Modeling of the Question Answering Task in the YodaQA System

Petr Baudiš and Jan Šedivý baudipet@fel.cvut.cz

Department of Cybernetics, Czech Technical University, Prague

Goal: General approach to answer naturally phrased factoid questions, using both structured and unstructured knowledge bases.

Contribution: A universal framework that allows integration of diverse approaches within a common pipeline and easy domain adaptation.

Dataset: We lack a good QA benchmark dataset. Both TREC and WebQuestions have issues. How to do reproducible QA research?

Background

Question Answering

Unstructured user query → narrow text snippet answering the query.

... vs. linked data graph search: requires a precisely structured user query.

... vs. a search engine: returns a whole document or passage.

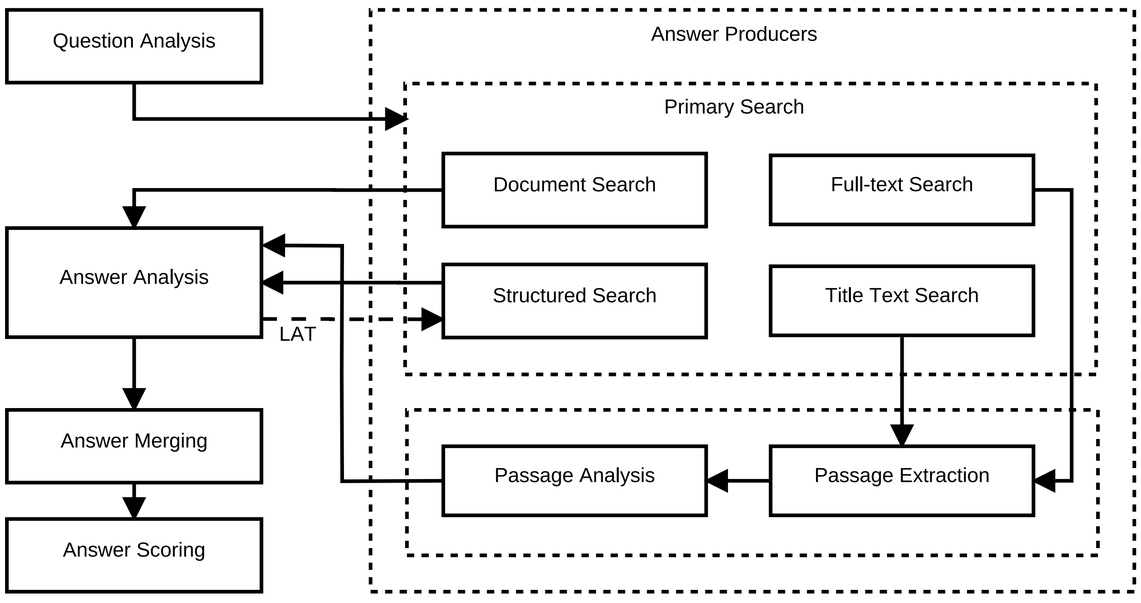
The Question Answering task is already part of e.g. the Google Search interface, and with the high profile IBM Watson Jeopardy! matches it has became a benchmark of progress in AI research.

We emphasize both open domain factoids and domain adaptation.

Previous Work

When querying structured knowledge bases, typically using the RDF paradigm and accessible via SPARQL, the problem can be either semantic parsing from free-text to a logical form (representable by SPARQL or KB-specific template subgraph), or using more fuzzy graph information retrieval.

When relying on unstructured knowledge bases, some offload the information retrieval on an external web search engine; we avoid this to keep domain flexibility and reproducibility of results.



BioASQ Challenge

Participated in **BIOASQ Task 3B phase B** (QA with information retrieval already performed). *Ideal answers* and *yes/no questions* not implemented.

Eventually, only very limited time was available for participation. We at least display baseline performance of general QA system with minimal domain adaptation:

- Questions with imperative words.
- Replaced default IR with input data.
- LAT also using GeneOntology.

Pipeline	AP Rcl.	Prec@1	MRR
BioASQ final	33.0%	10.0%	0.132
w/o G.O.	33.0%	8.0%	0.120
w/o G.O., CRF	33.0%	5.5%	0.114
w/o yes/no q.	43.5%	10.1%	0.148

Text	List causative genes for autosomal recessive forms of monogenic Parkinson's disease
Q. Analysis	**Focus:** genes; **LAT:** gene (by Wordnet hypernym: *sequence*, ..., group)
Clues	causative genes, autosomal recessive forms, monogenic Parkinson's disease
Snippets	Mutations in the Parkin gene (PARK2) are the major cause of autosomal recessive early-onset parkinsonism. The gene responsible for AR-JP was recently identified and designated parkin.
PARK2	DBp. LAT *protein*, GeneOnt LAT *gene*, *protein*, *gene product* Successful, "sharp"**!** type coercion match**!** **occurences:** 3, **origins:** noun phrase, **other:** adjecent to a clue mention, LAT from DBpedia and G.O. **!**
mutations	no LAT**!**; **occurences:** 8, **origins** noun phrase, **other:** adjecent$^+$ to clue mentions.
Final Answers	PINK1, ..., PARK2 (0.99), PD, mutations (0.93), we (0.8), AR-JP, ...

