

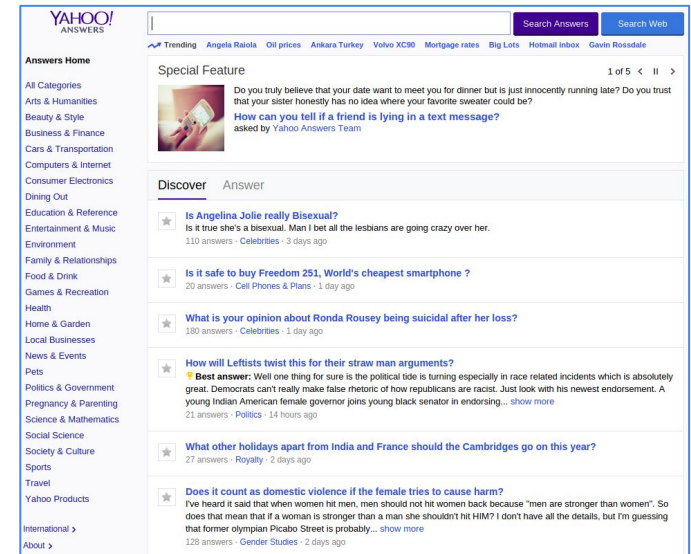
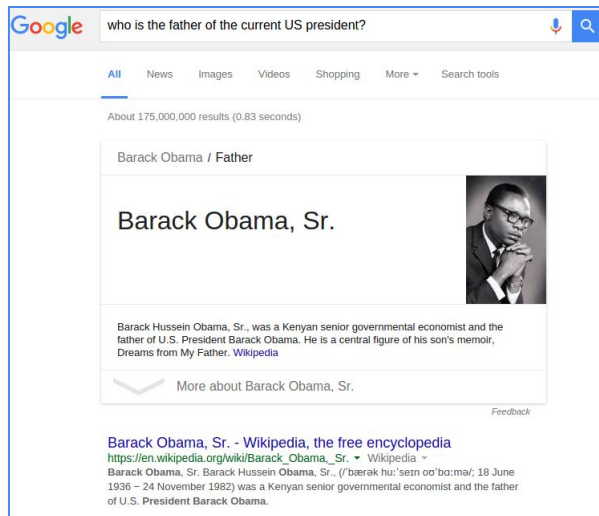
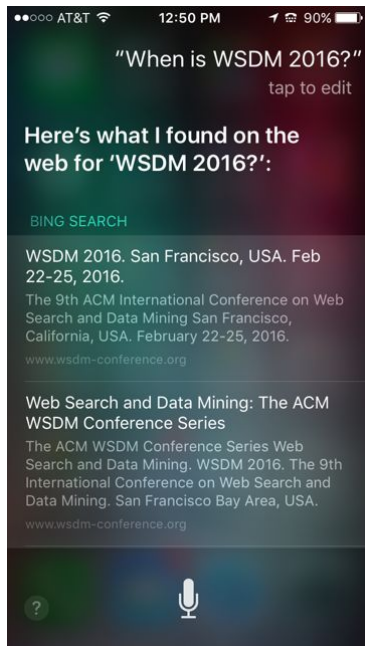
When a Knowledge Base is not Enough

Question Answering over Knowledge Bases with
External Text Data

Denis Savenkov
Emory University
dsavenk@emory.edu

Eugene Agichtein
Emory University
eugene@mathcs.emory.edu

Percentage of question search queries is growing^[1]



[1] "Questions vs. Queries in Informational Search Tasks", Ryen W. White et al, WWW 2015

Automatic Question Answering works relatively well for simple factoid questions

Google how to play go?

About 1,410,000,000 results (0.79 seconds)

The rules

1. A game of Go starts with an empty board. ...
2. Players take turns, placing one of their stones on a vacant point at each turn, with Black playing first.
3. Diagram 1 shows the position at the end of a game on a 9 by 9 board, during which Black captured a white stone at a.

More items...

[How to Play | British Go Association](http://www.britgo.org/intro/intro2.html)
www.britgo.org/intro/intro2.html

[How to Play | British Go Association](http://www.britgo.org)
www.britgo.org > An introduction to Go
The rules: A game of Go starts with an empty board. Players take their stones on a vacant point at each turn, with Black playing first position at the end of a game on a 9 by 9 board, during which Black stone at a.

[How To Play Go - Introduction](http://www.pandanet.co.jp/English/learning_go/learning_go_1.html)
www.pandanet.co.jp/English/learning_go/learning_go_1.html
How to play Go - Introduction to the basic rules of Go.

who was the second person to walk on the moon?

Buzz Aldrin
Engineer

Buzz Aldrin is an American engineer and former second person to walk on the Moon. He was the Apollo 11, the first manned lunar landing in history on 03:15:16 on July 21, 1969, following his brother James A. Lovell and Michael Collins. He is a former U.S. Air Force officer.

[Wikipedia](#) [Twitter](#) [Facebook](#) [LinkedIn](#)

Born: Jan 20, 1930 (age 86) · Glen Ridge, NJ
Height: 5' 10" (1.78 m)
Net worth: \$10 million USD (2016)
Spouse: Lois Aldrin (1988 - 2012) · Beverly Van Dyke (1954 - 1972)
Space missions: Apollo 11 · Gemini 12
Space agency: NASA

WolframAlpha computational knowledge engine

Is it going to rain tomorrow?

Current weather summary for Atlanta, Georgia

Today	Tomorrow	Fri	Sat
overcast wind: E at 7mph humidity: 64% 64°F	64°F 51°F	58°F 47°F	51°F 44°F

Input interpretation: precipitation forecast tomorrow

Result: rain (Thursday, April 14, 2016)

Current forecast:

Wed	Thu	Fri	Sat	Sun
rain	rain	rain	rain	rain
Apr 13	Apr 14	Apr 15	Apr 16	Apr 17

rain: 38.3% (2.8 days)

Precipitation rate (in/h):

Wed	Thu	Fri	Sat	Sun
0.08	0.04	0.02	0.01	0.01
Apr 13	Apr 14	Apr 15	Apr 16	Apr 17



(AP Photo/Jeopardy Productions, Inc.)

