

When a Knowledge Base Is Not Enough- Question Answering over Knowledge Bases with External Text Data

Date: 2016/11/08

Author: Denis Savenkov, Eugene Agichtein

Source: ACM SIGIR' 16

Advisor: Jia-ling Koh

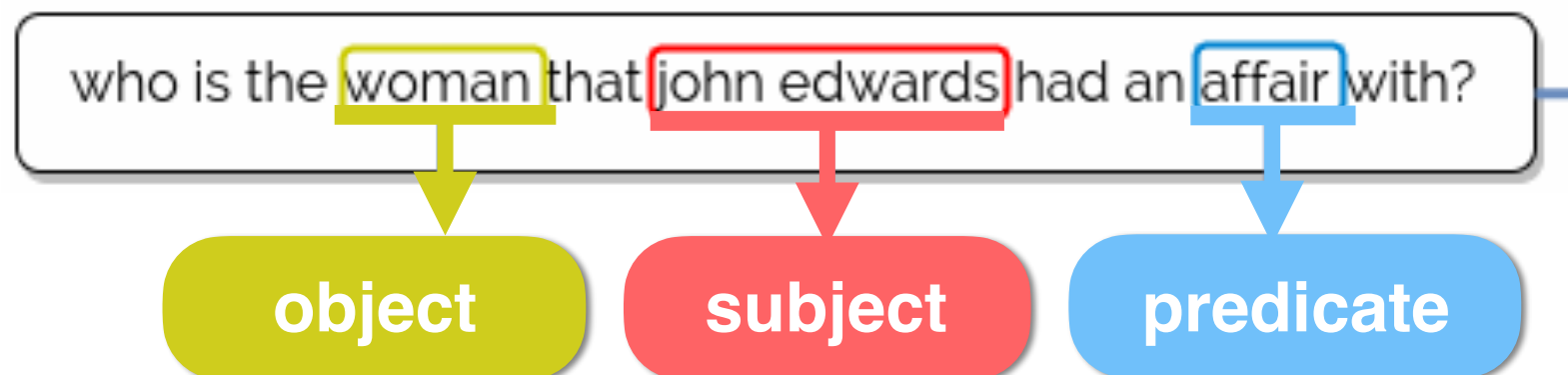
Speaker : Yi-hui Lee

Outline

- **Introduction**
- Approach
- Experiment
- Conclusion

Introduction

- Question Answering:
 - Text-centric, or Text-QA: use text document collections to retrieve passages relevant to a question and extract candidate answers
 - Knowledge base-centric, or KBQA
 - RDF triples [subject, predicate, object]



Introduction(cont.)

Aqqu KBQA system

Query Template

```
SELECT DISTINCT ?a_entity {
  <q_entity> <predicate> ?a_entity .
}
```

```
SELECT DISTINCT ?a_entity {
  <q_entity> <predicate_1> ?cvt_node .
  ?cvt_node <predicate_2> ?a_entity .
}
```

```
SELECT DISTINCT ?a_entity {
  <q_entity_1> <predicate_1> ?cvt_node .
  ?cvt_node <predicate_2> <q_entity_2> .
  ?cvt_node <predicate_3> ?a_entity .
}
```

mediator node

answer entity

question entity

Extension

Basic system extensions

Question

"what team did david beckham play for in 2011?"

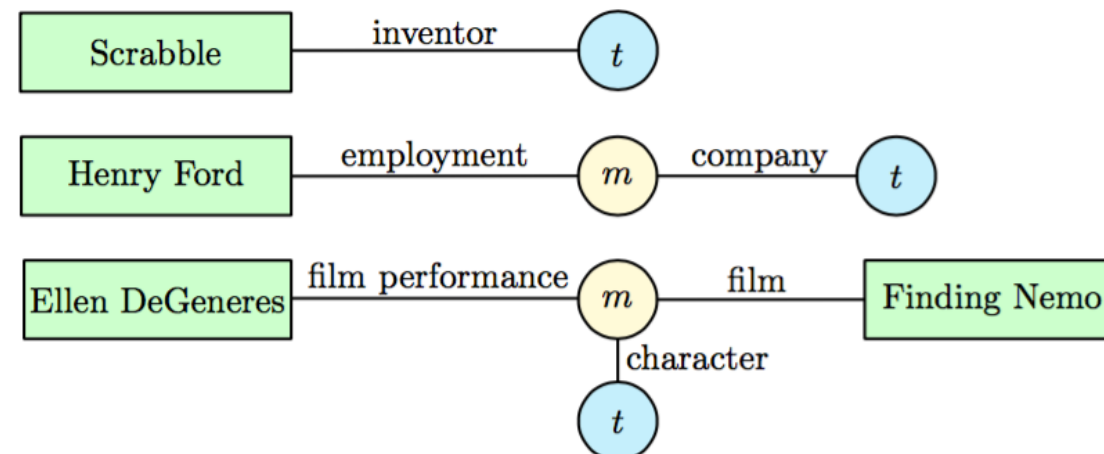
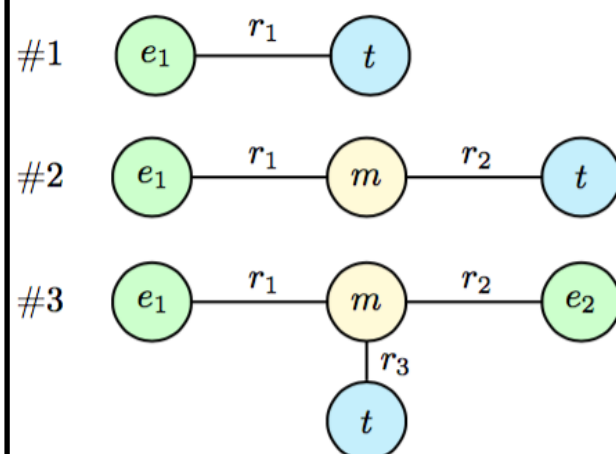
Query Template

```
SELECT DISTINCT ?a_entity {
  <q_entity_1> <predicate_1> ?cvt_node .
  ?cvt_node <from_predicate> ?date_from .
  ?cvt_node <to_predicate> ?date_to .
  ?cvt_node <predicate_2> ?a_entity .
  FILTER ( <question_date> >= ?date_from AND
    <question_date> <= ?date_to )
}
```

Template

Example Candidate

Question



who invented scrabble?

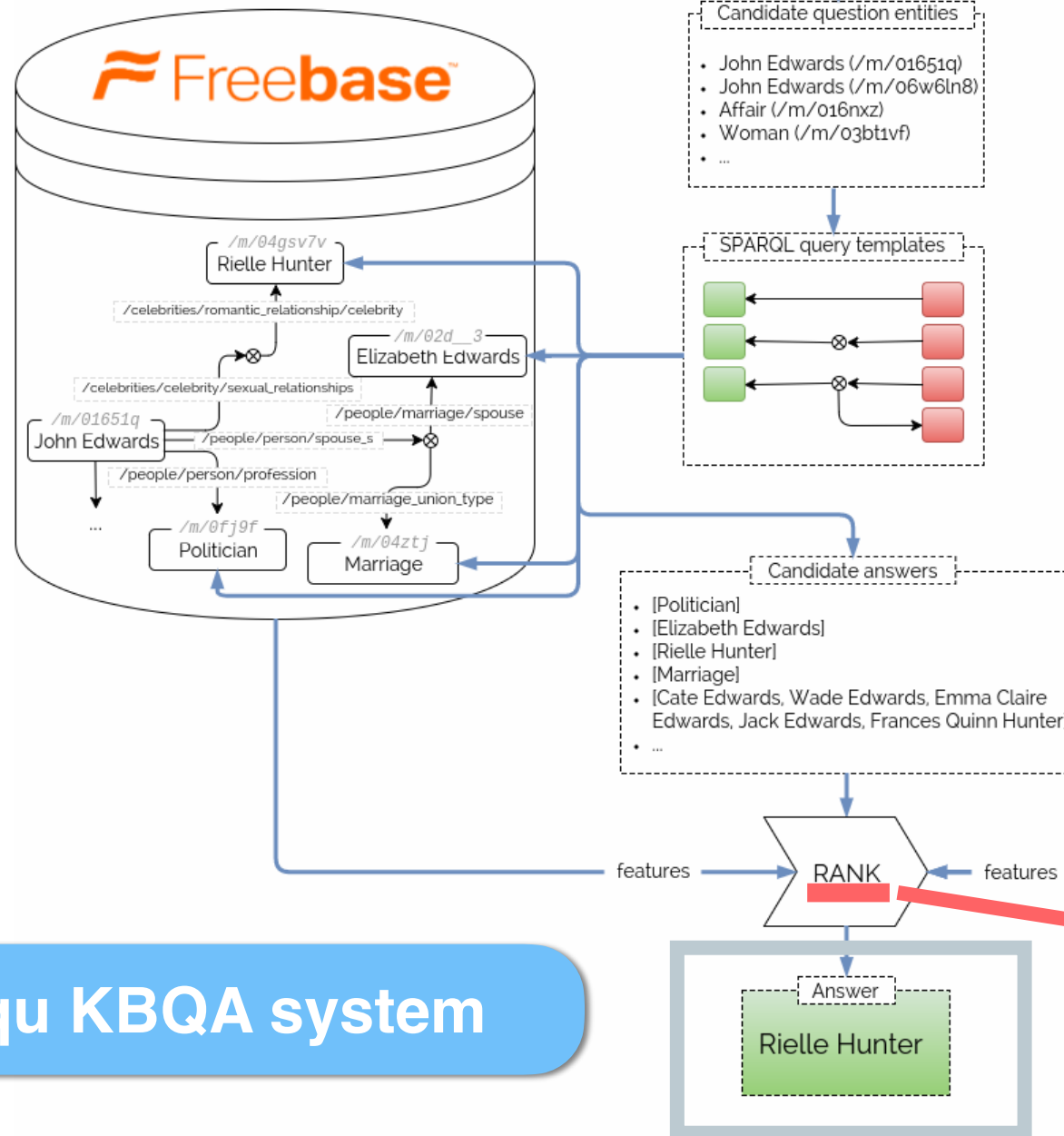
what company did henry ford work for?

what character does ellen play in finding nemo?

Introduction(cont.)

input Question

who is the woman that john edwards had an affair with?



