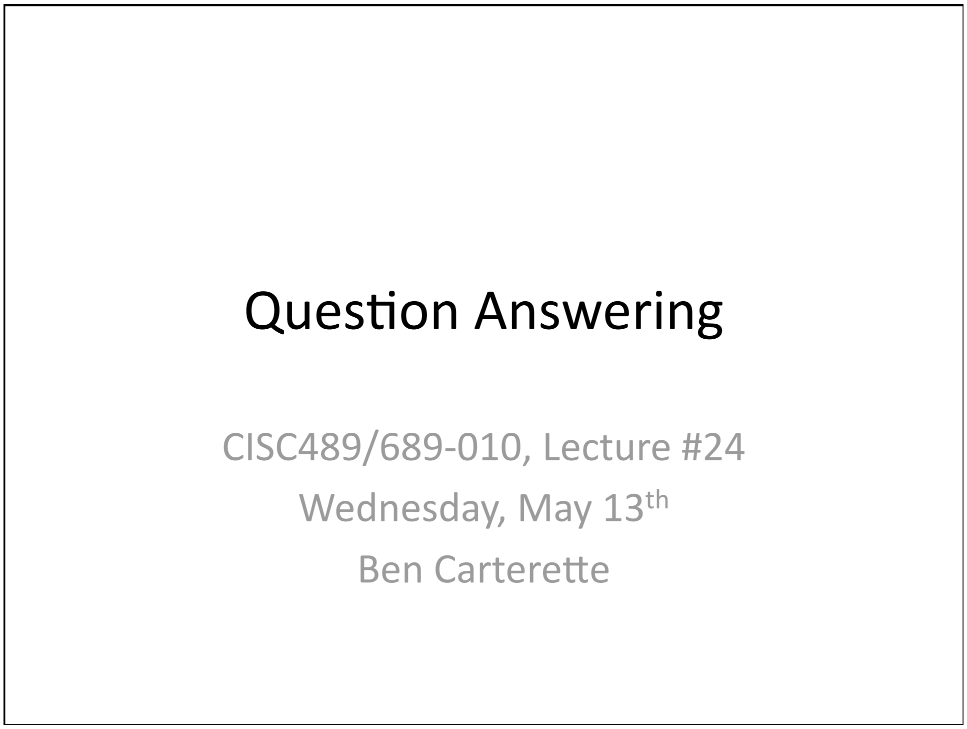
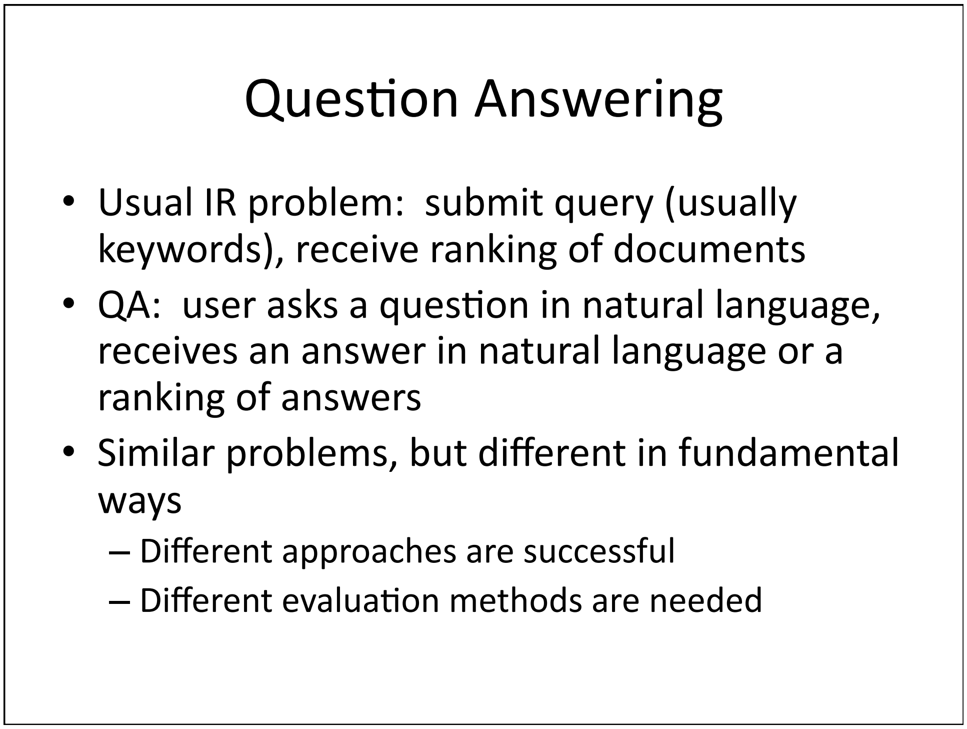