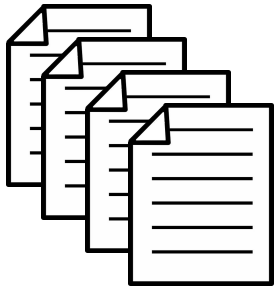
For many questions we still have to dig into “10 blue links”

The image displays four search engine results pages (SERPs) side-by-side, illustrating the concept of "10 blue links" for various queries. Each SERP shows a search bar with the query, navigation tabs (All, News, Images, Videos, Maps, News, Explore), and a list of search results. The queries and their corresponding top results are:

- Query 1:** "what ship did darwin sail on?"
 - Top result: AboutDarwin.com - www.aboutdarwin.com/v... The weather was now quite believe, did a vessel leave squall prevented the ship fr ability to change colors.
- Query 2:** "Which members of the Wu-Tang Clan..."
 - Top result: Wu-Tang Clan - Wikipedia https://en.wikipedia.org/wiki/Wu-Tang_Clan... Enter the Wu-Tang (36 Chambers) group after the film Shaolin and V excerpts from the movie ...
- Query 3:** "Where is the highest point in Japan?"
 - Top result: What is the highest point of Japan? www.answers.com > ... > Count... What is the highest point of e would be Mount Fuji ... Mount O...
- Query 4:** "who did draco malfoy end up marrying?"
 - Top result: Draco Malfoy - Harry Potter Wiki - Wikia harrypotter.wikia.com/wiki/Draco_Malfoy... scaring Draco Malfoy by ... who was unable to discover exactly what Draco was up to despite ... Sometime after the end of the war, Draco married ... Scorpius Malfoy - Astoria Greengrass

* the questions are taken from different QA datasets (WebQuestions, QALD-5, Yahoo! Answers Webscope)

Different data sources are used for question answering



Text documents

San Francisco, California	
Consolidated city-county	
City and County of San Francisco Nickname(s): <i>The City by the Bay; Fog City; San Fran</i> ^[1] ; <i>Frisco</i> (locally disparaged) ^{[2][3][4][5]} ; <i>The City that Knows How</i> (past) ^[6] ; <i>Baghdad by the Bay</i> (past) ^[7] ; <i>The Paris of the West</i> (past) ^[8] Motto: <i>Oro en Paz, Fierro en Guerra</i> (English: "Gold in Peace, Iron in War") Coordinates: 37°47′N 122°25′W﻿ / ﻿37.783°N 122.417°W﻿ / 37.783; -122.417	
Country	United States
State	California
CSA	San Jose–San Francisco–Oakland
Metro	San Francisco–Oakland–Hayward
Mission Incorporated	June 29, 1776 ^[9] April 15, 1850 ^[10]
Founded by	José Joaquín Moraga Francisco Pádua St. Francis of Assisi
Named for	
Government	
• Type	Mayor-council
• Body	Board of Supervisors
• Mayor	Edwin M. Lee (D) ^[11]
• Supervisors ^[12]	List [show]
• Assembly members ^{[13][14]}	David Chiu (D) Phil Ting (D)
• State senator	Mark Leno (D) ^[12]
• United States Representatives ^{[15][17]}	Nancy Pelosi (D) Jackie Speier (D)

2014 rank	City	State	2014 estimate	2010 Census	Change	2014 land area	2010 population density	Location
1	New York ^[8]	New York	8,491,079	8,175,133	+3.86%	302.6 sq mi 783.8 km²	27,012 per sq mi 10,430 km ⁻²	40°46′43″N 73°03′05″W﻿ / ﻿40.77861°N 73.05139°W﻿ / 40.77861; -73.05139
2	Los Angeles	California	3,928,864	3,792,021	+3.59%	468.7 sq mi 1,213.9 km²	8,092 per sq mi 3,124 km ⁻²	34°03′04″N 118°14′08″W﻿ / ﻿34.05111°N 118.23556°W﻿ / 34.05111; -118.23556
3	Chicago	Illinois	2,722,389	2,695,998	+0.99%	227.6 sq mi 589.6 km²	11,842 per sq mi 4,572 km ⁻²	41°53′07″N 87°50′07″W﻿ / ﻿41.88528°N 87.83528°W﻿ / 41.88528; -87.83528
4	Houston ^[7]	Texas	2,239,558	2,100,263	+6.63%	599.6 sq mi 1,552.9 km²	3,501 per sq mi 1,352 km ⁻²	29°38′05″N 95°38′03″W﻿ / ﻿29.63472°N 95.63417°W﻿ / 29.63472; -95.63417
5	Philadelphia ^[6]	Pennsylvania	1,560,297	1,526,006	+2.25%	134.1 sq mi 347.3 km²	11,579 per sq mi 4,394 km ⁻²	40°00′04″N 75°13′33″W﻿ / ﻿40.00111°N 75.22583°W﻿ / 40.00111; -75.22583
6	Phoenix	Arizona	1,537,058	1,445,632	+6.32%	516.7 sq mi 1,338.3 km²	2,798 per sq mi 1,080 km ⁻²	33°07′22″N 112°08′00″W﻿ / ﻿33.12278°N 112.13333°W﻿ / 33.12278; -112.13333

Web tables & infoboxes



Knowledge bases

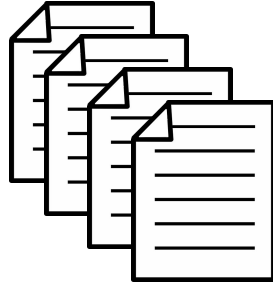
Unstructured data

Semi-structured data

Structured data

Data Sources have different advantages and problems

Text documents



Knowledge bases

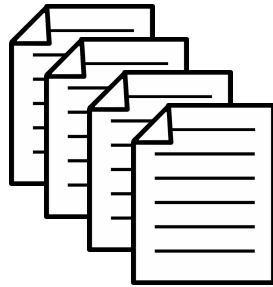


- + easy to match against question text
- + cover a variety of different information types
- each text phrase encodes a limited amount of information about mentioned entities

- + aggregate the information around entities
- + allow complex queries over this data using special languages (e.g. SPARQL)
- hard to translate natural language questions into special query languages
- incomplete (missing entities, facts and properties)

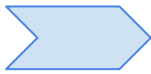
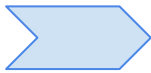
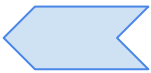
Advantages of one Data Source can compensate disadvantages of the other

Text documents



Knowledge bases



- + 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 
- hard to translate natural language questions into special query languages
 - incomplete (missing entities, facts and properties)
 - aggregate the information around entities

Knowledge Base Question Answering (KBQA)

- Goal: translate natural language question into structured KB query (e.g. SPARQL) to retrieve correct entity or attribute value

When did Tom Hanks win his first Oscar?

```
PREFIX fb: <http://rdf.freebase.com/ns/>
SELECT ?year WHERE {
  fb:/m/0bxtg fb:/award/award_winner/awards_won ?award .

  ?award fb:/award/award_honor/award fb:/m/0f4x7 .

  ?nomination fb:/award/award_honor/year ?year .
} ORDER BY ?year LIMIT 1
```


Knowledge Base Question Answering Challenges

1. Query analysis
 - How to identify question topic entity to anchor KB search?
2. Candidate generation
 - What predicates might correspond to words and phrases in the question?
 - What entities to include as candidate answers?
3. Evidence extraction
 - How to score correspondence between a certain candidate answer (e.g. involved predicates) and the question?
4. Answer selection
 - How to rank candidate answers to select the final response?