Aqqu KBQA system

output Answer

Query Template

```
SELECT DISTINCT ?a_entity {  
  <q_entity> <predicate> ?a_entity .  
}
```

```
SELECT DISTINCT ?a_entity {  
  <q_entity> <predicate_1> ?cvt_node .  
  ?cvt_node <predicate_2> ?a_entity .  
}
```

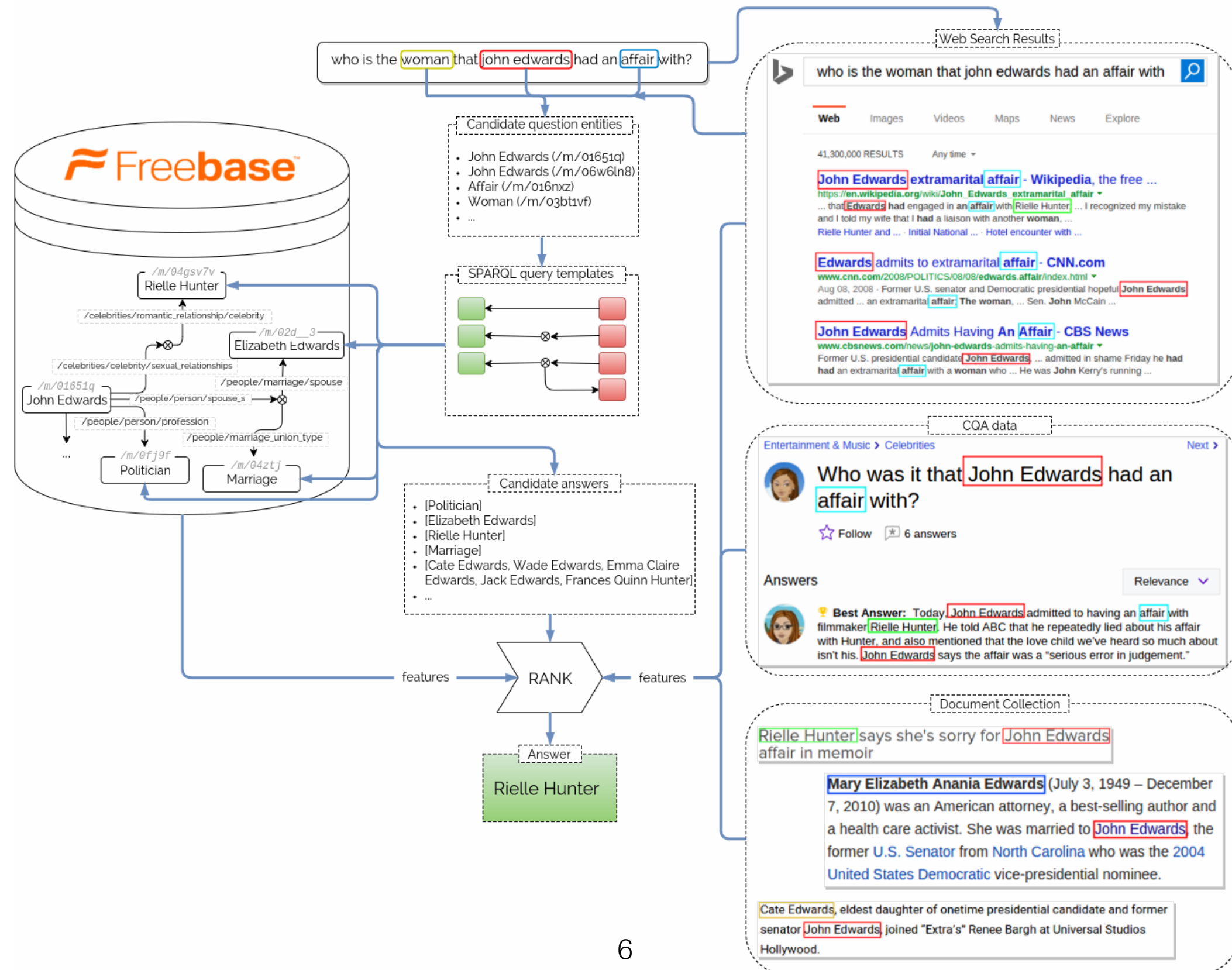
```
SELECT DISTINCT ?a_entity {  
  <q_entity_1> <predicate_1> ?cvt_node .  
  ?cvt_node <predicate_2> <q_entity_2> .  
  ?cvt_node <predicate_3> ?a_entity .  
}
```

```
SELECT DISTINCT ?a_entity {  
  <q_entity_1> <predicate_1> ?cvt_node .  
  ?cvt_node <from-predicate> ?date_from .  
  ?cvt_node <to-predicate> ?date_to .  
  ?cvt_node <predicate_2> ?a_entity .  
  FILTER ( <question-date> >= ?date_from AND  
    <question-date> <= ?date_to )  
}
```

pairwise learning to rank
using Random Forest Model

Introduction(cont.)

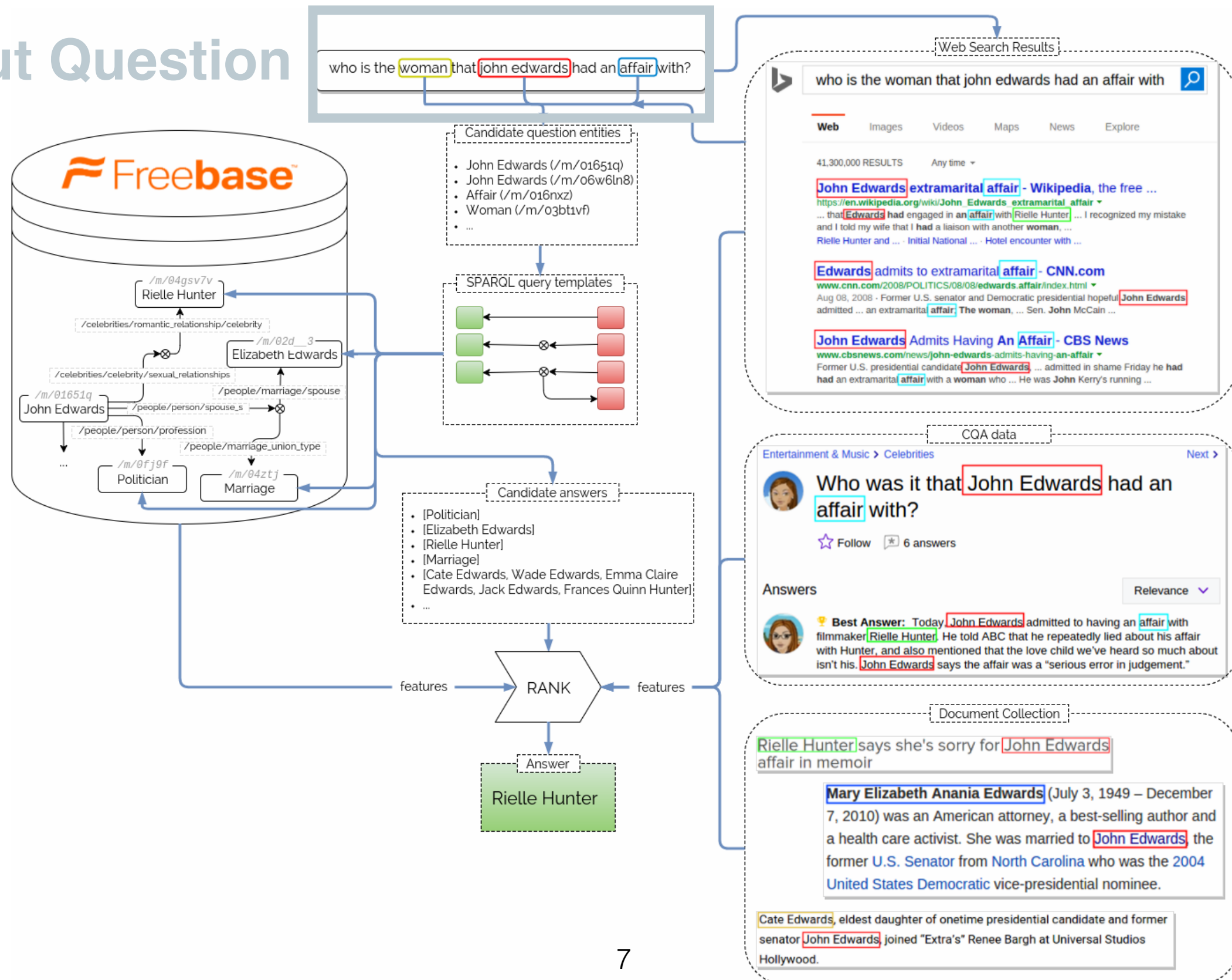
- Text2KB Question Answering system's architecture



Introduction(cont.)

