(TREC-style) Question Answering systems and Textual Inference

Christopher Manning

CS224N/Ling 280 2008

(includes slides borrowed from Sanda Harabagiu, Roxana Girju)

Modern QA from text

- An idea originating from the IR community
- With massive collections of full-text documents, simply finding *relevant documents* is of limited use: we want *answers* from textbases
- QA: give the user a (short) answer to their question, perhaps supported by evidence.
- The common person's view? [From a novel]
  - "I like the Internet. Really, I do. Any time I need a piece of shareware or I want to find out the weather in Bogota … I'm the first guy to get the modem humming. But as a source of information, it sucks. You got a billion pieces of data, struggling to be heard and seen and downloaded, and anything I want to know seems to get trampled underfoot in the crowd."
    - M. Marshall. *The Straw Men.* HarperCollins Publishers, 2002.

Sample TREC questions

1. Who is the author of the book, "The Iron Lady: A Biography of Margaret Thatcher"?
2. What was the monetary value of the Nobel Peace Prize in 1989?
3. What does the Peugeot company manufacture?
4. How much did Mercury spend on advertising in 1993?
5. What is the name of the managing director of Apricot Computer?
6. Why did David Koresh ask the FBI for a word processor?
7. What debts did Qintex group leave?
8. What is the name of the rare neurological disease with symptoms such as: involuntary movements (tics), swearing, and incoherent vocalizations (grunts, shouts, etc.)?

People *want* to ask questions…

Examples from AltaVista query log (late 1990s)

who invented surf music?
how to make stink bombs
where are the snowdens of yesteryear?
which english translation of the bible is used in official catholic liturgies?
how to do clayart
how to copy psx
how tall is the sears tower?

Examples from Excite query log (12/1999)

how can i find someone in texas
where can i find information on puritan religion?
what are the 7 wonders of the world
how can i eliminate stress
What vacuum cleaner does Consumers Guide recommend

Around 10% of early query logs

A Brief (Academic) History

- Question answering is not a new research area
- Question answering systems can be found in many areas of NLP research, including:
  - Natural language database systems
    - A lot of early NLP work on these: e.g., LUNAR system
    - There's still Microsoft English Query
  - Spoken dialog systems
    - Currently very active and commercially relevant

A Brief (Academic) History

- Focusing on open-domain QA is new focus
  - MURAX (Kupiec 1993): Encyclopedia answers
  - Hirschman: Reading comprehension tests
  - TREC QA competition: 1999–
- But not really new either: Simmons et al. 1965
  - Take an encyclopedia and load it onto a computer.
  - Take a question and parse it into a logical form
  - Perform simple information retrieval to get relevant texts, parse those into a logical form, match and rank
    - What do worms eat? Worms eat ???
      - Candidates
        - Worms eat grass
        - Grass is eaten by worms
        - Birds eat worms

Online QA Examples

- **LCC:** http://www.languagecomputer.com/demos/question_answering/index.html
- **AnswerBus** is an open-domain question answering system: www.answerbus.com
- **EasyAsk, AnswerLogic, AnswerFriend, Start, Quasm, Mulder, Webclopedia, ISI TextMap, etc.**
- **Google**

Question Answering at TREC

- Question answering competition at TREC consists of answering a set of 500 fact-based questions, e.g., *"When was Mozart born*?".
- For the first three years systems were allowed to return 5 ranked answer snippets (50/250 bytes) to each question.
  - Mean Reciprocal Rank (MRR) scoring:
    - 1, 0.5, 0.33, 0.25, 0.2, 0 for 1, 2, 3, 4, 5, 6+ doc
  - Mainly Named Entity answers (person, place, date, …)
- From 2002 the systems are only allowed to return a single *exact* answer and the notion of confidence has been introduced.

The TREC Document Collection

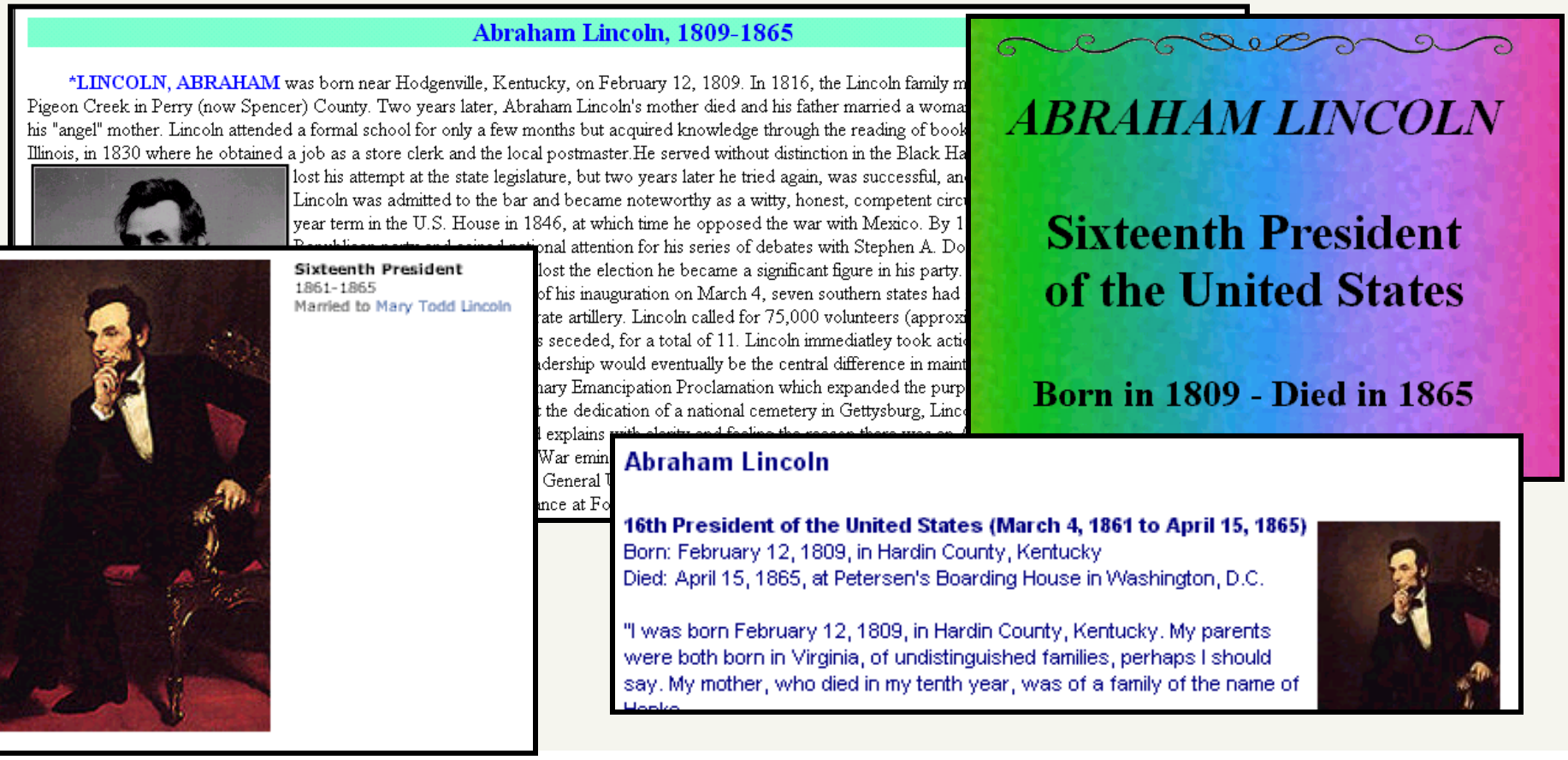
- The current collection uses news articles from the following sources:
  - AP newswire, 1998-2000
  - New York Times newswire, 1998-2000
  - Xinhua News Agency newswire, 1996-2000
- In total there are 1,033,461 documents in the collection. 3GB of text.
- This is a lot of text to process entirely using advanced NLP techniques so the systems usually consist of an initial information retrieval phase followed by more advanced processing.
- Many supplement this text with use of the web, and other knowledge bases

