Answering Elementary Science Questions by Constructing Coherent Scenes using Background Knowledge

Yang Li∗
UC Santa Barbara
Santa Barbara, CA, USA
yangli@cs.ucsb.edu

Peter Clark
Allen Institute for Artificial Intelligence
Seattle, WA, USA
peterc@allenai.org

Abstract

Much of what we understand from text is not explicitly stated. Rather, the reader uses his/her knowledge to fill in gaps and create a coherent, mental picture or “scene” depicting what text appears to convey. The scene constitutes an understanding of the text, and can be used to answer questions that go beyond the text.

Our goal is to answer elementary science questions, where this requirement is pervasive; A question will often give a partial description of a scene and ask the student about implicit information. We show that by using a simple “knowledge graph” representation of the question, we can leverage several large-scale linguistic resources to provide missing background knowledge, somewhat alleviating the knowledge bottleneck in previous approaches. The coherence of the best resulting scene, built from a question/answer-candidate pair, reflects the confidence that the answer candidate is correct, and thus can be used to answer multiple choice questions. Our experiments show that this approach outperforms competitive algorithms on several datasets tested. The significance of this work is thus to show that a simple “knowledge graph” representation allows a version of “interpretation as scene construction” to be made viable.

1 Introduction

Elementary grade science tests are challenging as they test a wide variety of commonsense knowledge that human beings largely take for granted, yet are very difficult for machines (Clark, 2015). For example, consider a question from a NY Regents 4th Grade science test:

∗Work was done while the author was an intern at Allen Institute for Artificial Intelligence.

Question 1 “When a baby shakes a rattle, it makes a noise. Which form of energy was changed to sound energy?” [Answer: mechanical energy]

Science questions are typically quite different from the entity-centric factoid questions extensively studied in the question answering (QA) community, e.g., “In which year was Bill Clinton born?” (Ferrucci et al., 2010; Yao and Van Durme, 2014). While factoid questions are usually answerable from text search or fact databases, science questions typically require deeper analysis. A full understanding of the above question involves not just parsing and semantic interpretation; it involves adding implicit information to create an overall picture of the “scene” that the text is intended to convey, including facts such as: noise is a kind of sound, the baby is holding the rattle, shaking involves movement, the rattle is making the noise, movement involves mechanical energy, etc. This mental ability to create a scene from partial information is at the heart of natural language understanding (NLU), which is essential for answering these kinds of question. It is also very difficult for a machine because it requires substantial world knowledge, and there are often many ways a scene can be elaborated.

We present a method for answering multiple-choice questions that implements a simple version of this. A scene is represented as a “knowledge graph” of nodes (words) and relations, and the scene is elaborated with (node,relation,node) tuples drawn from three large-scale linguistic knowledge resources: WordNet (Miller, 1995), DART (Clark and Harrison, 2009), and the Free-Association database (Nelson et al., 2004). These elaborations reflect the mental process of “filling in the gaps”, and multiple choice questions can then be answered by finding which answer option creates the most coherent scene.

The notion of NLU as constructing a most coherent scene is not new, and has been studied in several contexts including work on scripts (Schank and Abelson, 1977), interpretation as ab-
duction (Hobbs et al., 1988; Hobbs, 1979; Ovchin-
ikova et al., 2014), bridging anaphora (Asher and Lascarides, 1998; Fan et al., 2005), and para-
graph understanding (Zadrozny and Jensen, 1991; Harabagiu and Moldovan, 1997). These meth-
ods are inspiring, but have previously been limited by the lack of background knowledge to supply implicit information, and with the complexity of their representations. To make progress, we have chosen to work with a simple “knowledge graph” representation of nodes (words) and edges (relations). Although we lose some subtlety of expres-
sion, we gain the ability to leverage several vast resources of world knowledge to supply implicit information. The significance of this work is thus to show that, by working with a simple “knowledge graph” representation, we can make a viable version of “interpretation as scene construction”. Although the approach makes several simplifying assumptions, our experiments show that it outper-
forms competitive algorithms on several datasets of (real) elementary science questions.

2 Approach

The input to our question-answering system is a multiple choice question $Q$, a set of an-
ter options $a_k$, and one or more background knowledge base(s) each containing a set of $(word_i, relation, word_j)$ tuples, each denoting that $word_i$ is plausibly related to $word_j$ by relation. The output is a ranked list of the $K$ an-
ter options.

