Information Retrieval

Lecture 7: Question Answering

Computer Science Tripos Part II

UNIVERSITY OF CAMBRIDGE

Simone Teufel

Natural Language and Information Processing (NLIP) Group

sht25@cl.cam.ac.uk

Question Answering: Task definition in TREC-QA

- QA Track since TREC-1999: Open-domain factual textual QA
- Task requirements (in comparison with IR):
  1. Input: NL questions, not keyword-based queries
  2. Output: answers, not documents
- Rules:
  - All runs completely automatic
  - Frozen systems once questions received; answers back to TREC within one week
  - Answers may be extracted or automatically generated from material in document collection only
  - The use of external resources (dictionaries, ontologies, WWW) is allowed
  - Each returned answer is checked manually by TREC-QA (no comparison to gold standard)

TREC QA: Example questions

TREC-8	How many calories are there in a Big Mac? Where is the Taj Mahal?
TREC-9	Who invented the paper clip? How much folic acid should an expectant mother take daily? Who is Colin Powell?
TREC-10	What is an atom? How much does the human adult female brain weigh? When did Hawaii become a state?

Questions in TREC

- Type of question: reason, definition, list of instances, context-sensitive to previous questions (TREC-10)
- Source of question: invented for evaluation (TREC-8); since TREC-9 mined from logs (Encarta, Excite)
  - $\rightarrow$ strong impact on task: more realistic questions are harder on assessors and systems, but more representative for training
- Type of answer string: 250 Bytes (TREC-8/9, since TREC-12); 50 Bytes (TREC-8–10); exact since TREC-11
- Guarantee of existence of answer: no longer given since TREC-10

Examples of answer strings

What river in the US is known as the Big Muddy?

System A:	the Mississippi
System B:	Known as Big Muddy, the Mississippi is the longest
System C:	as Big Muddy , the Mississippi is the longest
System D:	messed with . Known as Big Muddy , the Mississip
System E:	Mississippi is the longest river in the US
System F:	the Mississippi is the longest river in the US
System G:	the Mississippi is the longest river(Mississippi)
System H:	has brought the Mississippi to its lowest
System I:	ipes.In Life on the Mississippi,Mark Twain wrote t
System K:	Southeast;Mississippi;Mark Twain;officials began
System L:	Known; Mississippi; US,; Minnessota; Cult Mexico
System M:	Mud Island,; Mississippi; “The; history; Memphis

Decreasing quality of answers

Manual checking of answers

- Systems return [docid, answer-string] pairs; mean answer pool per question judged: 309 pairs
- Answers judged in the context of the associated document
- ”Objectively” wrong answers okay if document supports them
  - Taj Mahal
- Considerable disagreement in terms of absolute evaluation metrics
- But relative MRRs (rankings) across systems very stable

Labels

- Ambiguous answers are judged as “incorrect”:

  What is the capital of the Kosovo?

  250B answer:

  protestors called for intervention to end the “Albanian uprising”. At Vucitrn, 20 miles northwest of Pristina, five demonstrators were reported injured, apparently in clashes with police. Violent clashes were also repo

- Answers need to be supported by the document context → the second answer is “unsupported”:

  What is the name of the late Phillippine President Marco’s wife?

  – Ferdinand Marcos and his wife Imelda... → [supported]

  – Imelda Marcos really liked shoes... → [unsupported]

List task (TREC-10, since TREC-12)

- 25 questions: retrieve a given target number of instances of something
- Goal: force systems to assemble an answer from multiple strings
  - `Name 4 US cities that have a ‘‘Shubert’’ theater`
  - `What are 9 novels written by John Updike?`
  - `What are six names of navigational satellites?`
  - `Name 20 countries that produce coffee.`
- List should not be easily located in reference work
- Instances are guaranteed to exist in collection
- Multiple documents needed to reach target, though single documents might have more than one instance
- Since TREC-12: target number no longer given; task is to find all

MRR: Mean reciprocal rank

- Task is precision-oriented: only look at top 5 answers
- Score for individual question i is the reciprocal rank r_i where the first correct answer appeared (0 if no correct answer in top 5 returns).

$$RR_i = \frac{1}{r_i}$$

- Possible reciprocal ranks per question: [0, 0.2, 0.25, 0.33, 0.5, 1]
- Score of a run (MRR) is mean over n questions:

$$MRR = \frac{1}{n} \sum_{i=1}^{n} RR_i$$

Example: Mean reciprocal rank

162: What is the capital of Kosovo?