Question Answering

CISC489/689-010, Lecture #24

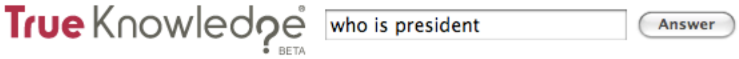
Wednesday, May 13th

Ben Carterette



Question Answering

- Usual IR problem: submit query (usually keywords), receive ranking of documents
- QA: user asks a question in natural language, receives an answer in natural language or a ranking of answers
- Similar problems, but different in fundamental ways
  - Different approaches are successful
  - Different evaluation methods are needed



who is president

Knowledge-Based Systems

- Create a database of known facts
- Reformulate question as a statement with a “blank”
  - Question: “who is president of the U.S.?”
  - Reformulation: “The president of the U.S. is ____”
- Find facts in database that match the statement
  - Fact in database: “The president of the U.S. is Barack Obama”
- If none match, find intermediate facts
  - For example, “The president of the U.S. is the former junior senator from Illinois”, “Barack Obama is the former junior senator from Illinois” → “The president of the U.S. is Barack Obama”

Reformulation

- Different reformulations may be possible depending on query type
  - “who” question: “___ is the president of the U.S.”, “The president of the U.S. is ___”, “___, who was born in ___, has been elected president of the U.S.”, …
  - “when” question: “Barack Obama took office in ___”, “Barack Obama was president of the U.S. from ___”, …
- Therefore it is useful to classify question by type
  - Who/what/where/when/how
  - Use type to generate reformulation patterns

Question Classification

- What are some features we could use for this?
  - Does the question contain “who”/”what”/”where”/”when”?
  - Entity types: person names might imply “who”, place names might imply “where”, etc
- This is not always as easy as it might seem
  - “Name the first private citizen to fly in space” – what features could we use to determine that is a “who” question?

Query Expansion

- Exact reformulation may not exist in database
  - Try different reformulations by query expansion/ rewriting
  - "The president of the U.S." → "The president of the United States"; "The president of the USA"; "President of America"; "American head of state"; …
- As usual, query expansion is noisy
- It is not always possible to find the right terms to expand with

Statistical Systems

- Knowledge-based systems have a lot of shortcomings
  - They require people to add facts to the database
  - Those facts have to be verified
  - Automatically matching a question to known facts is not easy
  - Identifying when a question does not match any fact is not easy
  - And figuring out that an answer can be deduced from other facts is definitely not easy
- Instead of relying on known facts, use large document corpora and text statistics to find likely answers

How To Do It?

- Locate documents that might contain answer
  - How?
- Locate parts of those documents that are most likely to contain answer
  - How?
- Reformulate those parts into natural-language answers
- Remove answers that seem to be duplicates

Document Retrieval for QA

- This is a task for which it might be useful to have entities tagged
- *Named entity recognition*: NLP task for finding elements in text and tagging them as belonging to predefined categories
- For example, "Jim bought 300 shares of Acme Corp. in 2006" might become:
  - <ENAMEX TYPE="PERSON">Jim</ENAMEX> bought <NUMEX TYPE="QUANTITY">300</NUMEX> shares of <ENAMEX TYPE="ORGANIZATION">Acme Corp.</ENAMEX> in <TIMEX TYPE="DATE">2006</TIMEX>.
- Given such output, we can index the content of these tags using the same methods used for indexing title words, etc

Document Retrieval with NEs

- With a named-entity-tagged corpus, we can transform the question into a query based on the question classification
- A “who” query like “who is the president of the U.S.” might become something like
  - #and(president U.S. #person.any)
  - Where #person.any tells the engine to match any document containing something tagged as a “person”
- Query expansion could be applied naturally using top-retrieved documents

Passage Retrieval

- QA systems often depend on retrieving short pieces of documents rather than full documents
- This is called *passage retrieval*
- Examples of passages: 50-word windows, 250-word windows, sentences, paragraphs, etc.
- Fixed- or variable-length passages can easily be retrieved if term positions have been indexed
- Sentences and paragraphs can be tagged, and tag information can be indexed just like entity or markup tags

Passage Reformulation

- It is possible to “learn” how to reformulate passages
- Idea: use training data (questions with known answers) to find the passages that contain the answers
- From those passages, learn patterns for the question type
- New questions can then be answered by finding passages that match the learned patterns and pulling the answers out

