SPAWAR
Systems Center
San Diego

The Tipster Question-and-Answer (Q&A) Summarization Task: Test Design

Beth Sundheim
sundheim@spawar.navy.mil
619/553-4145

Objective and Approach

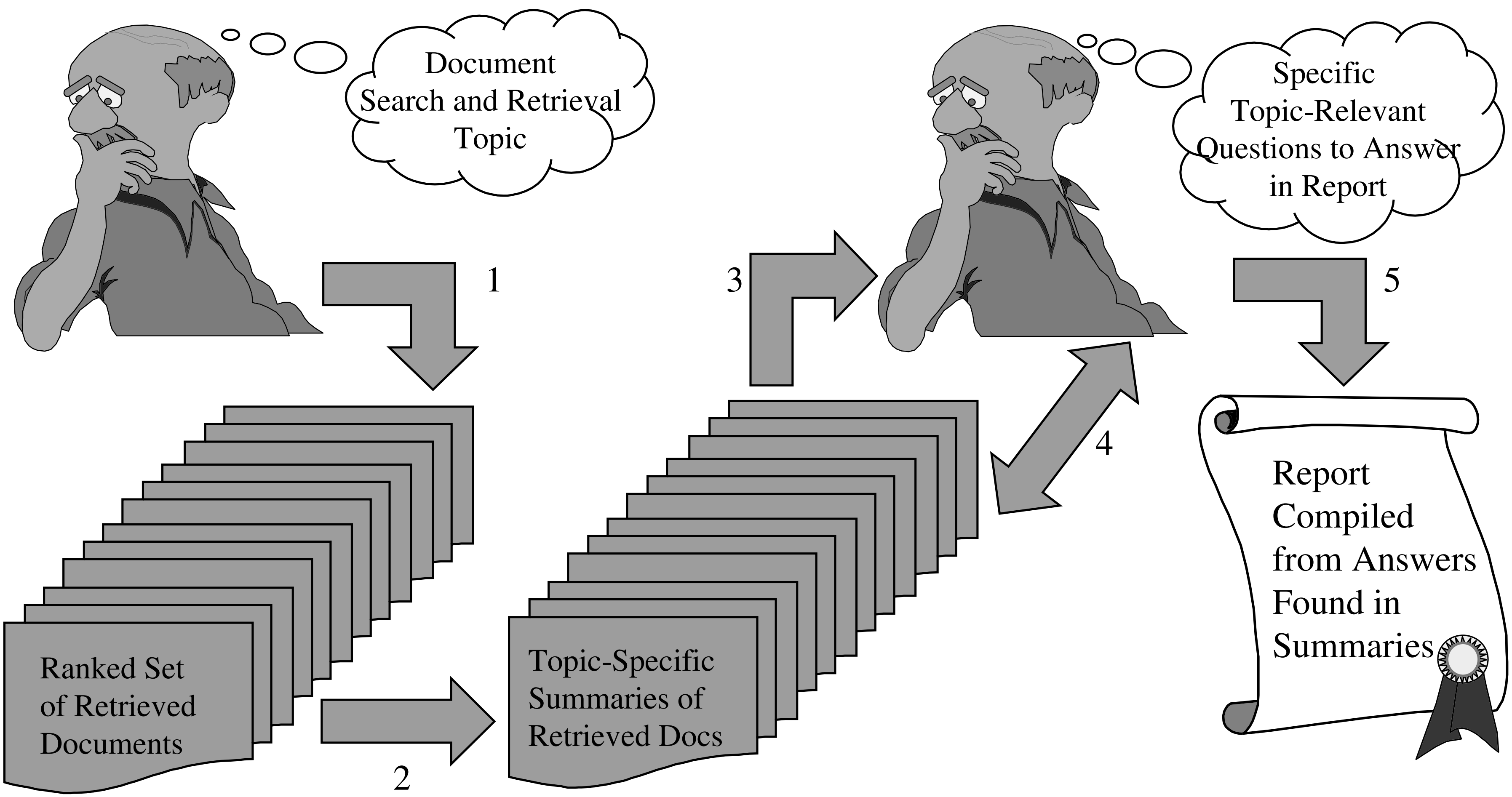
Systems Center
San Diego

- Objective: Find ways to determine to what extent a summary contains the gist of what a full document contains in relation to a topic.
- Approach: Define task that requires that summaries contain answers to questions derived from a detection need statement.

