COMP90042 LECTURE 12

QUESTION ANSWERING

WHAT IS QUESTION ANSWERING?

- The task of automatically determining an answer for a user-provided question
- Focus in the field (and this lecture) is on *factoid* QA
  - *Question:* Who said 'you will know a word by the company it keeps'?
  - *Answer: Firth*
- But there are other kinds of QA, e.g.
  - "episodic" QA: tell a story, ask a question about it
  - "community" QA: match a new question to an existing question on QA websites (StackExchange)

WHY DO QA?

- Why do people create QA systems?
  - Because we want quick access to specific information
  - More human-friendly than traditional web search
- Why are we learning about QA?
  - It's a richly challenging text processing task
  - Related to nearly every major topic in this class
  - The subject of the final project

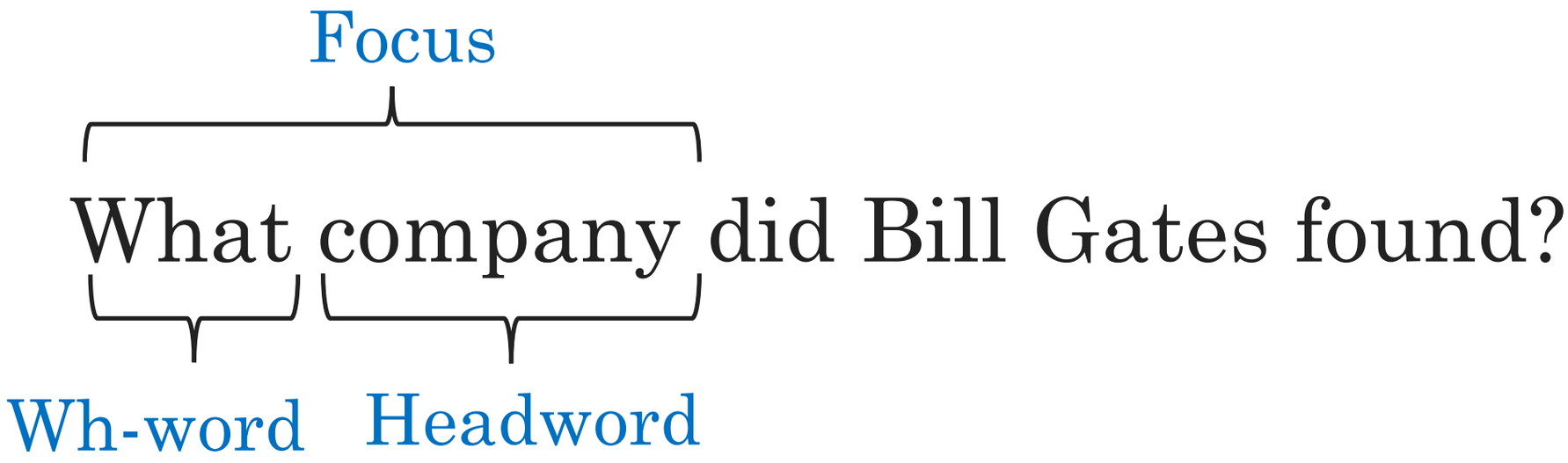
HOW DO WE DO QA?

- QA as mapping of question to a knowledge-base query
- QA as information retrieval
- QA as information extraction
- QA as sequential deep learning

QA SCOPE

- Restricted-domain
  - E.g. LUNAR, 50 year-old system for asking about lunar rock samples
- Open-domain
  - E.g. IBM Watson, modern system which won the Jeopardy! challenge

ANATOMY OF A FACTOID QUESTION



- **Focus** is the part of the question that "stands in" for the answer
  - Will usually disappear a full correct answer (*Bill Gates founded Microsoft*)
- Some ***Wh*-words** (*Who*, *where*, *how*, *when*) give general information about the type of answer required
- For *what* and *which*, the **headword** gives more info

LOGICAL FORMS

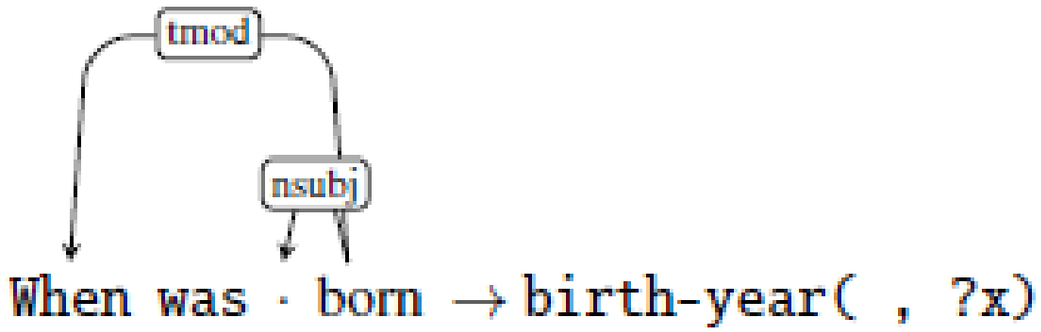
- In KB-driven QA, goal is to convert questions to a **logical form** appropriate for database query (semantic parsing)
- Logical forms typically include
  - Named entities, e.g. BILL_GATES
  - Predicates, e.g. founder(·, ·)
  - Variables, e.g. ?x, ?y
  - Logical connectives (^, v, ¬) and quantifiers (∀, ∃)