Top Performing Systems

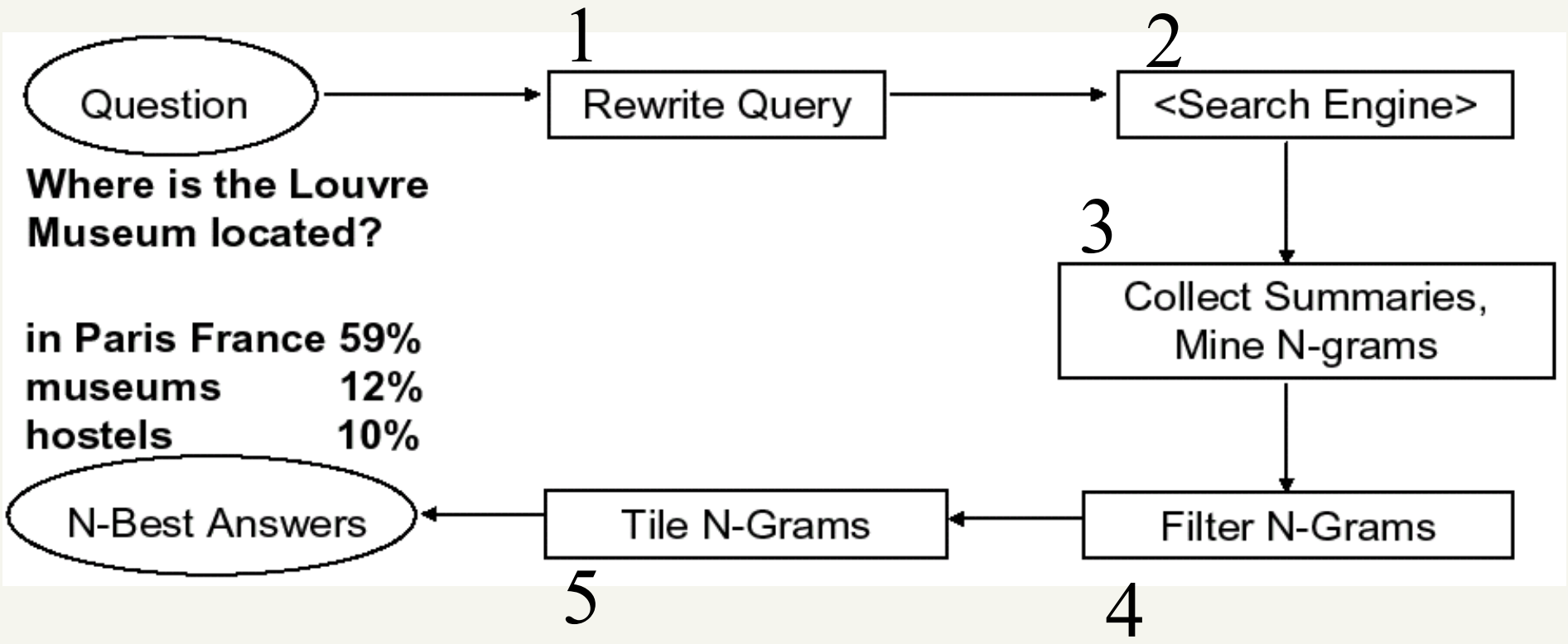
- Currently the best performing systems at TREC can answer approximately 70% of the questions !!!
- Approaches and successes have varied a fair deal
  - Knowledge-rich approaches, using a vast array of NLP techniques have been most successful
    - Notably Harabagiu, Moldovan et al. – SMU/UTD/LCC
  - AskMSR system stressed how much could be achieved by very simple methods with enough text (and now various copycats)

AskMSR: Shallow approach

- *In what year did Abraham Lincoln die?*
- Ignore hard documents and find easy ones



AskMSR: Details



Query rewriting: Ans is often similar to Ques

- Classify question into seven categories
  - **Who** is/was/are/were…?
  - **When** is/did/will/are/were …?
  - **Where** is/are/were …?

a. Category-specific transformation rules

eg "For Where questions, move 'is' to all possible locations"

"Where is the Louvre Museum located"

→ "is the Louvre Museum located"

→ "the is Louvre Museum located"

→ "the Louvre is Museum located"

→ "the Louvre Museum is located"

→ "the Louvre Museum located is"

Nonsense, but who cares? It's only a few more queries to Google.

b. Expected answer "Datatype" (eg, Date, Person, Location, …)

When was the French Revolution? → DATE

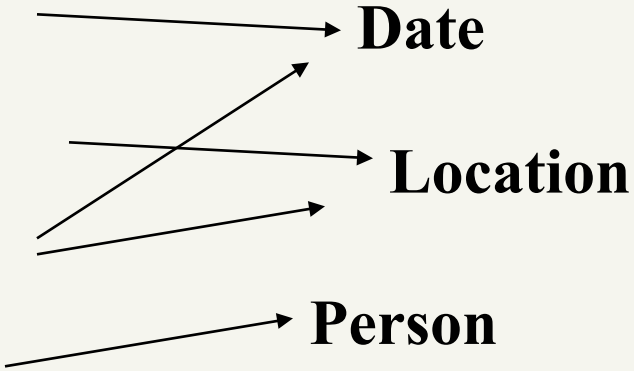
- Hand-crafted classification/rewrite/datatype rules (Could they be automatically learned?)

