The Summarization RoadMap

Breck Baldwin, Robert Donaway, Eduard Hovy,
Elizabeth Liddy, Inderjeet Mani, Daniel Marcu,
Kathleen McKeown, Vibhu Mittal, Marc Moens,
Dragomir Radev, Karen Sparck Jones, Beth Sundheim,
Simone Teufel, Ralph Weischedel, Michael White

Background

- TIDES
  - Long-term, DARPA-sponsored research program in Translingual Information Detection, Extraction, and **Summarization**
- Annual Workshops and Technical Sessions at ACL
- Increased interest in summarization both from the industry and academia.

How do we evaluate our work?

RoadMap Highlights

- Annual goals, gradually increasing in challenge through a progression from
  - Less to more demanding data
  - Intrinsic (system focused) to extrinsic (task specific) evaluations
- Ambitious plan to promote research in
  - Single/multiple document summarization
  - Natural language generation
  - Information clustering, theme identification, co-reference resolution, etc.

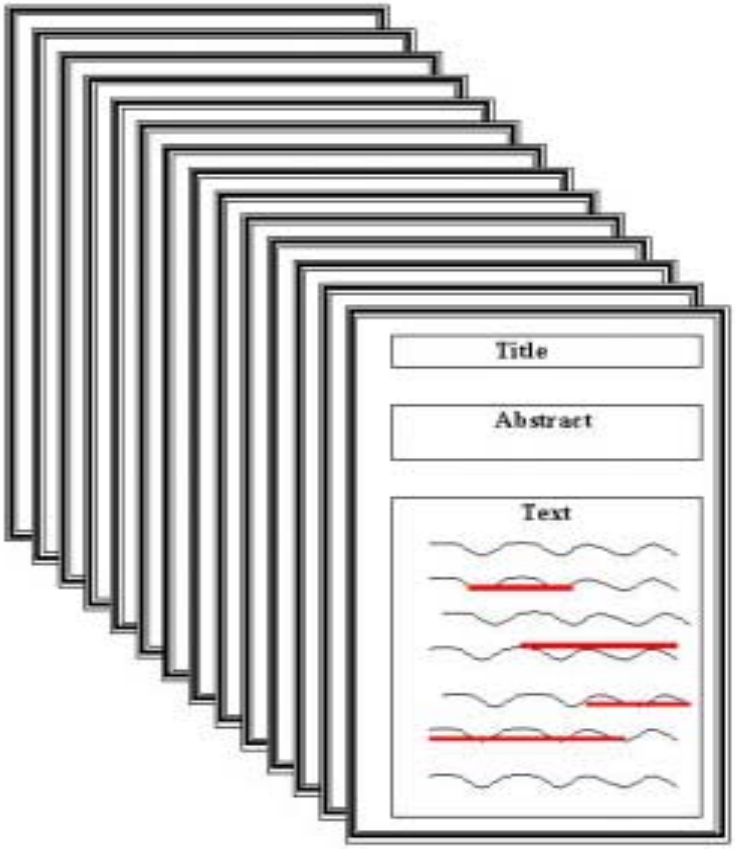
Initial Focus

- Intrinsic evaluations
  - Single and multiple document extraction techniques
  - Generation of coherent outputs

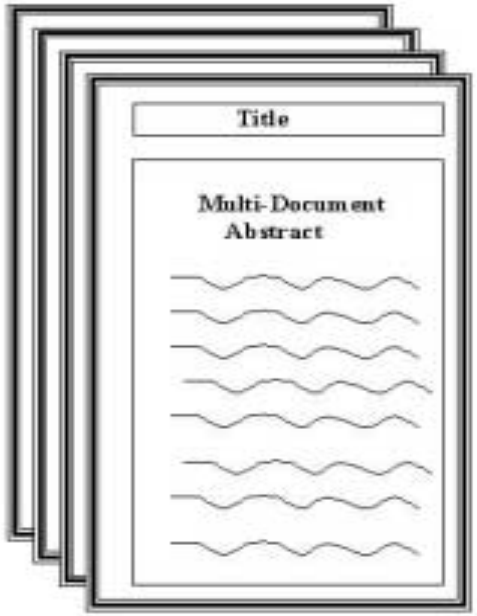
Plan for year 1

- NIST develops a summarization corpus
- NIST makes the training corpus available to the community
- Research groups prepare for evaluations on a number of summarization-specific tracks
- Research groups run their systems on test data
- NIST evaluates the results
- Report results at DUC-2001

Training corpus: 30 doc collections



Documents on a given Topic/Event.



Multi-document abstracts of the set of documents. The multi-document abstracts are approximately 50, 100, 200, and 500 words long.

