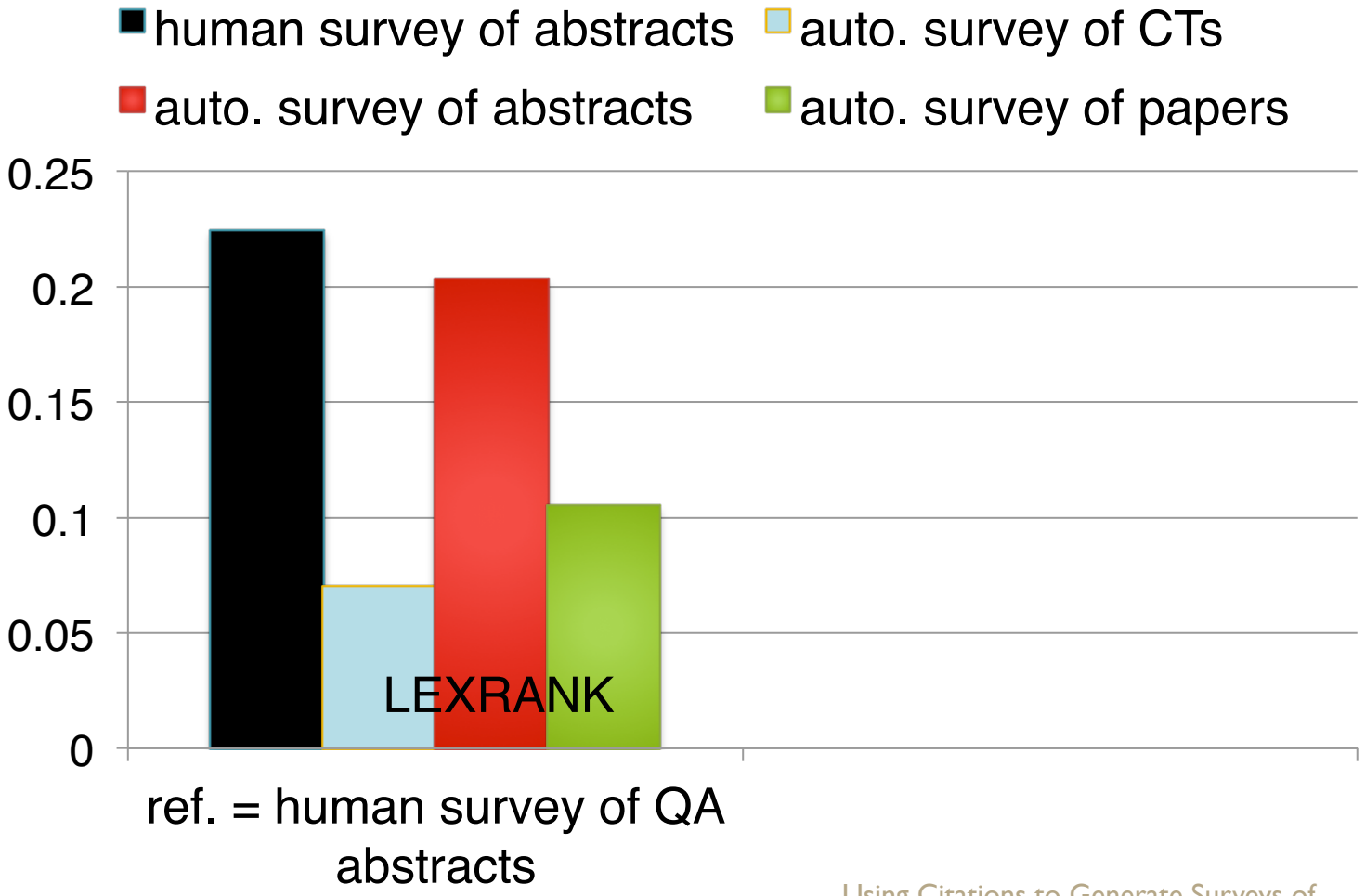
ROUGE evaluation

- Calculated ROUGE scores (ROUGE-1 and ROUGE-2) of the automatic summaries w.r.t. the reference summaries.
  - We report results for ROUGE-2.
  - ROUGE-1 followed a similar pattern.

ROUGE-2 of human and best auto-matic surveys: using CT refs.



ROUGE-2 of human and best automatic surveys: using abstract refs.



Example QA survey from CTs

To better facilitate user information needs, recent trends in QA research have shifted towards complex, context-based, and interactive question answering (Voorhees, 2001; Small et al, 2003; Harabagiu et al, 2005). Damerau (1981) summarizes experience with Transformational Question Answering System TQA during first full year of operation, 1978. Saquete et al (2004) focused on decomposition of complex question into several sub-questions. Yang et al, 2003, structured information extraction, and answer validation Magnini et al, 2002…

To sum up

- Investigated usefulness of directly summarizing citation texts.
- Generated automatic surveys of QA and DP papers, abstracts, and citation texts.
- Used both a nugget-based pyramid approach and ROUGE to evaluate the surveys.
- Showed that citation texts have useful survey-worthy information that is not present in, or difficult to extract from, abstracts or papers alone.

Future work

- Creating even better surveys by:
  - combining information from citation texts and abstracts
  - identifying citation sentences that express sentiment
- Creating a summarization workbench to aid humans in writing surveys by providing:
  - a unified framework with information from multiple sources and in multiple modes;
  - automatic summaries from different perspectives;
  - effective visualizations of collaborations and citations.