Plans for DUC 2005 and beyond

Past tasks at DUC

- DUC 2001 – generic summarization
  - Single documents at fixed length of 100 words
  - Across multiple documents at lengths of 50,100,200,400
- DUC 2002 – repeat except different lengths for multiple

- DUC 2003 – generic and query-based summarization
  - Single documents (generic) at length 10 words
  - Across multiple documents at length 100
    - Generic, focused by TDT cluster or viewpoint
    - Query-based using TREC topics as query
- DUC 2004 – repeat except use Who is questions as query
  - Also some cross-language (Arabic to English) summarization

2004 Roadmap committee summary

Useful in building community and good work on generic summarization of news, BUT

1) Problems in designing extrinsic evaluations:
   This comes from the question of where generic summaries are useful, and in particular how to PROVE this
2) Problems with extremely high baselines for news data
3) Problems with evaluation, in particular with the imprecision of using EDUs for judgment of coverage

2004 roadmap set new goals

1) Find some real need for summarization and motivate/define the evaluation framework from the point of view of one or more realistic task scenarios
2) Move away from generic summaries of news to summaries of additional genre with respect to broad subject areas (but overlap in some ways with previous source types and tasks)
3) Allow partial participation (by component, source type)

More goals

4) Continue working on evaluation
   a) Continue to support development, use and testing of tools for automatic evaluation, such as ROUGE
   b) Continue to explore better ways of coverage evaluation, such as the Columbia pyramid suggestions
5) Be open to evolution of goals in the nature of the task (fusion, extraction, Q&A), input (not just text), and output (lists, outlines, timelines, etc)

A Real Application

Higher-level reports that fuse information from many sources

Specific examples: World Health Organization situation reports on diseases, situation reports on natural disaster relief status (UN, Relief Web, OCHA, etc.)

Taipei Earthquake Situation Report No. 3 Oct. 4, 1999

- **Urgently needed relief supplies continued to arrive in the earthquake affected areas and have been distributed…Poor weather conditions…thousands of people remain in tents.**
- The Context
  - It has now been confirmed by authorities in Taiwan that 2,192 people were killed, 8,735 were seriously injured, 97 remain missing or trapped in collapsed building, and approximately 100,000 were left homeless as a result of the earthquake, measuring 7.6 on the Richter scale, which struck Taiwan in the early morning hours of Tuesday, 21 September
- Latest events
  - It still remains quite unclear how many people have sought shelter in tents outside their homes….Electricity is rationed…main water reservoirs in Nantou and Taichung counties were damaged…
- Red Cross/Red Crescent action
  - Taiwanese Red Cross Organisation provided USD1.6 million to support..

Types of data found

- News data from AQUAINT
- Longer UN reports on this event
- Reports from other governments about offering of aid, etc.
- Thread of messages from soc.culture.taiwan discussing this
- Page of links to reports from US Geologic Survey including BGS report, scientific articles on earthquakes, maps, photos

DUC 2005 scenario

Based on situation reports for natural disasters such as WHO, other UN

- What happened
- Geographic area affected, including infrastructure such as bridges, roads, land)
- Populations affected (mortality, homelessness)
- Main needs
- Local/national response
- Regional/international response
- Social/political/geographical constraints
- Expected developments

Thoughts on Data, Tasks

- NIST would supply documents drawn from AQUAINT newspaper/wire data
- NIST would also locate appropriate reports, maybe some scientific papers??, etc.
- Other groups would contribute additional data sources, including more English text material, and other items such as speech data, non-English sources, etc.
- In a similar vein, additional tasks could be proposed that would fit into this scenario, such as “headlines” for easy click-down
- All of the data would be available, but participants could decide what they would use in DUC 2005

Thoughts on Evaluation

- 2005 --- Feasibility is the main goal
- Distributed evaluation --- NIST + others
- NIST's role
  - Serve as central distribution, gathering location
  - Create X (maybe 25??) scenarios around natural disasters
  - Evaluate the 8 major outline summary boxes, creating reference summaries (and possibly derived "infolists") to allow manual "coverage" evaluation for the English text data; run automatic evaluation software
  - Do a pseudo-extrinsic evaluation looking at the time to complete judgments
- Groups contributing other types of data or other tasks would be responsible for evaluation of those

Thoughts on Metrics

- Need a joint effort to develop better units for the model summary (and for automatic eval??)
- One way is to start with the Columbia pyramid pilot
  - Columbia develop written guidelines
  - Others try to use those guidelines
  - Several small pilot studies using DUC 2004 reference summaries in order to perfect guidelines
  - NIST then try to use these guidelines in a final pilot
- Other ideas
  - Work on other units
  - Work on automatic use of these guidelines

Possible Timeline

Starting NOW – an organized set of pilot projects looking into these new units for judgment

Dec 2004 – NIST provides several examples using the suggested template and timeframe (1998-2000)

- Groups look for additional data, tasks, work on guidelines for their use, plan evaluation, etc.
- Groups train systems for new task

June 1, 2005 – NIST provides list of test events

By June 15 --- groups select additional documents, send to NIST for distribution; June 15 NIST distributes all data

July 1 – results due at NIST; July 30 – evaluation done

October?? – DUC 2005 meeting at HLT

Some issues

- THIS IS A PILOT TASK!!
- Importance of working on the evaluation metrics
- News data is easy to get and to fit into scenario, other data is harder and not as clear how its content would fit into this scenario
- How can groups train for this new task
- Additional issues???