Open Question Answering Over Curated and Extracted Knowledge Bases

Anthony Fader, Luke Zettlemoyer, Oren Etzioni
What is illegal in the US but legal in Mexico?
About 198,000,000 results (0.68 seconds)

**Illegal immigration to the United States** - Wikipedia, the free encyclopedia
en.wikipedia.org/.../Illegal_immigration_to_the_United_States  Wikipedia
Jump to **Mexican federal and state government assistance** - [edit]. The US Department of Homeland Security and some advocacy groups have criticized a ...

**Gun politics in Mexico** - Wikipedia, the free encyclopedia
A common misconception is that firearms are **illegal** in Mexico and that no .... Decree for the Liberty of **Mexican America** in Article 81 **prohibited** appearing at Vestry .... outside the home (in public) was no longer a right **but** a privilege federal **law** ...

**Immigration to the United States** - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Immigration_to_the_United_States  Wikipedia
Jump to **Illegal immigration** - The majority of the illegal immigrants are from Mexico. ...
"But we're also a nation of **laws**, that's part of our tradition," he ...
Question Answering

What is illegal in the US but legal in Mexico?

Find a concept $x$ that satisfies:
- $x$ illegal in US
- $x$ legal in Mexico

$x =$ Cuban cigars, absinthe
Previous Approaches to QA

Semantic Parsing for QA
How many states have a higher point than the highest point of the state with the largest capital city in the US?
Zelle & Mooney, ‘96
Zettlemoyer & Collins, ‘05
Liang et al., ‘11

Answer Retrieval from Documents
Who is Tom Cruise married to?
Tom Cruise is married to Katie Holmes, also an actress.
Voorhees & Tice, ‘00
Ravichandran & Hovy '02

Open QA: The Best of Both Worlds
Open-Domain KBs

Knowing is half the battle!

Don’t require: Normalization, Canonicalization, Ontologization

(banana, source of, potassium)
(mushroom, is-a, pizza topping)
(quinoa, compatible with dietary restrictions, gluten-free diet)
## Knowledge Base Statistics

<table>
<thead>
<tr>
<th>Source</th>
<th>Type</th>
<th># Triples</th>
<th># Relation Phrases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Freebase</td>
<td>Curated</td>
<td>300M</td>
<td>18K</td>
</tr>
<tr>
<td>Open IE</td>
<td>Extracted</td>
<td>500M</td>
<td>6M</td>
</tr>
<tr>
<td>Probase</td>
<td>Extracted</td>
<td>200M</td>
<td>1</td>
</tr>
<tr>
<td>NELL</td>
<td>Extracted</td>
<td>2M</td>
<td>300</td>
</tr>
</tbody>
</table>
OQA Contributions

- Combine extracted and curated KBs for QA
- Learn to interpret questions from KBs, question corpus, and QA pairs
- Empirical evaluation on multiple question sets
Comparison to Sempre
(Berant et al., 2013)
Effect of Knowledge Sources

- WebQuestions
  - All KBs
  - No Open IE
  - No Freebase
  - No Probase
  - No NELL
- TREC
  - All KBs
  - No Open IE
  - No Freebase
  - No Probase
  - No NELL
- WikiAnswers
  - All KBs
  - No Open IE
  - No Freebase
  - No Probase
  - No NELL

F1 scores:
- WebQuestions: [0, 0.35]
- TREC: [0, 0.29]
- WikiAnswers: [0, 0.08]
Future Work

Answerability

Compositional Analysis of Questions

Question-Guided Information Extraction
Conclusion - OQA

- Multiple knowledge sources
- Learns from KB, question corpus, QA pairs
- Has high performance across question sets
Thanks!

github.com/afader/oqa
Experiments

- How does OQA compare to state-of-the-art QA systems?
- How do different knowledge sources affect performance?
- How do different system components affect performance?
Question Sets

WebQuestions  Berant et al., 2013
where was nicki minaj born?

TREC  Vorhees and Tice, 2000
What other countries do kurds live in ?

WikiAnswers  Held-out from corpus
Who is a retired gay nfl player?
Learning and Inference

Learning w/QA pairs: Clarke et al., ‘10, Berant et al., ‘13, Kwiatkowski et al., ‘13
Latent-Variable Structured Perceptron: Liang et al., 2006; Sun et al., 2009

How can I tell if I have the flu?

\[ w = w + f(... \rightarrow \text{the chills}) - f(... \rightarrow \text{virus}) \]

