Testing a Knowledge Inquiry System on Question Answering Tasks

Kyriaki D. Zafeiroudi, Leah Eckman, and Rebecca J. Passonneau

The Pennsylvania State University

kzafeiroudi@cse.psu.edu
Outline

❖ Motivation
❖ InK
   ❖ Incremental Grounding
   ❖ Radiating
   ❖ Traversing
❖ Experiments
❖ Conclusion
Motivation

大多数问答系统接受自然语言的问题，并在知识库中找到答案后返回结果。

但如果没有好的答案可用怎么办？

我们介绍InK，一个初步的询情系统，它专注于在没有好答案可用时提出有帮助的回应。
Motivation

- InK is not tailor-made to a particular knowledge base:
  - Agnostic to the terminology of the KB.
  - Transparency of the knowledge it has access to.

- Semantic parsing not always possible:
  - Arbitrariness of KB terminology.
    - What’s the river beneath Brooklyn Bridge?
    - DBpedia ontology concept: crosses
  - Cannot always much linguistic structure to KB.
    - What color are Liza Kennedy’s eyes?
    - DBpedia ontology concept: eyeColor
Q: In which country is Solaize located?

- InK identifies *country* and *Solaize*.
- Gradually matches the two entity mentions to the KB.
- Retrieves relevant information for each entity mention.
- Finds connections between the entity mentions.
- Returns a response.
InK

Testing a Knowledge Inquiry System on Question Answering Tasks
Incremental Grounding

- Determining what a user’s question is about is a two step process:
  - Extracting entity mentions.
  - Grounding the extracted entity mentions to nodes in a KB.

- The following entity mention types are extracted using various toolkits:
  - Noun phrase
  - Named entity
  - Proper noun
  - Noun-noun compound
  - Noun
Incremental Grounding

- In order to identify the KB nodes that correspond to the extracted entity mentions, exact string matching is used. Many cases arise where more than one match is available.

- For every question it is preferable to have at least two entity mention matches; further preprocessing of the question takes place otherwise:
  - The main verb of the question is extracted and lemmatized.
  - Using WordNet’s synsets, the most probable morphological variant of the verb as a noun is extracted and matched to the KB.

- After the candidate concepts are identified, radiating and traversing apply.
Radiating

- A radiating result for an entity is the set of triples where this entity occurs as subject or object.
- The resulting information consist of a subgraph of diameter one centered on a node that matches an entity mention extracted from the input question.
- Thicker edges indicate more highly ranked predicates.
Radiating

A radiating result for an entity is the set of triples where this entity occurs as subject or object.

The resulting information consist of a subgraph of diameter one centered on a node that matches an entity mention extracted from the input question.

Thicker edges indicate more highly ranked predicates.
Radiating

To rank the informativeness of predicates $pr$ we use:

$$- \log p(pr)$$

where $p(pr) = \frac{\text{# of triples $pr$ is the predicate}}{\text{# of triples}}$
Traversing

- The extracted entity mentions along with their matched nodes to the KB usually consist of a big group in need of disambiguation.
- Traversing finds the shortest path that connects two entity mentions extracted from a NL question – if one exists.
- The return of this procedure is not a subgraph surrounding one core entity, but sequential RDF triples, where the object of the first triple serves as the subject of the next triple, etc.
Traversing

In cases where traversing returns any paths, at least one of them will often correspond to a sensible interpretation of the entity mentions, and can even correspond to an answer.

After extracting *language* and *Romania*, traversing returns 22 short paths, 18 of which consist the right answer to the question.
Pilot User Study: Traversing

- InK aims to balance relevance, informativeness and succinctness.
- A user study was designed to test whether users would recognize and value these three aspects of InK.
- Sixty questions were gathered from different QA datasets (Free917, WebQuestions & SQuAD).
- This initial pilot study investigated traversing responses only.
- Traversing could apply to 36 of the 60 questions, while only 22 of these returned results.
Pilot User Study: Traversing

For each question and traversing result, subjects were given the five Likert-scale questions below.

Questions A-B assess whether the questions make sense, and whether respondents are already aware of the answer.

Questions C-E roughly correspond to intelligibility of the triples in the response, their relevance to the question, and how informative they are.