- Text2KB Question Answering system's architecture

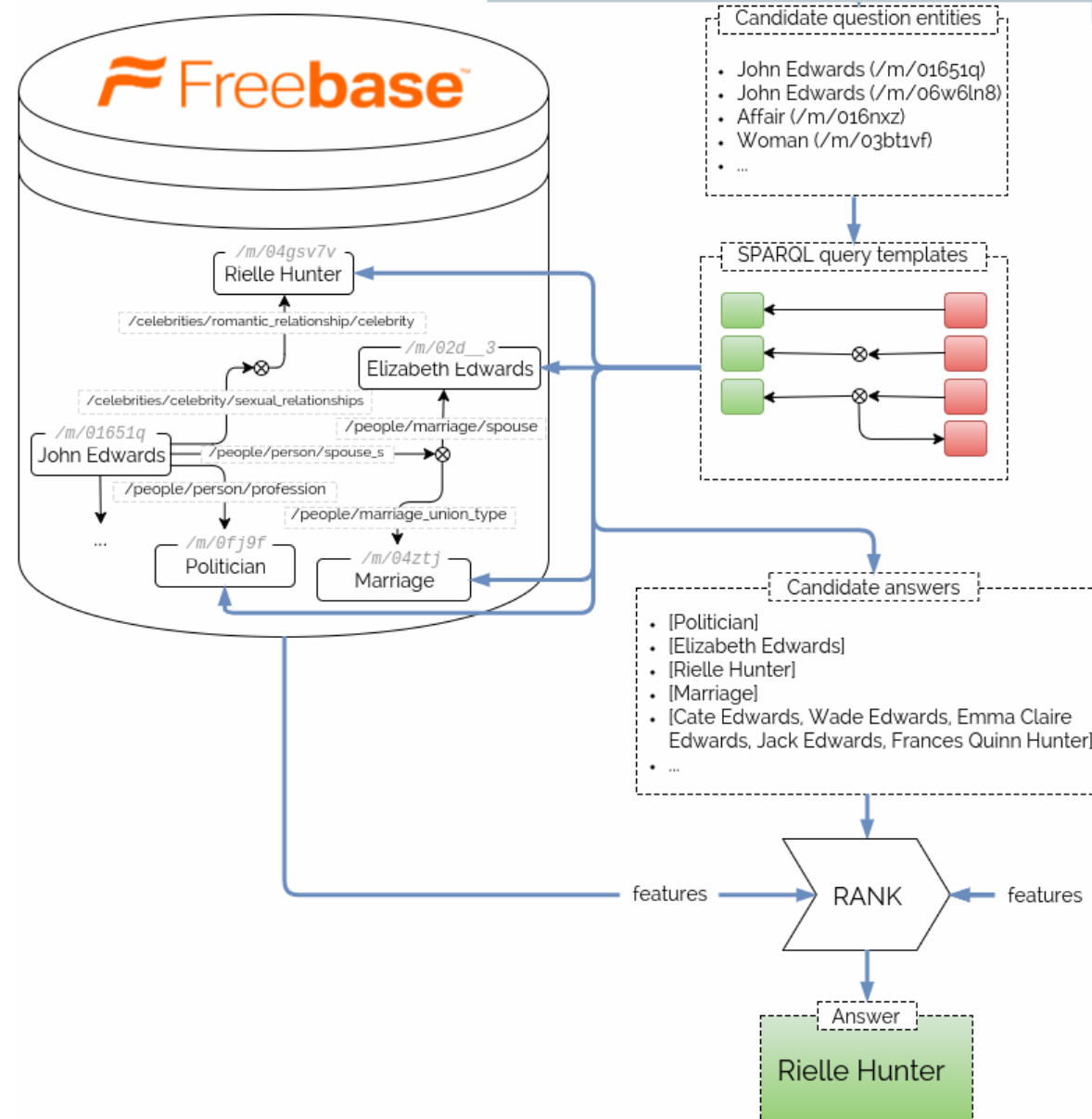
input Question



Introduction(cont.)

- Text2KB Question Answering system's architecture

input Question



Web Search Results

The screenshot shows search results for the question "who is the woman that john edwards had an affair with". The results include a Wikipedia entry titled "John Edwards extramarital affair - Wikipedia, the free ..." and a CNN article titled "Edwards admits to extramarital affair - CNN.com". The Wikipedia entry mentions that John Edwards had engaged in an affair with Rielle Hunter. The CNN article mentions that Edwards admitted to having an affair with a woman, who was later identified as Rielle Hunter.

step 1

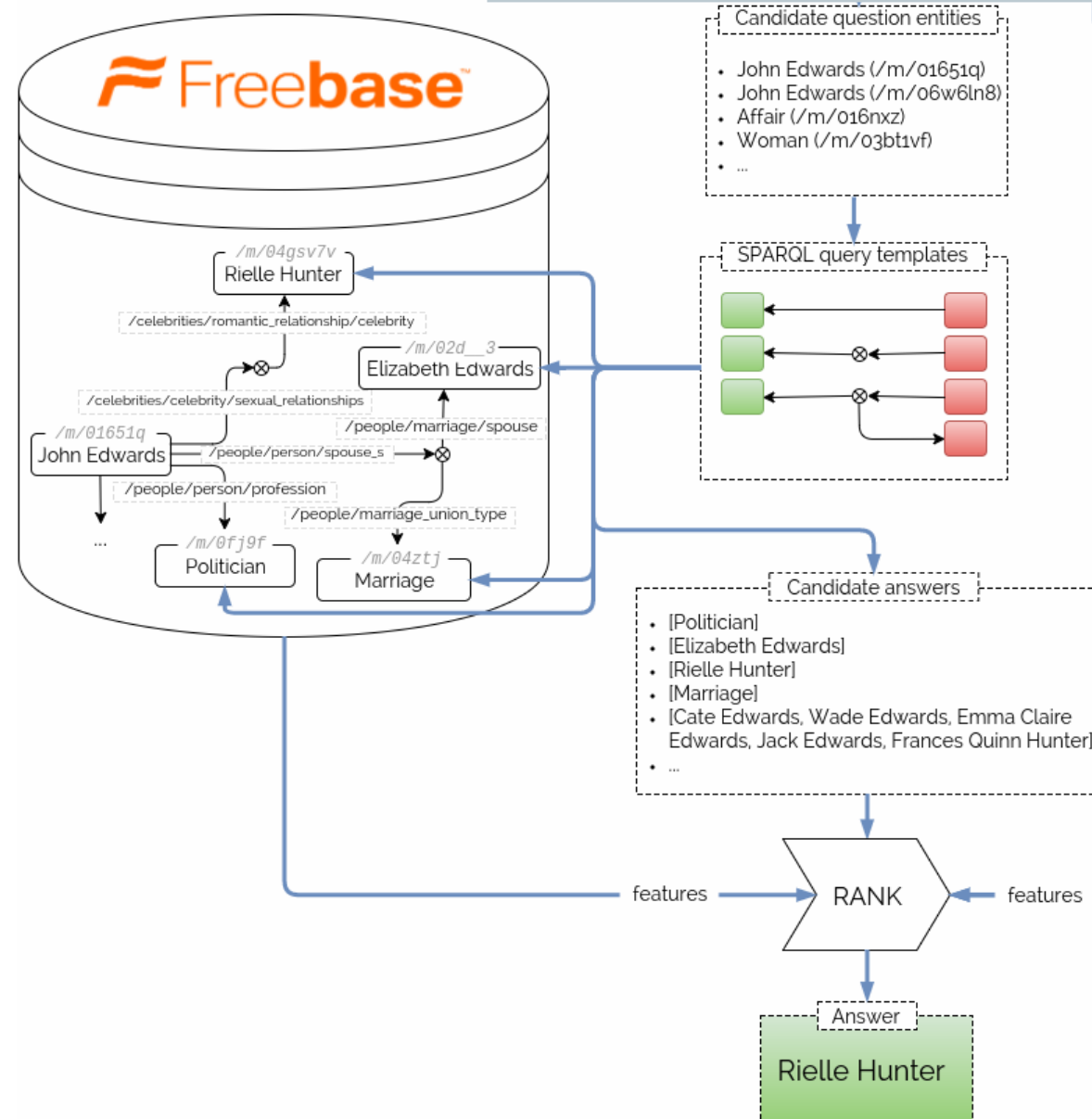
The screenshot shows a Q&A page for the question "Who was it that John Edwards had an affair with?". The page has a "Follow" button and "6 answers". The "Best Answer" is from a user named "Cate Edwards" and states: "Today John Edwards admitted to having an affair with filmmaker Rielle Hunter. He told 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 screenshot shows a "Document Collection" for the question "Who was it that John Edwards had an affair with?". The collection includes a document titled "Rielle Hunter says she's sorry for John Edwards affair in memoir" and another document titled "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." The collection also includes a snippet about Cate Edwards, the eldest daughter of onetime presidential candidate and former senator John Edwards, who joined "Extra's" Renee Bargh at Universal Studios Hollywood.

Introduction(cont.)

