Answering Multiple Questions on a Topic from Heterogeneous Resources

MIT Computer Science and Artificial Intelligence Laboratory

Boris Katz, Matthew Bilotti, Sue Felshin, Aaron Fernandes, Wesley Hildebrandt
Roni Katzir, Jimmy Lin, Daniel Loreto, Gregory Marton, Federico Mora, Ozlem Uzuner

Project Goals

Answering in Context

- Topic analysis: separating name from restrictive modifiers
  "philanthropist Alberto Vilar"
- Reference resolution in questions
  "What is Vilar's nationality?"
  "What is his nationality?"
  "What is the philanthropist's nationality?"
  ⟹ "What is Alberto Vilar's nationality?"
  "Name other famous Americans with Cuban ancestry."
  ⟹ use "Alberto Vilar" in retrieval
- Duplicate removal through diversity-based reranking

Expanding Knowledge Sources

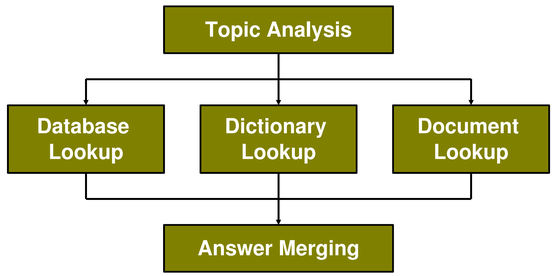
- Factoid: added Teoma search engine for Knowledge Mining
- List: 20 times more expected answer type lists than last year
- Definition:
  - new patterns give greater variety of information
  - can match multiple adjacent patterns
  - added simple anaphora resolution

Results

Task	best	CSAIL04-1	CSAIL04-2	median
Factoid	0.770	0.313	0.313	0.170
List	0.622	**0.119**	0.113	0.094
Definition	0.460	0.184	**0.186**	0.184
Total		0.232	0.231	

	CSAIL04-1	CSAIL04-2
List Questions:		
How reliable is answer type filtering? Include top 25% of guesses?	always guess	guess only if no list exists for answer type
Definition Questions:		
Reference Resolution:	none	names and personal pronouns

Definition Questions



Database Lookup

Using surface patterns, pre-compile every knowledge nugget about every entity into a database. Sample patterns:

copular: A **fractal** is a pattern that is irregular, but self-similar at all scales
appositive: The **Aga Khan**, Spiritual Leader of the Ismaili Muslims,
occupation: singer-choreographer **Fred Durst**
became: **Althea Gibson** became the first black tennis player to win Wimbledon

relative clause: **Dean**, who died at age 24 in a 1955 car crash, is...
such as: tested various nutrients such as **calcium**, **magnesium**, and **potassium**
also called: established a non-governmental organization called **Global Village**
age: **Adams**, 30, a convicted murderer, was fatally shot...

Multiple patterns: **Adams**, 30, a convicted murderer, was fatally shot...
Pronominal anaphora: **He [Franz Kafka]** *was* one of the best-known Czech authors

Answers are sentences containing matched patterns.

Dictionary Lookup

Project webster.com definition onto AQUAINT documents:

What is the vagus nerve?
Dictionary definition: either of the 10th pair of cranial nerves that arise from the medulla and supply chiefly the viscera…
Projected answer: The vagus nerve is sometimes called the 10th cranial nerve. It runs from the brain…

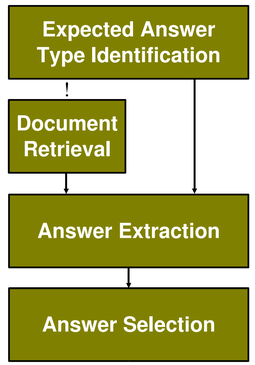
Document Lookup

Use topic as a search query, and retain sentences containing it, subject to passage-retrieval scoring

Answer Merging

Order nuggets by
- quality of **topic** match,
- quality of the definition's **source,**
- **novelty** of answer contents over previous definitions and all question answer pairs for the topic.

List Questions



Select known lists whose noun-phrase annotations match the question focus

Fetch relevant document chunks from the AQUAINT corpus

Extract exact answers from candidate document chunks

Remove near-duplicate answers

Select best answers and guesses

Expected Answer Types

Guess answer candidates are generic (named entity, noun phrase)
List answer candidates come from pre-compiled categories

A list is chosen if its noun-phrase annotation matches the question focus.

New List Sources

3300+ fixed lists with annotations, up from 150+ last year!

From AQUAINT+WordNet:
Category names are taken from "What *X*" context
WordNet hyponyms projected onto the corpus.
"Italian region": Abruzzi, {Basilicata,Lucania}, Calabria,...

From WorldBook Encyclopedia first sentence:
Sydney Hillman, an American labor leader, was the guiding spirit of the Amalgamated Clothing Workers of America.
"labor leader"/*"leader"*: Sidney Hillman, Leonard Woodcock,...

From online resources, manually compiled:
`http://www.wirelessadvisor.com/hardware.cfm`
"cell phone manufacturers": Nokia, Samsung, Motorola,...

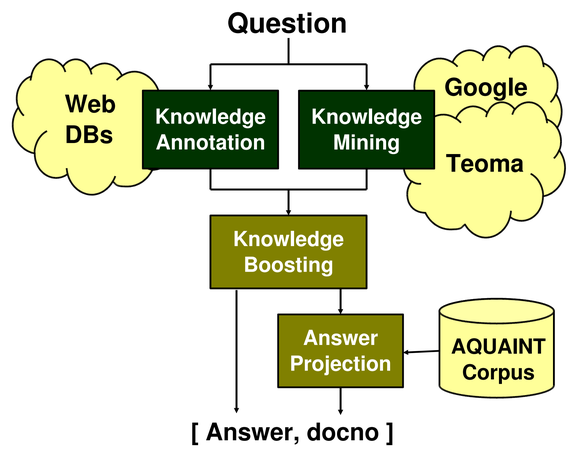
Answer Extraction

Select all List candidates from top-ranked retrieved passages

Condition 1:
Use top 25% of generic "guess" answers for every question

Condition 2:
Use top 25% "guess" answers only if no list answer type was identified.

Factoid Questions



The Aranea System

Knowledge Annotation: Make use of freely available structured and semi-structured knowledge on the Web using wrapper and database techniques.
"When was James Dean born?"
⟹ 8 February, 1931 (IMDB.com)

Knowledge Mining: Leverage the tremendous amount of free text available on the Web with redundancy-based statistical techniques.
Google top 100 document titles and snippets
New this year: Teoma top 100 titles and snippets
`http://www.teoma.com/`

Knowledge Boosting: Heuristics for filtering out bad answers and promoting good answers.

Answer Projection

- Retrieve candidate documents using the question.
- Score documents using the candidate answers in passage-retrieval techniques.
- Select the highest ranked document as the supporting document.