WorldTree: A Corpus of Explanation Graphs for Elementary Science Questions supporting Multi-Hop Inference

Peter A. Jansen*, Elizabeth Wainwright†, Steven Marmorstein‡, Clayton T. Morrison*
*School of Information, †Department of Linguistics, ‡Department of Computer Science
University of Arizona, Tucson, USA
pajansen@email.arizona.edu

Abstract

Developing methods of automated inference that are able to provide users with compelling human-readable justifications for why the answer to a question is correct is critical for domains such as science and medicine, where user trust and detecting costly errors are limiting factors to adoption. One of the central barriers to training question answering models on explainable inference tasks is the lack of gold explanations to serve as training data. In this paper we present a corpus of explanations for standardized science exams, a recent challenge task for question answering. We manually construct a corpus of detailed explanations for nearly all publicly available standardized elementary science question (approximately 1,680 3^{rd} through 5^{th} grade questions) and represent these as "explanation graphs" – sets of lexically overlapping sentences that describe how to arrive at the correct answer to a question through a combination of domain and world knowledge. We also provide an explanation-centered tablestore, a collection of semi-structured tables that contain the knowledge to construct these elementary science explanations. Together, these two knowledge resources map out a substantial portion of the knowledge required for answering and explaining elementary science exams, and provide both structured and free-text training data for the explainable inference task.

Keywords: question answering, explanations, explainable inference

1. Introduction

Question answering (QA) is a high-level natural language processing task that requires automatically providing answers to natural language questions. The approaches used to construct QA solvers vary depending on the questions and domain, from inference methods that attempt to construct answers from semantic, syntactic, or logical decompositions, to retrieval methods that work to identify passages of text likely to contain the answer in large corpora using statistical methods. Because of the difficulty of this task, overall QA task performance tends to be low, with generally between 20% and 80% of natural (non-artificially generated) questions answered correctly, depending on the questions, the domain, and the knowledge and inference requirements.

Standardized science exams have recently been proposed as a challenge task for question answering (Clark, 2015), as these questions have very challenging knowledge and inference requirements (Clark et al., 2013; Jansen et al., 2016), but are expressed in simple-enough language that the linguistic challenges are likely surmountable in the near-term. They also provide a standardized comparison of modern inference techniques against human performance, with individual QA solvers generally answering between 40% to 50% of multiple choice science questions correctly (Khot et al., 2015; Clark et al., 2016; Khashabi et al., 2016; Khot et al., 2017; Jansen et al., 2017, inter alia), and top-performing ensemble models nearly reaching a passing grade of 60% on middle school (8^{th} grade) science exams during a recent worldwide competition of 780 teams sponsored by the Allen Institute for AI (Schoenick et al., 2017).

One of the central shortcomings of question answering models is that while solvers are steadily increasing the proportion of questions they answer correctly, most solvers generally lack the capacity to provide human-readable explanations or justifications for why those answers are correct. This "explainable inference" task is seen as a limitation of current machine learning models in general (e.g. Ribeiro et al., (2016)), but is critical for domains such as science or medicine where user trust and detecting potentially costly errors are important. More than this, evidence from the cognitive and pedagogy literature suggests that explanations (when tutoring others) and self-explanations (when engaged in self-directed learning) are an important aspect of learning, helping humans better generalize the knowledge they have learned (Roscoe and Chi, 2007; Legare, 2014; Rittle-Johnson and Loehr, 2016). This suggests that explainable methods of inference may not only be desirable for users, but may be a requirement for automated systems to have human-like generalization and inference capabilities.

Building QA solvers that generate explanations for their answers is a challenging task, requiring a number of inference capacities. Central among these is the idea of *information aggregation*, or the idea that explanations for a given question are rarely found in a contiguous passage of text, and as such inference methods must generally assemble many separate pieces of knowledge from different sources in order to arrive at a correct answer. Previous estimates (Jansen et al., 2016) suggest elementary science questions require an average of 4 pieces of knowledge to answer and explain those answers (here our analysis suggests this is closer to 6), but inference methods tend to have difficulty aggregating more than 2 pieces of knowledge from free-text together due to the semantic or contextual "drift" associated with this aggregation (Fried et al., 2015). Because of the difficulty in assembling training data for the information aggregation task, some have