Mining N-Grams

- Send query to search engine; use result snippets
- Enumerate all N-grams in all retrieved snippets
  - Use hash table and other fancy footwork to make this efficient
- Weight of an n-gram: occurrence count, each weighted by "reliability" (weight) of rewrite that fetched the document
- Example: "Who created the character of Scrooge?"
  - Dickens - 117
  - Christmas Carol - 78
  - Charles Dickens - 75
  - Disney - 72
  - Carl Banks - 54
  - A Christmas - 41
  - Christmas Carol - 45
  - Uncle - 31

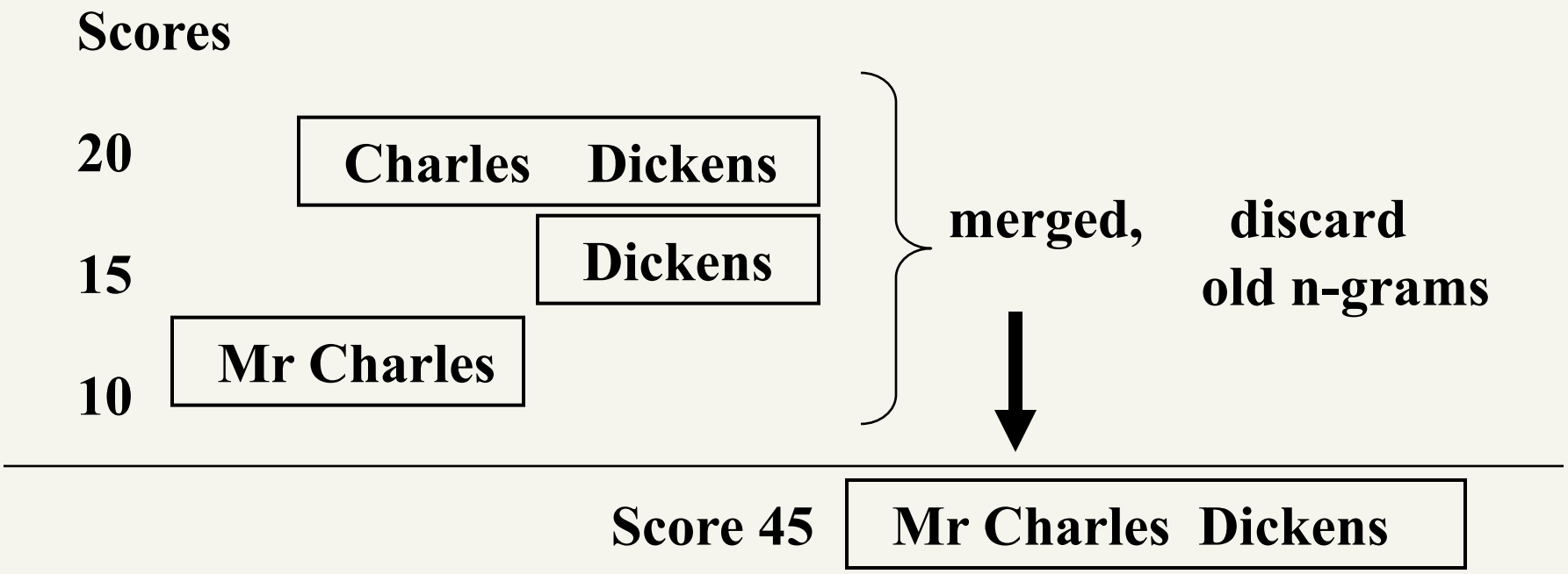
Filtering N-Grams

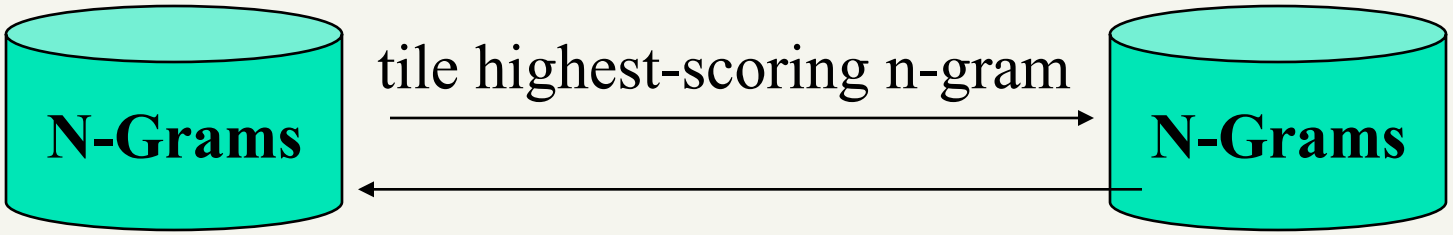
- Each question type is associated with one or more "**data-type filters**" = regular expression
- When…
- Where…
- What …
- Who …



- Boost score of n-grams that do match regexp
- Lower score of n-grams that don't match regexp
- Details omitted from paper….

