TREC Question Answering Track

Seminar Information Extraction

15th January 2014

Jennifer Nist

Albert-Ludwigs-Universität Freiburg

UNI FREIBURG

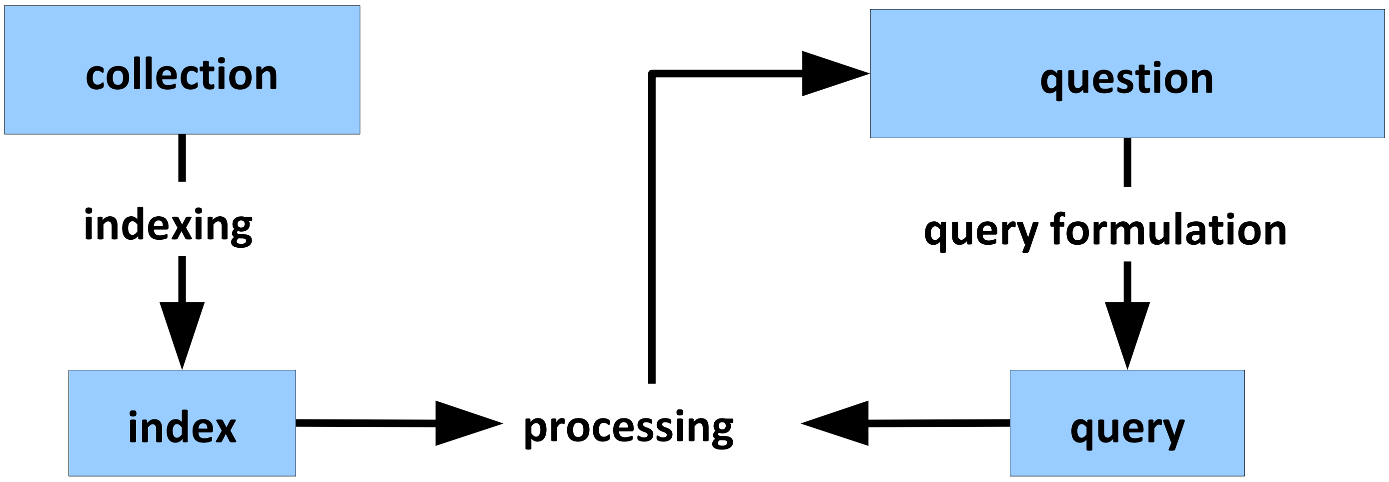
TREC - Text REtrieval Conference

- A workshop series (annually)
  - that provides the infrastructure for information retrieval technology
    - test collections,
    - open forum

- Sponsored by:
  - National Institute of Standards and technologies (NIST)
  - U.S. Department of Defence

- Goals:
  - encourage research in information retrieval
  - increase communication among industry, academia, and government

Document retrieval



- Document retrieval system
  - find relevant documents to user queries
  - evaluate and sort them according to relevance