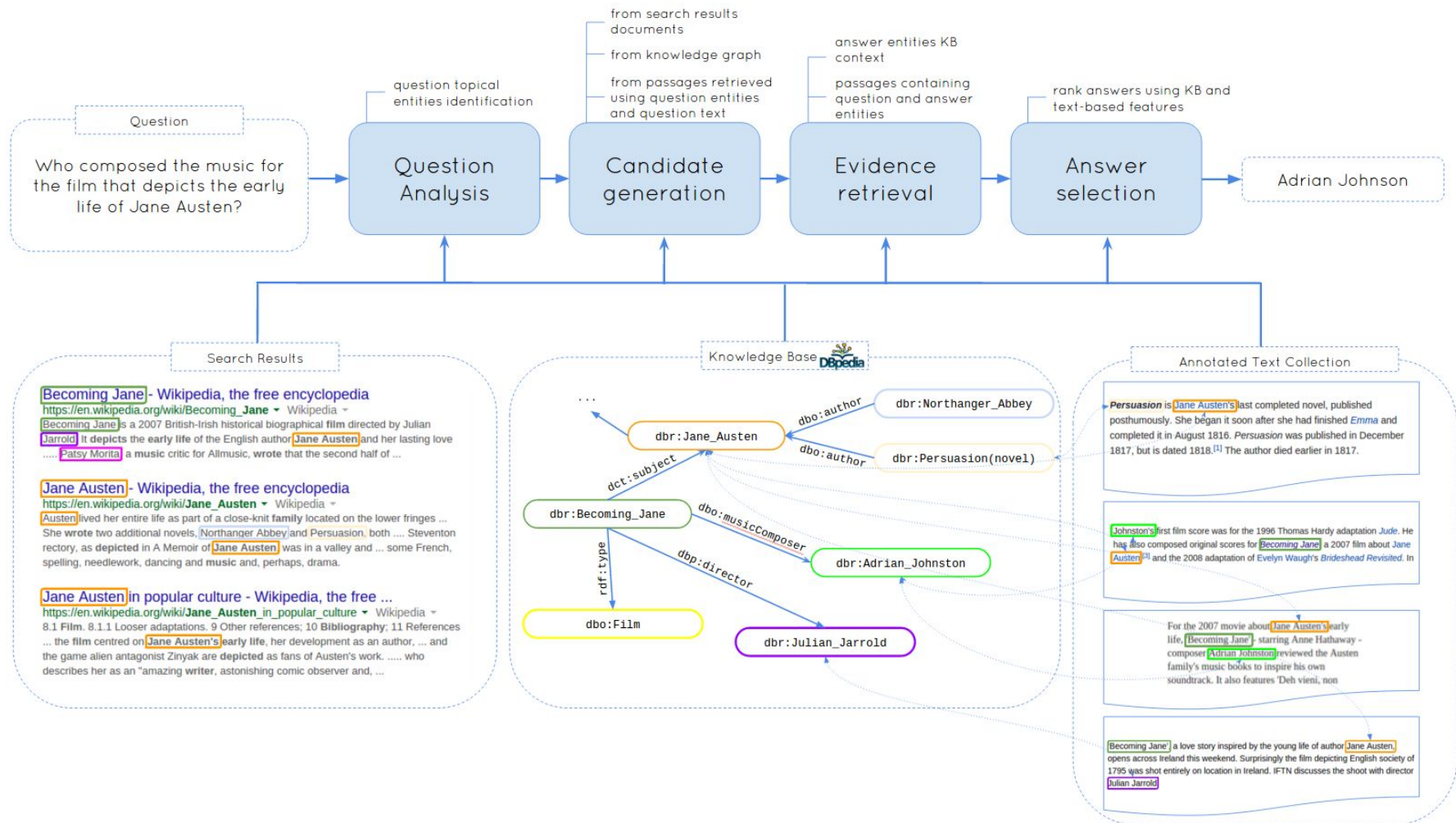
Existing Text-KB hybrid approaches

- ✓ **Open QA** [A.Fader et al. 2014]
 - Use Open Information Extraction to build semi-structured KB from text
 - Joint QA over extracted and curated KB
- ✓ **Extended Knowledge Graphs** [S. Elbassuoni et al 2009, M.Yahya et al 2016]
 - Extend triples in knowledge base with keywords
 - SPARQL query relaxation techniques to use keyword matches
- ✓ **“Open Domain Question Answering via Semantic Enrichment”** [H.Sun et al 2015]
 - Annotate text with entity mentions
 - Use entity types and textual KB descriptions to improve text-based QA
- ✓ **“Question Answering on Freebase via Relation Extraction and Textual Evidence”** [K. Xu et al. 2016]
 - Using text documents to refine answers, generated by KBQA system
- ✓ **Memory Networks** [A. Bordes et al 2015]
 - encode curated and OpenIE triples into NN memory