We define a scene $S$ as a “knowledge graph” of nodes (words) and edges (relations between words), where all $(word_i, relation, word_j)$ edges are sanctioned by (contained in) at least one of the background knowledge bases. Each scene node has an associated measure of coherence (described shortly), denoting how well-connected it is. The question-answering objective is, for each answer option $a_k$, to find the most coherent scene containing (at least) the question keywords $kw_i \in Q$ and answer option $a_k$, and then return the an-
ter option with the overall highest coherence score. Our implementation approximates this ob-
jective using a simple elaborate-and-prune algo-
rithm, illustrated in Figure 1 and now described.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure1.png}
\caption{(1) Question keywords are extracted to form the initial scene. (2) The scene is elaborated with background knowledge to add plausible relationships. (3) For each answer option, it is added into the scene and connected with additional relations.
shions. Then the scene is pruned. (4) A score is derived from the final scene, reflecting confidence that the answer option is correct.}
\end{figure}

2.1 Question Analyzer

The initial scene is simply the keywords (non-stop words) $KW = \{kw_i\}$ in the question $Q$, along with a measure of importance $IS(kw_i)$ for each word $kw_i$. For our purposes we compute importance by sending the question to Google, grouping the top 20 result snippets into a document $d$, and computing:

$$IS(kw) = \frac{tf_d(kw)}{df_Q(kw)},$$

where $tf_d(kw)$ is the term frequency of $kw$ in doc-
ument $d$, and $df_Q(kw)$ is the document frequency of $kw$ in question set $Q$ containing all the available elementary science questions. The intuition here is
that the words frequently mentioned in variations of the question should be important (reflected by "tf"), while the descriptive words (e.g. "following", "example") which are widely used in many questions should be penalized (reflected by "idf"). Other methods could equally be used to compute importance.

### 2.2 Builder

In this step our goal is to inject implicit knowledge from the background KBs to form an elaborated knowledge graph. To do this, we first fetch all the triples \((kw, relation, w)\) that are directly connected with any keyword \(kw \in KW\) from the background KBs, as well as all \((kw_i, relation, kw_j)\) triples between keywords. In our experiments we use three background knowledge bases to supply implicit knowledge, although in principle any triple store could be used: WordNet (Miller, 1995), DART (Clark and Harrison, 2009), and the FreeAssociation database (Nelson et al., 2004). Table 1 shows examples of each.

<table>
<thead>
<tr>
<th>KB</th>
<th>Size (# tuples)</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>WordNet</td>
<td>235k</td>
<td>(dog, isa, animal)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(sunlight, isa, energy)</td>
</tr>
<tr>
<td>DART</td>
<td>2.3M</td>
<td>(nutrients, in, food)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(animal, eat, food)</td>
</tr>
<tr>
<td>FreeAssoc</td>
<td>64k</td>
<td>(car, relate_to, tire)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(ice, relate_to, cold)</td>
</tr>
</tbody>
</table>

Table 1: Knowledge Bases Used

These triples introduce new nodes \(w\) into the graph. As we may get a large number of such nodes, we score them and retain only the top scoring ones (and all edges connecting to it). Informally, a new word \(w\) is preferred if it is connected to many important keywords \(kw\) with strong relationships. Formally, the scoring function is:

\[
score(w) = \sum_{kw \in K} IS(kw) \times rel(kw, w)
\]

where \(IS(kw)\) is the importance score of keyword \(kw\) and \(rel(kw, w)\) is the relatedness score between \(kw\) and \(w\). In this work we use the cosine similarity between word2vec (Mikolov et al., 2013) word vectors to measure two words’ relatedness, as a rough proxy for the strength of relatedness in the KBs (the KBs themselves do not provide meaningful strengths of relationship). After the ranking, the top \(N \times |KW|\) neighbor words \(w\) are retained\(^3\), along with their edges to keywords \(kw\) and each other.

Note that at this point the elaboration process is independent of any answer option; rather, the graph depicts the question scenario.

### 2.3 Elaborate and Prune

To score the \(K\) different answer options, the system now builds a different alternative elaborations of the scene so far, each one with answer option \(a_k\) added, and assesses the coherence of the addition. The answer option \(a_k\) that fits “most coherently” with the scene is returned as the answer to the question.