Example

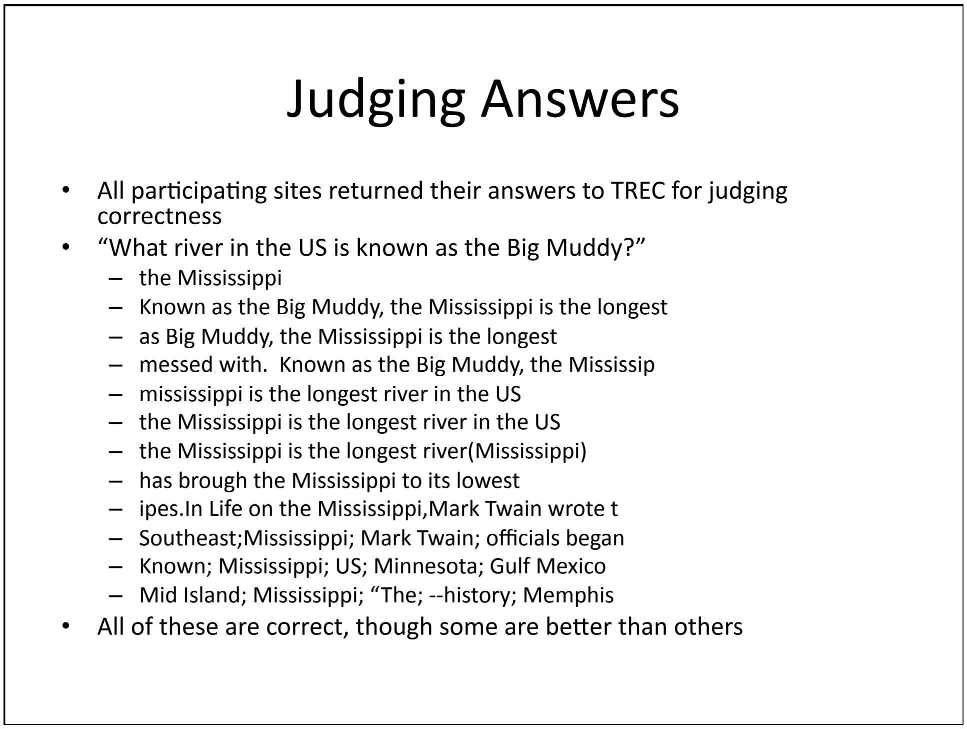
- “When was Bill Clinton elected president?” – 1992
- Passages that match the question and answer:
  - Bill Clinton was elected president in 1992
  - The election was won by Bill Clinton in 1992
  - Clinton defeated Bush in 1992
  - Clinton won the electoral college in 1992
- Take the most common of these and turn them into general patterns
  - #person was elected president in #year
  - The election was won by #person in #year
  - #person defeated Bush in #year (how useful is this?)
- Then new questions can be answered by finding passages that match the pattern
- “When was Barack Obama elected president?”

QA Experiments

- TREC ran a QA track from 1999 through 2003
- The track has changed a lot over time:
  - Question types, evaluation, document corpus
- The first track used factual questions that definitely had answers in a collection of news articles
  - Subsequent tracks have included definition questions and list questions, and not all questions have answers in the document set
- Systems answer questions with short text passages
  - 50 or 250 bytes that answer the question and support the answer
- Evaluation based on reciprocal rank of first correct answer

