CENTRE FOR LANGUAGE TECHNOLOGY

MACQUARIE UNIVERSITY - SYDNEY

AnswerFinder in TREC-QA 2003
How did it Go?

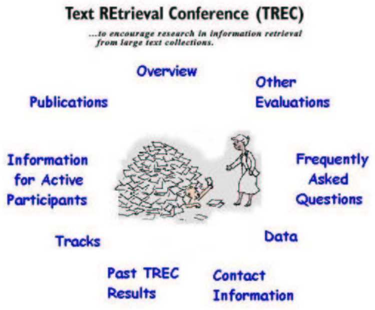
Diego Mollá-Aliod

13 October 2003

Outline

- TREC-QA 2003
- AnswerFinder

Text REtrieval Conference



Text REtrieval Conference (TREC)

...to encourage research in information retrieval from large text collections.

http://trec.nist.gov/

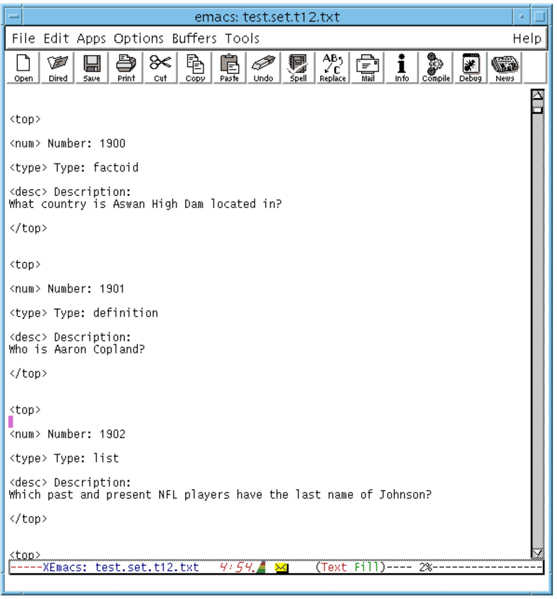
- NIST = National Institute of Standards and Technology
- DARPA = Defense Advanced Research Projects Agency
- ARDA = Advanced Research and Development Activity

TREC-QA 2003 Specifications

- 500 questions in total
- All runs must be fully automatic
- The system cannot be modified after the training phase
- Each question must be processed from the same machine state
- Three different types of questions
  - factoid questions
  - list questions (25-50 questions)
  - definition questions (25-50 questions)
- Different evaluation criterion for each question type
  - All evaluations combined into a unique score

final = 1/2*factoid-score + 1/4*definition-score + 1/4*list-score

Sample of Questions



emacs: test.set.t12.txt
File Edit Apps Options Buffers Tools Help
Open Dired Save Print Cut Copy Paste Undo Spell Replace Mail Info Compile Debug News

```
<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>


<top>
-----XEmacs: test.set.t12.txt     4:54.      (Text Fill)---- 2%----------------
```

Evaluation of Factoid Questions

- The system to return one answer only (or "nil") per question and a supporting document:

  1395000 exampleRun NYT19990326.0303 Nicole Kidman
- Each question/answer pair assessed by one (or several?) judge(s):
  - incorrect: the answer-string does not contain a correct answer or the answer is not responsive;
  - unsupported: the answer-string contains a correct answer but the document returned does not support that answer;
  - non-exact: the answer-string contains a correct answer and the document supports that answer, but the string contains more than just the answer (or is missing bits of the answer);
  - correct: the answer-string consists of exactly a correct answer and that answer is supported by the document returned.
- The final score is the fraction of questions judged correct

Evaluation of Factoid Questions

- The judges will decide if the answer is correct and exact
- *What is the longest river in the United States?*
  - Correct and exact answers:
    - *Mississippi*
    - *the Mississippi*
    - *the Mississippi River*
    - *Mississippi River*
    - *mississippi*
  - Incorrect or inexact answers:
    - *At 2,348 miles the Mississippi River is the longest river in the US.*
    - *2,348 miles; Mississippi*
    - *the river Mississippi*
    - *Missipp*
    - *Missouri*

Evaluation of List Questions

- The question does not indicate the target number of answers
- The response to a list question is a non-null, unordered, and unbounded set of [answer-string, docid] pairs
- An individual instance is interpreted as for factoid questions and will be judged in the same way
- The final answer set for a list question will be created from the union of the distinct, correct responses returned by all participants plus the set of answers found by the NIST assessor during question development