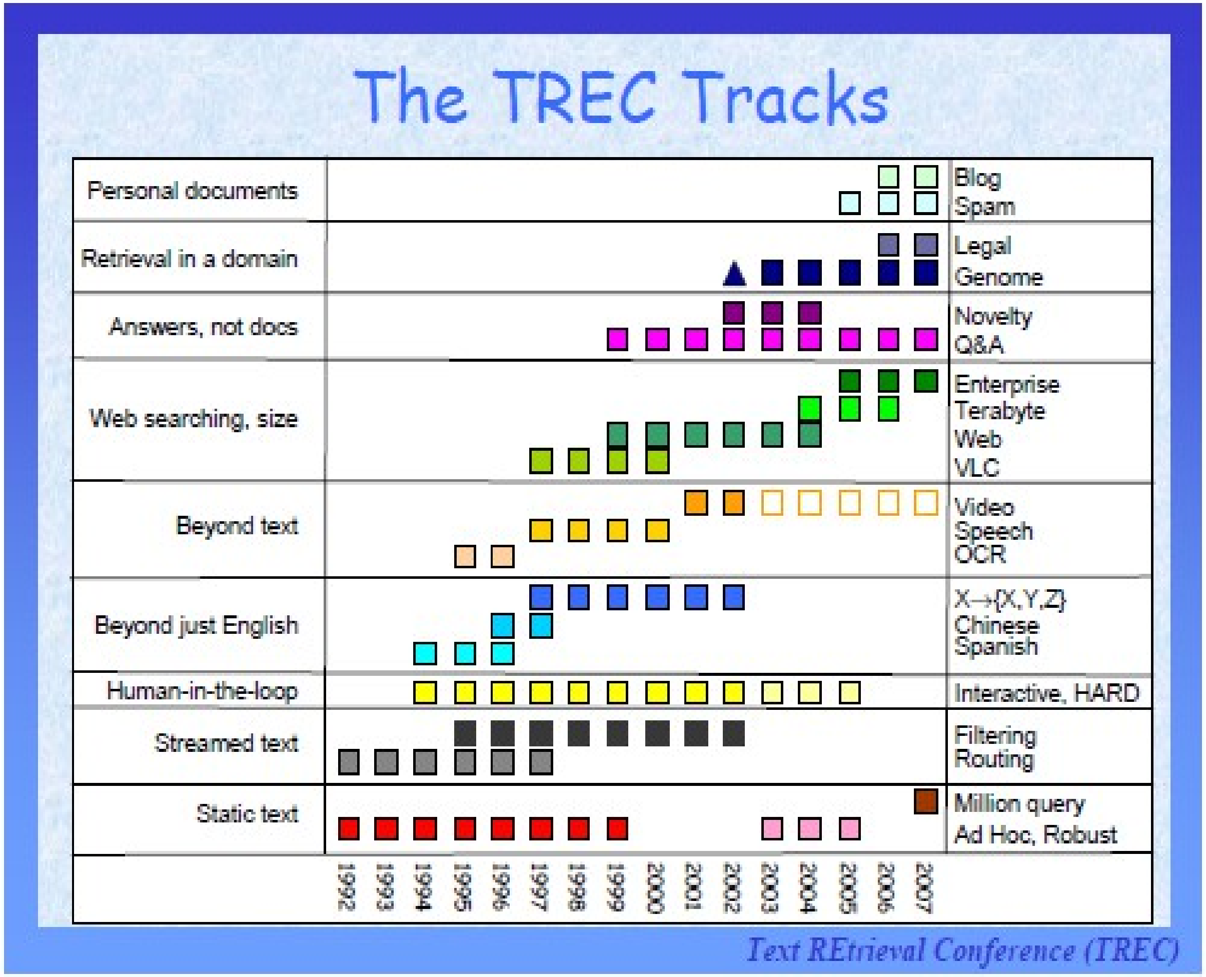
TREC cycle

- NIST (program committee)
  - provides a test set of documents
- Participants
  - run their retrieval systems on the data and
  - return a list of the top-ranked documents to NIST
- NIST
  - judges the documents for correctness, and evaluates the results
- Workshop for all participants to share their experiences

TREC Tracks

- Definition: Task that focuses on a particular sub problem of information retrieval.

- a workshop consists of a set of tracks


The TREC Tracks
Personal documents
Blog
Spam
Retrieval in a domain
Legal
Genome
Answers, not docs
Novelty
Q&A
Web searching, size
Enterprise
Terabyte
Web
VLC
Beyond text
Video
Speech
OCR
Beyond just English
X→{X,Y,Z}
Chinese
Spanish
Human-in-the-loop
Interactive, HARD
Streamed text
Filtering
Routing
Static text
Million query
Ad Hoc, Robust
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
Text REtrieval Conference (TREC)