- Text2KB Question Answering system's architecture

input Question

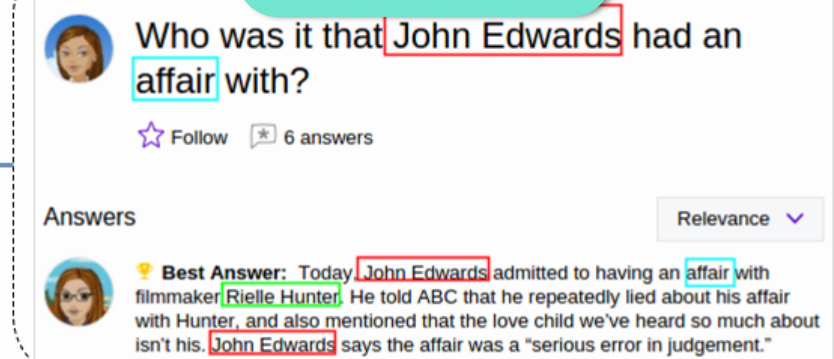


Web Search Results

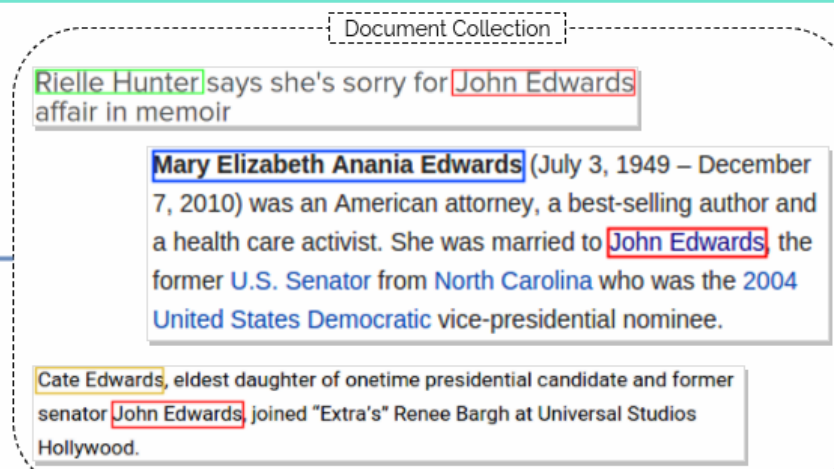


step 1

CQA data



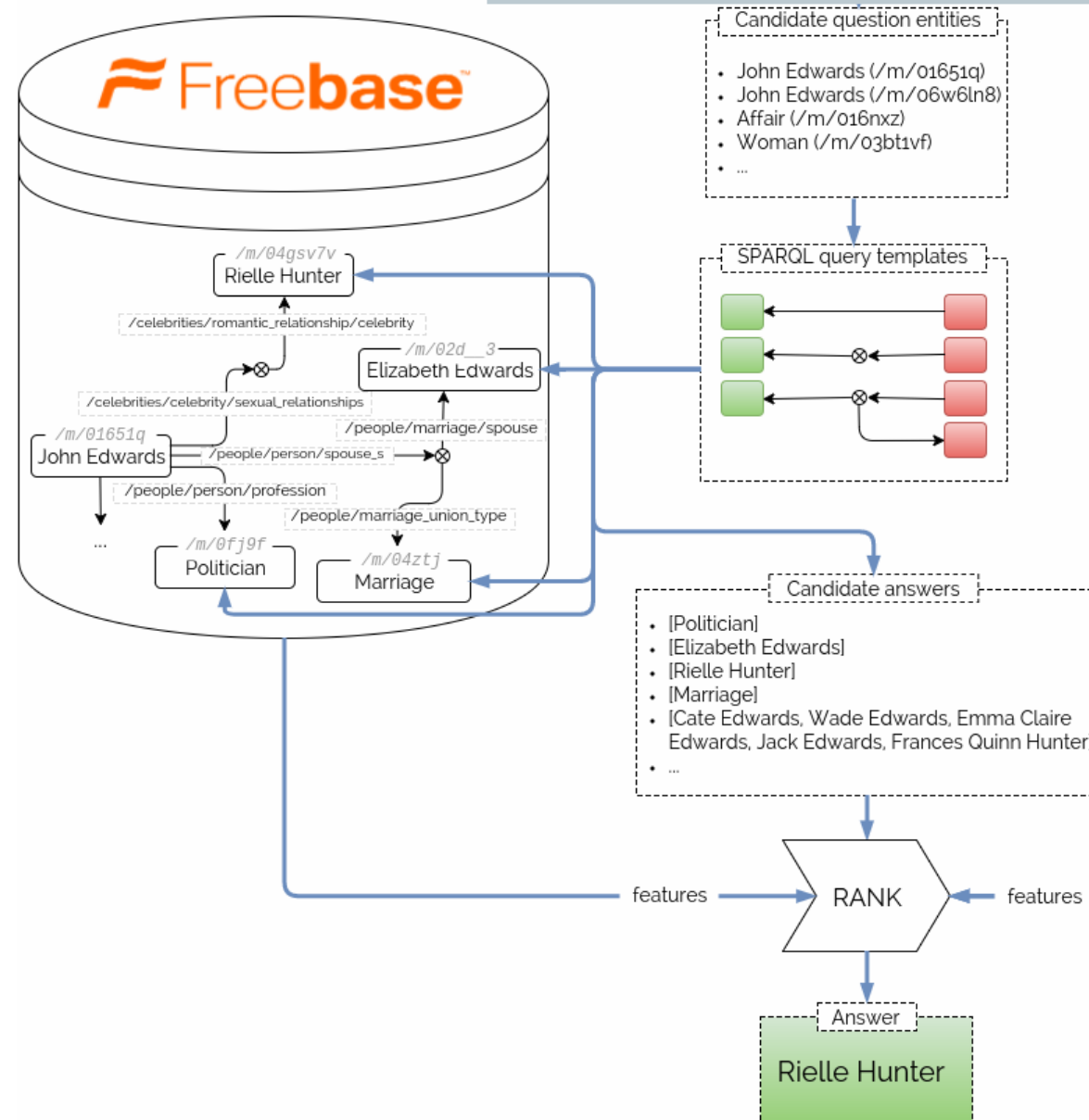
step 2



Introduction(cont.)

- Text2KB Question Answering system's architecture

input Question

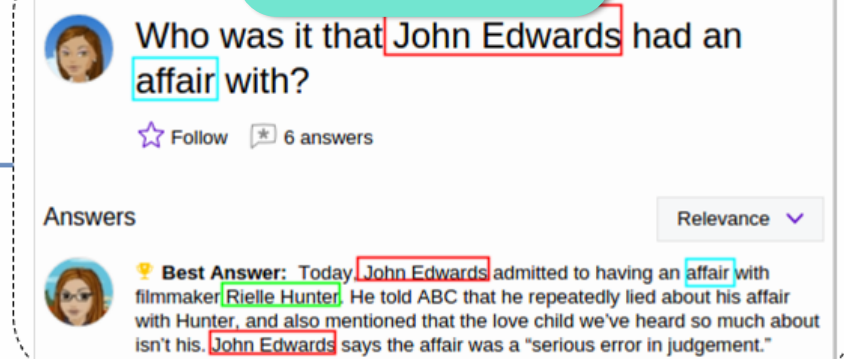


Web Search Results



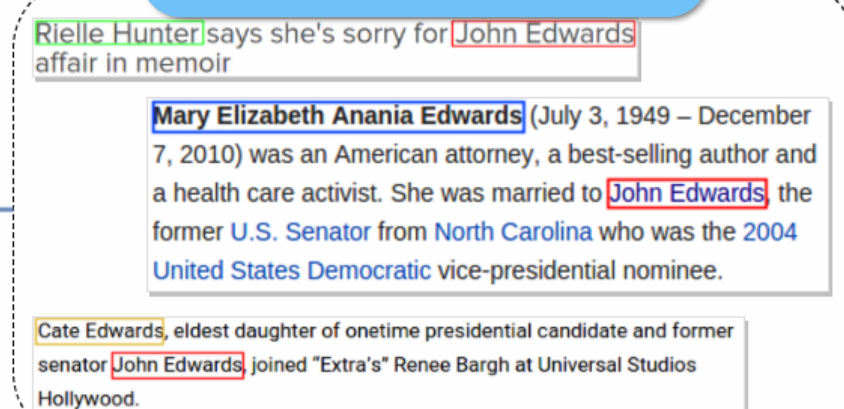
step 1

CQA data



step 2

Document Collection

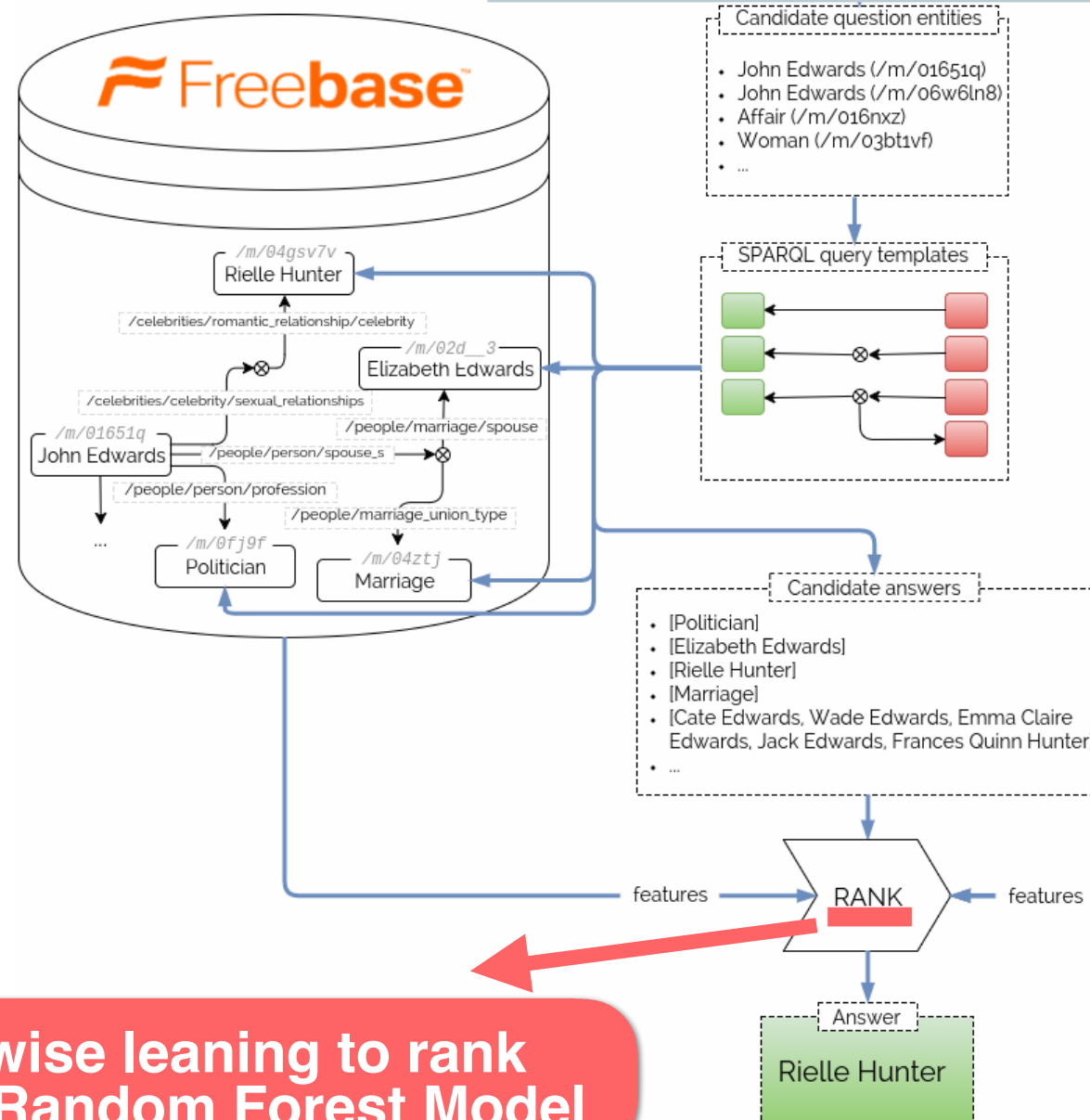


step 3

Introduction(cont.)