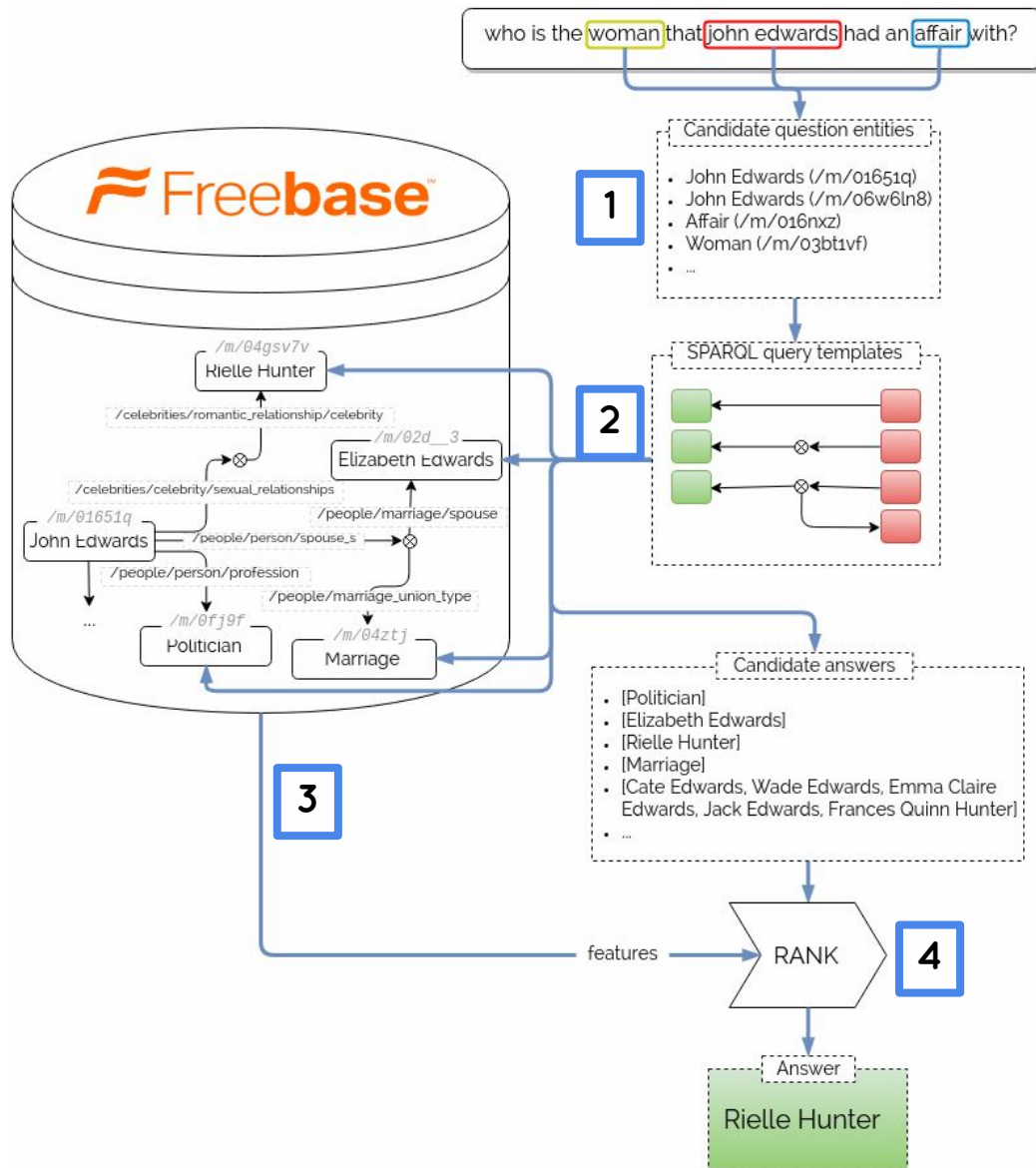
Text2KB: main idea

- ✓ Improve different stages in Knowledge Base Question Answering using various textual data
 - query analysis
 - ✓ question topic entity identification using web search results
 - candidate generation
 - ✓ Mine associations patterns between question terms and predicates from CQA data
 - evidence extraction
 - ✓ build language model for candidate question-answer entity pairs based on annotated corpus of text documents
 - answer selection
 - ✓ Score answer candidates using a combination of KB and text-based features

Text2KB: Incorporating Text in Answering Process

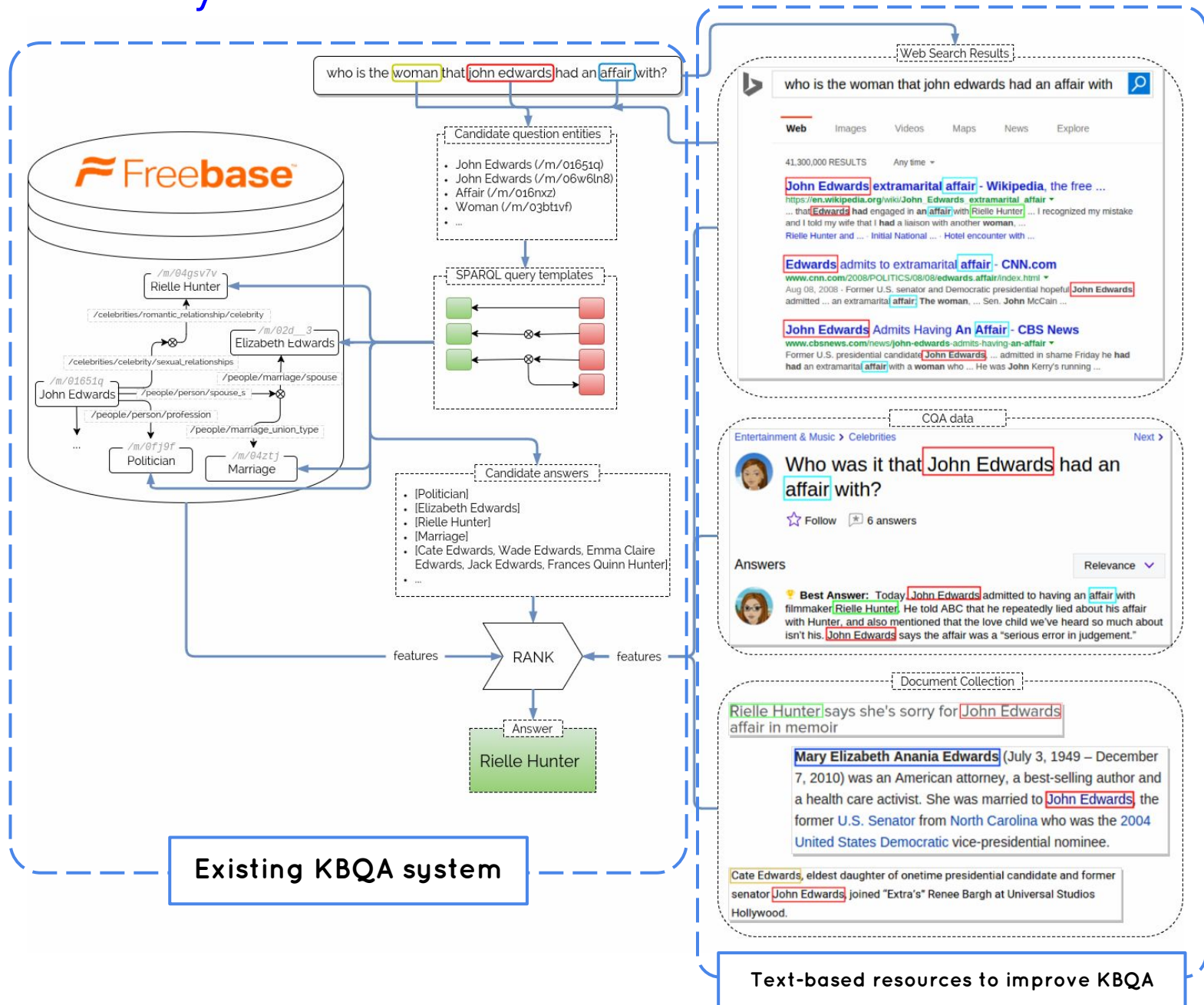


Baseline system architecture*



1. **Detecting question topic entity:** multiple candidates are detected using dictionary of names and aliases
2. **Answer candidate generation:** instantiate candidate SPARQL queries from the neighborhood of question entities using a set of template queries
3. **Evidence generation:** each candidate is represented with a set of features, describing the detected topic entity, predicates on KB path connecting topic and answer entities, etc.
4. **Answer selection:** candidate answers are ranked using a trained ranking model and top scoring one is returned as the answer

Text2KB System Architecture



Question Analysis: Entity Linking



what year did tut became king?

