Question Answering using Structured and Unstructured Data

Denis Savenkov
denis.savenkov@emory.edu
Emory University

advisor: Eugene Agichtein
eugene@emory.edu
Emory University

Intelligent Information Access Lab
http://ir.mathcs.emory.edu

Problem

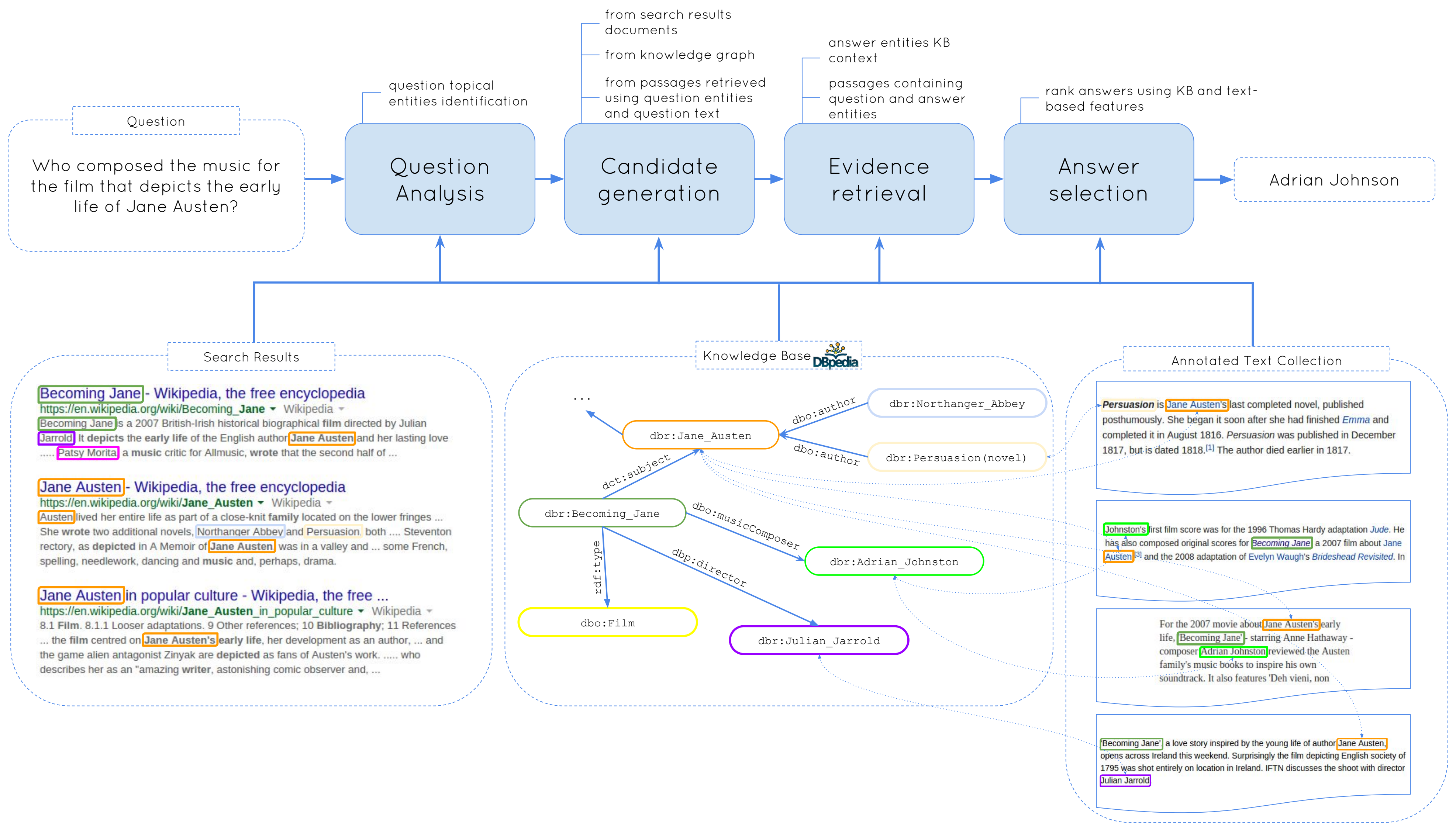
➔ This thesis investigates methods for combining structured and unstructured data for answering a variety of user questions