1	18 April, 1995, UK GMT Kosovo capital
2	Albanians say no to peace talks in Pr
3	0 miles west of Pristina, five demon
4	Kosovo is located in south and south
5	The provincial capital of the Kosovo

$$\rightarrow RR_{162} = \frac{1}{3}$$

23: Who invented the paper clip?

1	embrace Johan Vaaler, as the true invento
2	seems puzzling that it was not invented e
3	paper clip. Nobel invented many useful th
4	modern-shaped paper clip was patented in A
5	g Johan Valerand, leaping over Norway, in

$$\rightarrow RR_{23} = 1$$

2: What was the monetary value of the Nobel Peace Prize in 1989?

1	The Nobel poll is temporarily disabled. 1994 poll
2	perience and scientific reality, and applied to socie
3	Curies were awarded the Nobel Prize together with Beqc
4	the so-called beta-value. $40,000 more than expected
5	that is much greater than the variation in mean value

$$\rightarrow RR_{2} = 0$$

$$\rightarrow MRR = \frac{\frac{4}{3}}{3} = .444$$

Other QA evaluation metrics used in TREC

- Average accuracy since 2003: only one answer per question allowed; accuracy is $\frac{\textit{Answers correct}}{\textit{Total Answers}}$
- Confidence-weighted score: systems submit one answer per question and order them according to the confidence they have in the answer (with their best answer first in the file)

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

(Q being the number of questions). This evaluation metric (which is similar to Mean Average Precision) was to reward systems for their confidence in their answers, as answers high up in the file participate in many calculations.

Results

- In TREC-8, 9, 10 best systems returned MMR of .65–.70 for 50B answers, answering around 70–80% of all questions
- In 55% of the cases where answer was found in the first 5 answers, this answer was in rank 1
- Accuracy of best system in TREC-10's list task had an accuracy of .75
- The best confidence-weighted score in TREC-11 achieved was .856 (NIL-prec .578, NIL recall .804)
- Best performance in TREC-12 (exact task) was an accuracy of .700

QA systems

- Overview of three QA systems:
- Cymphony system (TREC-8)
  - NE plus answer type detection
  - Shallow parsing to analyse structure of questions
- SMU (TREC-9)
  - Matching of logical form
  - Feedback loops
- Microsoft (TREC-10)
  - Answer redundancy and answer harvesting
  - Claim: “Large amounts of data make intelligent processing unnecessary.”

Overall algorithm

- Question Processing
  - Shallow parse
  - Determine expected answer type
  - Question expansion
- Document Processing
  - Tokenise, POS-tag, NE-index
- Text Matcher (= Answer production)
  - Intersect search engine results with NE
  - Rank answers

Named entity recognition

- Over 80% of 200 TREC-8 questions ask for a named entity (NE)
- NE employed by most successful systems in TREC (Verhees and Tice, 2000))
- MUC NE types: person, organisation, location, time, date, money, percent
- Textract covers additional types:
  - frequency, duration, age
  - number, fraction, decimal, ordinal, math equation
  - weight, length, temperature, angle, area, capacity, speed, rate
  - address, email, phone, fax, telex, www
  - name (default proper name)
- Textract subclassifies known types:
  - organisation $\rightarrow$ company, government agency, school
  - person $\rightarrow$ military person, religious person

Expected answer type

Who won the 1998 Nobel Peace Prize?

Expected answer type:	PERSON
Key words:	won, 1998, Nobel, Peace, Prize

Why did David Koresh ask the FBI for a word processor?

Expected answer type:	REASON
Key words:	David, Koresh, ask, FBI, word, processor

Question Expansion:

Expected answer type:	[because | because of | due to | thanks to | since | in order to | to VP]
Key words:	[ask|asks|asked|asking, David, Koresh, FBI, word, processor]

FST rules for expected answer type

R1: Name NP(city | country | company) → CITY|COUNTRY|COMPANY

VG[name] NP[a country] that VG[is developing] NP[a magnetic levitation railway system]

R2: Name NP(person_w) → PERSON

VG[Name] NP[the first private citizen] VG[to fly] PP[in space]

("citizen" belongs to word class `person_w`).

R3: CATCH-ALL: proper noun

Name a film that has won the Golden Bear in the Berlin Film Festival.