Question Answering Track

- TREC-8 (1999) – TREC 2007

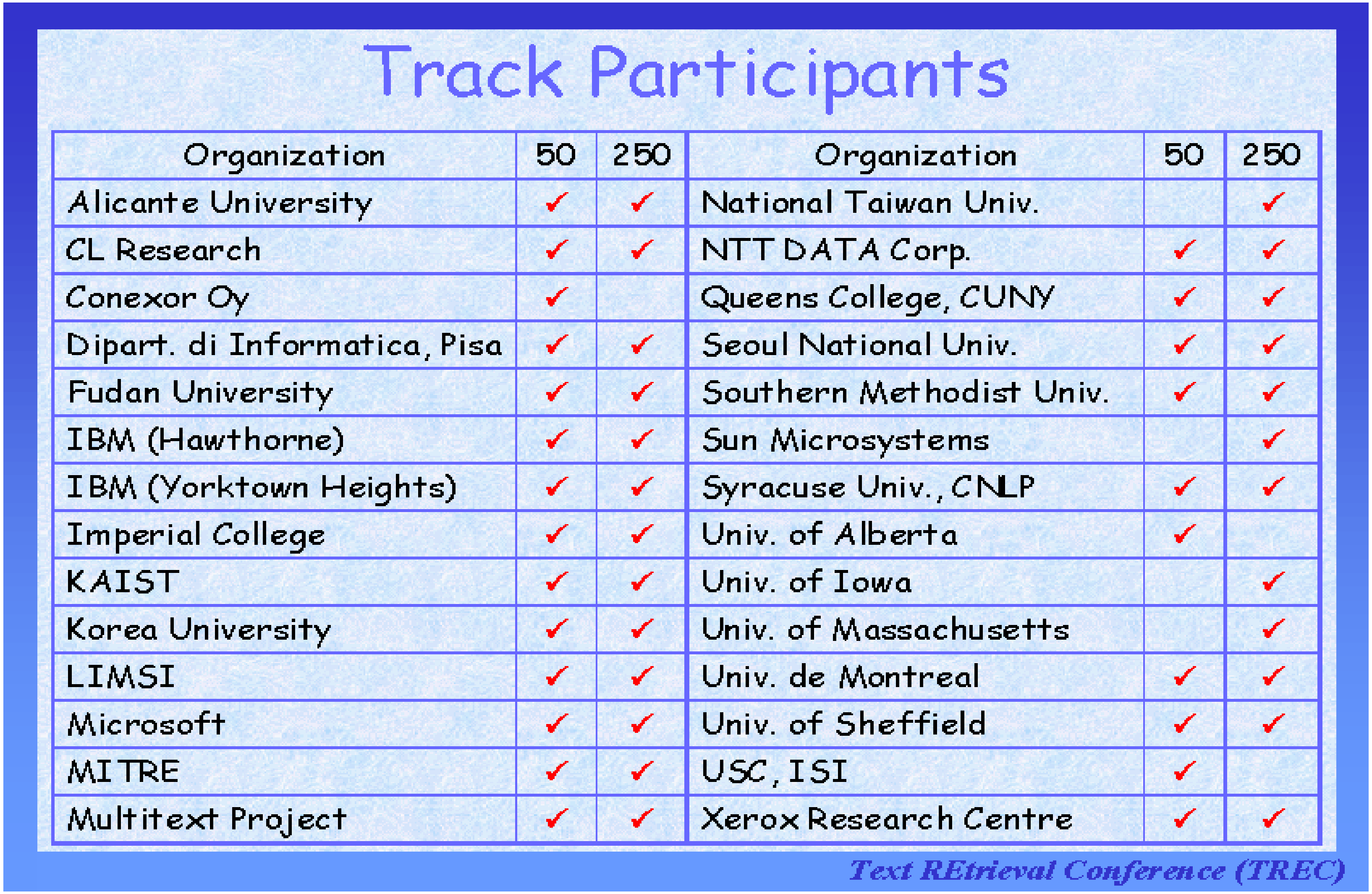
- Response to a question:
  - Traditionally: focused on returning a ranked list of documents
  - QA: returning the answers themselves

    Question: When was Beethoven born?

    Answer: Ludwig van Beethoven was born on December 16, 1770.

- Assumption:
  - users would prefer to be given the answer rather than find the answer themselves in a document