What company did Bill Gates found? →

?x s.t. founder(BILL_GATES, ?x) ^ company(?x)

SEMANTIC PARSING

- For simple questions in closed-domains, rules work well
- With supervised data, can learn general mappings between dependency parse fragments and logical forms



- Problem: natural language is too variable
  - learn how to paraphrase using machine translation; or
  - align existing KBs with results of large-scale unsupervised relation extraction

OPEN INFORMATION EXTRACTION

Bill Gates is a founder of Microsoft, a software company based in Redmond, Washington.

< *Bill Gates*, *be a founder of* , *Microsoft* >
< *Microsoft*, *based in*, *Redmond, Washington*>

- Purely unsupervised
- Use POS regex (chunking) and normalization
- Main difficulty: Lots of junk!
- After entity linking, different possible realisations can be clustered together by known relations in database

SOME HARDER (TO PARSE) QUESTIONS

What is the city where the Eiffel Tower is located? →

located-in(EIFFEL_TOWER,?x) ^ *city*(?x)

What kind of mammal lays eggs? →

Is-a(?x, MAMMAL)^egg-laying(?x)

When was the Magna Carta signed? →

year(MAGNA_CARTA_SIGNING, ?x)

How many countries are there in the United Nations? →

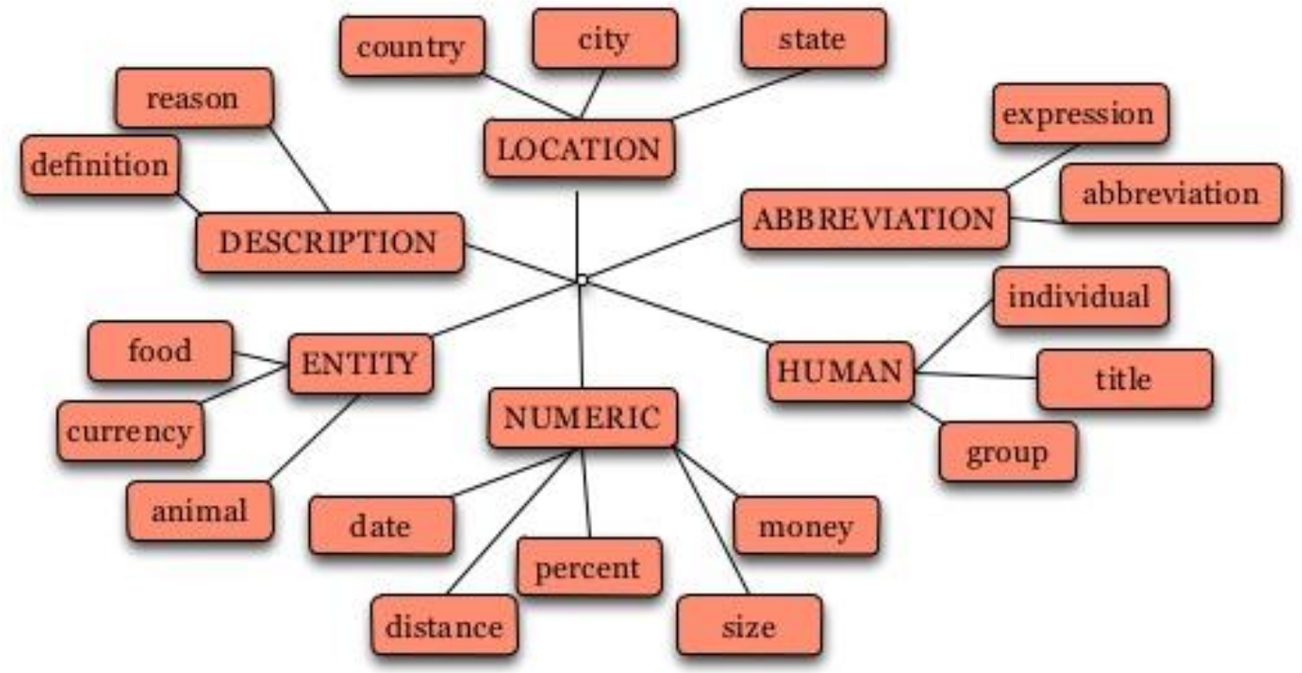
Count(∀ ?x s.t. member-of(?x, UNITED_NATIONS))