Direct matching of question words

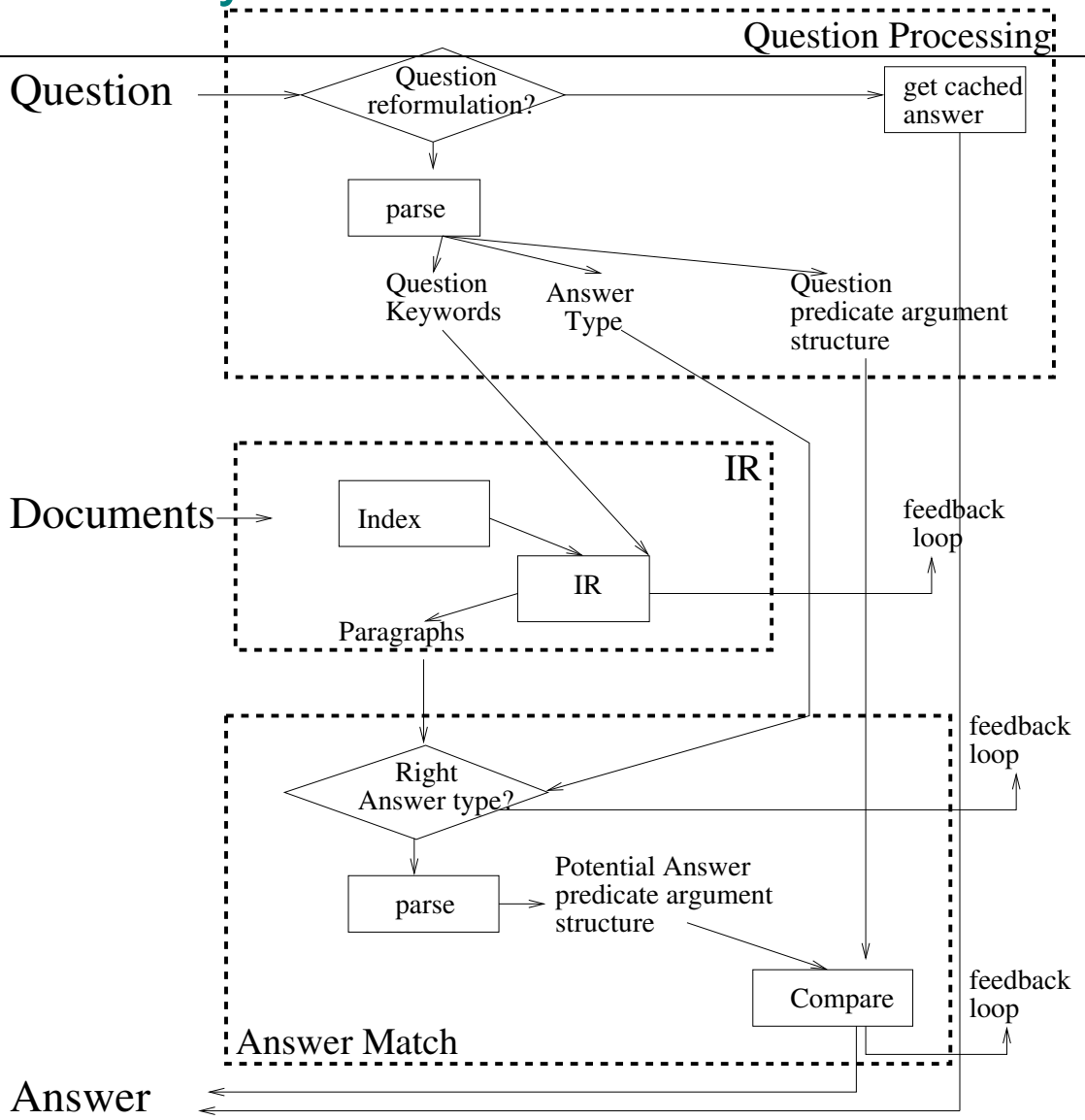
who/whom →	PERSON
when →	TIME/DATE
where/what place →	LOCATION
what time (of day) →	TIME
what day (of the week) →	DAY
what/which month →	MONTH
how often →	FREQUENCY
...	

This classification happens only if the previous rule-based classification did not return unambiguous results.