To do this for a given option \(a_k\), we add \(a_k\) to the graph along with all triples \((w, relation, a_k)\) in the KBs that relate any node \(w_i\) in the graph to \(a_k\). Now that the focus \(a_k\) of the question is known, some of the earlier added nodes \(w\) in the graph may be only weakly relevant to the question and answer, and so we add a pruning step to remove these nodes. The goal of this pruning is to find a dense subgraph (i.e. the coherent scene) that would ideally contain all the question keywords \(kw\), the answer option \(a_k\), and extra words \(w_k\) that are highly connected with them.

Inspired by Sozio et al’s work (Sozio and Gionis, 2010) on finding strongly interconnected subgroups in social networks, we have developed an iterative node removal algorithm for extracting this subgraph. We define the coherence of a node as the summed weight of its incident edges:

\[
coherence(w) = \sum_{w' \in \{(w, r, w')\}} rel(w, w')
\]

where \(rel(w, w')\) is the weight of edge \((w, r, w')\) in the graph, again computed using cosine similarity between \(w\) and \(w'\). We then iteratively remove the non-keyword node (and attached edges) with least coherence until the answer option \(a_k\) is about to removed. The resulting graph is thus maximally pruned, subject to the constraint it must still describe the question plus answer option.

Finally, we use the coherence of the answer option \(a_k\) in this final scene as the confidence that \(a_k\) is the correct answer. The system repeats this

\(^3\)The optimal \(N\) (here 6) was selected using an independent set of training questions
for all $K$ answer options and selects the $a_k$ with highest confidence.

## 3 Evaluation

The system was developed using a dataset of natural (unedited) elementary science exam questions, and then tested on three similar, unseen (hidden) datasets. Its performance was compared with two other state-of-the-art systems for this task. As our system only fields questions where the answer options are all single words, we evaluate it, and the other systems, only on these subsets. These subsets are in general easier than other questions, but this advantage is the same for all systems being compared so it is still a fair test.

### 3.1 Evaluation Datasets

The datasets used are the non-diagram, multiple-choice questions with single-word answer options drawn from the following exams:

- **Dev (System Development):** New York Regents 4th Grade Science [4](http://www.nysedregents.org/Grade4/Science/home.html) (47 questions in 6 years)
- **Test1:** New York Regents 4th Grade Science (23 questions in 3 years)
- **Test2:** Additional 4th Grade Science (from multiple States) (26 questions)
- **Test3:** 5th Grade Science (from multiple States) (197 questions)

Although these datasets are small (real exam questions of this type are in limited supply), the numbers are large enough to draw conclusions.

### 3.2 Experiments

We compared our system (called SceneQA) with two other state-of-the-art systems for this task:

- **LSModel** (Lexical semantics model): SVM combination of several language models (likelihood of answer given question) and information retrieval scores (score of top retrieved sentence matching question plus answer), trained on a set of questions plus answers. (An expanded version of Section 4.3 of [Jansen et al., 2014])
- **A*Rules:** “Prove” the answer option from the question by applying lexical inference rules automatically extracted from science texts. Select the option with the strongest “proof”. (Clark et al., 2014)

<table>
<thead>
<tr>
<th></th>
<th>Dev</th>
<th>Test1</th>
<th>Test2</th>
<th>Test3</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSModel</td>
<td>65.96</td>
<td>58.70</td>
<td>28.85</td>
<td>30.08</td>
</tr>
<tr>
<td>A*Rules</td>
<td>65.96</td>
<td>67.00</td>
<td>47.00</td>
<td>29.22</td>
</tr>
<tr>
<td>SceneQA</td>
<td><strong>83.51</strong></td>
<td><strong>66.30</strong></td>
<td><strong>65.38</strong></td>
<td><strong>55.20</strong></td>
</tr>
</tbody>
</table>

Table 2: SceneQA outperforms two competitive systems on two of the three test sets. The highlighted improvements are statistically significant.

The results (% scores, Table 2) show SceneQA significantly outperforms the two other systems on two of the three test sets, including the largest (Test3, 197 questions), suggesting the approach has merit.