Evaluations (1)

- Intrinsic evaluation of single-doc sentence/sentence fragment extractive summaries.
  - **Input:** one document
  - **Output:** extract of document
  - **Evaluation:** compare manual and automatically generated extracts
- Metric ??
  - Recall and precision
  - Utility [Radev]
  - Content overlap [Donaway et al.]
  - Humans? (to determine which of the above metrics correlates best with human judgments)

Evaluations (2)

- Intrinsic evaluation of single document abstracts.
  - Track 1:
    - **Input:** one document
    - **Output:** abstract of document
    - **Evaluation:** compare manual and automatically generated abstracts and assess quality of automatically generated abstracts
  - Track 2
    - **Input:** one document with important sentence fragments identified
    - **Output:** abstract of document
    - **Evaluation:** compare manual and automatically generated abstracts and assess quality of automatically generated abstracts
- Metric ?? Human judges assess
  - Adequacy of the semantic content selected in an abstract
  - Readability and coherence of an abstract

Evaluations (3)

- Intrinsic evaluation of multi-document abstracts.
  - Track 1:
    - **Input:** one collection of documents
    - **Output:** abstract of document collection
    - **Evaluation:** compare manual and automatically generated multidoc abstracts and assess quality of generated abstract
  - Track 2
    - **Input:** one collection of docs with important sentence fragments identified
    - **Output:** abstract of document collection
    - **Evaluation:** compare manual and automatically generated abstracts and assess quality of generated abstract
- Metric ?? Human judges assess
  - Adequacy of the semantic content selected in the abstract
  - Readability and coherence of the abstract

Adequacy of semantic content

- Recall and precision
- Utility [Radev]
- Content overlap [Donaway et al.]
- Question-answering game [Hovy and Lin]
  - Analysts prepare a set of questions whose answers should be found in the abstract (single or multidoc)
  - Human judges determine how many questions they can answer
    - Before reading the abstract
    - After reading the abstract
    - After reading the whole doc/collection of docs

Then move on to bigger things

- More complex texts; multiple compression rates
- Evolving summaries
- Task specific summaries (TDT, biographical summaries)
- Summaries that answer multiple questions
- Multiple language summaries
- Integrated q&a and summarization evaluation

Strengths

- Potential to automate partially the construction of the corpus
- Modularity [single vs. multiple docs; extraction vs. generation]
- Complexity [importance, co-reference, topic-fusion, generation]
- Incremental
- Specificity [no constraints on the framework]
- Exploration of summary space [long vs. short, summaries for various purposes, etc.]

Limitations

- Not clear what evaluation criteria are the best (ongoing refinement)
- Expensive
- Focuses only on textual data

Goals of this meeting

- Reports from various research groups concerning their experience in **evaluating** single- and multi-doc summaries
- Decide
  - What evaluation metrics we will use next year
  - When we will hold the next DUC
  - How we get feedback from the community
  - Who does what

Summarization at ISI

- Webclopedia [Hovy, Lin]
- ReWrite [Knight, Marcu]
  - Statistical methods for machine translation and sentence/text compression

- The documentation is typical of Epson quality: excellent.
- Documentation is excellent.

- Reach's E-mail product, MailMan, is a message-management system designed initially for VINES LANs that will eventually be operating system-independent.
- MailMan will eventually be operating system-independent.

- Although the modules themselves may be physically and/or electrically incompatible, the cable-specific jacks on them provide industry-standard connections.
- Cable-specific jacks provide industry-standard connections.

How do we evaluate such compressions?

Sentence compression evaluation

- Ask humans to assess on a 1 to 5 scale
  - the grammaticality of the compressed sentences
  - the ability of a system to select the most salient information in a sentence
- Method
  - Compared the output of our algorithms against a baseline and a "topline"
  - Enabled humans to see all compressions/alternatives at any given time

Web-based evaluation tool

12. the chemical etching process used for glare protection is effective and will help if your office has the fluorescent–light overkill that 's typical in offices .

the chemical etching process used for glare protection is effective and will help if your office has the fluorescent–light overkill that 's typical in offices .

(1) ◇ ◇ ◇ ◇ ◇ (5)

glare protection is effective .

(1) ◇ ◇ ◇ ◇ ◇ (5)

the etching process used for glare is effective and will help if your office has the fluorescent–light overkill that 's in .

(1) ◇ ◇ ◇ ◇ ◇ (5)

the process used for glare protection is and will help if your office has the overkill

(1) ◇ ◇ ◇ ◇ ◇ (5)

Evaluating the coherence of abstracts

- Evaluate comparatively the abstracts generated by the DUC participants.
- Include in the evaluation set
  - a baseline abstract:
    - first 100 words of the most recent document
    - first 100 words of a random document
  - a human generated abstract
  - the abstracts generated by the systems

Will this scale up to, let's say, 30 participants?

Thank you!

How to we evaluate our work?

- TIDES initiative
  - Vision committee:
    - Jaime Carbonell, Donna Harman, Eduard Hovy, Steve Maiorano, John Prange, and Karen Sparck-Jones
  - Summarization RoadMap committee:
    - Breck Baldwin, Robert Donaway, Eduard Hovy Elizabeth Liddy, Inderjeet Mani, Daniel Marcu, Kathleen McKeown, Vibhu Mittal, Marc Moens, Dragomir Radev, Karen Sparck Jones, Beth Sundheim, Simone Teufel, Ralph Weischedel, Michael White