King Tut - King - Biography.com
www.biography.com/people/king-tut-9512446 ▾
Video embedded · Synopsis. Born circa 1341 B.C.E., **King Tut** was the 12th **king** of the 18th Egyptian dynasty, in power from approximately 1332 to 1323 B.C.E. During his ...

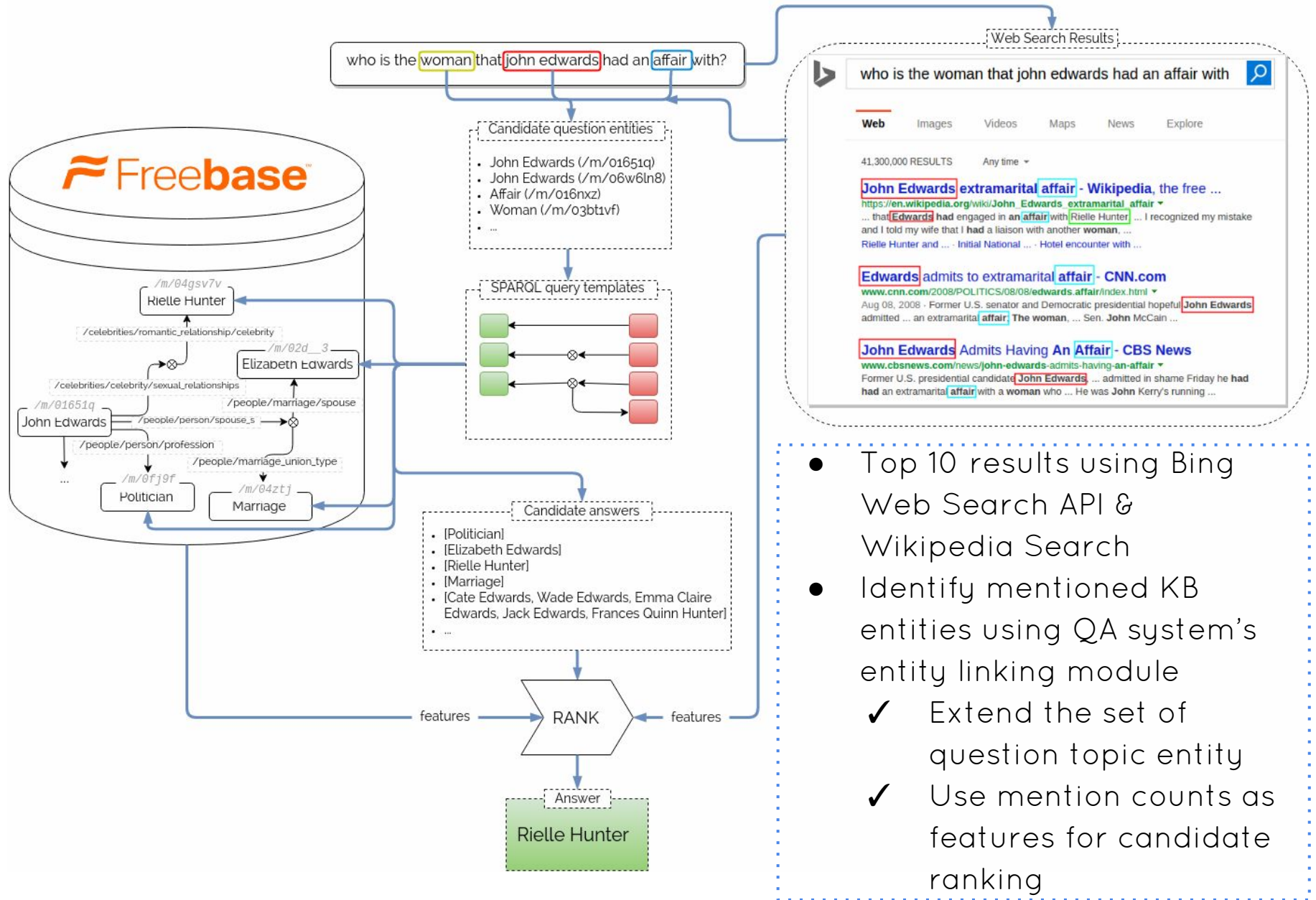
Tutankhamun - Ancient Egypt for Kids
resources.woodlands-junior.kent.sch.uk/homework/tut.html ▾
At what age **did Tutankhamun become** a Pharaoh? ... but also the **king's** regal **year** when each wine was laid down. The highest recorded date is **Year 9**, ...

In which year did Tutankhamun become king? - ask.com
www.ask.com > History > Ancient History > Ancient Egypt ▾
Tutankhamun became king of Egypt in 1332 BC, ... In which **year did Tutankhamun become king?** A: ... What **did King Tut** do during his reign? A:

Tutankhamun - Wikipedia, the free encyclopedia
https://en.wikipedia.org/wiki/King_Tut ▾
He is colloquially referred to as **King Tut**. ... "**King Tut**" became the name of products, businesses, and even the pet dog of U.S. President Herbert Hoover.
[Life](#) · [Significance](#) · [Tomb](#) · [Reuse of ...](#) · [Legacy](#) · [In popular culture](#)

- ✓ Web Search Results can help entity linking and provide textual evidence to answer candidates
- ✓ Contains multiple mentions of the question topic entity, often in variations, which might help entity linking
- ✓ Search results often contain the answer to the question itself, which is exploited by text-based question answering systems


Text2KB System Architecture: web search results





- Top 10 results using Bing Web Search API & Wikipedia Search
- Identify mentioned KB entities using QA system's entity linking module
 - ✓ Extend the set of question topic entity
 - ✓ Use mention counts as features for candidate ranking

Community Question Answering data can help map question phrases to predicates


Society & Culture > Other - Society & Culture Next >

 **Where martin luther king was born?**

i think he wasborn in south afriaca because that is one place wher there are black peoplebut i think it is somewhere in africa

 Follow  5 answers

Answers Relevance ▾

 **Best Answer:** Martin Luther King (Sr) was born in Stockbridge, Georgia in the U. S. A. on December 19, 1899. In the early 1930's he changed his name from Michael King to Martin Luther King (after Martin Luther, the German theologian who started the Protestant Reformation).