<table>
<thead>
<tr>
<th>Item</th>
<th>Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>How well do you understand the question?</td>
</tr>
<tr>
<td>B</td>
<td>How certain are you that you knew the answer before you saw the response?</td>
</tr>
<tr>
<td>C</td>
<td>How easily can you interpret the triples shown here?</td>
</tr>
<tr>
<td>D</td>
<td>How closely related are the triples shown here to the question?</td>
</tr>
<tr>
<td>E</td>
<td>How well do the triples shown here answer the question?</td>
</tr>
</tbody>
</table>
For each item, the average item score is reported, along with the standard deviation.

Questions were considered largely understandable due to items A-B.

For the 22 questions where traversing returned triples, the scores for items C-E indicate that the output was fairly clear, fairly to somewhat related to the question, and to some degree answered the question.

<table>
<thead>
<tr>
<th>Item</th>
<th>All</th>
<th>Free917</th>
<th>WebQ</th>
<th>SQuAD</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>4.51</td>
<td>4.51</td>
<td>4.66</td>
<td>4.38</td>
</tr>
<tr>
<td>B</td>
<td>2.08</td>
<td>1.93</td>
<td>2.28</td>
<td>2.06</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Twenty-two Questions with Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
</tr>
<tr>
<td>D</td>
</tr>
<tr>
<td>E</td>
</tr>
</tbody>
</table>
Experiments

- In order to test our system on a QA task, we used the QALD datasets. The goal of the QALD Question Answering over Linked Data challenges held every year is to benchmark QA systems for RDF datasets. We focus on the datasets where the user’s question is in English and the answer can be retrieved by posing a SPARQL query against DBpedia.

- InK proved to be quite efficient; average runtime in seconds is 2.53 for parsing, 1.30 for incremental grounding, 11.84 for radiating, and 42.08 for traversing.
Experiments

- We used recall as a metric:

\[
Recall(Q) = \frac{\text{# of correct answers for } Q \text{ occurring in InK response}}{\text{# of gold standard answers for } Q}
\]

- Recall is calculated separately for radiating and traversing, and their combination.

<table>
<thead>
<tr>
<th>Dataset</th>
<th># of Questions</th>
<th>Radiating</th>
<th>Traversing</th>
<th>InK</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td># of Questions</td>
<td>train</td>
<td>test</td>
<td>train</td>
</tr>
<tr>
<td>QALD-1</td>
<td>28</td>
<td>0.52</td>
<td>0.73</td>
<td>0.62</td>
</tr>
<tr>
<td>QALD-2</td>
<td>53</td>
<td>0.69</td>
<td>0.56</td>
<td>0.62</td>
</tr>
<tr>
<td>QALD-7</td>
<td>79</td>
<td>-</td>
<td></td>
<td>0.69</td>
</tr>
<tr>
<td>All</td>
<td>160</td>
<td>0.66</td>
<td>0.61</td>
<td>0.64</td>
</tr>
</tbody>
</table>
Experiments

- InK’s performance was compared against other systems from QALD-1 and QALD-2 challenges.
- For fair comparison, we recalculated InK recall over the same sets of questions the other systems evaluated on.
- No systems competed on QALD-7 using the same QA dataset we use here.
- InK has higher recall except in comparison to PowerAqua from QALD-1.

<table>
<thead>
<tr>
<th>System</th>
<th>Questions</th>
<th>Recall</th>
<th>InK Matched Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>PowerAqua</td>
<td>23</td>
<td>0.79</td>
<td>0.77</td>
</tr>
<tr>
<td>FREyA</td>
<td>23</td>
<td>0.72</td>
<td>0.77</td>
</tr>
<tr>
<td>SemSeK</td>
<td>54</td>
<td>0.56</td>
<td>0.64</td>
</tr>
<tr>
<td>MHE</td>
<td>53</td>
<td>0.49</td>
<td>0.63</td>
</tr>
<tr>
<td>QAKis</td>
<td>26</td>
<td>0.42</td>
<td>0.73</td>
</tr>
</tbody>
</table>
Conclusion

- InK grounds Natural Language questions in knowledge graph concepts through a set of procedures, agnostically to the semantics of the knowledge graph.
- Contentful phrases and words are independently matched to concepts in the knowledge graph, and a succession of queries searches for triples that contain these concepts.
- A set of triples is assembled into a response that aims for relevance, informativeness and concision.
- In the process, potential ambiguity of concept grounding is often resolved.
- An evaluation of the recall on InK responses to factual questions shows that the responses often contain the correct answer.
Thank you for your time!

Questions?