YodaQA: A Modular Question Answering System Pipeline

Petr Baudiš baudipet@fel.cvut.cz

Department of Cybernetics, Czech Technical University, Prague

Goal: Answer naturally phrased factoid questions, using both structured (e.g. Freebase) and unstructured (e.g. Wikipedia) knowledge bases.

Contribution: A universal framework that allows integration of diverse state-of-art approaches within a common pipeline.

Keywords: Question answering, information retrieval, information extraction, linked data, natural language processing, Apache UIMA, software engineering.

Ask for a live demo! (live.ailao.eu)

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 the Google Search interface or personal assistants like Apple Siri, and with the high profile IBM Watson Jeopardy! matches it has became a benchmark of progress in AI research.

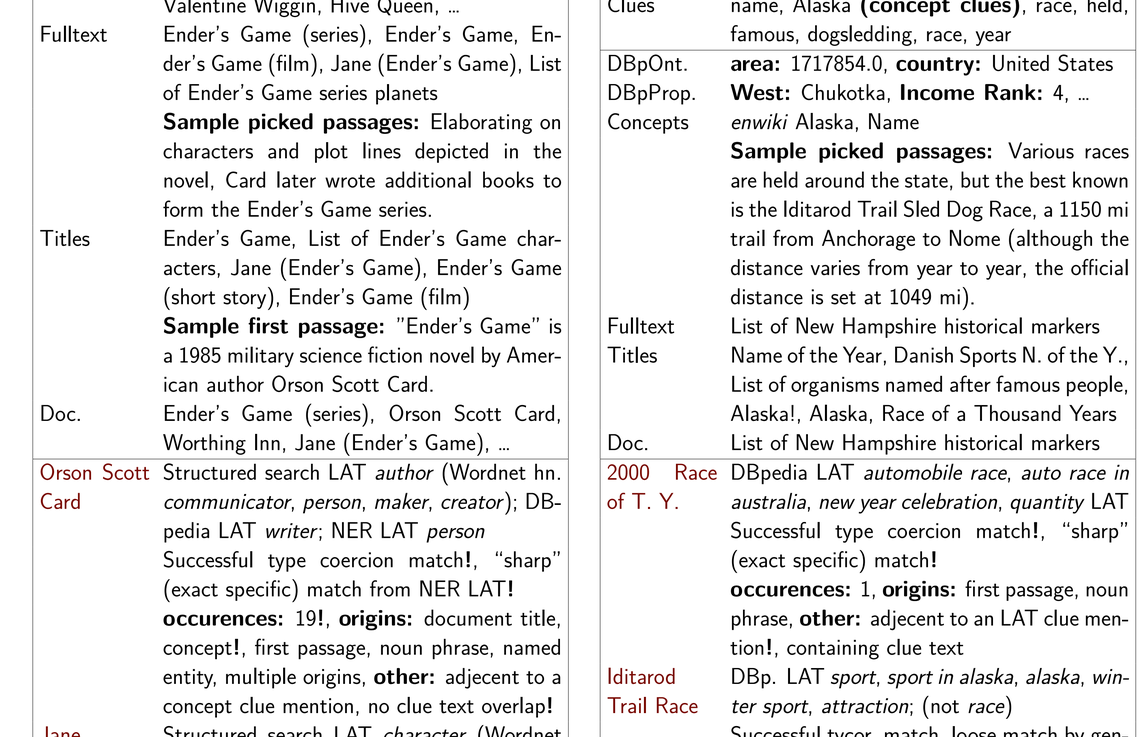
As we are interested in a general purpose QA system, we will consider an "open domain" factoid question answering, rather than domain-specific applications (though we have domain flexibility as one of our goals).

Previous Work

The most popular approach in QA research has been restricting the task to querying structured knowledge bases, typically using the RDF paradigm and accessible via SPARQL. The problem can be then rephrased as machine translation from free-text user query to a structured query (SPARQL, λ-expr).

When relying on unstructured knowledge bases, a common strategy is to offload the information retrieval on an external high-quality web search engine like Google or Bing; we avoid this for the sake of domain flexibility and reproducibility of results.

Notable open source systems: OpenEphyra, OAQA, WatsonSim, Jacana, OpenQA.