  IR = # instances judged correct & distinct/ | final answer set |

  IP = # instances judged correct & distinct/# instances returned

  F = (2*IP*IR)/(IP+IR)
- The final score is the mean of the F scores

Evaluation of Definition Questions

- The response is like that of a list question, but the evaluation is different
- For each definition question, the assessor will create a list of acceptable information nuggets about the target from the union of the returned responses and the information discovered during question development.
- Two types of information nuggets
  - Essential information
  - Acceptable information

NR = # essential nuggets returned in response/# essential nuggets

NP is defined using

allowance = 100*(# essential+acceptable nuggets returned)

length = total # non-white-space characters in answer strings

NP = 1 if length < allowance
else 1-[(length-allowance)/length]

F = (26*NP*NR)/(25*NP + NR)

- The final score is the mean of the F scores

The Passage Task

- A simplification of the main task
- Only factoid questions are used
  - 413 questions in TREC 2003
- The system returns a passage of length 250 characters or less containing the answer

TREC-QA 2003 Corpus

- Data: AQUAINT corpus
  - Newswire text in English
    - Xinhua News Service (People's Republic of China)
    - New York Times News Service
    - Associated Press Worldstream News Service
  - Over 3 Gb when uncompressed
    - 1,033,461 documents
  - SGML tagged

Corpus – Sample

```
<DOC>
<DOCNO> NYT19980601.0001 </DOCNO>
<DOCTYPE> NEWS STORY </DOCTYPE>
<DATE_TIME> 1998-06-01 00:02 </DATE_TIME>
<HEADER>
A7753 &Cx1f; taf-z
u a &Cx13;  &Cx11;  BC-OBIT-LENIHAN-NYT &LR;      06-01 0290
</HEADER>
<BODY>
<SLUG> BC-OBIT-LENIHAN-NYT </SLUG>
<HEADLINE>
KENNETH J. LENIHAN, SOCIOLOGIST WHO STUDIED CAUSES OF
     RECIDIVISM,
DIES AT 69
</HEADLINE>
  (sw)
 By WOLFGANG SAXON
 c.1998 N.Y. Times News Service
<TEXT>
<P>
  NEW YORK _ Kenneth Joseph Lenihan, a New York research
sociologist who helped refine the scientific methods used in
criminology, died May 25 at his home in Manhattan. He was 69.
</P>
<P>
  The cause was a heart attack, his family said.
</P>
<P>
  Lenihan retired in 1995 as an associate professor of sociology
at John Jay College of Criminal Justice. He had joined the faculty
in 1980, after earlier stints as a researcher at Columbia
University's Bureau of Applied Social Research, the Vera Institute
of Justice in New York and the Bureau of Social Science Research in
Washington.
</P>
<P>
  He brought his expertise to the study of recidivism rates among
criminal offenders. He conducted a study in Baltimore, called the
Life Project, for the U.S. Department of Labor in the early 1970s.
</P>
```

Corpus – Sample

```
<P>
  A large research project, it measured whether and how giving
jobs or money to recently released offenders would affect the
chances of their becoming repeaters. That project and further
studies formed the basis of a standard work in the field, which he
wrote with P. Rossi and D. Berk, ``Money, Work and Crime''
(Academic Press, 1980).
</P>
<P>
  Lenihan was born in Queens, and graduated from Columbia's
     School
of General Studies in 1960. He also earned his M.A. and Ph.D. in
sociology at Columbia, the latter in 1974.
</P>
<P>
  Lenihan is survived by two sons, Andrew of Miami, and William of
Manhattan; a daughter, Jean Lenihan of Seattle; four sisters,
Eileen McEwan of Houston, Moira Earhart of North Carolina, Jean
Dobson of Bay Shore, N.Y., and Sue Adams of Cape May, N.J.; and
three grandchildren.
</P>
</TEXT>
</BODY>
<TRAILER>
NYT-06-01-98 0002EDT &QL;
</TRAILER>
</DOC>
```

Tags Used in the Corpus

- apw
  - (DOC (DOCNO) (DOCTYPE*) (DATE_TIME*) (HEADER) (BODY (SLUG) (HEADLINE*) (TEXT (SUBHEAD*))) (TRAILER))
- nyt/1998:
  - (DOC (DOCNO) (DOCTYPE) (DATE_TIME) (HEADER) (BODY (SLUG*) (HEADLINE*) (TEXT (ANNOTATION*))) (TRAILER))
- xie/* (all years):
  - (DOC (DOCNO) (DATE_TIME) (BODY (HEADLINE) (TEXT)))