His son Martin Luther King Jr., whose birthday was celebrated this month in the United States, was born on January 15, 1929 in Atlanta, Georgia.

So which ever one you meant, he wasn't born in Africa.

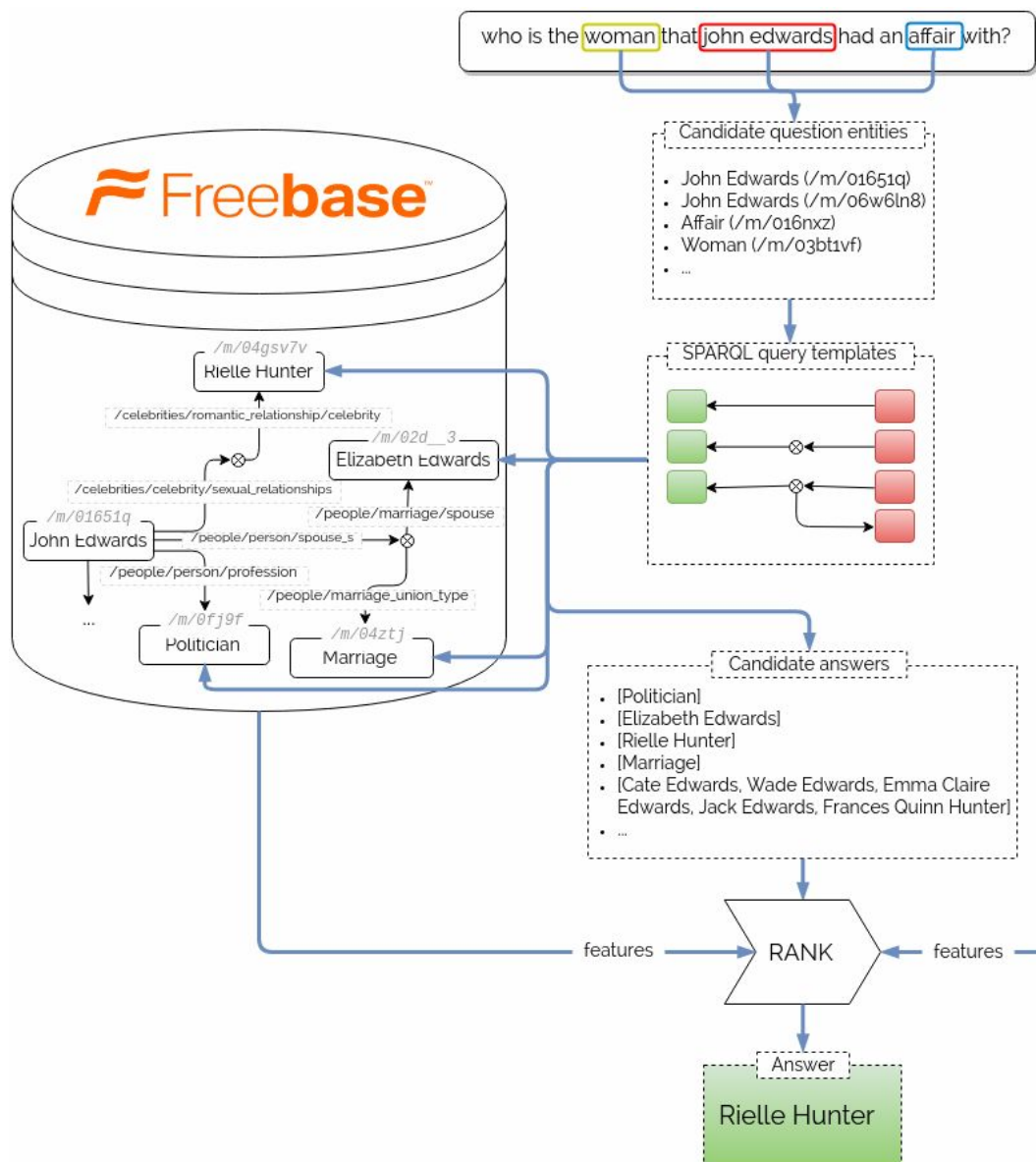
- ✓ Huge number of question-answer pairs, but noisy (most of the questions aren't factoid, answers are verbose and contain redundant information)
- ✓ Can be helpful to learn associations between the language of a question and KB predicates using distant supervision assumption

Examples of term-predicate associations computed using CQA data

Term	Predicate	PMI score
born	people.person.date_of_birth	3.67
	people.person.date_of_death	2.73
	location.location.people_born_here	1.60
kill	people.deceased_person.cause_of_death	1.70
	book.book.characters	1.55
currency	location.country.currency_formerly_used	5.55
	location.country.currency_used	3.54
school	education.school.school_district	4.14
	people.education.institution	1.70
	sports.school_sports_team.school	1.69
win	sports.sports_team.championships	4.11
	sports.sports_league.championship	3.79

- ✓ Despite the noisy distant supervision labeling, top scoring predicates are indeed related to the corresponding word

Text2KB System Architecture: CQA data



- Distant supervision to label question-answer pairs from Yahoo! Answers WebScope collection with KB predicates
- Learn associations between question terms and predicates using PMI scores
 - Use these PMI scores as features to score candidate answer predicates

The screenshot shows a Yahoo! Answers page for the question "Who was it that John Edwards had an affair with?". The question is highlighted in blue. The page shows 6 answers. The best answer is highlighted in yellow and contains the text: "Today John Edwards admitted to having an affair with filmmaker Rielle Hunter. He told ABC that he repeatedly lied about his affair with Hunter, and also mentioned that the love child we've heard so much about isn't his. John Edwards says the affair was a 'serious error in judgement.'" The names "John Edwards" and "Rielle Hunter" are highlighted in red and green respectively, corresponding to the entities in the diagram.

Text around mentions of pairs of entities in documents help explain relationships between the entities

Rielle Hunter says she's sorry for John Edwards affair in memoir

Mary Elizabeth Anania Edwards (July 3, 1949 – December 7, 2010) was an American attorney, a best-selling author and a health care activist. She was married to John Edwards the former U.S. Senator from North Carolina who was the 2004 United States Democratic vice-presidential nominee.