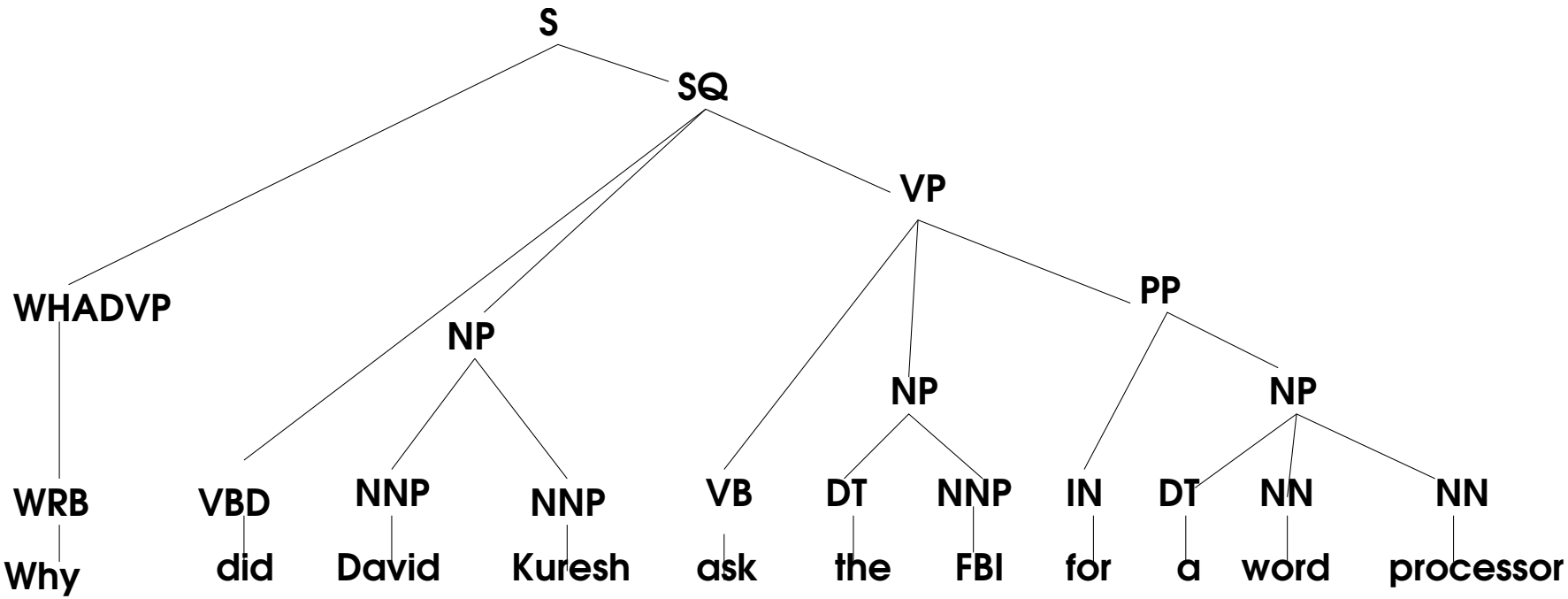
The Southern Methodist University (SMU) system (Harabagiu et al.)

- Example of a deep processing system which has been extremely successful in TREC-QA (clear winner in most years)
- Machinery beyond answer type determination:
  1. Variants/feedback loops: morphological, lexical, syntactic, by reasoning
  2. Comparison between answer candidate and question on basis of logical form
- Deep processing serves to
  - capture semantics of open-domain questions
  - justify correctness of answers