CAVEAT: The current evaluation design is highly experimental, and the test corpus is very small. No statistical significance tests were run on the results.

Using Summaries as Surrogates for Full Documents in Report Writing

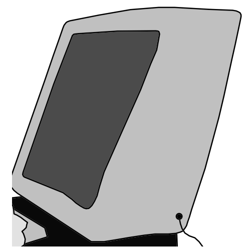
Systems Center San Diego



SPAWAR
Systems Center
San Diego

Basic Steps in the Q&A Test

1. Understand Selected TREC Topic and TREC Relevance Judgments
2. Compose Questions on "Obligatory Aspects" of Topic
3. Create Answer Key: Annotate Each Doc in Test Set to Identify Answer(s) to the Questions



4. Run Test: Summarization Systems Get Topic (NOT Questions) and Original Documents as Inputs

5. Assess Summaries for Number and Quality of Answers Found (in comparison to key)

6. Analyze Results in Terms of "Answer Recall" in Relation to Compression Ratio

Q&A Evaluation Support

SPAWAR
Systems Center
San Diego

Data Prep and Assessment	Design, Coord. & Analysis
Leo Obrst, Mitre	David House, Mitre
Michael Valenzuela, UCSD	Inderjeet Mani, Mitre
Sandra Wagner, DoD	Leo Obrst, Mitre

Participating Sites

UPenn	GE CRD
SRA	ISI
Carnegie Group	Cornell
TextWise	NMSU CRL

Training and Test Corpora

Systems Center San Diego

- Test was on three TREC topics
  - One requests statistical data, one requests event data, one requests subject/trend data
  - Summarization systems got topic title, description and narrative as input
- One test set of articles per topic
  - Summarization systems got 90 articles per topic as input
  - Evaluation done on subset of 30 relevant articles per topic
- Training corpus consisted of one topic, set of 5 questions, 50 relevant articles and manually prepared answer keys

Test Topics and Test Questions

Systems Center San Diego

Number: 257

Topic: Cigarette Consumption

Description: What data is available on cigarette consumption by country?

Narrative: If cigarette smoking is a causative factor in lung cancer, then countries with higher cigarette consumption per capita might experience a higher incidence of lung cancer. This topic would provide basic data for such a comparison.

Test Questions:

1) How many cigarette smokers are reported; include date if available.

2) How many cigarettes are reported to have been consumed; include date if available.

3) What is the location for which cigarette consumption data is reported?

4) How many people suffer health effects related to cigarette consumption?

Test Topics and Test Q’s (Cont’d)

Systems Center San Diego

Number: 258

Topic: Computer Security

Description: Identify instances of illegal entry into sensitive computer networks by non-authorized personnel.

Narrative: Illegal entry into sensitive computer networks is a serious and potentially menacing problem. Both 'hackers' and foreign agents have been known to acquire unauthorized entry into various networks. Items relative this subject would include but not be limited to instances of illegally entering networks containing information of a sensitive nature to specific countries, such as defense or technology information, international banking, etc. Items of a personal nature (e.g. credit card fraud, changing of college test scores) should not be considered relevant.

Test Questions:

1) **Who is the known or suspected hacker accessing a sensitive computer or computer network?**
2) **How is the hacking accomplished or putatively achieved?**
3) **Who is the apparent target of the hacker?**
4) **What did the hacker accomplish once the violation occurred? What was the purpose in performing the violation?**
5) **What is the time period over which the break-ins were occurring?**

Test Topics and Test Q's (Cont'd)

Systems Center San Diego

Number: 271

Topic: Solar Power

Description: To what extent is solar power used as an alternative to fossil fuels in various countries worldwide?

Narrative: Although the development of solar power as a major energy source has progressed slowly, in some parts of the world it is used extensively. Where and for what purposes?