Step 5: Tiling the Answers





Repeat, until no more overlap

Results

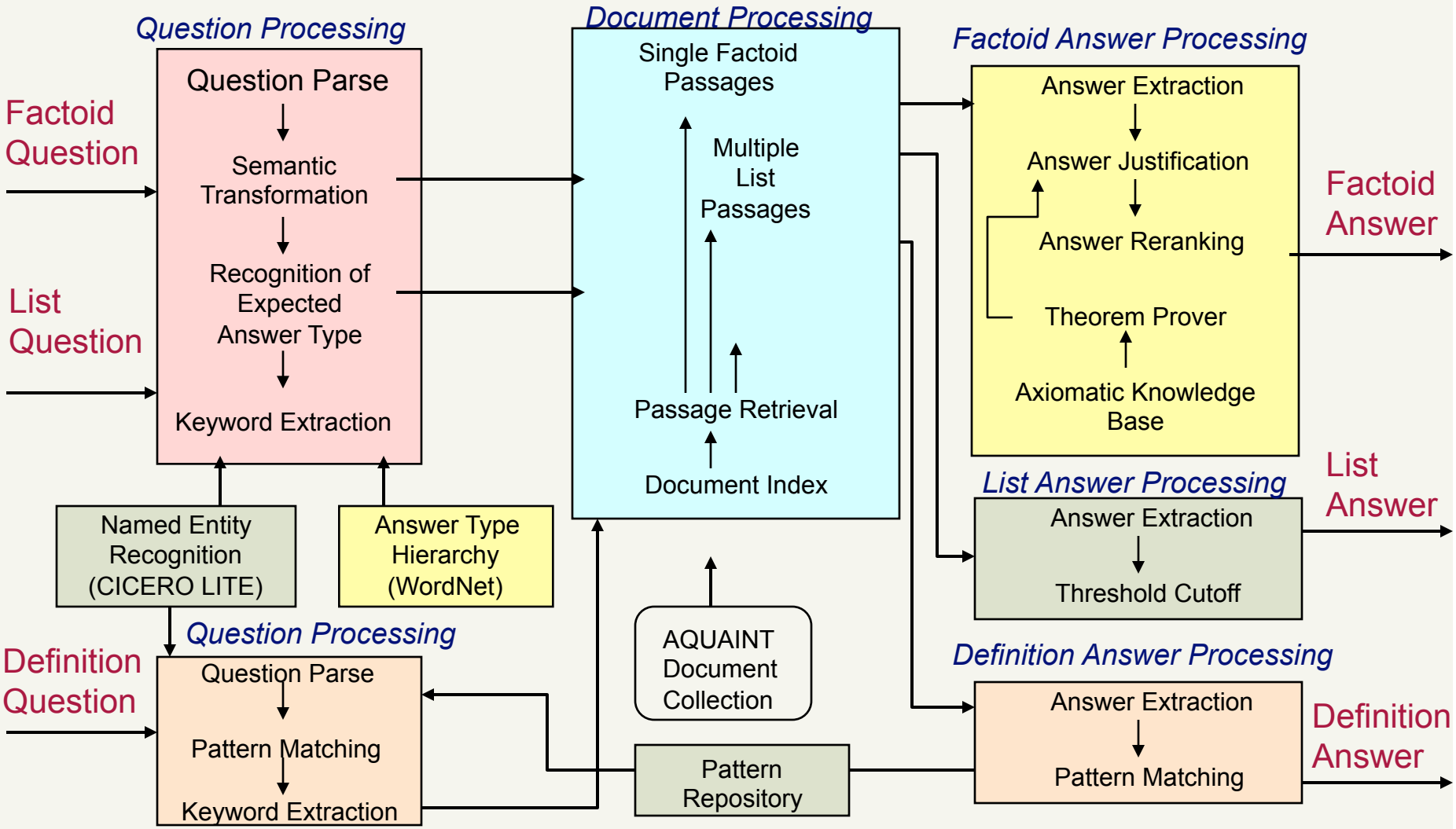
- Standard TREC contest test-bed:
  ~1M documents; 900 questions
- Technique doesn't do so well (though would have placed in top 9 of ~30 participants!)
  - MRR = 0.262
    - Right answer ranked about #4-#5 on average
  - Why? Because it relies on the enormity of the Web!
- But using the Web as a whole, not just TREC's 1M documents: MRR = 0.42
  - On average, right answer is ranked about #2-#3

Limitations

- In many scenarios (e.g., monitoring an individuals email…) we only have a small set of documents
- Works best/only for “Trivial Pursuit”-style fact-based questions
- Limited/brittle repertoire of
  - question categories
  - answer data types/filters
  - query rewriting rules

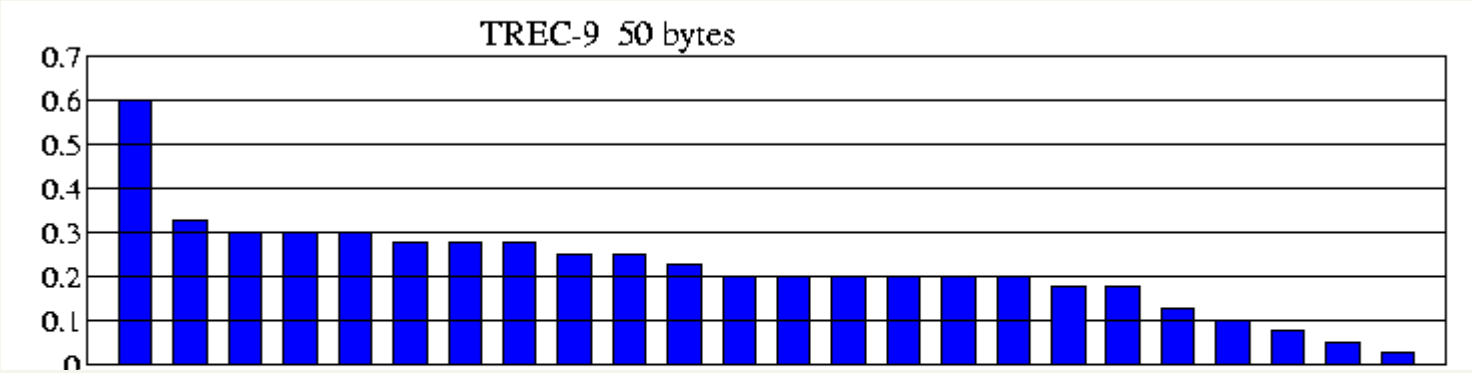
Full NLP QA: LCC (Harabagiu/Moldovan)

[below is the Architecture of LCC's QA system circa 2003]