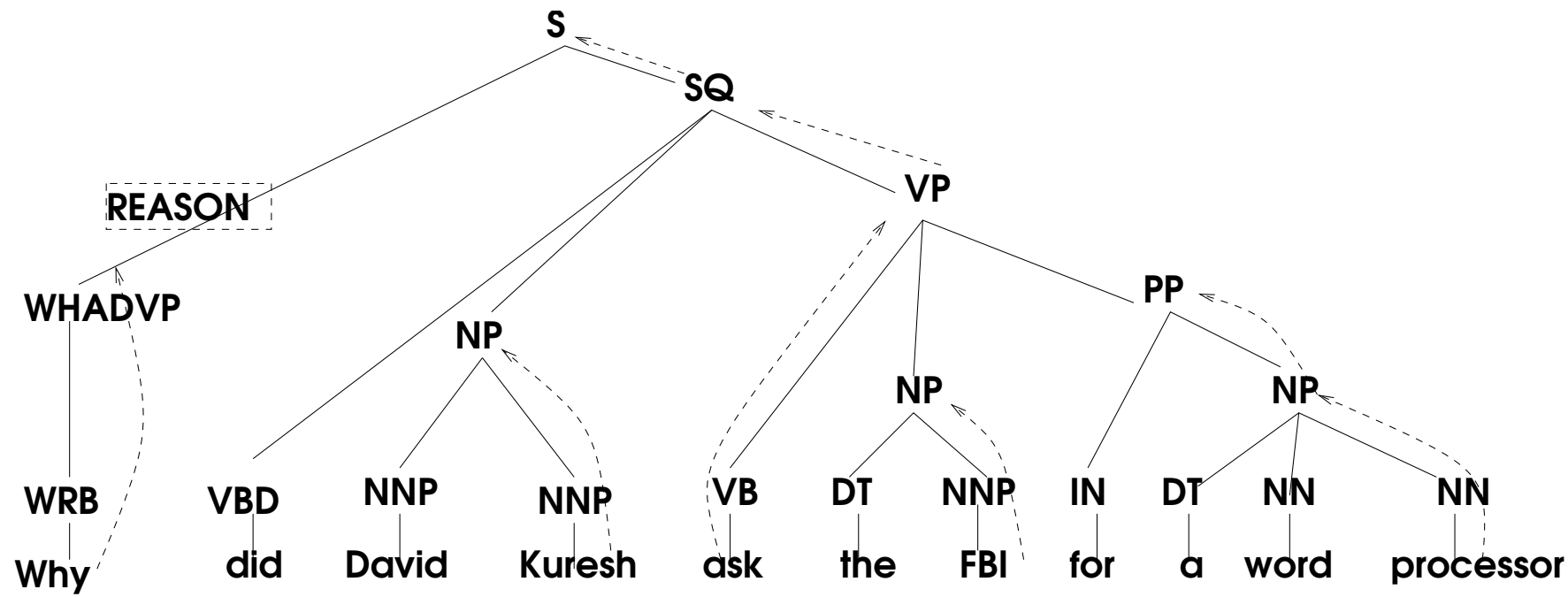
Overview of SMU system

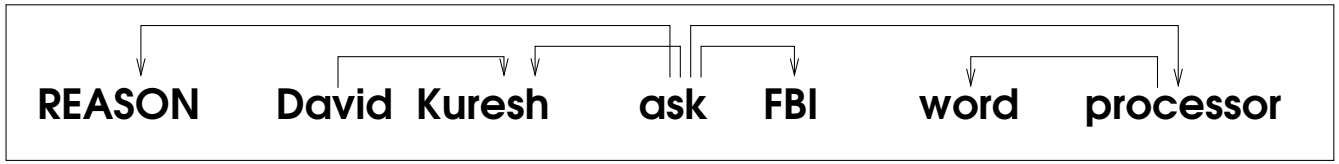


SMU: Derivation of logical forms



SMU: Derivation of logical forms





SMU: Variants (“Feedback loops”)

- Morphological (+40%):
  - *Who invented the paper clip?* — Main verb “invent”, ANSWER-TYPE “who” (subject) → add keyword “inventor”
- Lexical (+52%; used in 129 questions):
  - *How far is the moon?* — “far” is an attribute of “distance”
  - *Who killed Martin Luther King?* — “killer” = “assassin”
- Semantic alternations and paraphrases, abductive reasoning (+8%; used in 175 questions)
  - *How hot does the inside of an active volcano get?*
  - Answer in “lava fragments belched out of the mountain were as hot as 300 degrees Fahrenheit”
  - Facts needed in abductive chain:
    - volcano IS-A mountain; lava PART-OF volcano
- Combination of loops increases results considerably (+76%)

At the other end of the spectrum: the Microsoft system

- Circumvent difficult NLP problems by using more data
- The web has 2 billion indexed pages
- Claim: deep reasoning is only necessary if search ground is restricted
- The larger the search ground, the greater the chance of finding answers with a simple relationship between question string and answer string:

Who killed Abraham Lincoln?