Text	Who wrote Ender's Game?
Q. Analysis	**Focus:** who; **SV:** wrote; **LAT:** person
Clues	Ender's Game **(concept clue)**, wrote
DBpOnt.	**author:** Orson Scott Card, **pub.:** Tor Books
Freebase	**Author** Orson Scott Card, **Characters** Valentine Wiggin, Hive Queen, ...
Fulltext	Ender's Game (series), Ender's Game, Ender's Game (film), Jane (Ender's Game), List of Ender's Game series planets **Sample picked passages:** Elaborating on characters and plot lines depicted in the novel, Card later wrote additional books to form the Ender's Game series.
Titles	Ender's Game, List of Ender's Game characters, Jane (Ender's Game), Ender's Game (short story), Ender's Game (film) **Sample first passage:** "Ender's Game" is a 1985 military science fiction novel by American author Orson Scott Card.
Doc.	Ender's Game (series), Orson Scott Card, Worthing Inn, Jane (Ender's Game), ...
Orson Scott Card	Structured search LAT *author* (Wordnet hn. *communicator*, *person*, *maker*, *creator*); DBpedia LAT *writer*; NER LAT *person* Successful type coercion match**!**, "sharp" (exact specific) match from NER LAT**!** **occurences:** 19**!**, **origins:** document title, concept**!**, first passage, noun phrase, named entity, multiple origins, **other:** adjecent to a concept clue mention, no clue text overlap**!**
Jane	Structured search LAT *character* (Wordnet hn. *imaginary being*, *creativity*, *person*, *message* and 36 others); NER LAT *person* Successful type coercion match**!**, "sharp" (exact specific) match from NER LAT**!** **occurences:** 4, **origins** document title, first passage, noun phrase, named entity, multiple origins, **other:** no clue text overlap**!**
Final Answers	Orson Scott Card (0.99), Neal Shusterman (0.96), American author O. S. Card (0.96), List of Ender's Game series planets (0.94), Gavin Hood (0.94), Jane (0.91), ...

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*, *biological group*, *canal* and 9 others)
Clues	name, Alaska **(concept clues)**, race, held, famous, dogsledding, race, year
DBpOnt.	**area:** 1717854.0, **country:** United States
DBpProp.	**West:** Chukotka, **Income Rank:** 4, ...
Concepts	*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 (although the distance varies from year to year, the official distance is set at 1049 mi).
Fulltext	List of New Hampshire historical markers
Titles	Name of the Year, Danish Sports N. of the Y., List of organisms named after famous people, Alaska!, Alaska, Race of a Thousand Years
Doc.	List of New Hampshire historical markers
2000 Race of T. Y.	DBpedia LAT *automobile race*, *auto race in australia*, *new year celebration*, *quantity* LAT Successful type coercion match**!**, "sharp" (exact specific) match**!** **occurences:** 1, **origins:** first passage, noun phrase, **other:** adjecent to an LAT clue mention**!**, containing clue text
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), ...

The YodaQA Framework

Paradigm: We are interested in combining different approaches, using different question representations, answer sources and scoring features. Our baseline is domain flexible and we strongly prefer machine learning to hand-crafted heuristics.

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

Availability: Publicly available free software under the Apache licence at https://github.com/brmson/yodaqa.

The Baseline QA Pipeline

The basic pipeline flow is much inspired by the DeepQA model of IBM Watson. Throughout the flow, answer features are gradually accumulated.

Question Analysis

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

Outcome: Question representation

Answer Production

- Passage-yielding enwiki search
  - *Fulltext:* Full-text and title search, passages containing clues are considered
  - *Title-in-clue:* Title search for clues, initial passage is considered
  - Passages are parsed, NEs and NPs are answers
- Full-text enwiki search for clues, document titles are answers
- Structured search (DBpedia, Freebase), triple objects are answers

Outcome: Set of candidate answers

Answer Analysis

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

Outcome: Ordered set of Answers

Performance Analysis

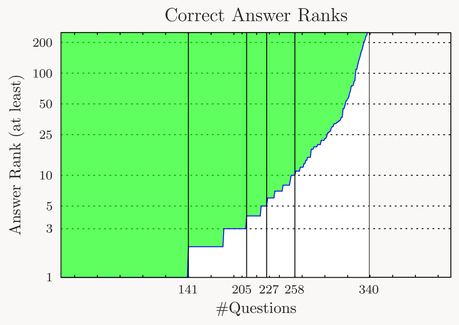
Dataset: 430+430 trivia factoid questions (TREC 2001, 2002 + IRC).

Recall: Whether a correct answer has been generated (with any score)

Accuracy-at-one: Whether the correct answer has been returned as the top answer by the system.

Acc-at-1 32.6%, but Acc-at-5 52.7%

Pipeline	Recall	Acc-at-1	time
default	79.3%	32.6%	28.8s
full-text scaling ($6 \to 12$ results)	82.3%	34.0%	50.0s
¬ type coercion	79.3%	22.1%	30.0s
¬ concept clues	67.9%	23.0%	25.6s



Future Work

- Better, larger dataset
- Insightful web interface
- Real-world domains
- B-I-O answer extraction
- Tree alignment features
- Smarter scoring model
- Question representation
- Text entailment

Acknowledgements

This research is supervised by Dr. Jan Šedivý and Dr. Petr Pošík, and supported by the CTU grant SGS14/194/OHK3/3T/13.