TREC-9: Participants



Track Participants

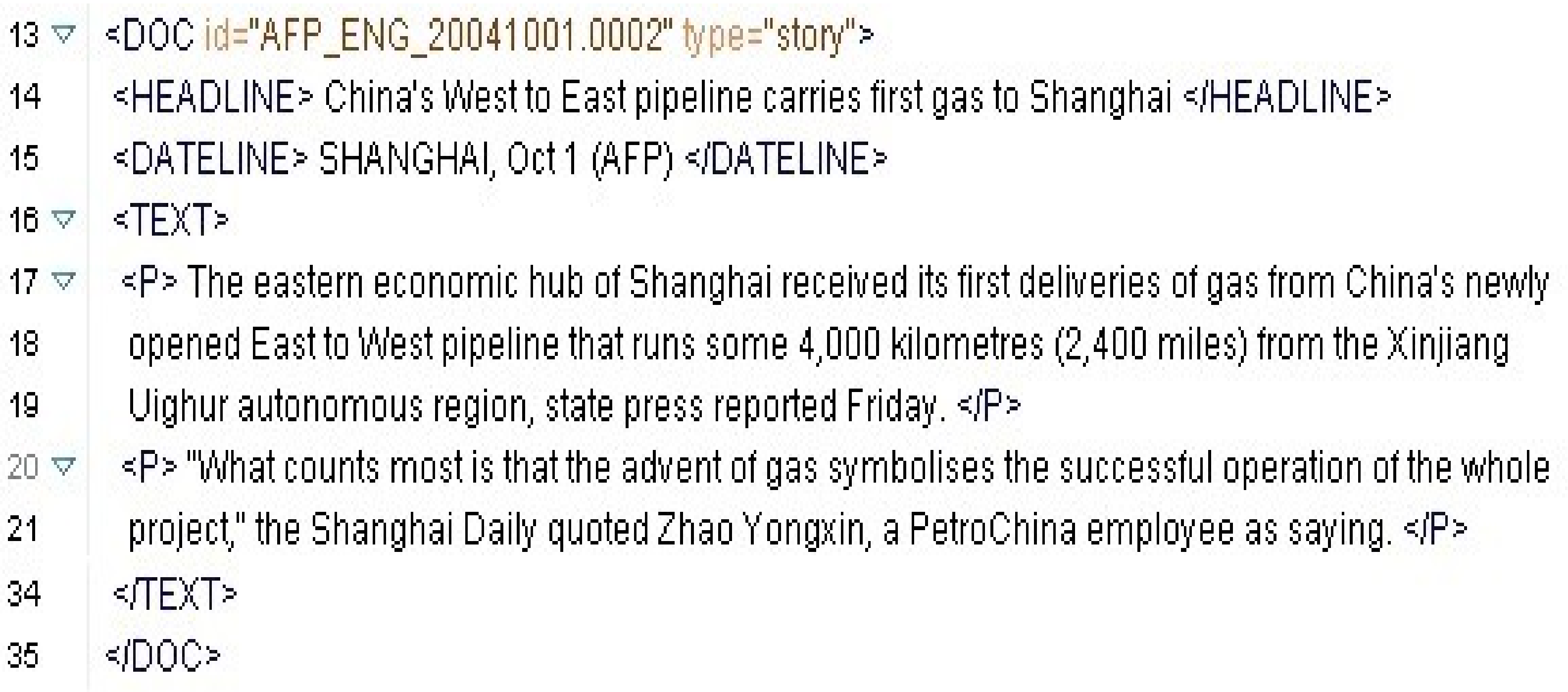
Organization	50	250	Organization	50	250
Alicante University	✓	✓	National Taiwan Univ.		✓
CL Research	✓	✓	NTT DATA Corp.	✓	✓
Conexor Oy	✓		Queens College, CUNY	✓	✓
Dipart. di Informatica, Pisa	✓	✓	Seoul National Univ.	✓	✓
Fudan University	✓	✓	Southern Methodist Univ.	✓	✓
IBM (Hawthorne)	✓	✓	Sun Microsystems		✓
IBM (Yorktown Heights)	✓	✓	Syracuse Univ., CNLP	✓	✓
Imperial College	✓	✓	Univ. of Alberta	✓	
KAIST	✓	✓	Univ. of Iowa		✓
Korea University	✓	✓	Univ. of Massachusetts		✓
LIMSI	✓	✓	Univ. de Montreal	✓	✓
Microsoft	✓	✓	Univ. of Sheffield	✓	✓
MITRE	✓	✓	USC, ISI	✓	
Multitext Project	✓	✓	Xerox Research Centre	✓	✓

Text REtrieval Conference (TREC)

Working material: Document collection

- various set of news articles (1999-2001)
  - Wall Street Journal, Los Angeles Times...etc.
- AQUAINT Corpus of English News Text (2002-2007)
  - approx. 3 GB of text
    - New York Times News Service, Associated Press Worldstream News Service and Xinhua News Service
- AQUAINT-2 Corpus of English News Text (2007)
  - approx. 2,5 GB of text
    - Agence France Press, Associated Press, Central News Agency English Service, Los Angeles Times-Washington Post News Service, New York Times and Xinhua News Agency
- Blog06 corpus (2007)
  - Polling 100,649 RSS and Atom feeds