Test Questions:

1. **What is full company name or affiliate who developed the product or procedure of a solar powered energy source, else the name of the person(s) who developed the item or procedure and those who test it, modify it, or give the money to further its use or development?**
2. **What is full name of the continent, country, state, province, region, city, or other location on Earth or off Earth (i.e. in orbit around Earth, on the moon/Mars/Venus etc.) where solar power is being used either for an experimental purpose or a commercial one?**
3. **What is the reason(s) or idea(s) behind why solar power is preferred over conventional energy sources such as fossil fuels or nuclear power, or what is the benefit or advantage of using solar over other energy sources (e.g., renewable or non-polluting)?**
4. **In what way does solar energy manifest itself? Is it in a device that delivers the energy to another source like a solar panel or a solar cell which could heat the water of a home or power a radio or a calculator, or is it like a car or other mechanical device that is powered by solar energy which charges batteries that allow the device to run on a source other than fossil fuels?**
5. **Is this solar device or mechanism completed? If so, what is the cost? If not, when will it be completed and what is that cost? What level or stage of development is the solar energy source with regard to use in the public domain? Is it in the prototype stage, is it in the experimental stage is it in the commercial stage where it can be used by all?**

Q-and-A Test: Setup

Systems Center
San Diego

- For each topic, one set of 4 or 5 Q's was written for the entire set of 30 relevant articles
  - Each Q reflects an "obligatory aspect" of the topic. (An obligatory aspect is something that must be conveyed by the document for the document to be considered relevant to the topic.)
  - NB: A given article may not answer all Q's, even though it is relevant
- Each topic was assigned to a different annotator/assessor, who prepared the Q's, keys and model summaries, and scored the automatic summaries in accordance with guidelines prepared by SSC
  - Each person kept notes on interpretation of Q's and special cases encountered during scoring
  - A small-scale study comparing stringency of scoring among the three assessors was conducted
- Total time spent on preparing keys and model summaries varied greatly: ~20 hrs (Leo, topic 257), ~60 hrs (Michael, topic 271), ~225 hrs (Sandy, topic 258)

Scoring

Systems Center
San Diego

- Scored “best” response summaries
  - Model summaries serve to indicate degree of compression achievable without sacrificing recall
- Scoring categories: Correct, Partially Correct, Missing
  - No score assigned for a Q if a given article contains no answer
- Decisions re amount of credit for the A to a given Q were made by comparison of quality as well as quantity of pertinent text segments in the key and response
  - Summary had to provide enough context to make a matching text segment recognizable as an A (scored as Missing if context insufficient)
- Subjectivity of scoring is increased when the response does not include as many text segments as are marked as A’s in the key
  - Multiple annotations for a given Q for a given article could indicate redundant A’s, differing degrees of specificity of a single A, or multiple distinct A’s (different instances, examples, etc.)
- Total time spent on scoring (per assessor) was probably ~20 hrs

Metrics

Systems Center
San Diego

- Answer recall (generous version)

  = ((num correct A's in response summary) + (.5 X num partially correct A's in response summary)) / (num A's in key)
- Answer recall (harsh version)