QUERY FORMULATION FOR IR QA

- Convert the question to information retrieval query
  - Used when searching for the answer in a text corpus
- Remove stopwords
  - Including *wh*-word and common verbs
- Query expansion using morphological variants, synonyms, or similar words
  - E.g. for *found*, include *founder*, *start a company*
- Re-order and rephrase so query structured like declarative answer
  - E.g. "Where is X?" → "X is in/at/on"

ANSWER TYPE RECOGNITION

- Select the best answer type from an answer type taxonomy, e.g.



- Rule-based classification using *wh*-word/headword question pattern
- Supervised classification with BOW, POS, NE, and WordNet information

ANSWER TYPE EXAMPLES

What kind of cheese is most common on pizza?

Answer type: food

What's the capital of Nepal?

Answer type: city

How long did World War 2 last?

Answer type: duration

Which company was founded by Bill Gates?

Answer type: group/organization

What does perspicacity mean?

Answer type: definition

PASSAGE RETRIEVAL

- Information retrieval of large documents of limited use for Factoid QA
  - Too much reading involved
  - Highly relevant documents might not contain answer
- Passage retrieval: IR on small documents within a larger text collection
  - E.g. section, paragraph, sentence
  - Can be done by unsupervised VSM model
  - But often considered a supervised *ranking* task

FEATURES FOR PASSAGE RETRIEVAL

- Count of named entities corresponding to answer type
- Count of query words
- Longest overlapping sequence of query words
- Proximity of query words to each other
- N-gram overlap between original question and passage

ANSWER EXTRACTION

- Possibly trivial, if only one named entity of correct type in relevant passage
- But sometimes multiple entities
- Or answer isn't a (named) entity at all, e.g. definition
- Harder cases can be addressed with regex patterns

What is Microsoft?

"he found a job with Microsoft, a software company best known for creating the Windows operating system."

<QP>, a <AP>

FEATURES FOR ANSWER EXTRACTION

- Matches answer type
- Matches regex pattern
- Number of matched question keywords
- Contains novel words
- Distance from keywords
  - Word distance
  - Syntactic distance using parse
- Followed by punctuation

OPEN-DOMAIN CORPORA FOR QA

- TREC (IR shared task)
  - Series of Factoid QA datasets with guaranteed answer in corpus of newswire
- WEBQUESTIONS
  - ~6k constrained questions asked by netizens, hand-written answers drawn from Freebase page
- SQuAD
  - Latest and greatest in QA, reading comprehension based on paragraphs inWikipedia, annotated by crowdsourcing 100000+ question answer pairs

MEAN RECIPROCAL RANK

- Reciprocal rank: inverse rank of the first correct answer
  - Zero if correct answer not ranked
  - Averaged across all test instances
- Only appropriate for ranking models
- Gives partial credit for near-misses
- In non-ranking situations, f-score most common

END-TO-END REAL QA SYSTEM: WATSON

- Watson beat Jeopardy! grandmaster in 2011
- In Jeopardy! question and answer are reversed
  - "Questions" tend to be simple, no definitions/descriptions
  - But otherwise extremely open-domain

Robert Redford and Paul Newman starred in this depression-ear grifter flick.

What is *The String?*

QUESTION PROCESSING

- Preprocessing: Parsing and NER
- Identify focus, e.g. *this depression-era grifter flick*
- Relation extraction, e.g. starred-in(Robert_Redford, focus)
- Identify answer type(s)
  - Watson has some 5000 total answer types!
  - Often identified based on headword of focus (e.g. *flick*)
  - Or the category of the clue
  - But may require co-reference resolution

He was a bank clerk in the Yukon before he published "Songs of a Sourdough" in 1907.

CANDIDATE ANSWER GENERATION

- For extracted relations (logical forms), query databases like IMBD or DBpedia to find potential answers
  - I.e. find all movies starring Robert Redford
- Search text collections with weighted query
  - For Wikipedia, take titles of high ranked documents
  - For other texts, identify anchor texts or Wikipedia document titles

ANSWER SELECTION

- Look at overlap of words in question with texts which contain possible answers
- Use WordNet to see if candidate is a potential instance of identified answer type
- Encourage temporal consistency
- Remove redundant candidates
- Apply logistic regression classifier to choose final answer
  - Rank using output probability

A FINAL WORD

- QA a complex problem: many approaches, many steps
- Requires a system with full linguistic competence: morphology, syntax, semantics, and discourse
- More on discourse next week…

FURTHER READING

- J&M3, Ch 28