We also performed some case studies to identify what kinds of questions SceneQA does well on, relative to the baselines. In general, SceneQA performs well when the question words and the (correct) answer can be tightly related by background knowledge, including through intermediate nodes (words). For example, in Question 2 below:

**Question 2**

Which type of energy does a person use to pedal a bicycle? (A) light (B) sound (C) mechanical (D) electrical

the KB relates the correct answer "mechanical" to the question words "energy", "pedal", "bicycle", and the intermediate node "power" forming a tight graph. In contrast, the other algorithms select the wrong answer "light" due to frequent mentions of "bicycle lights" in their supporting text corpora that confuses their algorithms.

### 3.3 Ablations

We also performed ablations to assess which parts of our method are contributing the most:

- **-NewNodes:** Only add edges but no new nodes $w$ during the Build step.
- **-Prune:** Do not prune nodes during the Elaborate and Prune step.
- **-NewNodes-Prune:** No new nodes, no pruning

<table>
<thead>
<tr>
<th></th>
<th>Dev</th>
<th>Test1</th>
<th>Test2</th>
<th>Test3</th>
</tr>
</thead>
<tbody>
<tr>
<td>SceneQA</td>
<td><strong>83.51</strong></td>
<td><strong>66.30</strong></td>
<td><strong>65.38</strong></td>
<td><strong>55.20</strong></td>
</tr>
<tr>
<td>-NewNodes</td>
<td>65.96</td>
<td>69.57</td>
<td>42.31</td>
<td>51.78</td>
</tr>
<tr>
<td>-Prune</td>
<td>70.74</td>
<td>57.61</td>
<td>47.12</td>
<td>50.13</td>
</tr>
<tr>
<td>-NewNodes-Prune</td>
<td>59.57</td>
<td>65.22</td>
<td>42.31</td>
<td>50.25</td>
</tr>
</tbody>
</table>

Table 3: SceneQA outperforms all the ablations on two of the three test sets. The highlighted improvements are statistically significant.

The results (% scores, Table 3) suggest that the two most important algorithmic features - adding
concepts implied but not explicitly stated in the text (NewNodes), and later removing implied information that is of low relevance to the answer (Prune) - are important for answering the questions correctly. (The small gain without adding NewNodes on Test1 is not statistically significant).

3.4 Error Analysis

We also examined cases where SceneQA gave the wrong answer. Two problems were particularly common:

(1) There were two answer options with opposite meanings, and one of them was correct. For example:

**Question 3** An animal that has a backbone is called a(n) (A) invertebrate (B) vertebrate (C) exoskeleton (D) sponge

Since the relatedness measure we use (i.e. word2vec) cannot distinguish words with similar distributional semantics (a common property of antonyms), our method cannot confidently identify which of the opposites (e.g., here, vertebrate vs. invertebrate) is correct.

(2) The word ordering in the question is particularly important, e.g., questions about processes or sequences. For example:

**Question 4** The process that changes a gas to liquid is called (A) condensation (B) melting (C) evaporation (D) vaporization

Because our method ignores word order (the knowledge graph is initially populated with keywords in the question), the representation is inherently incapable of capturing sequential information (e.g., here, gas to liquid vs. liquid to gas). As a result, it struggles with such questions.

4 Discussion and Conclusion

Our goal is to answer simple science questions. Unlike entity-centric factoid QA tasks, science questions typically involve general concepts, and answering them requires identifying implicit relationships in the question. Our approach is to view question-answering as constructing a coherent scene. While the notion of scene construction is not new, our insight is that this can be done with a simple “knowledge graph” representation, allowing several massive background KBs to be applied, somewhat alleviating the knowledge bottleneck. Our contribution is to show this works well in the elementary science domain.

Despite this, there are clearly many limitations with our approach: we are largely ignoring syntactic structure in the questions; the KBs are noisy, contributing errors to the scenes; the graph representation has limited expressivity (e.g., no quantification or negation); the word2vec measure of relationship strength does not account for the question context; and contradictions are not detected within the scene. These all contributed to QA failures in the tests. However, the approach is appealing as it takes a step towards a richer picture of language understanding, the empirical results are encouraging, and there are many ways these initial limitations can be addressed going forward. We are confident that this is a rich and exciting space, worthy of further exploration.

References


Peter Clark. 2015. Elementary school science and math tests as a driver for ai: Take the aristo challenge! In *Twenty-Seventh IAAI Conference*.


Xuchen Yao and Benjamin Van Durme. 2014. Information extraction over structured data: Question answering with freebase. In Proceedings of ACL.