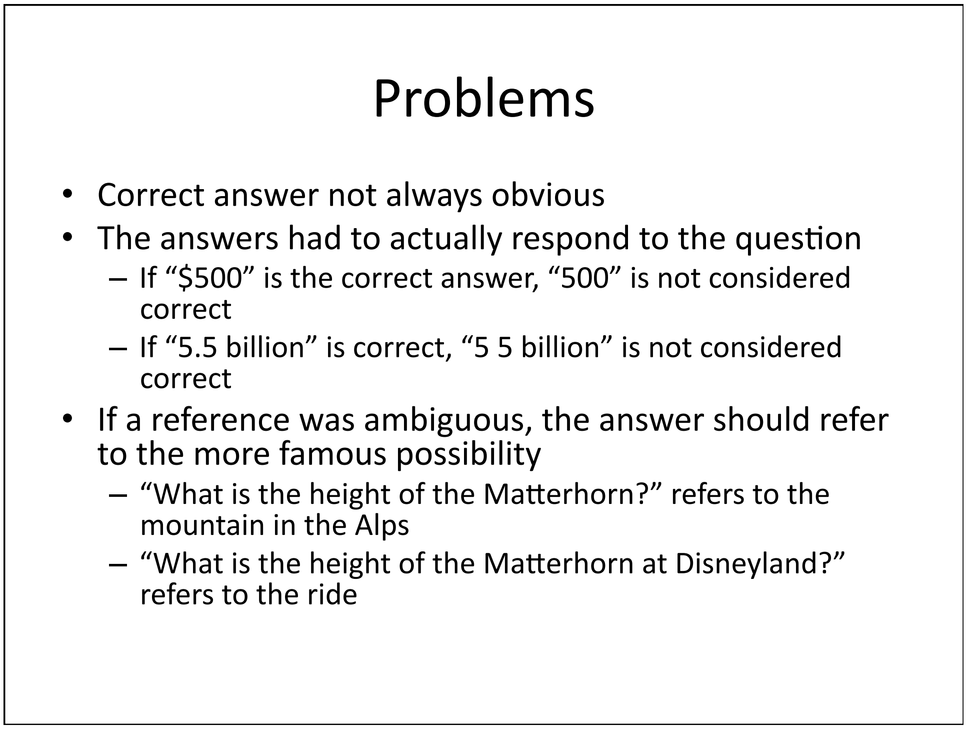
QA at TREC in 1999

- 200 questions given to participants
  - How many calories are there in a Big Mac?
  - What two US biochemists won the Novel Prize in medicine in 1992?
  - Who is the voice of Miss Piggy?
  - What river in the US is known as the Big Muddy?
  - Who is the 16th president of the United States?
  - Name a film in which Jude Law acted.
- Participants must use 528,000 news documents to come up with 50 or 250 bytes to answer



Judging Answers

- All participating sites returned their answers to TREC for judging correctness
- “What river in the US is known as the Big Muddy?”
  - the Mississippi
  - Known as the Big Muddy, the Mississippi is the longest
  - as Big Muddy, the Mississippi is the longest
  - messed with. Known as the Big Muddy, the Mississip
  - mississippi is the longest river in the US
  - the Mississippi is the longest river in the US
  - the Mississippi is the longest river(Mississippi)
  - has brough the Mississippi to its lowest
  - ipes.In Life on the Mississippi,Mark Twain wrote t
  - Southeast;Mississippi; Mark Twain; officials began
  - Known; Mississippi; US; Minnesota; Gulf Mexico
  - Mid Island; Mississippi; “The; --history; Memphis
- All of these are correct, though some are better than others



Problems

- Correct answer not always obvious
- The answers had to actually respond to the question
  - If “$500” is the correct answer, “500” is not considered correct
  - If “5.5 billion” is correct, “5 5 billion” is not considered correct
- If a reference was ambiguous, the answer should refer to the more famous possibility
  - “What is the height of the Matterhorn?” refers to the mountain in the Alps
  - “What is the height of the Matterhorn at Disneyland?” refers to the ride

Problems

- "Answer stuffing"
  - Participants could include a bunch of different words in the answer hoping that one is correct
  - "Known; Mississippi; US; Minnesota; Gulf Mexico"
  - Clearly not very useful to a user, so not considered correct
- No justification for answer
  - "Who was the 16th President of the US?"
  - Right answer: Abraham Lincoln
  - A document about the Gettysburg Address that only contained Lincoln's name, but no mention that he was president, would be considered correct

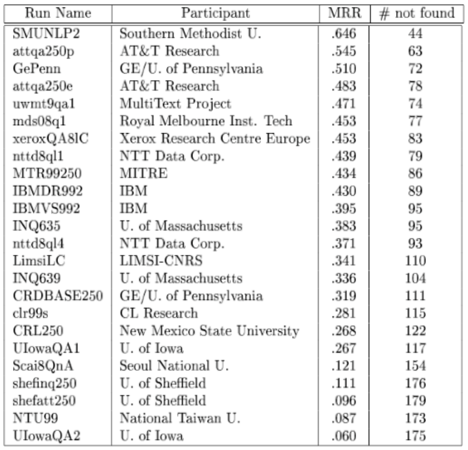
QA at TREC in 2000

- Change from 1999 to 2000:
  - More documents: 979,000 from 528,000
  - More questions: 693 = 500 + 193 variants
  - "Real" questions: 500 drawn from query logs and are more difficult than the 1999 "fake" questions
  - Stricter judging: answers had to include some justification
    - The "Abraham Lincoln" example no longer considered correct

Question Variants

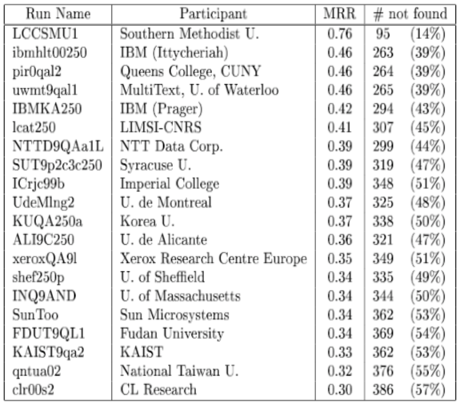
- 193 of the questions were variants of the first 500
- Restatements to test whether systems were robust to different ways of expressing the same information need
  - “What is the tallest mountain?”
  - “What is the world’s highest peak?”
  - “What is the tallest mountain in the world?”
  - “Name the highest mountain.”
  - “What is the name of the tallest mountain in the world?”