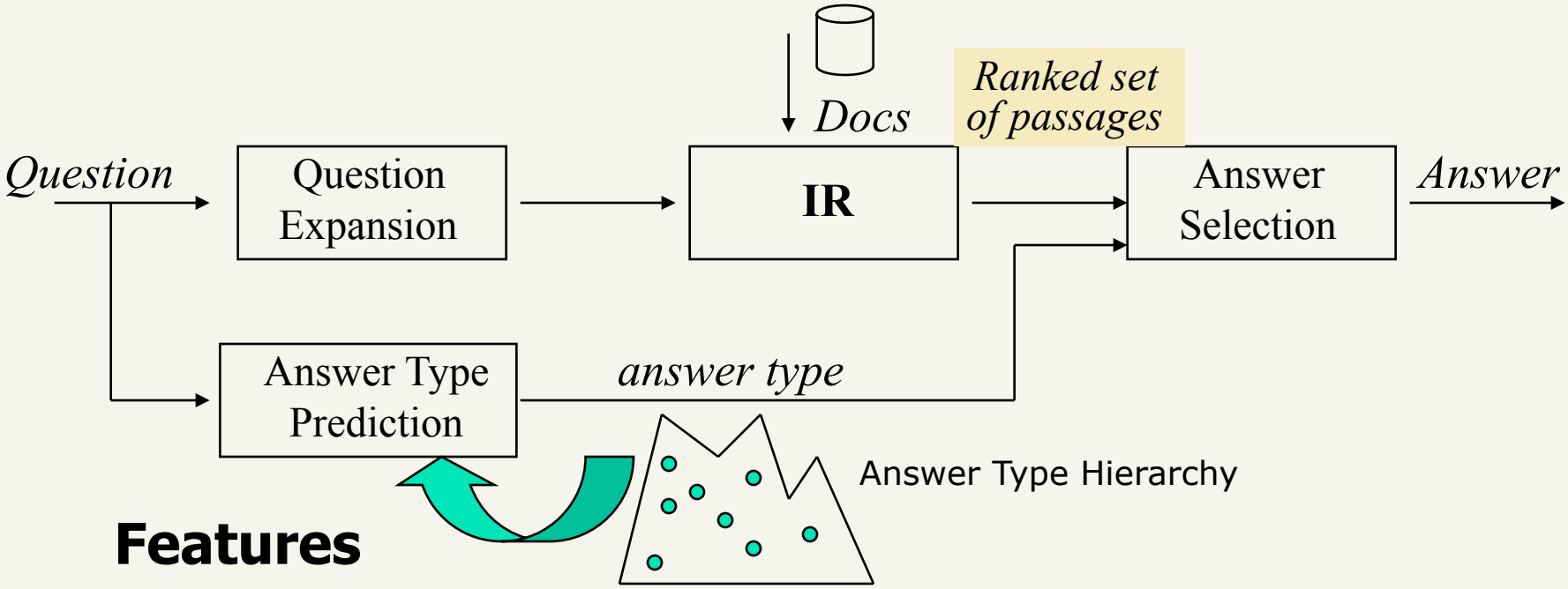


Value from sophisticated NLP – Pasca and Harabagiu (2001)

- Good IR is needed: SMART paragraph retrieval
- Large taxonomy of question types and expected answer types is crucial
- Statistical parser used to parse questions and relevant text for answers, and to build KB
- Further value comes from deeper NLP and inferencing



Answer types in State-of-the-art QA systems



Features

- Answer type
  - Labels questions with answer type based on a taxonomy
    - Person, location, weight, temperature, year, vehicle
  - Classifies questions (e.g. by using a maximum entropy model)

Answer Types

- Of course, determining the answer type isn't that easy…
  - Who questions can have organizations as answers
    - Who sells the most hybrid cars?
  - Which questions can have people as answers
    - Which president went to war with Mexico?

Lexical Terms Extraction as input to Information Retrieval

- Questions approximated by sets of unrelated words (lexical terms)
- Similar to bag-of-word IR models: but choose nominal non-stop words and verbs

Question (from TREC QA track)	Lexical terms
Q002: What was the monetary value of the Nobel Peace Prize in 1989?	monetary, value, Nobel, Peace, Prize, 1989
Q003: What does the Peugeot company manufacture?	Peugeot, company, manufacture
Q004: How much did Mercury spend on advertising in 1993?	Mercury, spend, advertising, 1993

Keyword Selection Algorithm

1. Select all non-stopwords in quotations
2. Select all NNP words in recognized named entities
3. Select all complex nominals with their adjectival modifiers
4. Select all other complex nominals
5. Select all nouns with adjectival modifiers
6. Select all other nouns
7. Select all verbs
8. Select the answer type word

Passage Extraction Loop

- Passage Extraction Component
  - Extracts passages that contain all selected keywords
  - Passage size dynamic
  - Start position dynamic
- Passage quality and keyword adjustment
  - In the first iteration use the first 6 keyword selection heuristics
  - If the number of passages is lower than a threshold ⇒ query is too strict ⇒ drop a keyword
  - If the number of passages is higher than a threshold ⇒ query is too relaxed ⇒ add a keyword

Passage Scoring

- Passage ordering is performed using a sort that involves three scores:
  - The number of words from the question that are recognized in the same sequence in the window
  - The number of words that separate the most distant keywords in the window
  - The number of unmatched keywords in the window

Rank candidate answers in retrieved passages

Q066: Name the first private citizen to fly in space.

- Answer type: Person
- Text passage:

  "Among them was Christa McAuliffe, the first private citizen to fly in space. Karen Allen, best known for her starring role in "Raiders of the Lost Ark", plays McAuliffe. Brian Kerwin is featured as shuttle pilot Mike Smith..."

- Best candidate answer: Christa McAuliffe

Extracting Answers for Factoid Questions: NER!

- In TREC 2003 the LCC QA system extracted 289 correct answers for factoid questions
- The Name Entity Recognizer was responsible for 234 of them
  - Current QA is largely based on the high accuracy recognition of a large variety of Named Entity types

QUANTITY	55	ORGANIZATION	15	PRICE	3
NUMBER	45	AUTHORED WORK	11	SCIENCE NAME	2
DATE	35	PRODUCT	11	ACRONYM	1
PERSON	31	CONTINENT	5	ADDRESS	1
COUNTRY	21	PROVINCE	5	ALPHABET	1
OTHER LOCATIONS	19	QUOTE	5	URI	1
CITY	19	UNIVERSITY	3		

Semantics and Reasoning for QA: Predicate-argument structure

- **Q336:** *When was Microsoft established?*
- This question is **difficult** because Microsoft tends to establish lots of things…

