SUMMARIZATION OF SCIENTIFIC LITERATURE

Lucy Vanderwende, Microsoft Research

Anita de Waard, Elsevier Labs

PRIMARY JUSTIFICATION FOR SUMMARIZATION

… information overload …

… paper deluge …

… too much data …

Specific example: PubMed: 2k new articles per day, or 2 per minute. A few years ago, annual growth was 500k, now > 700k

PRIMARY QUESTION SHOULD BE …

- Who are the users?
  - Previously, in DUC and TAC, unnamed analysts
  - Unspecified how summarization was going to be used
- Who will be the users?
  - Ourselves!
  - And then others who seek information from text: If we develop tools that are useful for ourselves, we can hope that the tools will be useful for others as well – 'eating our own dog food'
- We propose to envision summarization as an end-user task

WHEN/WHY DO YOU NEED A SUMMARY? DOCUMENT-CENTRIC

- When planning research – ideally ☺
  - Starting from full-text corpus, identify:
    - Summary of specific article (measured against the abstract and/or Research Highlights, see later slide)
    - Summary of past work, drawn from this text and all texts it refers to recursively
    - Summary of methods/procedures used
- When writing:
  - Properly acknowledge prior work
  - Differentiate current contributions from prior work

WHEN/WHY DO YOU NEED A SUMMARY? IDEA-CENTRIC

- You have a question/hypothesis: “Is machine learning useful in summarization?”
- Identify past/present/future work that confirms/contradicts/hypothesizes
- Growing supporting corpora: RTE, CoNLL-2010 hedging, i2B2 assertion detection
- Summarization offers support for discovery:
  Starting from what you know, you can discover/see what you don’t already know, rather than presenting a full list of papers to be absorbed

WHY DO WE THINK THIS IS INTERESTING?

- We can use this form of summarization for our own research
- We know that the life sciences in particular are eager for innovation beyond keyword search
- However, pilots in our own field will be easier to judge
- You will care a lot when summarization is used to speed the rate of discovery (cf Alzheimer's, heart disease …)
- Computers at their best: tireless and neutral:
  They will review all the papers, not only those from brand name universities

WHAT SHOULD IT LOOK LIKE?

- We propose several pilots …
  - Summary of the main contributions of a paper
  - Summary of the main contributions of group of papers
  - Fact-based summary:
    will need iteration because it is less like pre-existing summaries

WHAT MAKES A GOOD TASK?

- Available data: see slide on Elsevier Research Highlights
- Meaningful evaluation
  - This may lend itself to Pyramid/nugget based evaluation
  - Possible extrinsic evaluation: “Should I cite this paper, given only the summary?”
- Aligned with funded initiatives
  - We will cast our net wide in the next few weeks contacting researchers with possible related funding: FUSE, Machine Reading, BOLT, NIH (?) …

AVAILABLE DATA FROM ELSEVIER

- Journal content in XML:
  - Full-text
  - Abstracts
  - Research Highlights *(see next slide for details)*
  - Pretty much any domain, any number of papers, including medical
    - Lots of CL in Artifical Intelligence that we can use for a pilot
- Books:
  - Reference works in life sciences, earth sciences, several other fields
  - Methods books/reference work in the life sciences
  - Methods lexicon
- Reference data (Scopus) in XML:
  - Heads (Title/authors/abstract)
  - Tails (references with DOI)
  - Deduplicated for author name
- Databases – manually curated:
  - Drug database (Reaxys)
  - Side effect database (Pharmapendium).

(RESEARCH) HIGHLIGHTS:

- Starting mid-2010, now implemented for > 1200 journals => available for appr. 60,000 papers so far
- 3-5 bullet points convey 'the core findings' of the article
- Up to the author to decide what that is
- For experimental fields (e.g. biology): 'Research Highlights'; for other domains (e.g. computer science): 'Highlights'
- Authors submit (Research) Highlights, at article submission stage
- Freely available with full text, abstract and keywords in XML

DEGREE CENTRALITY FOR SEMANTIC ABSTRACTION SUMMARIZATION OF THERAPEUTIC STUDIES

JOURNAL OF BIOMEDICAL INFORMATICS, 44(5) 2011, PP830-838

HAN ZHANG, MARCELO FISZMAN, DONGWOOK SHIN, CHRISTOPHER M. MILLER, GRACIELA ROSEMBLAT, THOMAS C. RINDFLESCH

Example Abstract vs. Highlights

- **Abstract:**
  Automatic summarization has been proposed to help manage the results of biomedical information retrieval systems. Semantic MEDLINE, for example, summarizes semantic predications representing assertions in MEDLINE citations. Results are presented as a graph which maintains links to the original citations. Graphs summarizing more than 500 citations are hard to read and navigate, however. We exploit graph theory for focusing these large graphs. The method is based on degree centrality, which measures connectedness in a graph. Four categories of clinical concepts related to treatment of disease were identified and presented as a summary of input text. A baseline was created using term frequency of occurrence. The system was evaluated on summaries for treatment of five diseases compared to a reference standard produced manually by two physicians. The results showed that recall for system results was 72%, precision was 73%, and F-score was 0.72. The system F-score was considerably higher than that for the baseline (0.47).

- **Keywords:** Automatic summarization; Natural language processing; Graph theory; Degree centrality; Semantic processing; Disease treatment

- **Highlights:**
  - Graph theory is exploited to extend a semantic abstraction method for summarizing multiple biomedical texts.
  - Degree centrality is effective in selecting information crucial for summarizing research on treatment of disease.
  - The system performs significantly better than a frequency-based method in identifying salient information.

POINTS FOR DISCUSSION

- Bulleted list of author-generated highlights
- If bulleted list is our new target, where does that leave readability?
- Are there other tasks where we ourselves are the end-users?
- Suggestions for initiatives to align to?
- Domain focus?
- Document focus or fact/proposition focus?
- Single/multi-document?

NEXT STEPS

- Incorporate your feedback
- Email discussions with TAC summarization program committee
- Email discussions with possible funding organizations
- Design the pilot of a pilot:
  - We will send to the TAC alias for help creating the summaries / judging summaries
  - Lather – rinse – repeat until satisfactory definition
- Inform TAC alias of schedule as soon as known