Deadlines

- The deadlines for 2003 are:
  - August 4: Testing questions and top ranked documents available
    - System to be frozen before downloading the questions
    - Possible to arrange a different (earlier) time
  - August 11: Submission of results to NIST
    - One week of time
  - October 1: Evaluated results from NIST
  - October 29: notebook papers sent to NIST
  - November 18-21: Conference
  - ... : Final papers due

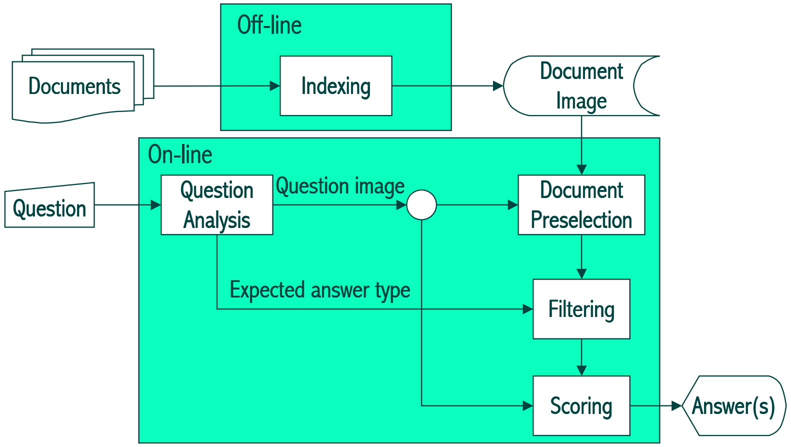
Outline

- TREC-QA 2003
- AnswerFinder

The Proposed System in a Nutshell

- Limited time and resources available to build the system
  - Only a few hours per week during about 3 months
    - Eventually the system was developed in about 55 hours + machine execution time
- Participate in the passage task only
- Implement a simple and functional system
- Use third-party modules whenever possible
- Experiment on overlap scores
  - dependencies, GRs, MLFs, …

Architecture of our QA System



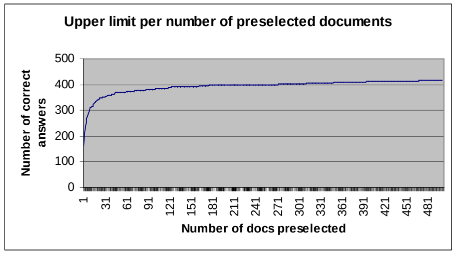
Document Indexing

- Source:
  - 1,033,461 documents
  - Over 3Gb of data
- No time nor resources to build complex document images
  - e.g. finding the named entities of all documents may have taken months of run time
- NIST provides the top 1000 documents of each question
- Decision: No document indexing

Document Preselection

- NIST provides the 1000 top-ranking documents, ordered by relevance
  - IR engine used: PRISE
- Still, do we need to use all the 1000 documents per question?
- Based on TREC 2002 data...

# Documents	Upper Limit
1	31.1%
5	53.8%
10	62%
20	68.2%
50	74%
100	76.4%
1000	83%

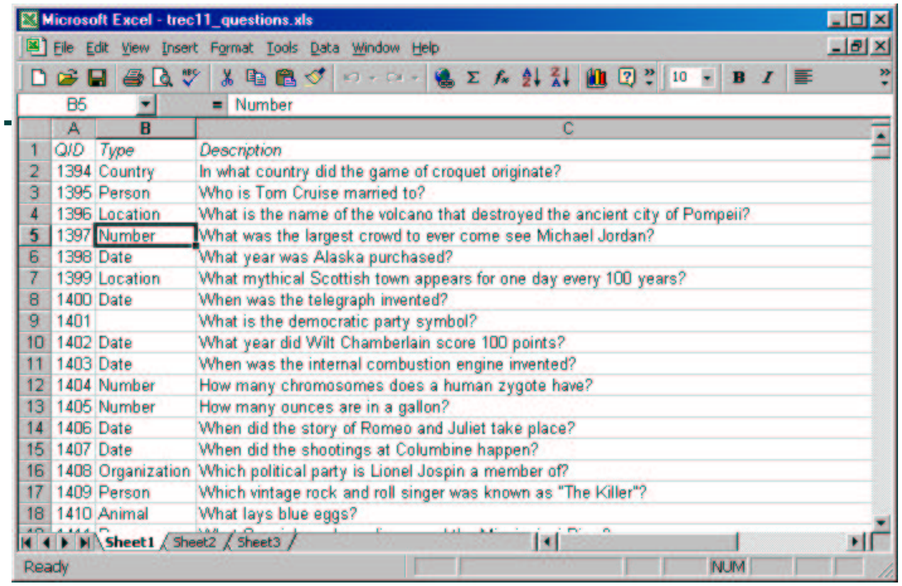


Filtering

- Proposal
  - Use a third-party NE recogniser
  - Develop a simple question classifier
  - Determine the answer type for every type of question
  - Give a high score to sentences with compatible entities
    - we want to find likely answers even if there are no sentences with compatible entities