- Text2KB Question Answering system's architecture

input Question



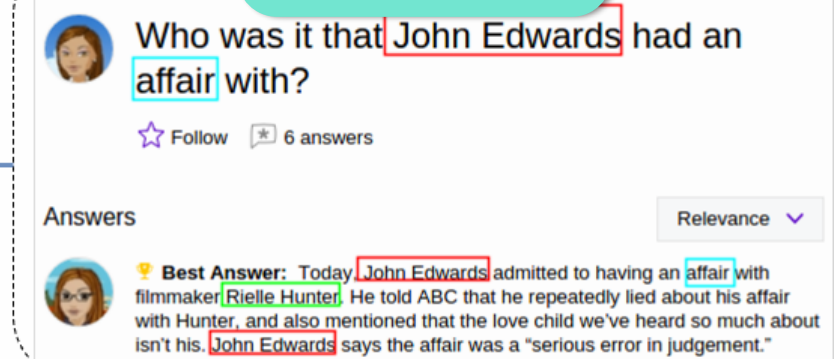
pairwise learning to rank
using Random Forest Model

Web Search Results



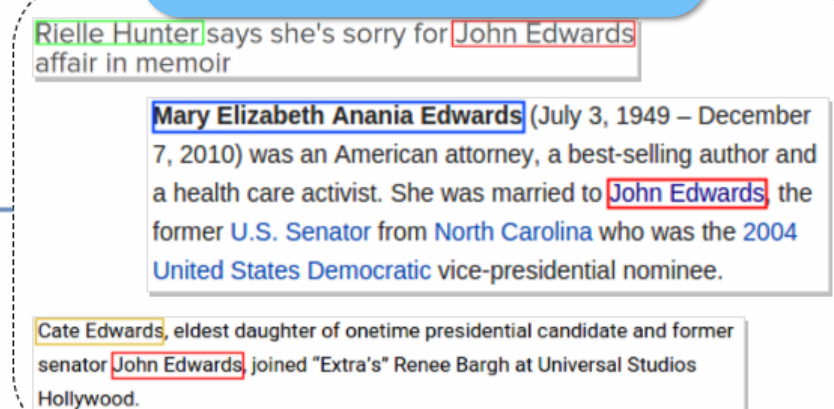
step 1

CQA data



step 2

Document Collection

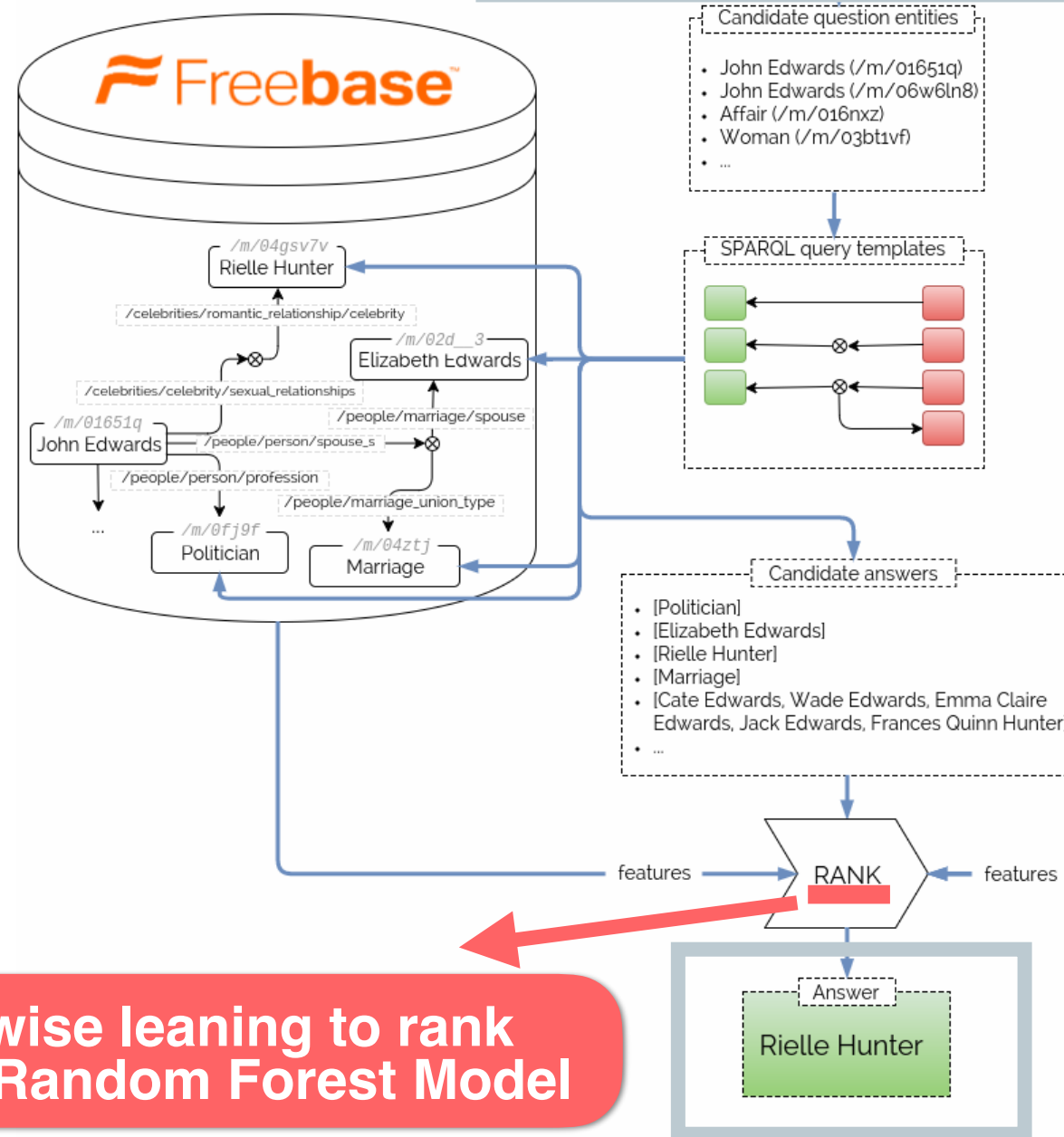


step 3

Introduction(cont.)