Derivations not available in training data
## Example Errors

<table>
<thead>
<tr>
<th>Input</th>
<th>What animal represents California?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paraphrase</td>
<td>What are California’s symbols?</td>
</tr>
<tr>
<td>Parse</td>
<td>?x: (california, symbols, ?x)</td>
</tr>
<tr>
<td>Execute</td>
<td><strong>CWT:</strong></td>
</tr>
<tr>
<td></td>
<td>(California Water Service, Trading symbol, CWT)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Input</th>
<th>What actor first portrayed James Bond?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parse</td>
<td>?x: (?x, is-a, actor first)</td>
</tr>
<tr>
<td></td>
<td>(?x, portrayed, James Bond)</td>
</tr>
<tr>
<td>Execute</td>
<td><strong>Daniel Craig:</strong></td>
</tr>
<tr>
<td></td>
<td>(Daniel Craig, is-a, first class actor)</td>
</tr>
<tr>
<td></td>
<td>(Danny Craig, portrays, James Bond)</td>
</tr>
</tbody>
</table>
Comparison to Paralex

WebQuestions
- OQA
- PARALEX

TREC
- OQA
- PARALEX

WikiAnswers
- OQA
- PARALEX

Precision vs. Recall graphs for WebQuestions, TREC, and WikiAnswers comparing OQA, PARALEX, and QOA.
Effect of System Components

- **WebQuestions**
  - Full Model: 0.35
  - No Weight Learning: 0.25
  - No Paraphrases: 0.20
  - No Query Rewrites: 0.15

- **TREC**
  - Full Model: 0.29
  - No Weight Learning: 0.20
  - No Paraphrases: 0.15
  - No Query Rewrites: 0.10

- **WikiAnswers**
  - Full Model: 0.08
  - No Weight Learning: 0.06
  - No Paraphrases: 0.05
  - No Query Rewrites: 0.04
## Mining Query Rewrite Operators

<table>
<thead>
<tr>
<th>arg1</th>
<th>arg2</th>
<th>sign of</th>
<th>symptom$^{-1}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>cough</td>
<td>cold</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>jealousy</td>
<td>love</td>
<td>✔</td>
<td></td>
</tr>
<tr>
<td>diziness</td>
<td>Meniere's</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>chills</td>
<td>flu</td>
<td></td>
<td>✔</td>
</tr>
</tbody>
</table>

74 million \{r_1, r_2\} pairs with ≥ 10 shared arguments

DIRT (Lin and Pantel, 2001)
Cherry-Picked Rewrite Rules

<table>
<thead>
<tr>
<th>Source Relation</th>
<th>Target Relation</th>
</tr>
</thead>
<tbody>
<tr>
<td>(?x, children, ?y)</td>
<td>(?y, was born to, ?x)</td>
</tr>
<tr>
<td>(?x, birthdate, ?y)</td>
<td>(?x, date of birth, ?y)</td>
</tr>
<tr>
<td>(?x, headquartered in, ?y)</td>
<td>(?y, is based in, ?x)</td>
</tr>
<tr>
<td>(?x, invented, ?y)</td>
<td>(?y, was invented by, ?x)</td>
</tr>
<tr>
<td>(?x, is the language of, ?y)</td>
<td>(?y, languages spoken, ?x)</td>
</tr>
</tbody>
</table>
WikiAnswers Paraphrase Corpus

- How can I tell if I have strep throat?
- What are signs of strep throat?
- Am I sick with strep throat?
- What are strep throat symptoms?

20 million user-created clusters
### Mining Paraphrase Templates

<table>
<thead>
<tr>
<th>$t_1$</th>
<th>$t_2$</th>
<th>count($t_1, t_2$)</th>
<th>count($t_1$)</th>
<th>count($t_2$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>How can I tell if I have _</td>
<td>What are signs of _</td>
<td>8</td>
<td>21</td>
<td>45</td>
</tr>
<tr>
<td>How can I tell if I have _</td>
<td>Am I sick with _</td>
<td>11</td>
<td>21</td>
<td>52</td>
</tr>
<tr>
<td>How can I tell if I have strep _</td>
<td>What are signs of strep _</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

5 million $\{t_1, t_2\}$ pairs with $\text{count}(t_1, t_2) \geq 5$
# Cherry-Picked Paraphrase Templates

<table>
<thead>
<tr>
<th>Template 1</th>
<th>Template 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>How does _ affect your body?</td>
<td>What body system does _ affect?</td>
</tr>
<tr>
<td>What is the latin name for _?</td>
<td>What is _’s scientific name?</td>
</tr>
<tr>
<td>Why do we use _?</td>
<td>What did _ replace?</td>
</tr>
<tr>
<td>What to use instead of _?</td>
<td>What is a substitute for _?</td>
</tr>
<tr>
<td>Was _ ever married?</td>
<td>Who has _ been married to?</td>
</tr>
</tbody>
</table>
Task

Input = Question
Output = Most confident answer (or none)

Precision = correct / returned answers
Recall = correct / questions
Baseline Systems

Paralex *(Fader et al., 2013)*
Learns from WikiAnswers corpus
Lexicalized feature representation
Queries all knowledge sources

Sempre *(Berant et al., 2013)*
Learns from QA pairs
Lexicalized feature representation
Queries Freebase only