Question Classifier

- Very simple set of regular expressions
  - 29 rules
  - Based on TREC2002 data
- Question types:
  - person, date, location, money, number, city, date, organization, location, percent,country, state, river, name, unknown
- Accuracy: 393 out of 500 questions (TREC2002 data)
  - not tested with new data


Microsoft Excel - trec11_questions.xls
File Edit View Insert Format Tools Data Window Help
10
B5
= Number
Sheet1
Sheet2
Sheet3
Ready
NUM

	A	B	C
1	QID	Type	Description
2	1394	Country	In what country did the game of croquet originate?
3	1395	Person	Who is Tom Cruise married to?
4	1396	Location	What is the name of the volcano that destroyed the ancient city of Pompeii?
5	1397	Number	What was the largest crowd to ever come see Michael Jordan?
6	1398	Date	What year was Alaska purchased?
7	1399	Location	What mythical Scottish town appears for one day every 100 years?
8	1400	Date	When was the telegraph invented?
9	1401		What is the democratic party symbol?
10	1402	Date	What year did Wilt Chamberlain score 100 points?
11	1403	Date	When was the internal combustion engine invented?
12	1404	Number	How many chromosomes does a human zygote have?
13	1405	Number	How many ounces are in a gallon?
14	1406	Date	When did the story of Romeo and Juliet take place?
15	1407	Date	When did the shootings at Columbine happen?
16	1408	Organization	Which political party is Lionel Jospin a member of?
17	1409	Person	Which vintage rock and roll singer was known as "The Killer"?
18	1410	Animal	What lays blue eggs?

Named Entity Recogniser – GATE

- Entities recognised:
  - person, location, date, money, organization
- Mapping of unknown entities:
  - country, city, state, river → location
  - percent → number
  - name → person OR organization OR location
  - any other question type yields same answer type
- Java API
  - Java programming: steep learning curve
  - Initial attempts failed
    - the program would crash unexpectedly
- Decision: do not use named entities
  - question classifier wasn't used either

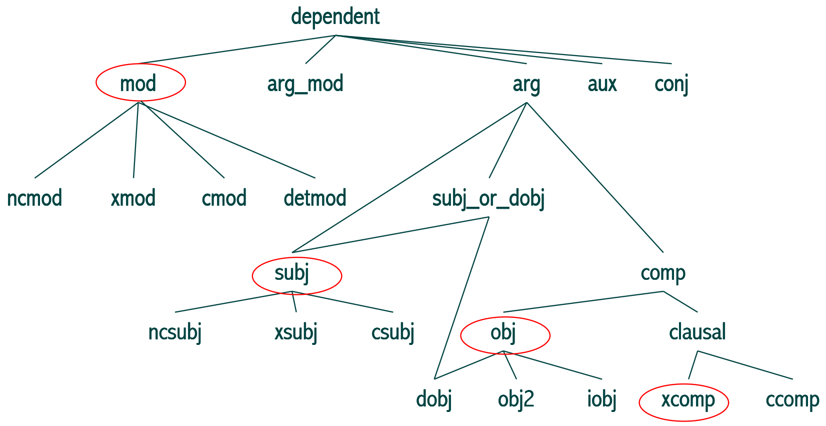
Filtering Revisited

- Split the text into sentences
  - Leading blanks are removed:
    '\s*'
  - Sentence endings determined by punctuation marks and XML tags:
    '(?:\.|\?|!|;|<.*?>)+'
- Rank sentences according to word overlap
  - Word forms
  - Stop words removed
    http://www-fog.bio.unipd.it/waishelp/stoplist.html
  - Repeated words in the answer do not count
- Return the N top-scoring sentences
  - N = 100

Scoring

- Simplest score: Word Overlap
  - Use N=1 in the filtering module
  - Result: 14.8%
- Other possible scores:
  - Grammatical Relations
  - Minimal Logical Forms

Grammatical Relations



Grammatical Relations

- *The man that came ate bananas and apples with a fork.*

  (detmod _ man the) (cmod that man come) (ncsubj come man _) (ncsubj eat man _) (dobj eat banana _) (dobj eat apple _) (conj and banana apple) (ncmod fork eat with) (detmod _ fork a)