- Text2KB Question Answering system's architecture

input Question



pairwise learning to rank
using Random Forest Model

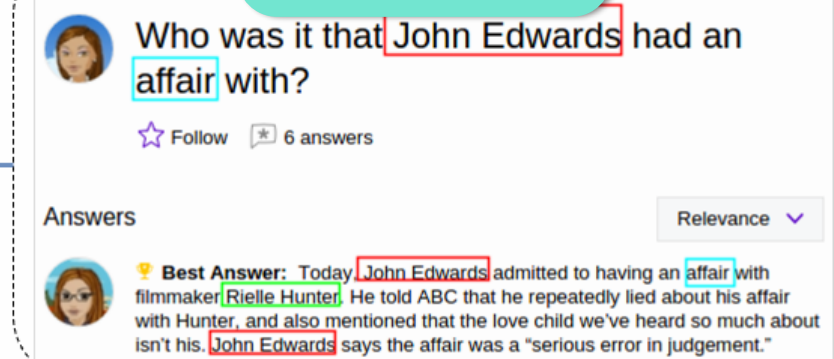
output Answer

Web Search Results



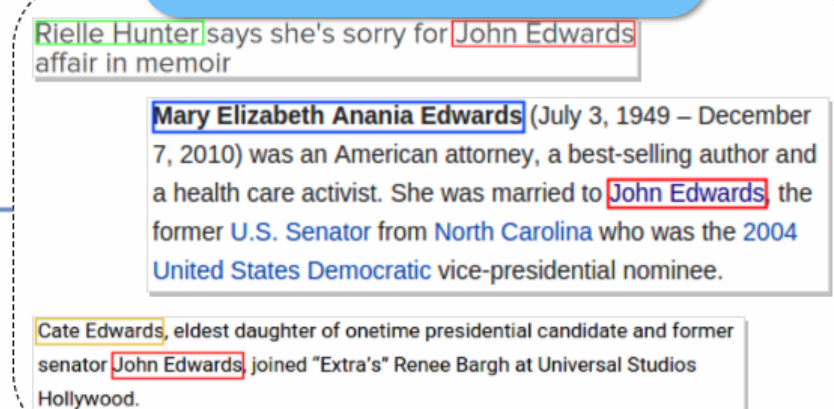
step 1

CQA data



step 2

Document Collection



step 3

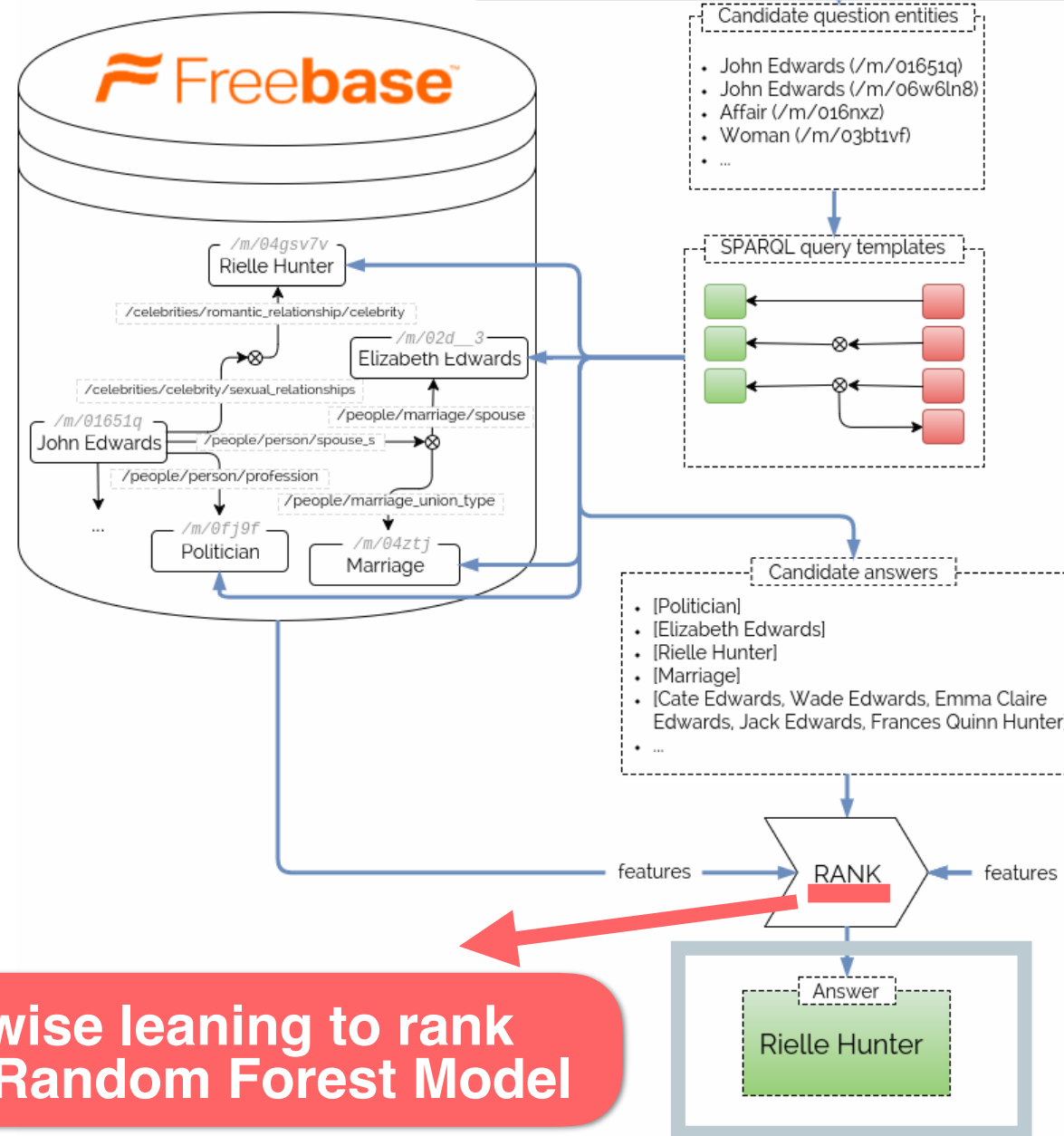
Outline

- Introduction
- **Approach**
- Experiment
- Conclusion

Approach

- Text2KB Question Answering system's architecture

input Question



pairwise learning to rank
using Random Forest Model

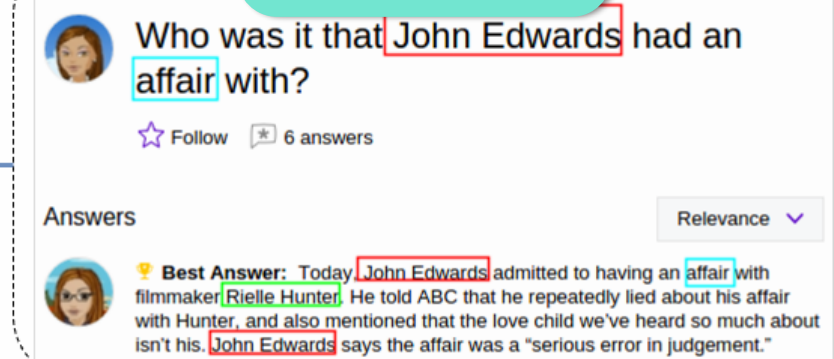
output Answer

Web Search Results



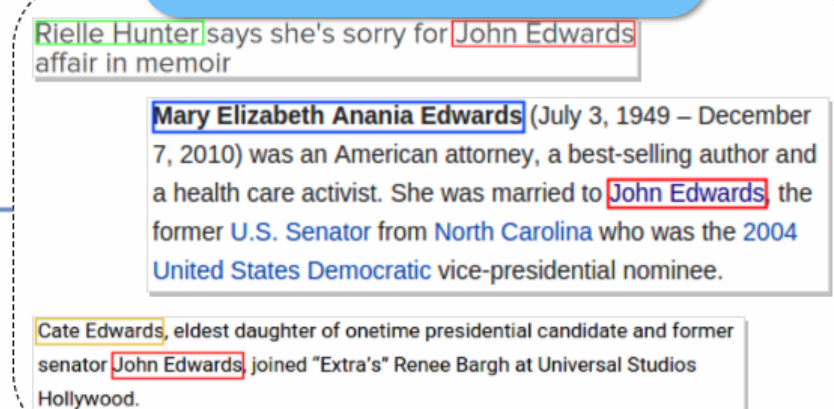
step 1

CQA data

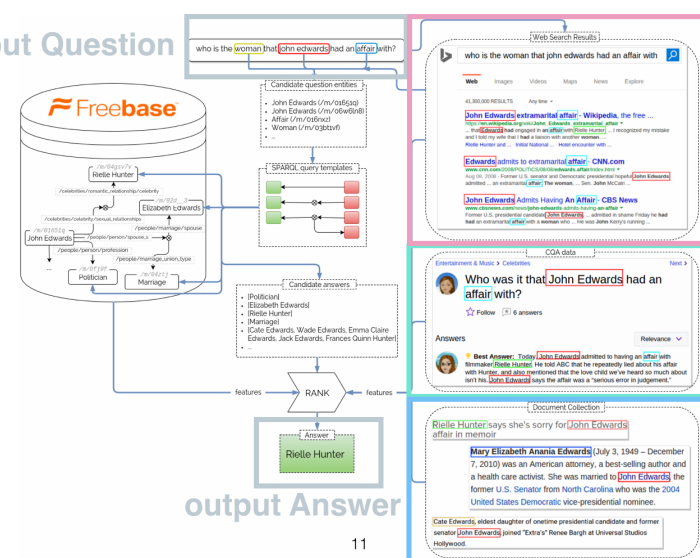


step 2

Document Collection



step 3



Approach(cont.)

Jaro-Winkler string distance

• Web search results for KBQA:

• Question entity identification:

$$\max_{e_t \in M \setminus Stop, q_t \in Q \setminus Stop} 1 - \text{dist}(e_t, q_t) \geq 0.8$$

The Jaro distance d_j of two given strings s_1 and s_2 is

$$d_j = \begin{cases} 0 & \text{if } m = 0 \\ \frac{1}{3} \left(\frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m-t}{m} \right) & \text{otherwise} \end{cases}$$

Where:

- m is the number of *matching characters* (see below);
- t is half the number of *transpositions* (see below).

$$d_w = d_j + (\ell p(1 - d_j))$$

Given the strings s_1 MARTHA and s_2 MARHTA we find:

- $m = 6$
- $|s_1| = 6$
- $|s_2| = 6$
- There are mismatched characters T/H and H/T leading to $t = \frac{2}{2} = 1$

We find a Jaro score of:

$$d_j = \frac{1}{3} \left(\frac{6}{6} + \frac{6}{6} + \frac{6-1}{6} \right) = 0.944$$

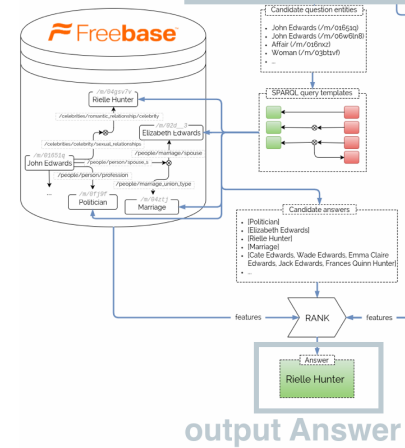
$$\ell = 3$$

Thus:

$$d_w = 0.944 + (3 * 0.1(1 - 0.944)) = 0.961$$

The screenshot shows Google search results for the query 'who is the woman that john edwards had an affair with'. The top results are:

- John Edwards extramarital affair - Wikipedia, the free ...**
https://en.wikipedia.org/wiki/John_Edwards_extramarital_affair
... that Edwards had engaged in an affair with Rielle Hunter ... I recognized my mistake and I told my wife that I had a liaison with another woman, ...
Rielle Hunter and ... Initial National ... Hotel encounter with ...
- Edwards admits to extramarital affair - CNN.com**
www.cnn.com/2008/POLITICS/08/08/edwards.affair/index.html
Aug 08, 2008 · Former U.S. senator and Democratic presidential hopeful John Edwards admitted ... an extramarital affair. The woman, ... Sen. John McCain ...
- John Edwards Admits Having An Affair - CBS News**
www.cbsnews.com/news/john-edwards-admits-having-an-affair
Former U.S. presidential candidate John Edwards, ... admitted in shame Friday he had had an extramarital affair with a woman who ... He was John Kerry's running ...



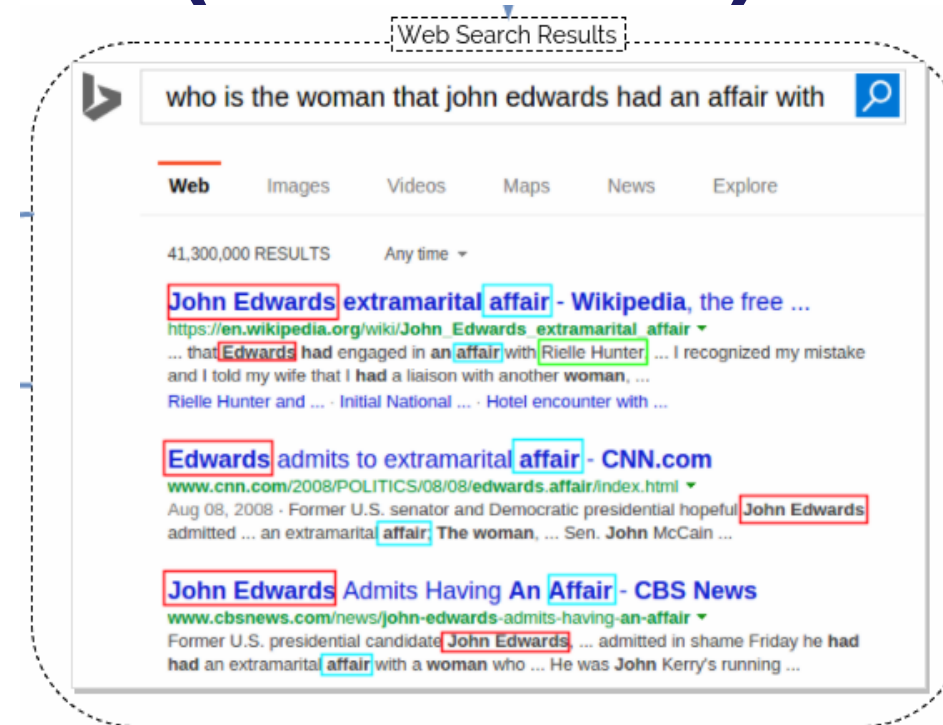
11

Approach(cont.)

step 1

step 2

step 3



• Web search results for KBQA:

• Answer candidate features:

Step 1. Precompute term and entity IDF scores

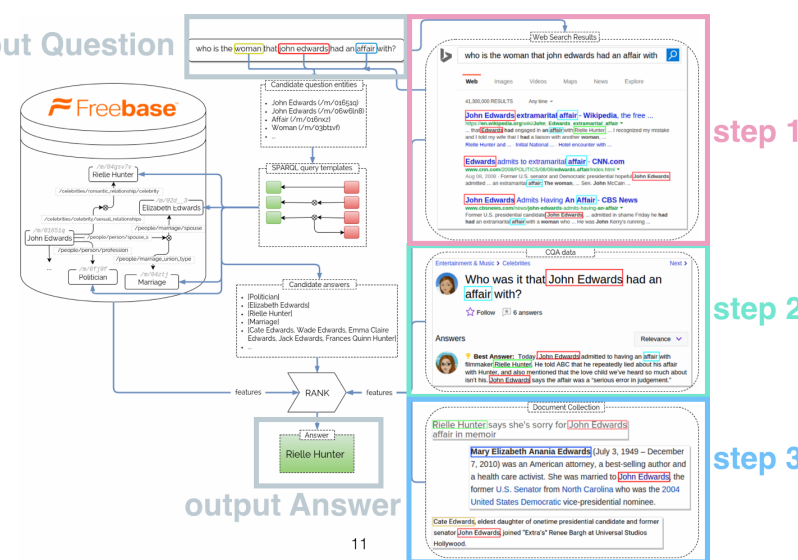
Step 2. Snippet and Document represent by TF-IDF vectors

Step 3. Combined token and entity vectors

Step 4. Answer candidate represent by TF-IDF vectors as well

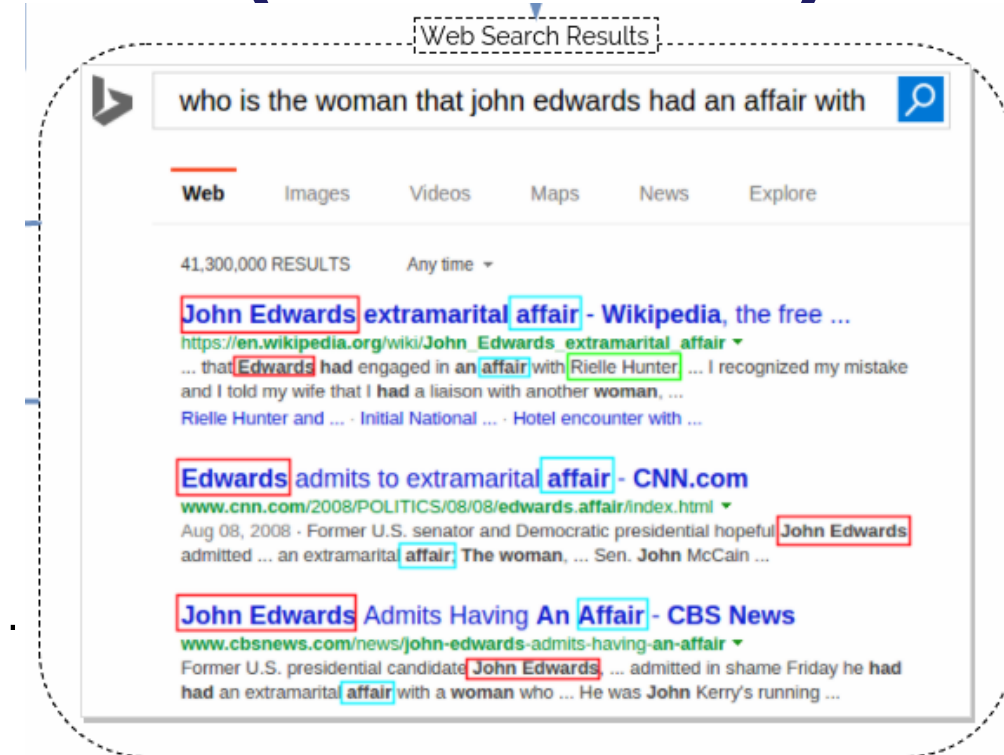
Step 5. Cosine similarities between answer and each of 10 snippet and document. Using average score and maximum score as features

Step 6. Compute Answer similarities as well



11

Approach(cont.)



• Web search results for KBQA:

• Answer candidate features:

Step 1. John Edwards: 0.7, affair: 0.3, Rielle Hunter: 0.7, ...

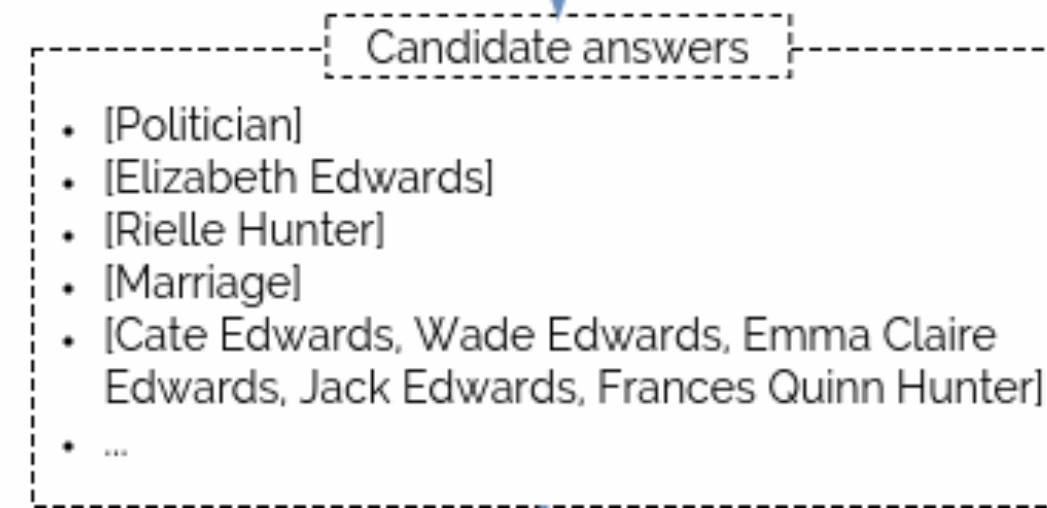
Step 2. $(6 \times 0.7, 5 \times 0.3, 1 \times 0.7) = (4.2, 1.5, 0.7)$, ...

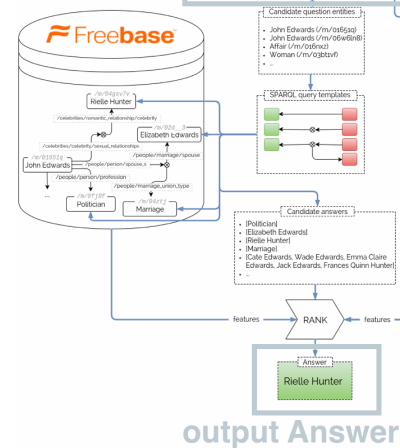
Step 3. Combined token and entity vectors

Step 4. [Politician](0, 0, 0), [Elizabeth Edwards](0, 0, 0), [Rielle Hunter](0, 0, 0.7), ...

Step 5. Cosine similarities between answer and snippet, $(4.2, 1.5, 0.7) \times (0, 0, 0.7) = 0.49$, ...

Step 6. Features: 0.49, ...





11

step 1

step 2

step 3

Approach(cont.)

- CQA data for Matching Questions to Predicates:

- Dataset: Yahoo! WebScope L6 dataset, question and answer texts were run through an entity linker.
- Label question-answer pair with predicates between entities mentioned in the question and in the answer.

- PMI
$$\text{pmi}(x; y) \equiv \log \frac{p(x, y)}{p(x)p(y)} = \log \frac{p(x|y)}{p(x)} = \log \frac{p(y|x)}{p(y)}$$

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.

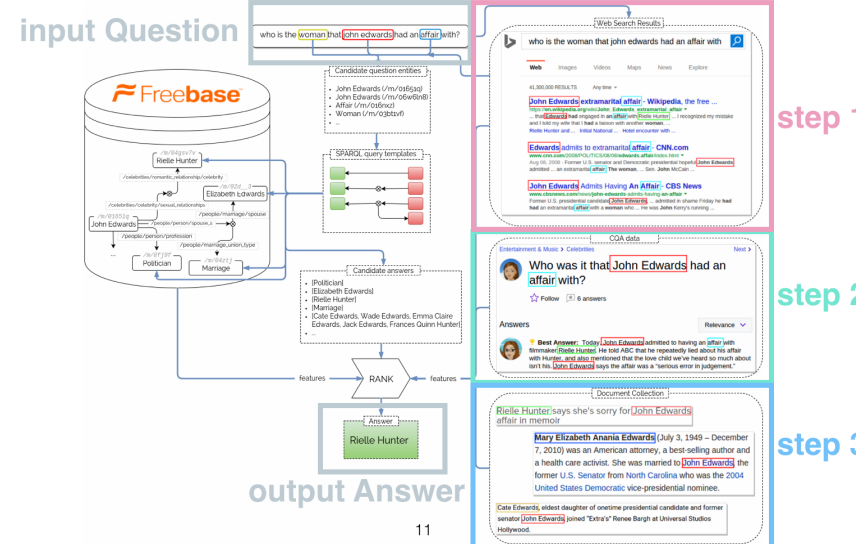
So which ever one you meant, he wasn't born in Africa.

Predicate	Term
people.person.date_of_birth people.person.date_of_death location.location.people_born_here	born



- CQA data for Matching Questions to Predicates:

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



Approach(cont.)

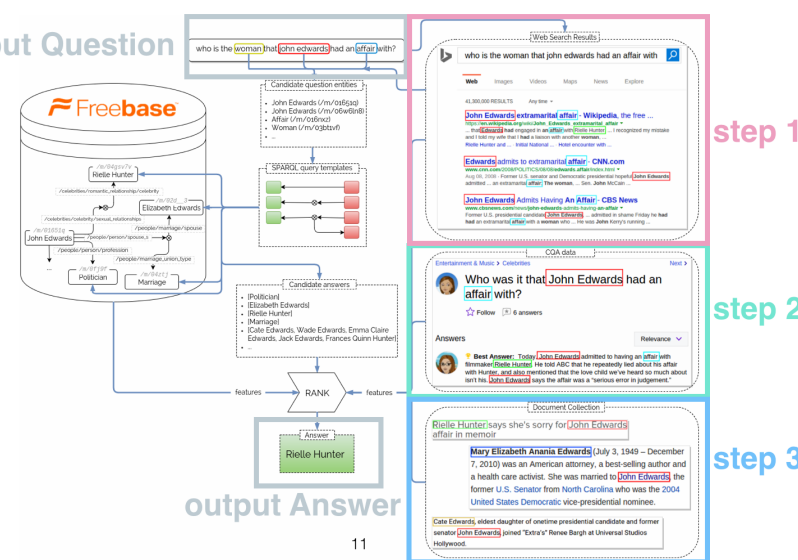
- Estimating Entity Associations: question and answer entities are likely to be mentioned together
- Ranking candidate answers through textual data(ClueWeb12 corpus)

Language Model score:
$$p(Q|e_1, e_2) = \prod_{t \in Q} p(t|e_1, e_2)$$

The diagram shows the components of the equation:
 - **term**: points to the variable t in the product.
 - **question entity**: points to e_1 .
 - **answer entities**: points to e_2 .