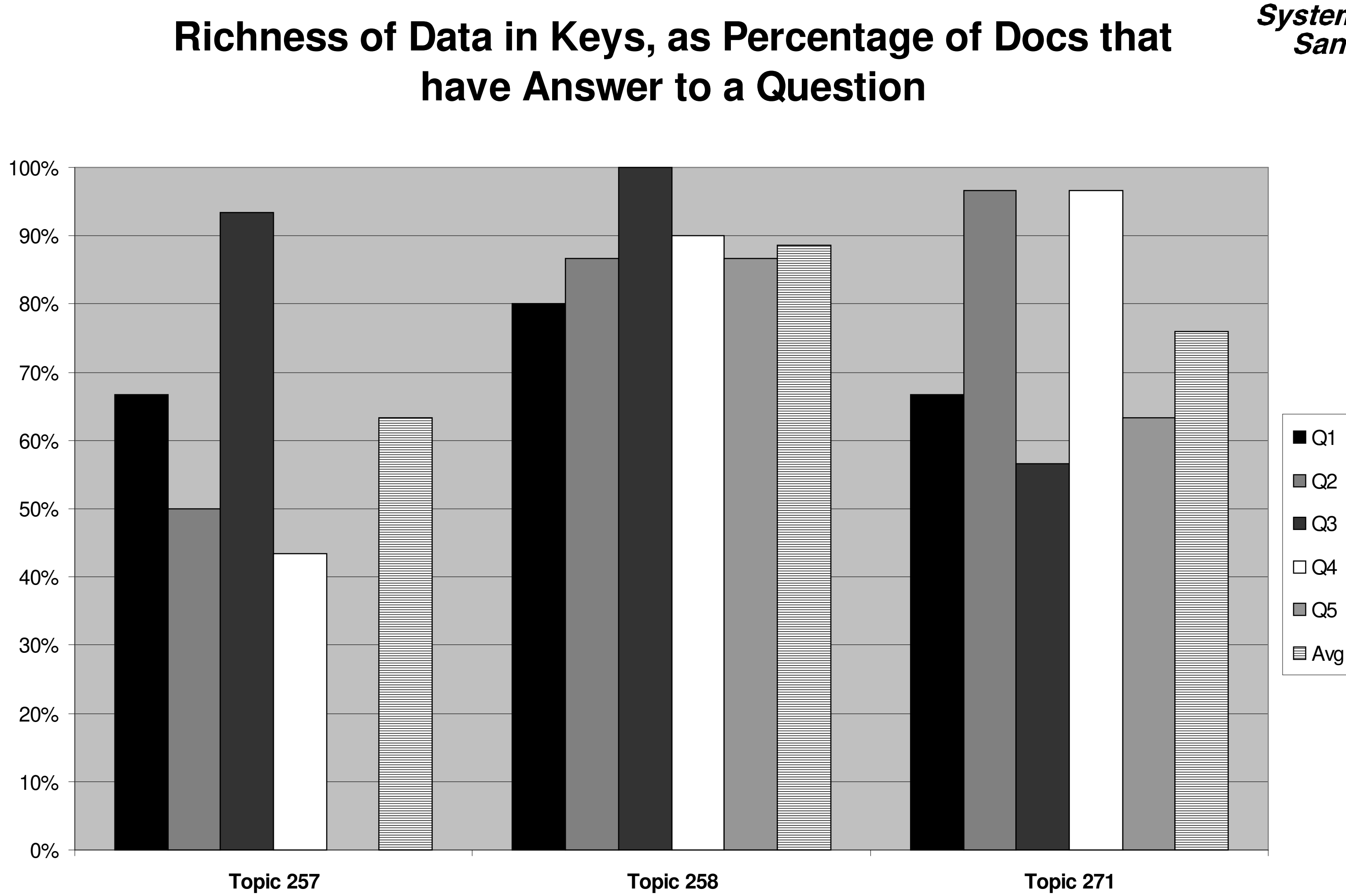
  = (num correct A's in response summary) / (num A's in key)
- Compression ratio

  = 1 - ((num S's in response summary) / (num S's in full doc))

where *response [summary]* is any summary under evaluation, and *key* is the annotated full text.

Note: Precision is not measured, since there is no clear notion of an *incorrect* or *spurious* answer

Richness of Data in Answer Keys



Richness of Data in Answer Keys (Cont'd)

Systems Center San Diego

1. Number answers per question per topic (maximum = num docs = 30)

	Topic 257	Topic 258	Topic 271
Q1	20	24	20
Q2	15	26	29
Q3	28	30	17
Q4	13	27	29
Q5	N/A	26	19
Total	76	107	95
Avg/Doc	2.5	3.6	3.2

2. Average number tagged segments (SGML tags) in key per answer

	Topic 257	Topic 258	Topic 271
Q1	1.3	2.7	5.4
Q2	3.5	3.2	3.2
Q3	3.5	4.4	2.7
Q4	2.2	3.6	3.3
Q5	N/A	2.6	3.5
Overall Avg	2.6	3.3	3.6

SPAWAR

Systems Center
San Diego

Answer Recall (ARA) vs Compression (1-Compression) for Each Topic and Site