	Unstructured data	Structured data
factoid questions	*Text collections* + easy to match against question text + cover a variety of different information types - each text phrase encodes a limited amount of information about mentioned entities	*Knowledge Bases (KB)* + aggregate all the information about entities + allow complex queries over this data using special languages (e.g. SPARQL) - hard to translate natural language questions into special query languages - KBs are incomplete (missing entities, facts and properties)
non-factoid questions	*Text collections* + contain relevant information to a big chunk of user needs - hard to match an answer to a question (lexical gap)	*Question-Answer pairs* + easy to find a relevant answer by matching the corresponding questions - cover a smaller subset of user information needs

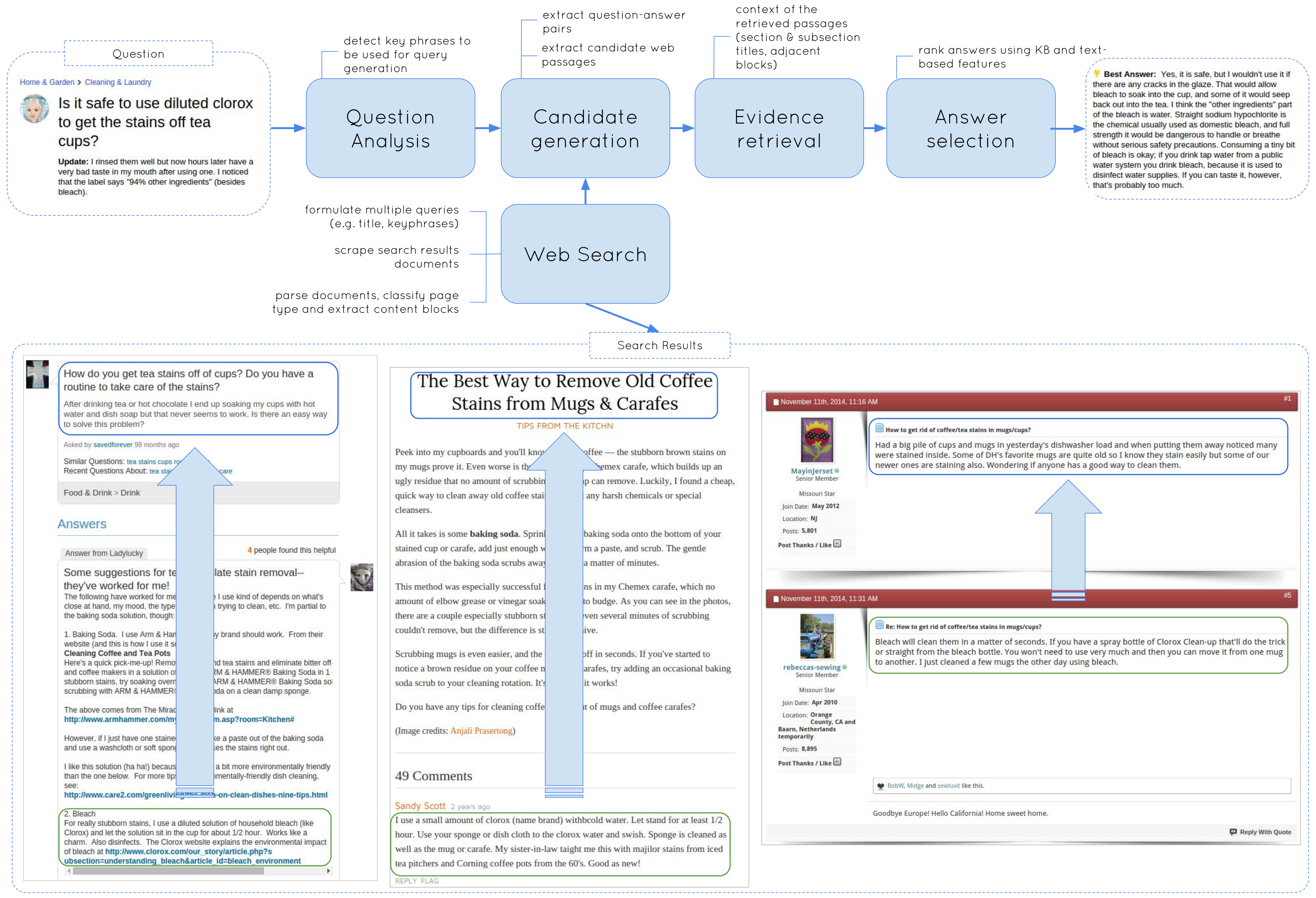
Research Questions

1. What types of questions can be answered using text, KB or a combination of both?
2. How does semantic annotation of unstructured data compare to information extraction for question answering?
   - information extraction for KB construction vs open information extraction vs unstructured data annotation
3. How does a combination of structured and unstructured data sources improve each of the main QA system components: question analysis, candidate generation, evidence extraction and answer selection?