Cate Edwards, eldest daughter of onetime presidential candidate and former senator John Edwards joined "Extra's" Renee Bargh at Universal Studios Hollywood.

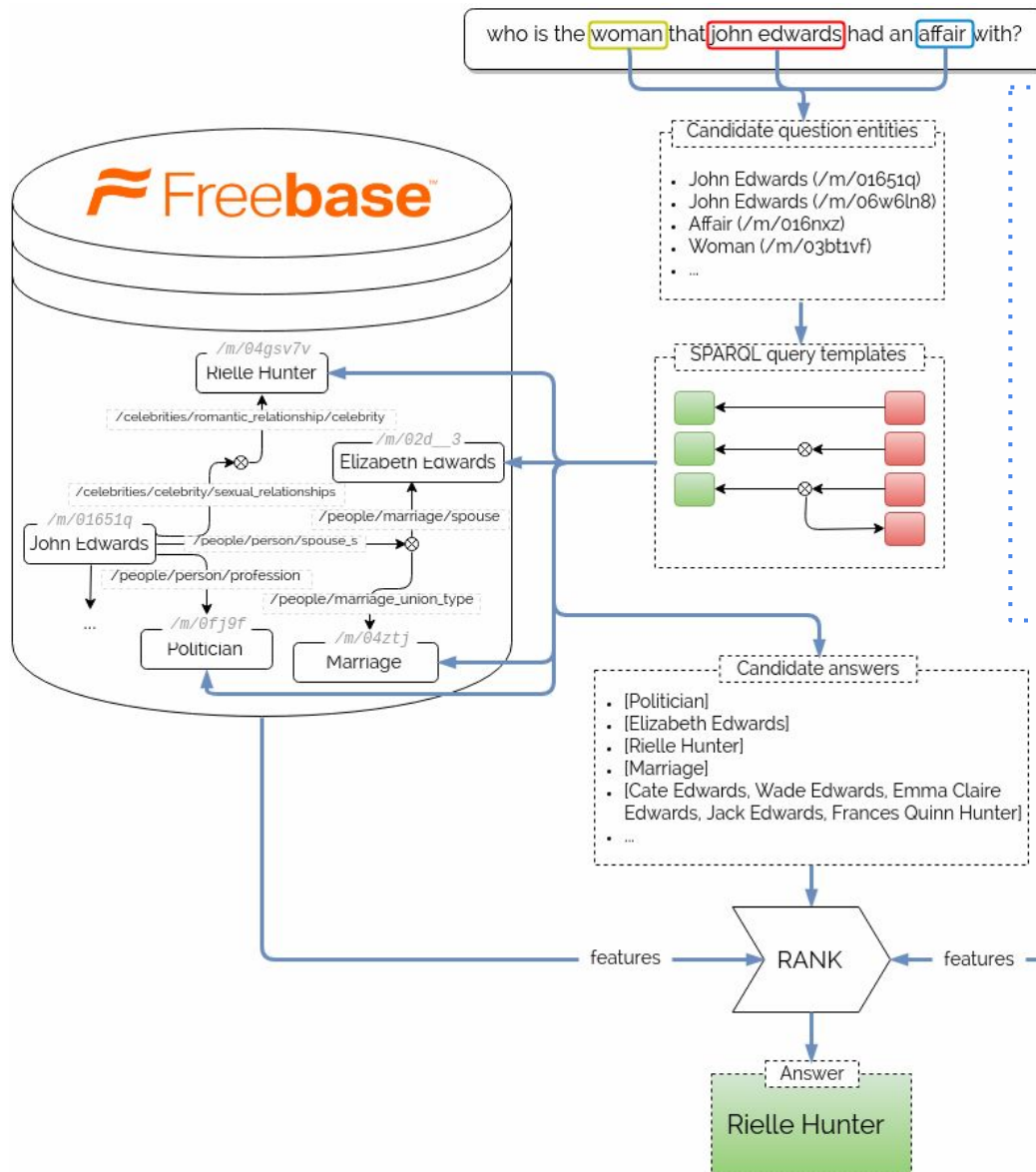
- ✓ Sentences and passages that mention multiple entities often express some facts about them
- ✓ Terms used in these passages can explain the relationships between the entities

Examples of entity pair language models

Entity 1	Entity 2	Term counts
John Edwards	Rielle Hunter	campaign, affair, mistress, child, former ...
John Edwards	Cate Edwards	daughter, former, senator, courthouse, greensboro, eldest ...
John Edwards	Elizabeth Edwards	wife, hunter, campaign, affair, cancer, rielle, husband ...
John Edwards	Frances Quinn	daughter, john, rielle, father, child, former, paternity...

- ✓ Terms most frequently used around mention of a pair of entities indeed shed some light on the relationship between the entities

Text2KB System Architecture: document collection



- Extract text around mentions of entity pairs in ClueWeb12
- Learn entity pair language model $p(\text{term} | \text{entity}_1, \text{entity}_2)$
 - ✓ Use language model scores as features for candidate answer ranking

Document Collection

Rielle Hunter says she's sorry for John Edwards affair in memoir

Mary Elizabeth Anania Edwards (July 3, 1949 – December 7, 2010) was an American attorney, a best-selling author and a health care activist. She was married to John Edwards the former U.S. Senator from North Carolina who was the 2004 United States Democratic vice-presidential nominee.

Cate Edwards, eldest daughter of onetime presidential candidate and former senator John Edwards joined "Extra's" Renee Bargh at Universal Studios Hollywood.

Evaluation

- ✓ WebQuestions dataset
 - 3,778 training and 2,032 test questions

- ✓ Metrics:

- Average F1: $avg F1 = \frac{1}{|Q|} \sum_{q \in Q} f1(a_q^*, a_q)$

$$f1(a_q^*, a_q) = 2 \frac{precision(a_q^*, a_q) recall(a_q^*, a_q)}{precision(a_q^*, a_q) + recall(a_q^*, a_q)}$$

- ✓ Methods compared:

- Aqqu (Bast et al, 2015) - our KB-only baseline
- STAGG (Yih et al, 2015) - SOTA at the moment of publication
- our Text2KB (Web search)
- our Text2KB (Wikipedia search)

Results

	Recall	Precision	F1
OpenQA [A.Fader et al 2014]	-	-	0.35
STAGG [H.Sun et al 2015]	0.607	0.528	0.525
Aqqu (baseline) [H.Bast et al 2015]	0.604	+5.7%	0.494
Text2KB (wikipedia search)	0.632	0.498	0.514
Text2KB (web search)	0.635	0.506	0.522