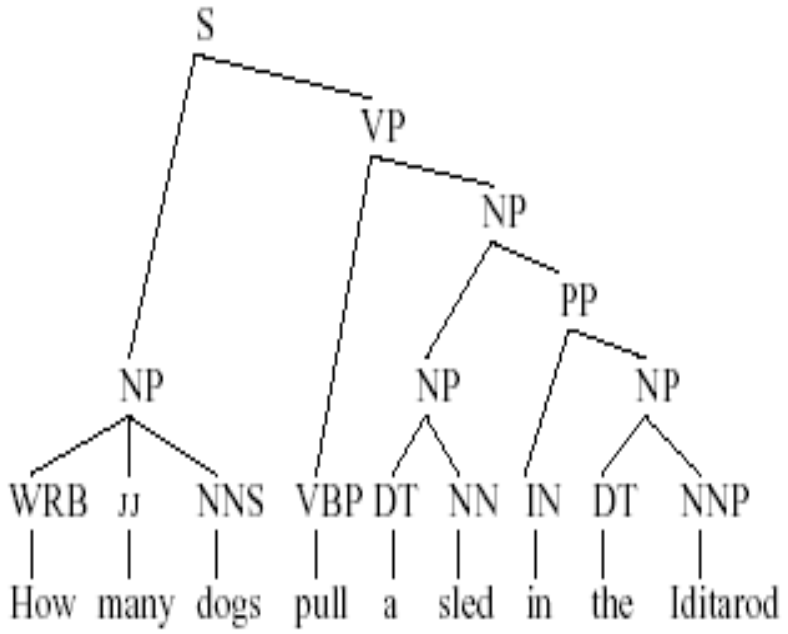
  Microsoft plans to establish manufacturing partnerships in Brazil and Mexico in May.

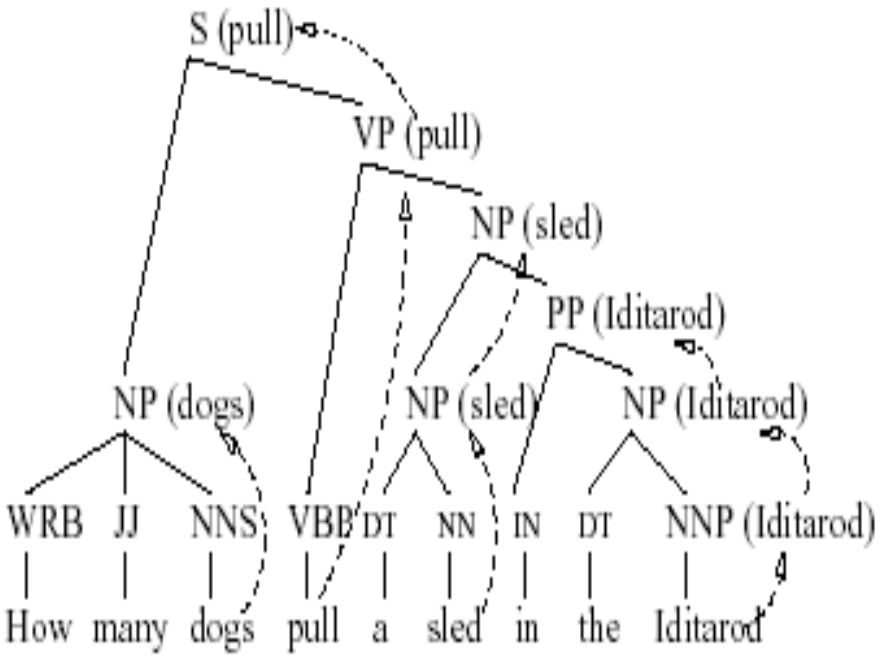
- Need to be able to detect sentences in which 'Microsoft' is **object** of 'establish' or close synonym.
- Matching sentence:

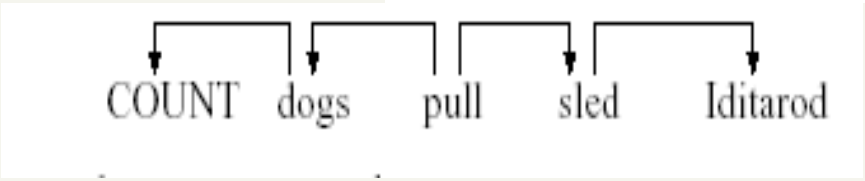
  Microsoft Corp was founded in the US in 1975, incorporated in 1981, and established in the UK in 1982.

- Requires analysis of sentence syntax/semantics!

Semantics and Reasoning for QA: Syntax to Logical Forms







- Syntactic analysis plus semantic => logical form
- Mapping of question and potential answer LFs to find the best match

Abductive inference

- System attempts inference to justify an answer (often following lexical chains)
- Their inference is a kind of funny middle ground between logic and pattern matching
- But very effective: 30% improvement
- *Q: When was the internal combustion engine invented?*
- *A: The first internal-combustion engine was built in 1867.*
- invent → create_mentally → create → build

Question Answering Example

Q: How hot does the inside of an active volcano get?

- get(TEMPERATURE, inside(volcano(active)))

A: "lava fragments belched out of the mountain were as hot as 300 degrees Fahrenheit"

- fragments(lava, TEMPERATURE(degrees(300)), belched(out, mountain))
  - volcano ISA mountain
  - lava ISPARTOF volcano
  - lava inside volcano
  - fragments of lava HAVEPROPERTIESOF lava
- The needed semantic information is in WordNet definitions, and was successfully translated into a form that was used for rough 'proofs'

Answer Validation motivates the Robust Textual Inference Task

- The task: Can systems correctly perform 'local textual inferences' [individual inference steps]?
  - On the assumption that some piece of text (T) is true, does this imply the truth of some other hypothesis text (H)?
    - *Sydney was the host city of the 2000 Olympics* →
    - *The Olympics have been held in Sydney* TRUE
  - The format could be used for evaluating extended inferential chains or knowledge
    - But, in practice, fairly direct stuff

The textual inference task

- Does text *T* justify an inference to hypothesis *H*?
  - Emphasis on variability of linguistic expression
- Robust, accurate textual inference would enable:
  - Semantic search: H: *lobbyists attempting to bribe U.S. legislators*

    T: *The A.P. named two more senators who received contributions engineered by lobbyist Jack Abramoff in return for political favors.*
  - Question answering: H: *Who bought J.D. Edwards?*

    T: *Thanks to its recent acquisition of J.D. Edwards, Oracle will soon be able…*
  - Customer email response
  - Relation extraction (database building)
  - Document summarization

Natural Examples: Reading Comprehension

- **(CNN Student News) -- *January 24, 2006***
- **Answer the following questions about today's featured news stories. Write your answers in the space provided.**
- 1. Where is the country of Somalia located? What ocean borders this country?