QA Results (1999 and 2000)



Run Name	Participant	MRR	# not found
SMUNLP2	Southern Methodist U.	.646	44
attqa250p	AT&T Research	.545	63
GePenn	GE/U. of Pennsylvania	.510	72
attqa250e	AT&T Research	.483	78
uwmt9qa1	MultiText Project	.471	74
mds08q1	Royal Melbourne Inst. Tech	.453	77
xeroxQA8lC	Xerox Research Centre Europe	.453	83
nttd8ql1	NTT Data Corp.	.439	79
MTR99250	MITRE	.434	86
IBMDR992	IBM	.430	89
IBMVS992	IBM	.395	95
INQ635	U. of Massachusetts	.383	95
nttd8ql4	NTT Data Corp.	.371	93
LimsiLC	LIMSI-CNRS	.341	110
INQ639	U. of Massachusetts	.336	104
CRDBASE250	GE/U. of Pennsylvania	.319	111
clr99s	CL Research	.281	115
CRL250	New Mexico State University	.268	122
UIowaQA1	U. of Iowa	.267	117
Scai8QnA	Seoul National U.	.121	154
shefinq250	U. of Sheffield	.111	176
shefatt250	U. of Sheffield	.096	179
NTU99	National Taiwan U.	.087	173
UIowaQA2	U. of Iowa	.060	175

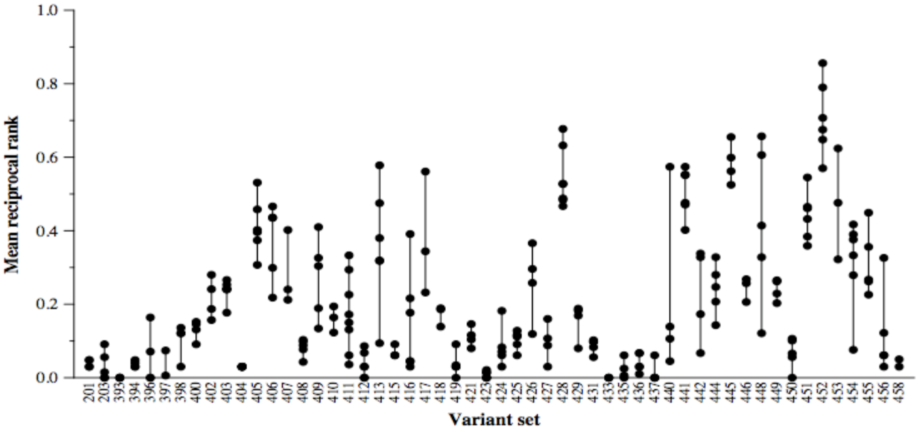
TREC 1999 QA results



Run Name	Participant	MRR	# not found
LCCSMU1	Southern Methodist U.	0.76	95 (14%)
ibmhlt00250	IBM (Ittycheriah)	0.46	263 (39%)
pir0qal2	Queens College, CUNY	0.46	264 (39%)
uwmt9qal1	MultiText, U. of Waterloo	0.46	265 (39%)
IBMKA250	IBM (Prager)	0.42	294 (43%)
lcat250	LIMSI-CNRS	0.41	307 (45%)
NTTD9QAa1L	NTT Data Corp.	0.39	299 (44%)
SUT9p2c3c250	Syracuse U.	0.39	319 (47%)
ICrjc99b	Imperial College	0.39	348 (51%)
UdeMlng2	U. de Montreal	0.37	325 (48%)
KUQA250a	Korea U.	0.37	338 (50%)
ALI9C250	U. de Alicante	0.36	321 (47%)
xeroxQA9l	Xerox Research Centre Europe	0.35	349 (51%)
shef250p	U. of Sheffield	0.34	335 (49%)
INQ9AND	U. of Massachusetts	0.34	344 (50%)
SunToo	Sun Microsystems	0.34	362 (53%)
FDUT9QL1	Fudan University	0.34	369 (54%)
KAIST9qa2	KAIST	0.33	362 (53%)
qntua02	National Taiwan U.	0.32	376 (55%)
clr00s2	CL Research	0.30	386 (57%)

TREC 2000 QA results

Performance on Variants



In general, systems are pretty robust to different phrasings.