DOC 1	John Wilkes Booth is perhaps America's most infamous assassin. He is best known for having fired the bullet that ended Abraham Lincoln's life.	TREC
DOC 2	John Wilkes Booth killed Abraham Lincoln.	web

The Microsoft system: Methods

1. Question processing is minimal: reordering of words, removal of question words, morphological variations
2. Matching done by Web query (google):
   - Extract potential answer strings from top 100 summaries returned
3. Answer generation is simplistic:
   - Weight answer strings (frequency, fit of match) – learned from TREC-9
   - Shuffle together answer strings
   - Back-projection into TREC corpus: keywords + answers to traditional IR engine
4. Improvement: Expected answer type filter (24% improvement)
   - No full-fledged named entity recognition

Query string generation

Rewrite module outputs a set of 3-tupels:

- Search string
- Position in text where answer is expected with respect to query string : LEFT|RIGHT|NULL
- Confidence score (quality of template)

Who is the world's richest man married to?

[+is the world's richest man married to	LEFT	5]
[the +is world's richest man married to	LEFT	5]
[the world's +is richest man married to	RIGHT	5]
[the world's richest +is man married to	RIGHT	5]
[the world's richest man +is married to	RIGHT	5]
[the world's richest man married +is to	RIGHT	5]
[the world's richest man married to +is	RIGHT	5]
[world's richest man married	NULL	2]
[world's AND richest AND married	NULL	1]

String weighting

- Obtain 1-grams, 2-grams, 3-grams from google short summaries
- Score each n-gram n according to the weight r_q of query q that retrieved it
- Sum weights across all summaries containing the ngram n (this set is called S_n)

$$w_n = \sum_{n \in S_n} r_q$$

w_n: weight of ngram n
S_n: set of all retrieved summaries which contain n
r_q: rewrite weight of query q

Answer string generation

- Merge similar answers (ABC + BCD $\rightarrow$ ABCD)
  - Assemble longer answers from answer fragments
  - Weight of new n-gram is maximum of constituent weights
  - Greedy algorithm, starting from top-scoring candidate
  - Stop when no further ngram tiles can be detected
  - But: cannot cluster “redwoods” and “redwood trees”
- Back-projection of answer
  - Send keywords + answers to traditional IR engine indexed over TREC documents
  - Report matching documents back as “support”
- Always return NIL on 5th position

The Microsoft system: Examples

- Time sensitivity of questions:
  Q1202: Who is the Governor of Alaska? → system returns governor in 2001, but TREC expects governor in 1989.
- Success stories:

Question	Answer	TREC document
What is the birth-stone for June?	Pearl	for two weeks during June (the pearl is the birth-stone for those born in that month)
What is the rainiest place on Earth?	Mount Wailaleale	and even Pago Pago, noted for its prodigious showers, gets only about 196 inches annually (The titleholder, according to the National Geographic Society, is Mount Wailaleale in Hawaii, where about 460 inches of rain falls each year).

Microsoft system: Discussion

- Results: mid-range (.347 MRR, 49% no answer)
- Development time of less than a month
- Produced “exact strings” before TREC-11 demanded it: average returned length 14.6 bytes
- Does this system undermine of QA as a gauge for NL understanding?
  - If TREC wants to measure straight performance on factual question task, less NLP might be needed than previously thought
  - But if TREC wants to use QA as test bed for text understanding, it might now be forced to ask “harder” questions
- And still: the really good systems are still the ones that do deep NLP processing!

Summary

- Open domain, factual question answering
- TREC: Source of questions matters (web logs v. introspection)
- Mean reciprocal rank main evaluation measure
- MRR of best systems 0.68 - 0.58
- Best systems answer about 75% of questions in the first 5 guesses, and get the correct answer at position 1.5 on avg ($\frac{1}{.66}$)
- System technology
  - NE plus answer type detection (Cymphony)
  - Matching of logical form, Feedback loops (SMU)
  - Answer redundancy and answer harvesting (Microsoft)

- Ellen Voorhees (1999): The TREC-8 Question Answering Track Report, Proceedings of TREC
- E.M. Voorhees (2003): Overview of the TREC 2003 Question Answering Track, electronic TREC proceedings
- S. Teufel (2005, To Appear): Chapter *IR and QA evaluation*. In: Evaluation Methods in Speech and NLP. Kluwer.
- R. Srihari and W. Li (1999): “Information-extraction supported question answering”, TREC-8 Proceedings
- S. Harabagiu et al (2000), “FALCON: Boosting Knowledge for Answer Engines, TREC-9 Proceedings”, TREC-9 Proceedings
- S. Harabagiu et al (2001), “The role of lexico-semantic feedback in open-domain textual question-answering”, ACL-2001
- E. Brill et al (2001), “Data intensive question answering”, TREC-10 Proceedings