AQUAINT-2



```
<DOC id="AFP_ENG_20041001.0002" type="story">
 <HEADLINE> China's West to East pipeline carries first gas to Shanghai </HEADLINE>
 <DATELINE> SHANGHAI, Oct 1 (AFP) </DATELINE>
 <TEXT>
  <P> The eastern economic hub of Shanghai received its first deliveries of gas from China's newly
   opened East to West pipeline that runs some 4,000 kilometres (2,400 miles) from the Xinjiang
   Uighur autonomous region, state press reported Friday. </P>
  <P> "What counts most is that the advent of gas symbolises the successful operation of the whole
   project," the Shanghai Daily quoted Zhao Yongxin, a PetroChina employee as saying. </P>
 </TEXT>
</DOC>
```

Working material: Ranked lists

- As a convenience for QA track participants:
  - available rankings of the top 200 (1000) documents per questions (targets)
    - question (target) as the query
    - PRISE document retrieval system
  - 1999-2007
  - Top 5 (AQUAINT-2, 2007):

216	1	NYT_ENG_20041029.0109	1.0
216	2	NYT_ENG_20041108.0120	0.875
216	3	NYT_ENG_20050825.0063	0.717103
216	4	NYT_ENG_20050909.0014	0.717103
216	5	NYT_ENG_20041130.0132	0.70710677

Working material: Questions

- Sources:
  - Participants, NIST (1999)
  - MSNSearch logs,
  - AskJeeves logs
  - AOL logs

→ Raw logs were automatically filtered:

contained a question word (where, when, which, etc.)

began with modals or the verb to be (are, can, define, etc:)

or ended with a question mark

→ NIST fixed: spelling, punctuation, grammar

Working material: Questions

- each question has
  - an guaranteed answer (1999-2000)
  - no guarantee to find an answer
    in the document collection (2001-2007)

- different types:
  - factoid, list, definition, Other

- XML format

Factoid questions

- fact-based
- short answer

<top>

<num> Number: 1406

<desc> Description:

When did the story of Romeo and Juliet take place?

</top>

List questions

- asking the same factoid question multiple times
- assemble a set of instances
  - information located in multiple documents
    - multiple instances in a single document
    - same instance repeated in multiple documents
  - specified number of instances (2002)
    - List 13 countries that export lobster.
  - no target number of instances (2003-2007)
    - List the names of chewing gums.

List questions

- TREC 2002:

  <top>

  <num> Number: 34

  <desc> Description:

  List 13 countries that export lobster.

  </top>

- TREC 2003:

  <top>

  <num> Number: 1915

  <desc> Description:

  List the names of chewing gums.

  </top>

Definition questions

- Have a target:
  - Persons
    - Jon Bon Jovi, Jay-Z (2007, No.: 271, 217)
  - Organisations
    - Ben & Jerry's (2006, No.: 172, 160)
  - Things
    - Avocados, Australian wine (2006, No.: 188; 2007, No.: 279)
  - Events
    - 1999 Sundance Film Festival (2006, No.: 215)

- Task scenario of the questioner:
  - adult, native speaker, “average” reader of US newspaper
  - looking for more information

Question series

- Questions for the Track:

  set of question series (2004 - 2007)

  - questions grouped into a series
    - factoid, list and a final other question

  - each series has a target

    → **target is**: person, organization, thing or event (2005)

→ abstraction of an information dialog

Question series

```
<target id="11" text="the band Nirvana">
    <qa>
        <q id="11.1" type="FACTOID"> Who is the lead singer/musician in
        Nirvana?</q>
    </qa>
    <qa>
        <q id="11.2" type="LIST"> Who are the band members?</q>
    </qa>
    <qa>
        <q id="11.4" type="FACTOID"> What is their biggest hit?</q>
    </qa>
    <qa>
        <q id="11.7" type="OTHER"> Other</q>
    </qa>
</target>
```

The “Other” question

- asked for additional information about a target
  - not covered by previous questions
- “Tell me other interesting things about this target I don't know enough to ask directly.”
- no explicit question

Responses

- [document-id, answer-string] or the string 'NIL'