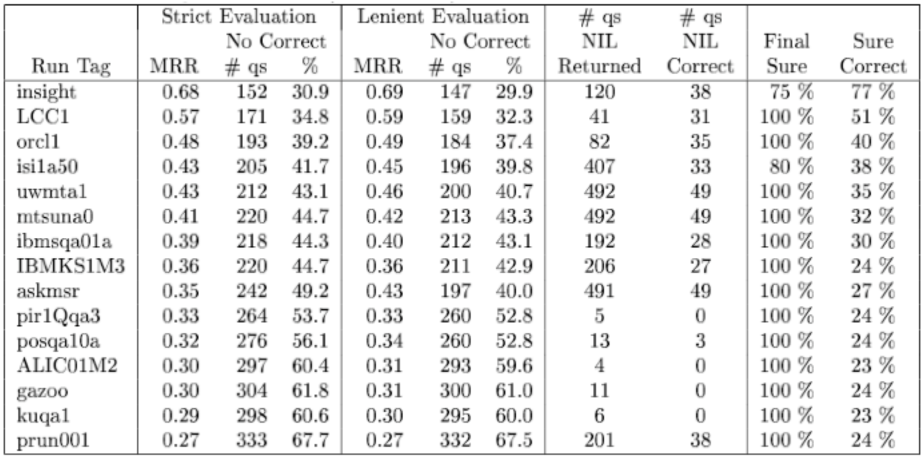
QA at TREC in 2001

- Changes from 2000:
  - Addition of a new task: *list questions*
    - Some questions had more than one answer in more than one document
    - The systems had to find all the answers
    - Examples:
      - "Name 4 U.S. cities that have a 'Shubert' theater"
      - "Name 30 individuals who served as a cabinet officer under Ronald Reagan."
  - Answers limited to 50 bytes
  - Some questions had no answer in the documents; the systems had to identify these and return "NIL"

Questions in QA 2001

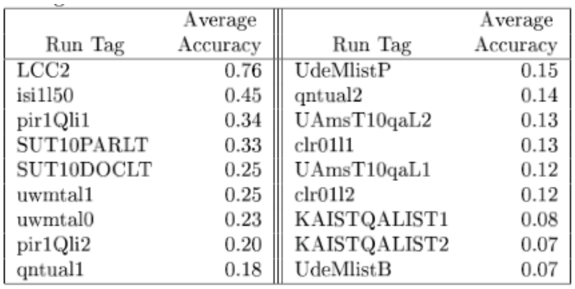
- There was no attempt to do quality control or type-balancing of questions
- Some of them would be very hard to answer in just 50 bytes
- "Definition questions":
  - "Who is Bill Clinton?"
  - Very hard to answer with no context in only 50 bytes
  - Many different answers could be considered correct

QA 2001 Results



	Strict Evaluation			Lenient Evaluation			# qs	# qs		
		No Correct			No Correct		NIL	NIL	Final	Sure
Run Tag	MRR	# qs	%	MRR	# qs	%	Returned	Correct	Sure	Correct
insight	0.68	152	30.9	0.69	147	29.9	120	38	75 %	77 %
LCC1	0.57	171	34.8	0.59	159	32.3	41	31	100 %	51 %
orcl1	0.48	193	39.2	0.49	184	37.4	82	35	100 %	40 %
isi1a50	0.43	205	41.7	0.45	196	39.8	407	33	80 %	38 %
uwmta1	0.43	212	43.1	0.46	200	40.7	492	49	100 %	35 %
mtsuna0	0.41	220	44.7	0.42	213	43.3	492	49	100 %	32 %
ibmsqa01a	0.39	218	44.3	0.40	212	43.1	192	28	100 %	30 %
IBMKS1M3	0.36	220	44.7	0.36	211	42.9	206	27	100 %	24 %
askmsr	0.35	242	49.2	0.43	197	40.0	491	49	100 %	27 %
pir1Qqa3	0.33	264	53.7	0.33	260	52.8	5	0	100 %	24 %
posqa10a	0.32	276	56.1	0.34	260	52.8	13	3	100 %	24 %
ALIC01M2	0.30	297	60.4	0.31	293	59.6	4	0	100 %	23 %
gazoo	0.30	304	61.8	0.31	300	61.0	11	0	100 %	24 %
kuqa1	0.29	298	60.6	0.30	295	60.0	6	0	100 %	23 %
prun001	0.27	333	67.7	0.27	332	67.5	201	38	100 %	24 %

List Task Results



Run Tag	Average Accuracy	Run Tag	Average Accuracy
LCC2	0.76	UdeMlistP	0.15
isi1l50	0.45	qntual2	0.14
pir1Qli1	0.34	UAmsT10qaL2	0.13
SUT10PARLT	0.33	clr01l1	0.13
SUT10DOCLT	0.25	UAmsT10qaL1	0.12
uwmtal1	0.25	clr01l2	0.12
uwmtal0	0.23	KAISTQALIST1	0.08
pir1Qli2	0.20	KAISTQALIST2	0.07
qntual1	0.18	UdeMlistB	0.07