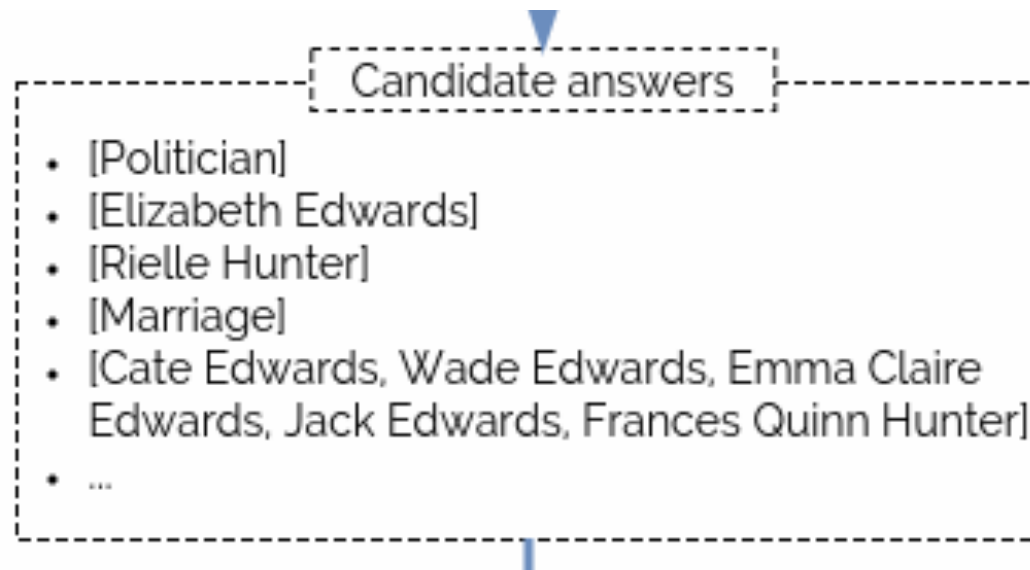
Entity 1	Entity 2	Term counts
John Edwards	Rielle Hunter	campaign, affair, mistress, child, former ...
John Edwards	Cate Edwards	daughter, former, senator, courthouse, greensboro, eldest ...
John Edwards	Elizabeth Edwards	wife, hunter, campaign, affair, cancer, rielle, husband ...
John Edwards	Frances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

use the minimum, average, maximum score over all answer entities as features



Approach(cont.)

- Pairwise learning to rank model:



(Politician, Elizabeth Edwards) → +1 (Politician is better than Elizabeth Edwards)

(Politician, Rielle Hunter) → -1 (Rielle Hunter is better than Politician)

(Politician, Marriage) → +1 (Politician is better than Marriage)

.....

- Classifier: Random Forest Model

Outline

- Introduction
- Approach
- **Experiment**
- Conclusion

Experiment

- Methods Compare:
 - Aqqu
 - Text2KB(Web search): Bing search
 - Text2KB(Wikipedia search)
 - STAGG: The current highest performing KBQA system as measured on the WebQuestion dataset.

Experiment(cont.)

- Datasets: standard evaluation procedure for the WebQuestions dataset
 - The original 70-30% train- test split (3,778 training and 2,032 test instances).
 - Within the training split, 10% was set aside for validation.

Experiment(cont.)

- Evaluation Metrics
 - WebQuestions dataset have primarily used the average F1-score as the main evaluation metric

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

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

$$precision(a_q^*, a_q) = \frac{|a_q^* \cap a_q|}{|a_q|} \text{ and } recall(a_q^*, a_q) = \frac{|a_q^* \cap a_q|}{|a_q^*|}$$

- a_q^* : correct answers
- a_q : given answers

Experiment(cont.)

- Methods Compare:

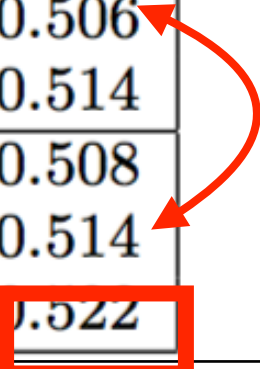
System	avg Recall	avg Precision	F1 of avg P and R	avg F1
OpenQA [16]	-	-	-	0.35
YodaQA [4]	-	-	-	0.343
Jacana [30]	0.458	0.517	0.486	0.330
SemPre [6]	0.413	0.480	0.444	0.357
Subgraph Embeddings [10]	-	-	0.432	0.392
ParaSemPre [7]	0.466	0.405	0.433	0.399
Kitt AI [28]	0.545	0.526	0.535	0.443
AgendaIL [8]	0.557	0.505	0.530	0.497
STAGG [31]	0.607	0.528	0.565	0.525
Aququ (baseline) [3]	0.604	0.498	0.546	0.494
Text2KB (Wikipedia search)	0.632* (+4.6%)	0.498	0.557* (+2.0%)	0.514* (+4.0%)
Text2KB (Web search)	0.635* (+5.1%)	0.506* (+1.6%)	0.563* (+3.1%)	0.522* (+5.7%)

Experiment(cont.)

- Datasource and Features Contribution:
 - T: notable type score model as a ranking feature
 - DF: date range filter-based query template
 - WebEnt: using web search result snippets for question entity identification
 - WikiEnt: using wikipedia search result snippets for question entity identification
 - Web: using web search results for feature generation
 - Wiki: using wikipedia search results for feature generation
 - CQA: using CQA-based [question term, KB predicate] PMI scores for feature generation
 - CW: features, computed from entity pairs language model, estimated on ClueWeb

Experiment(cont.)

System	R	P	F1
Aqqu	0.604	0.498	0.494
Text2KB (base) = Aqqu+DF+T	0.617	0.481	0.499
+Wiki+CQA+CL	0.623	0.487	0.506
+WikiEnt +Wiki+CQA+CL	0.632	0.498	0.514
+WebEnt	0.627	0.492	0.508
+Web+CQA+CL	0.634	0.497	0.514
+WebEnt +Web+CQA+CL	0.635	0.506	0.522



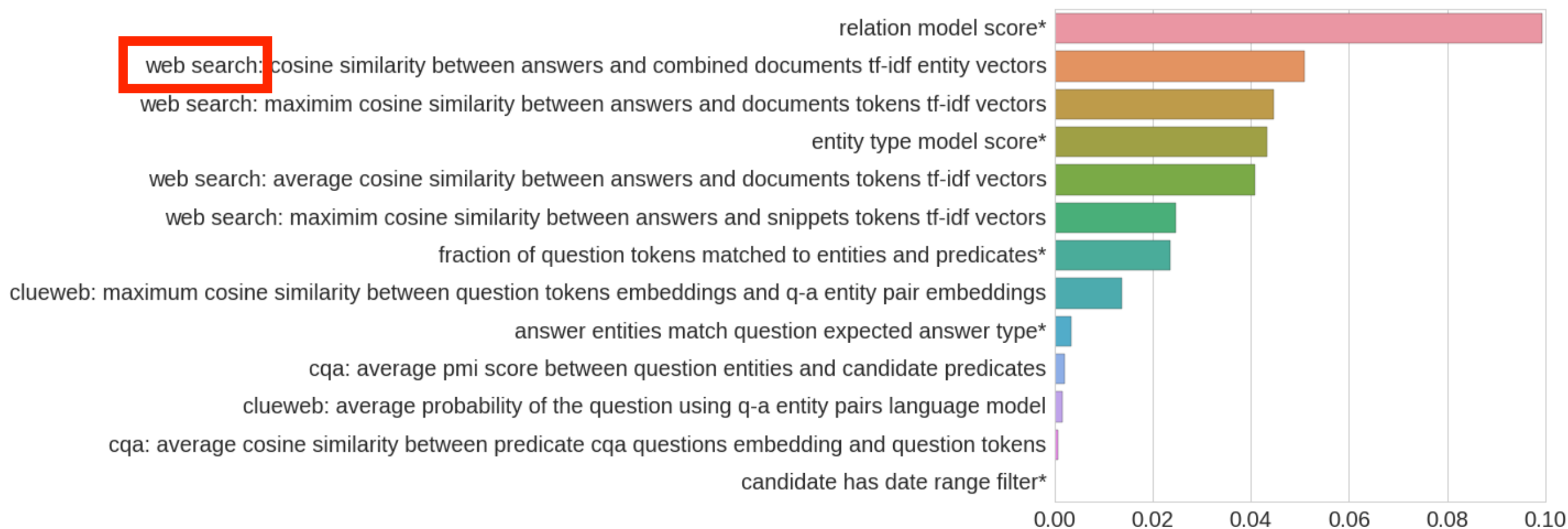
Average Recall (R), Precision (P), and F1 of Aqqu and Text2KB system with and without different components.

System	R	P	F1
Text2KB (Web search)	0.635	0.506	0.522
Text2KB -Web	0.633	0.496	0.513
Text2KB -CQA	0.642	0.499	0.519
Text2KB -CL	0.644	0.505	0.523
Text2KB -CQA-CL	0.642	0.503	0.522
Text2KB -Web-CQA	0.631	0.498	0.514
Text2KB -Web-CL	0.622	0.493	0.508

Average Recall (R), Precision (P), and F1 of Text2KB with and without features based on web search results, CQA data and ClueWeb collection.

Experiment(cont.)

- Feature Importance for Ranking:



A plot of Gini importances of different features of our answer ranking random forest model (features marked * are not text-based and are provided for comparison)

Outline

- Introduction
- Approach
- Experiment
- **Conclusion**

Conclusion

- Unstructured text resources can be effectively utilized for knowledge base question answering.
- Three particular techniques as follows:
 - Web search results for query understanding and candidate ranking.
 - Community question answering data for candidate generation
 - Text fragments around entity pair mentions for ranking

Conclusion(cont.)

- Future work:
 - Extend our work to the more open setup, similar to the benchmark QALD(Question Answering over Linked Data) hybrid task
 - Questions no longer have to be answered exclusively from the KB. This would require extending the described techniques, and creating new QA benchmarks.