- 2. Why did crew members from the USS Winston S. Churchill recently stop a small vessel off the coast of Somalia? What action did the crew of the Churchill take?

Verification of terms [Dan Roth]

- Non-disclosure Agreement

  WHEREAS Recipient is desirous of obtaining said confidential information for purposes of evaluation thereof and as a basis for further discussions with Owner regarding assistance with development of the confidential information for the benefit of Owner or for the mutual benefit of Owner and Recipient;

  THEREFORE, Recipient hereby agrees to receive the information in confidence and to treat it as confidential for all purposes. Recipient will not divulge or use in any manner any of said confidential information unless by written consent from Owner, and Recipient will use at least the same efforts it regularly employs for its own confidential information to avoid disclosure to others.

  Provided, however, that this obligation to treat information confidentially will not apply to any information already in Recipient's possession or to any information that is generally available to the public or becomes generally available through no act or influence of Recipient. Recipient will inform Owner of the public nature or Recipient's possession of the information without delay after Owner's disclosure thereof or will be stopped from asserting such as defense to remedy under this agreement.

  Each party acknowledges that all of the disclosing party's Confidential Information is owned solely by the disclosing party (or its licensors and/or other vendors) and that the unauthorized disclosure or use of such Confidential Information would cause irreparable harm and significant injury, the degree of which may be difficult to ascertain. Accordingly, each party agrees that the disclosing party will have the right to obtain an immediate injunction enjoining any breach of this Agreement, as well as the right to pursueany and all other rights and remedies available at law or in equity for such a breach.

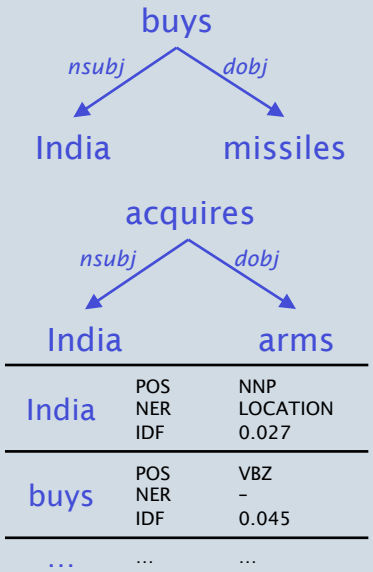
  Recipient will exercise its best efforts to conduct its evaluation within a reasonable time after Owner's disclosure and will provide Owner with its assessment thereof without delay. Recipient will return all information, including all copies thereof, to Owner upon request. This agreement shall remain in effect for ten years after the date of it's execution, and it shall be construed under the laws of the State of Texas.

- Conditions I care about:
  - All information discussed is freely shareable unless other party indicates in advance that it is confidential
  - TRUE? FALSE?

Stanford system three-stage architecture [MacCartney et al. 2006]

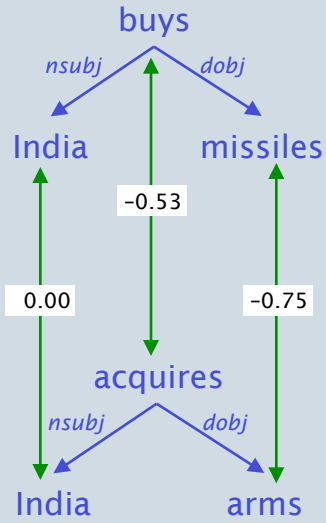
T: *India buys missiles.* ⊨
H: *India acquires arms.*

1. linguistic analysis



India	POS NER IDF	NNP LOCATION 0.027
buys	POS NER IDF	VBZ – 0.045
…	…	…

2. graph alignment



-1.28

3. features & classification

Feature	f_i	w_i
Structure match	+	0.10
Alignment: good	+	0.30

score = $\sum_i w_i f_i$ = -0.88

tuned threshold

yes

no

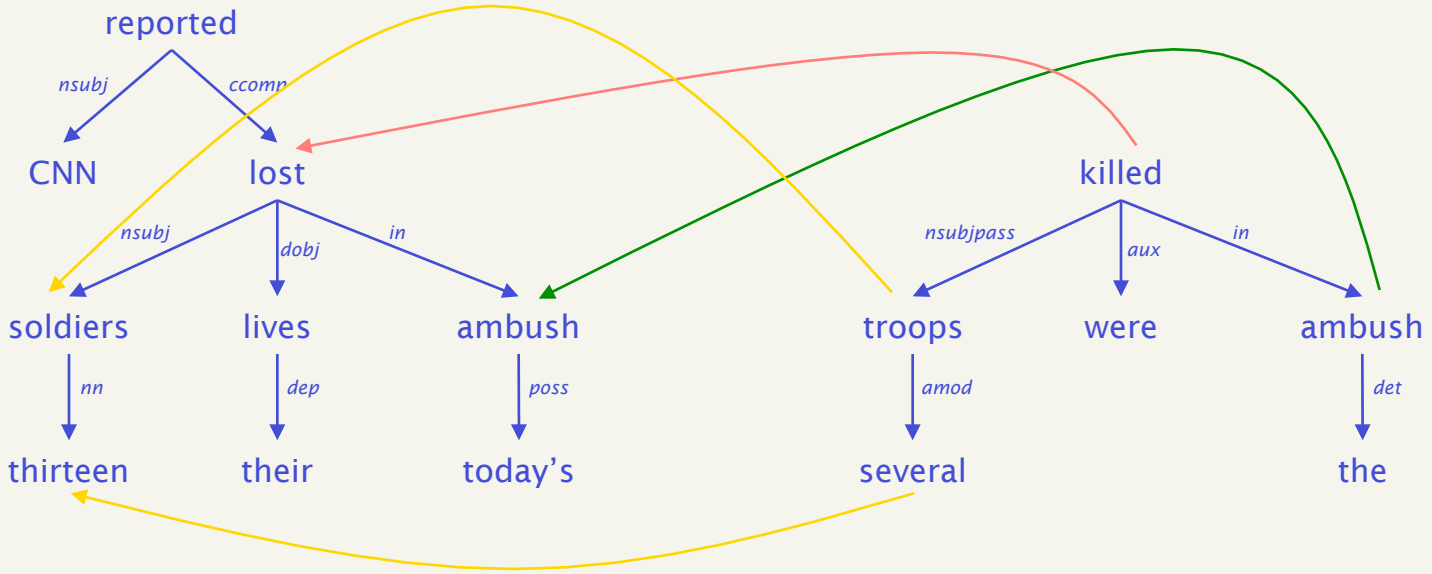
Textual inference as graph alignment

[Haghighi et al. 05, de Salvo Braz et al. 05]