Accuracy is the number of right answers found divided by total number of answers

QA at TREC in 2002

- Changes from 2001:
  - Exact answers – don't let the system just return a bunch of unrelated words
  - Judge system answer along with supporting document on 4-level scale:
    - Wrong – system answer does not contain right answer
    - Not supported – system answer is right, but document does not support it
    - Not exact – system answer is right and document supports it, but answer is either incomplete or contains other possible answers
    - Right – system answer is right and document supports it
  - Systems return their confidence in their answer
  - Larger corpus (1 million documents)

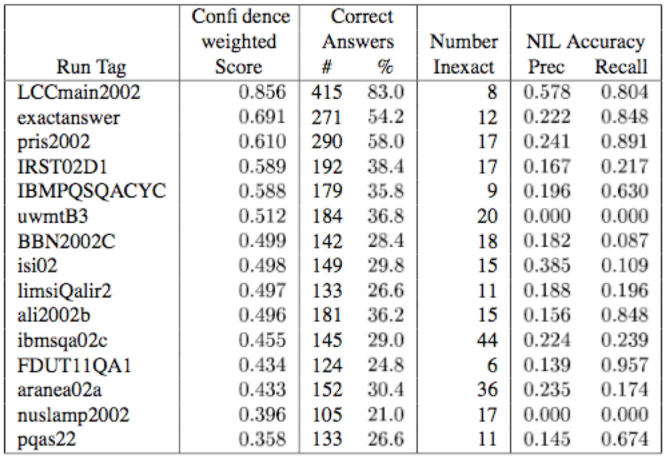
QA Evaluation in 2002

- Instead of reciprocal rank of first correct answer, a confidence-weighted average

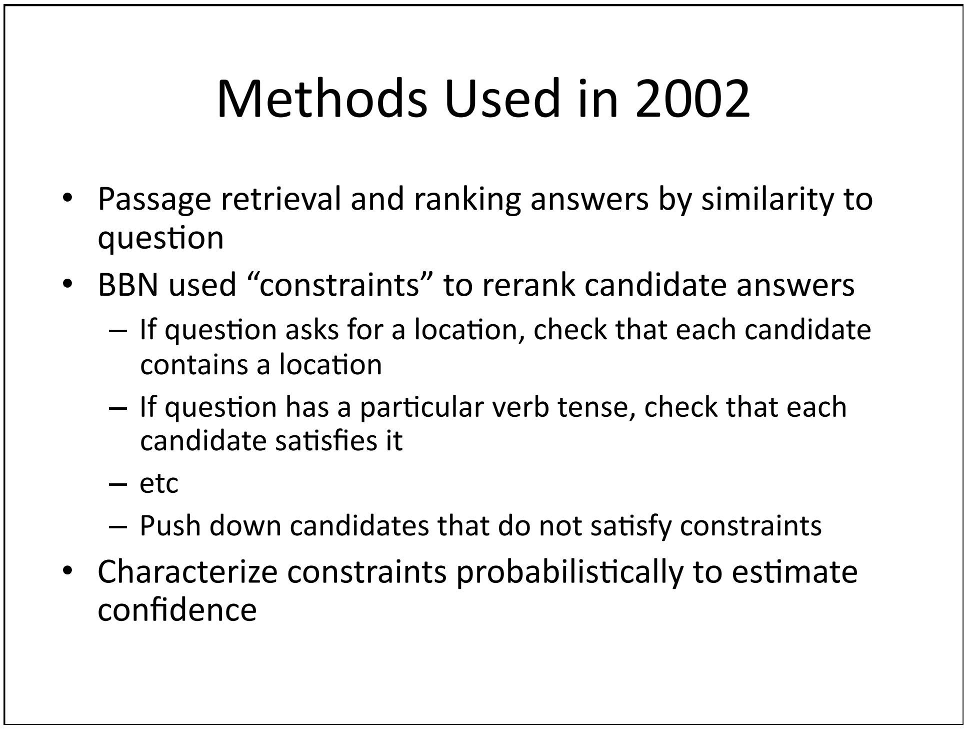
$$\frac{1}{Q}\sum_{i=1}^{Q}\frac{\#\ \text{correct in first } i \text{ ranks}}{i}$$

- Systems that do a better job of estimating their confidence will score higher
  - These systems are more useful to users, since they tell the user whether to trust it or not