Text	What is the name of the famous dogsledding race held each year in Alaska?
Q. Analysis	**Focus:** name; **SV:** held; **LAT:** race (by Wordnet hypernym: *contest*, *event*, *canal*, ...)
Clues	name, Alaska **(concept clues)**, race, held, famous, dogsledding, year
DBpOnt. Concepts	**area:** 1717854.0, **country:** United States *enwiki* Alaska, Name **Sample picked passages:** Various races are held around the state, but the best known is the Iditarod Trail Sled Dog Race, a 1150 mi trail from Anchorage to Nome ...
Fulltext Titles	List of New Hampshire historical markers Name of the Year, Danish Sports N. of the Y., List of organisms named after famous people, Alaska!, Alaska, Race of a Thousand Years
2000 Race of T. Y.	DBpedia LAT *automobile race*, *auto race in australia*, *new year celebration*, *quantity* LAT "sharp" (exact specific)**!** TyCor match**!** **occurences:** 1, **origins:** first passage, NP, **other:** near a clue mention**!**, clue text inside
Iditarod Trail Race	DBp. LAT *sport*, *sport in alaska*, *alaska*, *winter sport*, *attraction*; (not *race*) Successful tycor. match, loose match by generalization of *attraction* to *social event***!** **occurences:** 1, **origins** passage by various clues, noun phrase, **other:** suff. by clue text
Final Answers	The 2000 Race of a Thousand Years (0.97), –01-03 (0.94), List of New Hampshire historical markers (0.93), a binomial name, a "make" (manufacturer) and a "model", in addition to a model year, such as a 2007 Chevrolet Corvette (0.90), the Iditarod Trail Sled Dog Race (0.89), Various races (0.83)

Ask for a live demo! (live.ailao.eu)
Completely open source! (github.com/brmson)

The YodaQA Framework

Paradigm: We build a portfolio of representations, answer sources and scoring features, but avoiding hand-crafted heuristics.

Platform: Mainly Java, using the Apache UIMA framework and DKpro family of adapters to various NLP tools.

The Baseline QA Pipeline

Question Analysis

- **Focus**, **LAT** (Lexical Answer Type)
  - What was the first book written by Terry Pratchett?
  - The actor starring in Moon?
  - Where is Mount Olympus? location
- **Clues** (search keywords/phrases)
  - POS and constituent token whitelist
  - Named entities
  - **Concepts:** `enwiki` article titles (entity linking task)

Outcome: Question representation

Answer Production

- Passage-yielding enwiki search
  - *Fulltext:* passages containing clues
  - *Title-in-clue:* initial passage
  - Answers from NEs and NPs, as well as alignment CRF sequence tagging model
  - Document titles may be answers
- Structured search (DBpedia, Freebase), all triplet objects of concepts are answers; also, multi-label cfier estimates specific property paths based on question bag-of-words

Outcome: Set of candidate answers

Answer Analysis

- **LAT:** NE type, DBpedia concept type, WordNet relations, numerical hypernymy distance
- **Type coercion** of question and answer LATs: *Unspecificity* is WordNet
- Phrase origin, clue overlaps, LAT kinds, type coercion (⇒ 81 features)

Outcome: Ordered set of Answers

System Evaluation

Benchmark results on the *TREC2001, 2002* test:

System	Precision@1	F_1	MRR
LLCpass03 (hand-crafted system)	68.5%	—	—
OpenEphyra (hand-crafted OSS)	"above 25%"	—	—
JacanaIR (modern fully-learned OSS)	—	23.1%	—
YodaQA v1.1	26.4%	26.4%	0.325

Benchmark results on the *WebQuestions* test:

System	F_1@1	F_1 (Berant)
Sempre	35.7%	35.7%
JacanaFB	35.4%	33.0%
YodaQA v1.1	34.3%	—
STAGG (summer 2015, state-of-art)	—	52.5%

Dataset

- Stable (gold standard valid in 5 years)
- Focused (single question style, clean, without required inference)
- Realistic (typos, complex reasoning)
- Low bias towards any knowledge base
- Mixing questions with independent answering strategies (e.g. yes/no vs. factoid vs. paraphrasing)

We propose: A new hard factoid dataset **curated**, based on manually reviewed TREC 2001+2002 and YodaQA user data, mostly deducible from Wikipedia. 867 questions. ▪ A new easy factoid dataset **moviesC** based on Webquestions and YodaQA user data within the movies domain, answerable via Freebase. 777 questions.

Advantages: All questions should be temporally independent. Reference IR setup with fixed KB versions also available. Single style of questions (factoid).

Current work: Revised datasets for subtasks. Entity linking (!), answering sentence selection, matching subgraphs.

Open problem: Automatic answer verification, when we do not limit just to entity names as answers. The current approach of using regex patterns has many caveats.