- Same example with the selected gramrels

  (mod that man come) (subj come man _) (subj eat man _) (obj eat banana _) (obj eat apple _) (mod fork eat with)

Grammatical Relations

- *Failure to do this will continue to place a disproportionate burden on Fulton taxpayers.*

  (xcomp to failure do) (dobj do this _) (ncsubj continue failure _) (xcomp to continue place) (ncsubj place failure _) (dobj place burden _) (ncmod _ burden disproportionate) (iobj on place tax-payer) (ncmod _ tax-payer Fulton) (detmod _ burden a) (aux _ continue will)
- Same example with the selected gramrels

  (xcomp to failure do) (obj do this _) (subj continue failure _) (xcomp to continue place) (subj place failure _) (obj place burden _) (mod _ burden disproportionate) (obj on place tax-payer) (mod _ tax-payer Fulton)

Grammatical Relations

– *A man named Richard Sears has been playing a joke on shoppers.*
(xmod _ man name) (detmod _ man a)
(subj name man) (dobj name richard _)
(detmod _ joke a) (subj sear man _) (subj play sear _)
(aux _ play have) (aux _ play be) (ncmod _ play on)
(xcomp _ play joke)

– *Who played a joke on shoppers?*
(subj play who _) (dobj play joke _) (ncmod _ play on)
(detmod _ joke a)

Minimal Logical Forms

- Called Minimal Logical forms because they encode the minimum information required for AE
- Flat expressions that use reification
- Example: *cp will quickly copy files*

  holds(e4), object(cp,o1,[x1]), object(s_command,o2,[x1]), evt(s_copy,e4,[x1,x6]), object(s_file,o3,[x6]), prop(quickly,p3,[e4]).
- Example: *the man that came ate bananas and apples with a fork*

  holds(e1), object(s_man,o2,[x2]), evt(s_come,e4,[x2]), evt(s_eat,e5,[x7]), e6@<e7, e8@<e7, evt(s_eat,e5_1,[x6]), evt(s_eat,e5_2,[x8]), object(s_banana,o6,[x6]), object(s_apple,o8,[x8]), prop(with,p9,[e6]), object(s_fork,o11,[x11]).

Minimal Logical Forms

– *A man named Richard Sears has been playing a joke on shoppers.*
holds(v_o10), object('man',v_o2,[v_x2]),
evt('name',v_e3,[v_X3,v_x4,v_x2]),
object('joke_on',v_o10,[v_e5,v_x12]),
object('richard',v_o4,[v_x4]), evt('sear',v_e5,[v_x2]),
object('shopper',v_o12,[v_x12])

– *Who played a joke on shoppers?*
holds(v_e2), object('who',v_o1,[v_x1]),
evt('play_on',v_e2,[v_x1,v_x4,v_x6]),
object('joke',v_o4,[v_x4]), object('shopper',v_o6,[v_x6])

Results

- Basic Scores
  - Word Overlap: 14.8%
  - Grammical Relation Overlap: 9%
  - Minimal Logical Form Overlap: 10.8%
- Combination of Scores
  - 3*mo + wo: 13%
  - 3*wo + mo: 15.8%
  - 3*wo + gro: 16.2% (16.8%?) ← answfind1
  - 9*wo + 3*gro + mo: 16.2% (16.8%?) ← answfind2
  - 9*wo + 3*mo + gro: 15.6% ← answfind3

Results from NIST

Run	*Formula*	*2002*	*2003*
answfind1	3wo+gro	16.8%	19.1%
answfind2	9wo+3gro+mo	16.8%	18.6%
answfind3	9wo+3mo+gro	15.6%	18.2%

- Comparison with other runs (21 runs in total, including ours):
  - Best: 0.685
  - Median: 0.182
  - Worse: 0.085

Epilogue: What about the Named Entities?

- Managed to extract the NEs of the preselected documents (TREC QA 2002)
- Results:

Run	*Formula*	*Without*	*With NE*
answfind1	3wo+gro	16.8%	19.1%
answfind2	9wo+3gro+mo	16.8%	19.3%
answfind3	9wo+3mo+gro	15.6%	18.4%
	mo	10.8%	13%
	wo	14.8%	16.4%
	gro	9%	12.4%

What's Next?

- Error analysis
- Finalise NE integration
  - Use NEs for parsing and semantic interpretation
- Convert MLF overlap into a process of abduction
- Extract the exact answer
- List questions
- Definition questions
- ...