- ✓ Text2KB significantly improves upon the baseline Aqqu system (0.494 -> 0.522 avg F1 score)
- ✓ Text2KB reaches the performance of STAGG, best result at the moment of publication
 - but this work is orthogonal to improvements in STAGG and therefore can be combined

Component ablation

System	avg F1
Aqqu	0.494
+ Entity linking from search results	0.508
+ Search results, CQA and Clueweb features for ranking	0.514
Text2KB	0.522

System	avg F1
Aqqu	0.494
Text2KB (Web search)	0.522
- Web search data	0.513
- CQA data	0.519
- ClueWeb data	0.523
+ Web search data only	0.522
+ CQA data only	0.508
+ ClueWeb data only	0.514

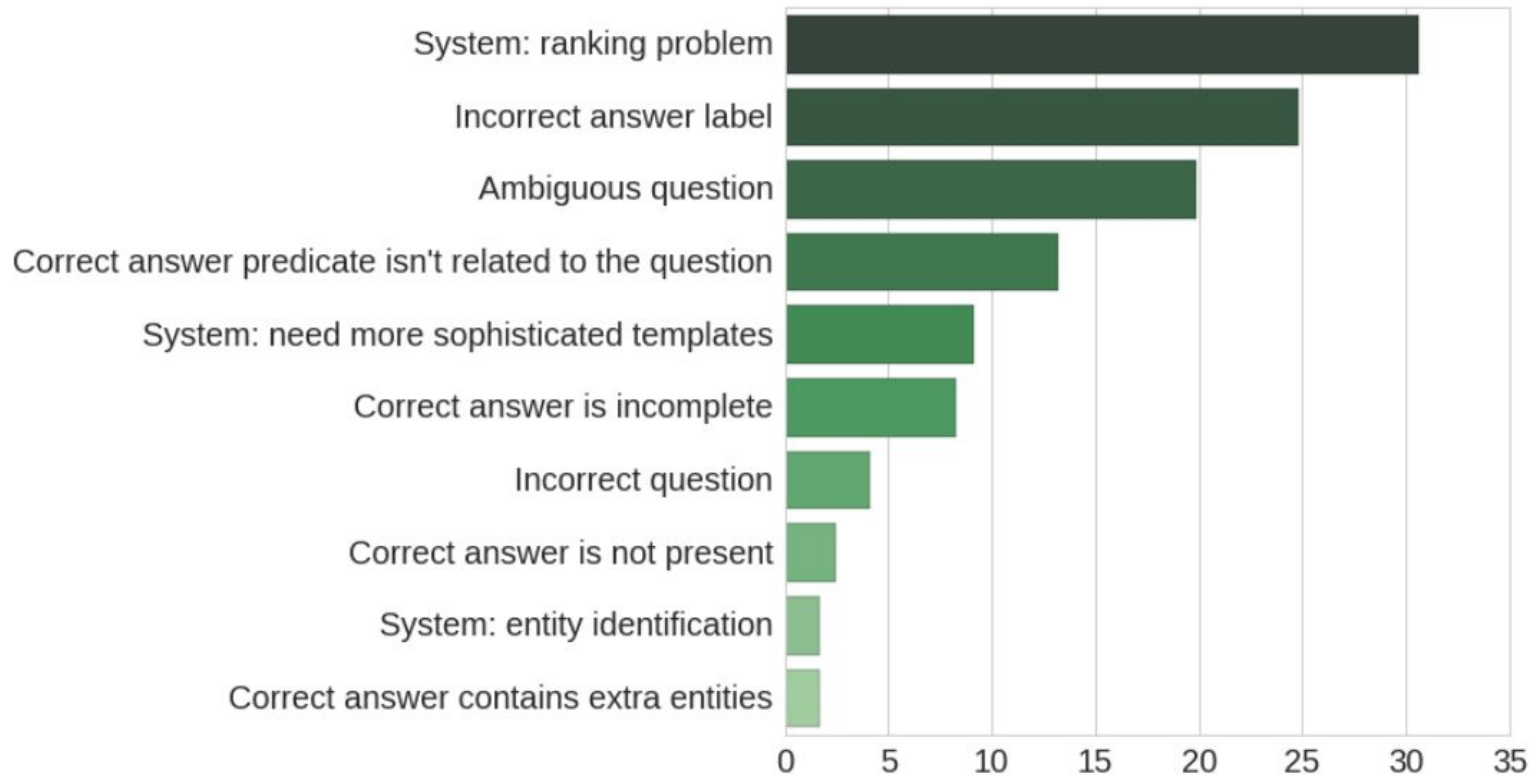
- ✓ Both entity linking using web search results and features for answer ranking contribute to improvements
- ✓ Search results have the largest contribution to the overall performance, but CQA and ClueWeb are also useful

Combining Text2KB & STAGG

System	avg F1
STAGG (Yih et al, 2015)	0.525
Text2KB + STAGG (takes STAGG answers if it has less entities)	0.532
Text2KB + STAGG (Oracle: chooses answer with higher F1 score)	0.606

- ✓ Combining results of Text2KB and STAGG suggests that our ideas could benefit it as well
 - Heuristic combination: take Text2KB or STAGG answer, which contains less entities
 - Oracle combination always choose the answer with higher F1

Error analysis



- ✓ Majority of errors ($F1 < 1$) are ranking errors
- ✓ But there are also many problems in questions and labels
- ✓ Check out the new WebQuestionsSP dataset:
<https://goo.gl/eQF0tM>

Current & Future work

- Overall, our system is most helpful:
 - Question topic entity is hard to identify (uncommon alias, misspelling)
 - Form of the question or ground truth predicate is less frequent in the training set
- Our system has the following problems:
 - Less effective for tail and abstract entities, whose mentions are harder to find in text. For example entity “Associated Press Male Athlete of the Year” isn’t linked correctly (unless mentioned exactly by name)
 - Our use of text doesn’t help much to solve KB incompleteness (e.g. missing facts or predicates)
- Future work:
 - Instead of improving KBQA, move to more open scenario
 - new hybrid model that will use all the information available in different data sources
 - new dataset of entity-centric factoid questions

Conclusions

- Textual data sources provide additional information, that can compensate disadvantages of structured knowledge bases
- Our Text2KB system uses a combination of structured and unstructured data to improve Knowledge Base Question Answering
 - Improve avg F1 on WebQuestions dataset: **0.494 -> 0.522**

Acknowledgements



Denis Savenkov is planning to defend in December 2016 and will be on the market for postdoc and industry research positions

dsavenk@emory.edu

Thank you!

This work was partially supported by the Yahoo Labs Faculty Research Engagement Program (FREP).