- document-id (doc-id):
  - document that supports the answer
- answer-string:
  - 1999-2001
    - small snippets of text that contain the answer (50, 250 bytes)
    - ranked list of up to five pairs per question
  - 2002-2007
    - exact answers, one pair per question
- 'NIL' (2001)
  - there is no correct answer in the collection

Set of judgements

- factoid, list and definition questions:
  - Incorrect (1999)
    - answer is wrong
  - unsupported (2000)
    - right answer, but not supported by the document
  - not exact (2002)
    - right answer, supported
    - string contains more than just the answer, or is missing bits
  - locally correct (2006)
    - right answer (regarding supporting document)
  - globally correct (2006)
    - right answer (regarding the whole document collection)

Responses TREC 2001/2002

- What river in the US is known as the Big Muddy?
- 2001: correct answer strings (50 bytes)
  - the Mississippi
  - known as Big Muddy, the Mississippi is the longest
  - southeast;Mississippi;Mark Twain;officials began
- 2002: correct, exact answers:
  - mississippi,
  - The Mississippi River
- 2002: not exact
  - 2,348 miles; Mississippi
  - Missipp

History: TREC-8 and TREC-9

- factoid questions
  - guarantee: answer is contained in the collection
- response
  - up to five [document-id, answer-string] pairs per question
  - small snippets of text (50, or 250 bytes)

- Challenges for QA systems:
  - find the correct answer in a document
  - limited to 50, or 250 bytes

  Answered questions:

    - TREC-8: 70%; TREC-9: 65%

History: TREC 2001, 2002 and 2003

- factoid, list and definition (2003) questions
  - no guarantee to find an answer for factoid questions
- response (2002):
  - only one response per question,
  - [document-id, answer-string] or the string 'NIL'
  - exact answers instead of text snippets

- Challenges for QA systems
  - recognizing that no answer exists
  - assemble an answer from information located in multiple documents, and detect duplicates
  - returning an exact answer

Questions TREC 2003

```
<top>
<num> Number: 1900
<type> Type: factoid
<desc> Description:
What country is Aswan High
Dam located in?
</top>
```

```
<top>
<num> Number: 1901
<type> Type: definition
<desc> Description:
Who is Aaron Copland?
</top>
```

```
<top>
<num> Number: 1902
<type> Type: list
<desc> Description:
Which past and present NFL
players have the last name of
Johnson?
</top>
```

History: TREC 2004, 2005 and 2006

- set of question series
- time frame: date of the last document in the collection (2006)

- Challenges for QA systems:
  - Other questions: don't repeat information already covered by earlier questions
  - give the most up-to-date answer
  - process series independently from another
  - process individual series in question order

History: TREC 2007

- additional Blog06 corpus

- Challenges for QA systems:
  - handle language that is not well-formed
  - discourse structures that are more informal and less reliable than newswire

Sources

- http://trec.nist.gov/
- http://start.csail.mit.edu/index.php
- http://www.yale.edu/lawweb/lawcrs/arc9798/ftext1.jpg
- http://catalog.ldc.upenn.edu/desc/addenda/LDC2008T25.jpg

Thank you for listening!

General strategy

1.) determine the expected answer type

- e.g. by question word

2.) receive documents likely to contain answers to the question

- using important question words and related terms as the query

3.) perform a match between the question words and receives documents/passages to extract the answer

Some results

- 250 bytes in a response was easier than limiting responses to 50 bytes

- definition questions were harder to answer than factoid questions

- Other questions were harder to answer than definition questions

- EVENTs (target) were harder to answer than PERSONs

Evaluation TREC 2007

- Factoid questions (# = number of times)
  - accuracy: fraction of questions judged to be globally correct
  - NIL precision: (# NIL returned and correct)/ (# NIL returned)
  - NIL recall: (# NIL returned and correct)/ (# NIL correct in the set (17))
- List questions
  - Instance precision (IP) = (# globally correct)/(# total of responses)
  - Instance recall (IR) = (# globally correct)/(# instances)
  - F-score = (2*IP*IR)/(IP+IR)
- Other questions
  - Information nuggets = vital (good info) or non-vital (don't care)
- Per-series Combined Weighted Score
  - WeightedScore = 1/3*Factoid + 1/3*List + 1/3*Other