QA 2002 Results

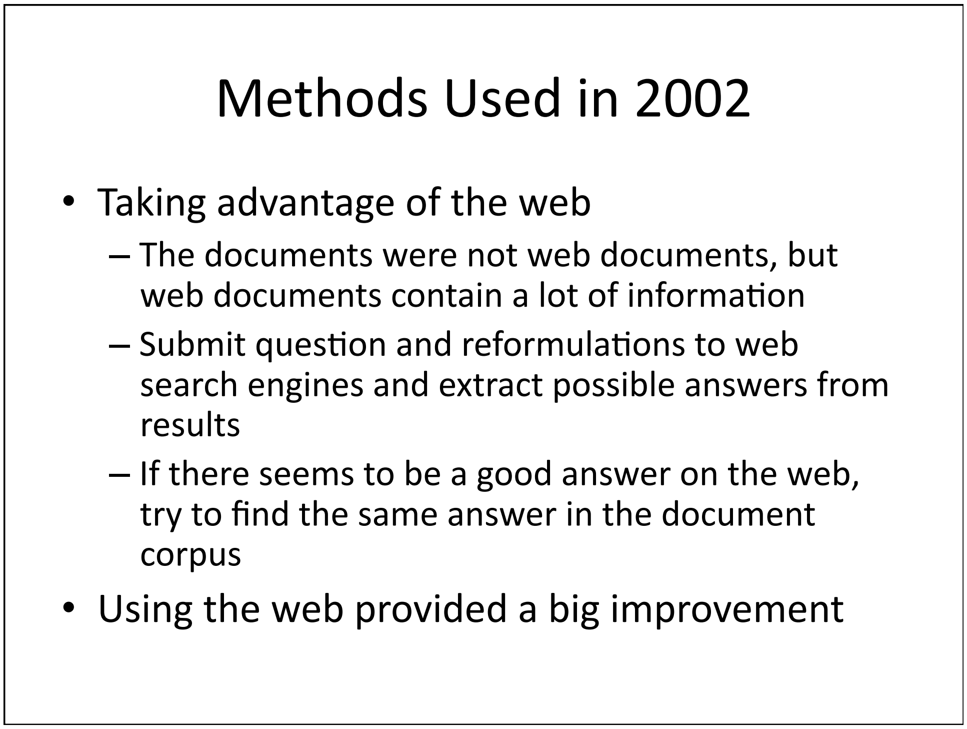


Run Tag	Confidence weighted Score	Correct Answers #	Correct Answers %	Number Inexact	NIL Accuracy Prec	NIL Accuracy Recall
LCCmain2002	0.856	415	83.0	8	0.578	0.804
exactanswer	0.691	271	54.2	12	0.222	0.848
pris2002	0.610	290	58.0	17	0.241	0.891
IRST02D1	0.589	192	38.4	17	0.167	0.217
IBMPQSQACYC	0.588	179	35.8	9	0.196	0.630
uwmtB3	0.512	184	36.8	20	0.000	0.000
BBN2002C	0.499	142	28.4	18	0.182	0.087
isi02	0.498	149	29.8	15	0.385	0.109
limsiQalir2	0.497	133	26.6	11	0.188	0.196
ali2002b	0.496	181	36.2	15	0.156	0.848
ibmsqa02c	0.455	145	29.0	44	0.224	0.239
FDUT11QA1	0.434	124	24.8	6	0.139	0.957
aranea02a	0.433	152	30.4	36	0.235	0.174
nuslamp2002	0.396	105	21.0	17	0.000	0.000
pqas22	0.358	133	26.6	11	0.145	0.674



Methods Used in 2002

- Passage retrieval and ranking answers by similarity to question
- BBN used “constraints” to rerank candidate answers
  - If question asks for a location, check that each candidate contains a location
  - If question has a particular verb tense, check that each candidate satisfies it
  - etc
  - Push down candidates that do not satisfy constraints
- Characterize constraints probabilistically to estimate confidence

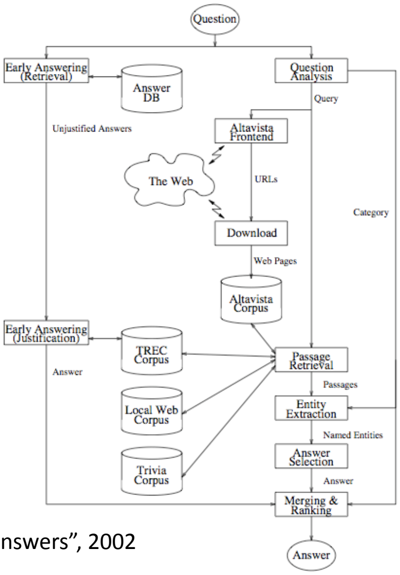


Methods Used in 2002

- Taking advantage of the web
  - The documents were not web documents, but web documents contain a lot of information
  - Submit question and reformulations to web search engines and extract possible answers from results
  - If there seems to be a good answer on the web, try to find the same answer in the document corpus
- Using the web provided a big improvement

Methods Used in 2002

- U. Waterloo used a mix of knowledge-based systems, the web, and traditional IR methods
- Many sources of evidence combined into a single candidate answer



From Clarke et al., “Statistical Selection of Exact Answers”, 2002

Methods Used in 2002

- IBM used translation, like in cross-language retrieval
- Translate “question words” to “answer words”
- Natural probabilistic method gives confidence scores