T: *CNN reported that thirteen soldiers lost their lives in today's ambush.*

$\models$

H: *Several troops were killed in the ambush.*



- Find least cost alignment of H to part of T, using locally decomposable cost model (lexical and structural costs)
- Assumption: good alignment ⇒ valid inference

Why we need sloppy matching (i.e., almost IR-style techniques)

- Passage: *Today's best estimate of giant panda numbers in the wild is about 1,100 individuals living in up to 32 separate populations mostly in China's Sichuan Province, but also in Shaanxi and Gansu provinces.*
- Hypothesis 1: *There are 32 pandas in the wild in China.* (FALSE)
- Hypothesis 2: *There are about 1,100 pandas in the wild in China.* (TRUE)
- We'd like to get this right, but **we just don't have the technology to fully infer** from *best estimate of giant panda numbers in the wild is about 1,100* to *there are about 1,100 pandas in the wild*
  - But with a little bit more than IR, we could do it.

Problem: graph embedding isn't sufficient

- To be tractable, alignment scoring must be local
- But valid inference can hinge on non-local factors:

T1: *The army acknowledged that interrogators had desecrated the Koran.* ⊨

H: *Interrogators desecrated the Koran.*

T2: *Newsweek retracted its report that the army had acknowledged that interrogators had desecrated the Koran.* ⊭

H: *Interrogators desecrated the Koran.*

Features of valid inferences

- After alignment, extract features of inference
  - Look for *global* characteristics of valid and invalid inferences
  - Features embody crude semantic theories
  - Feature categories: adjuncts, modals, quantifiers, implicatives, antonymy, tenses, structure, explicit numbers & dates
  - Alignment score is also an important feature
- Extracted features $\Rightarrow$ statistical model $\Rightarrow$ score
  - Can learn feature weights using logistic regression
  - Or, can use hand-tuned weights
- (Score $\geq$ threshold) ? $\Rightarrow$ prediction: yes/no
  - Threshold can be tuned

Structural (mis-)match features

T: *Ahmadinejad attacked the "threat" to bring the issue of Iran's nuclear activity to the UN Security Council by the US, France, Britain and Germany.*

H: *Ahmadinejad attacked the UN Security Council*. (FALSE)

We check particularly the main predicate of the hypothesis and its match in the text to try and assess compatibility using syntactic grammatical relations:

Object of *attack* in hypothesis is not related to object of *attack* in text

Factives & other implicatives

T: *Libya has tried, with limited success, to develop its own indigenous missile, and to extend the range of its aging SCUD force for many years under the Al Fatah and other missile programs.*

H: *Libya has developed its own domestic missile program.* (FALSE)

T: *Scientists have discovered that drinking tea protects against heart disease by improving the function of the artery walls.*

H: *Tea protects from some disease.* (TRUE)

- Evaluate governing verbs for implicativity
  - Unknown: *say*, *tell*, *suspect*, *try*, …
  - Fact: *know*, *wonderful*, …
  - True: *manage to*, …
  - False: *doubtful*, *misbelieve*, …
- Need to check for negative context

Restrictive adjuncts

- We can check whether adding/dropping restrictive adjuncts is licensed relative to upward and downward entailing contexts

  - In all, Zerich bought $422 million worth of oil from Iraq, according to the Volcker committee
  - $\not\models$ Zerich bought oil from Iraq during the embargo

  - Zerich didn't buy any oil from Iraq, according to the Volcker committee
  - $\models$ Zerich didn't buy oil from Iraq during the embargo

QA beyond TREC

- Answers to complex questions that require a longer answer
  - *What is a PIC Freeze?*
  - *Can I travel with Ameripass in Mexico?*
- Soricut and Brill 2006
  - Use the web (real FAQ websites)
- Otterbacher *et al*. 2005
  - Random walk model similar to PageRank
- Daume and Marcu 2006
  - Formal model for query expansion

Not all problems are solved yet!

- Where do lobsters like to live?
  - on a Canadian airline
- Where are zebras most likely found?
  - near dumps
  - in the dictionary
- Why can't ostriches fly?
  - Because of American economic sanctions
- What's the population of Mexico?
  - Three
- What can trigger an allergic reaction?
  - ..something that can *trigger* an allergic reaction

References

R. F. Simmons, Natural language question-answering systems: 1969. Communications of the ACM. Volume 13, 1970.

M. Banko, E. Brill, S. Dumais, and J. Lin. 202. AskMSR: Question Answering Using the Worldwide Web.

In Proceedings of 2002 AAAI SYMPOSIUM on Mining Answers from Text and Knowledge Bases, March 2002

S. Dumais, M. Banko, E. Brill, J. Lin, A. Ng. 2002. Web Question Answering: Is More Always Better? SIGIR 2002.

D. Ravichandran and E.H. Hovy. 2002. Learning Surface Patterns for a Question Answering System. ACL conference, July 2002.

M. Pasca and S. Harabagiu. 2001. High Performance Question/Answering. In *ACM SIGIR-2001*, New Orleans LA, pages 366-374.

References

L. Hirschman, M. Light, E. Breck and J. Burger. *Deep Read: A Reading Comprehension System*. In ACL 1999.

M. Light, G. Mann, E. Riloff and E. Breck. *Analyses for Elucidating Current Question Answering Technology.* Journal of Natural Language Engineering, Vol. 7, No. 4 (2001).

M. M. Soubbotin. *Patterns of Potential Answer Expressions as Clues to the Right Answer.* TREC 2001.

H. Daume and D. Marcu. Bayesian Query-Focused Summarization. ACL 2006

J. Otterbacher, G. Erkan, and D. Radev. Using random walks for question-focused sentence retrieval. In *Proceedings of HLT-EMNLP*, 2005

R. Soricut and E. Brill. 2006. Automatic Question Answering Using the Web: Beyond the Factoid. Journal of Information Retrieval - Special Issue on Web Information Retrieval, 9:191-206