Combination of Text & KB for factoid QA


Question
Who composed the music for the film that depicts the early life of Jane Austen?
question topical entities identification
Question Analysis
from search results documents
from knowledge graph
from passages retrieved using question entities and question text
Candidate generation
answer entities KB context
passages containing question and answer entities
Evidence retrieval
rank answers using KB and text-based features
Answer selection
Adrian Johnson
Search Results
Knowledge Base
DBpedia
...
dbr:Jane_Austen
dbo:author
dbr:Northanger_Abbey
dbo:author
dbr:Persuasion(novel)
dct:subject
dbr:Becoming_Jane
dbo:musicComposer
dbr:Adrian_Johnston
rdf:type
dbo:Film
dbp:director
dbr:Julian_Jarrold
Annotated Text Collection

Web page structure for non-factoid QA


Question
Home & Garden > Cleaning & Laundry
Is it safe to use diluted clorox to get the stains off tea cups?
Update: I rinsed them well but now hours later have a very bad taste in my mouth after using one. I noticed that the label says "94% other ingredients" (besides bleach).
detect key phrases to be used for query generation
Question Analysis
extract question-answer pairs
extract candidate web passages
Candidate generation
context of the retrieved passages (section & subsection titles, adjacent blocks)
Evidence retrieval
rank answers using KB and text-based features
Answer selection
Best Answer: Yes, it is safe, but I wouldn't use it if there are any cracks in the glaze. That would allow bleach to soak into the cup, and some of it would seep back out into the tea. I think the "other ingredients" part of the bleach is water. Straight sodium hypochlorite is the chemical usually used as domestic bleach, and full strength it would be dangerous to handle or breathe without serious safety precautions. Consuming a tiny bit of bleach is okay; if you drink tap water from a public water system you drink bleach, because it is used to disinfect water supplies. If you can taste it, however, that's probably too much.
formulate multiple queries (e.g. title, keyphrases)
scrape search results documents
parse documents, classify page type and extract content blocks
Web Search
Search Results
How do you get tea stains off of cups? Do you have a routine to take care of the stains?
The Best Way to Remove Old Coffee Stains from Mugs & Carafes
TIPS FROM THE KITCHN
How to get rid of coffee/tea stains in mugs/cups?

Proposed Experiments

Factoid questions

➔ Datasets:
- TREC QA [RQ1]
- WebQuestions [RQ1]
- Develop a new dataset [RQ1]
  - ✓ questions derived from Yahoo! Answers WebScope collection
  - ✓ answers are KB entities
  - ✓ heuristic-based filtering: single sentence, no personal pronouns, no comparative and superlative adjectives, not future questions, and no keywords (e.g. recommend, suggest, will, etc)
  - ✓ crowdsourcing for filtering to keep factoid, non-subjective questions and labeling answer entities, based on the answer text provided by a CQA user

➔ Compare against existing text-based, KB, Open IE and hybrid methods [RQ2, RQ3]

➔ Error analysis [RQ1, RQ3]

Non-factoid questions

➔ Datasets:
- TREC LiveQA 2015 and the new 2016 shared task

➔ Comparison with last year system that doesn't use the information on the structure of web pages and other LiveQA participants [RQ3]

Expected Results

1. new factoid question answering dataset
2. new approach to combining unstructured text and structured KB data, that:
   - ✓ improves QA precision
   - ✓ improves recall, including questions that could only be answered with a combination of text and KB data
3. new system for non-factoid question answering with improved performance due to better utilization of the information provided in web documents

Open Questions

→ How to incorporate other available sources of factual information, including semi-structured, e.g. tables, diagrams, etc?

→ Non-factoid questions often include unique context information, that makes reusing information non effective. How a system can generate the answer given all potentially useful extracted information?

→ How to construct KBs for non-factoid information needs?