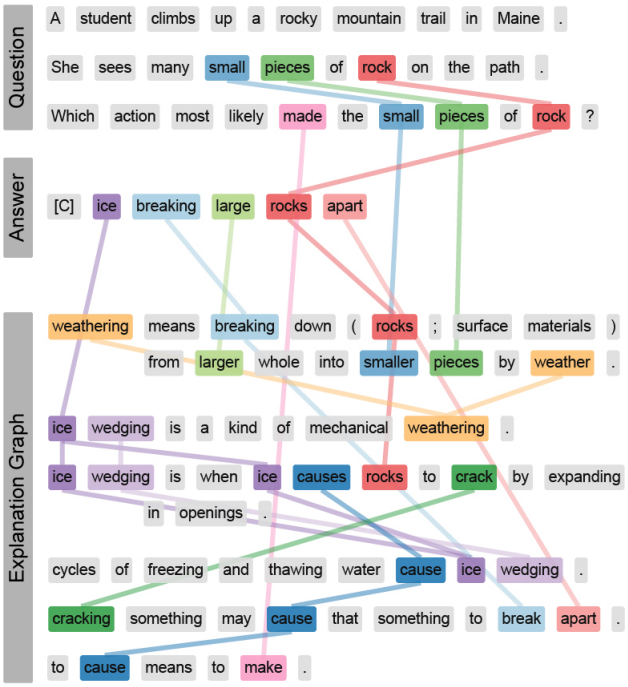

Figure 1: An example multiple choice science question, the correct answer, and a sample explanation graph for why that answer is correct. Here, the explanation graph consists of six sentences, each interconnected through lexical overlap with the question, answer, and other explanation sentences.

approached explanation generation as a distant supervision problem, with explanation quality modelled as a latent variable (Jansen et al., 2017; Sharp et al., 2017). While these techniques have had some success in constructing short explanations, semantic drift likely limits the viability of this technique for explanations requiring more than two pieces of information to be aggregated.

To address this, here we construct a large corpus of explanation graphs (see Figure 1) to serve as training data for explainable inference tasks. The contributions of this work are are:

We construct a set of explanations for 1,680 standardized elementary science exam questions, represented as both free-text, and as lexically-overlapping "explanation graphs" that provide training data for inference models by detailing explicit connections between knowledge in different sentences of an explanation.

We provide an explanation-centered "tablestore", a set of 62 semi-structured tables containing 4,950 rows that provide a substantial portion of the knowledge required to answer non-spatial, non-mathematical elementary science questions.

We provide an analysis of the knowledge growth and explanation overlap properties of this corpus, suggesting both requirements for inference algorithms to make use of explanation corpora, as well as methods of estimating the difficulty in constructing explanation corpora in other domains.

2. Related Work

In terms of question answering, the ability to provide compelling human-readable explanations for answers to questions has been proposed as a complementary metric to assess QA performance alongside the proportion of questions answered correctly. Jansen et al. (2017) developed a QA system for elementary science that answers questions by building and ranking explanation graphs built from aggregating multiple sentences read from free text corpora, including study guides and dictionaries. Because of the difficulty in constructing gold explanations to serve as training data, the explanations built with this system were constructed by modeling explanation quality as a latent variable machine learning problem. First, sentences were decomposed into sentence graphs based on clausal and prepositional boundaries, then assembled into multi-sentence "explanation graphs". Questions were answered by ranking these candidate explanation graphs, using answer correctness as well as features that capture the connectivity of key-terms in the graphs as a proxy for explanation quality. Jansen at al. (2017) showed that it is possible to learn to generate high quality explanations for 60% of elementary science questions using this method, an increase of 15% over a baseline that retrieved single continuous passages of text as answer justifications. Critically, in their error analysis Jansen et al. found that for questions answered incorrectly by their system, nearly half had successfully generated high-quality explanation graphs and ranked these highly, though they were not ultimately selected. They suggest that the process of building and ranking explanations would be aided by developing more expensive second-pass reranking processes that are able to better recognize the components and structure of high-quality explanations within a short list of candidates.

Knowledge bases of tables, or *"table stores"*, have recently been proposed as a semi-structured knowledge formalism for question answering that balances the cost of manually crafting highly-structured knowledge bases with the difficulties in acquiring this knowledge from free text (Yin et al., 2015; Sun et al., 2016; Jauhar et al., 2016). The methods for question answering over tables generally take the form of constructing chains of multiple table rows that lead from terms in the question to terms in the answer, while the tables themselves are generally either collected from the web, automatically generated by extracting relations from free text, or manually constructed.

At the collection end of the spectrum, Pasupat and Liang (2015) extract 2,108 HTML tables from Wikipedia, and propose a method of answering these questions by reasoning over the tables using formal logic. They also introduce the WikiTableQuestions dataset, a set of 22,033 question-answer pairs (such as *"Greece held its last Summer Olympics during which year?"*) that can be answered using these tables. Demonstrating the ability for collection at scale, Sun et al. (2016) extract a total of 104 million tables from Wikipedia and the web, and develop a model that constructs relational chains between table rows using a deep-learning framework.[1] Using their system and ta-

[1] Sun et al. (2016) note that the 99 million tables extracted from

ble store, Sun et al. demonstrate state-of-the-art performance on several benchmark datasets, including WebQuestions (Berant et al., 2013), a set of popular questions asked from the web designed to be answerable using the large structured knowledge graph Freebase (e.g. *"What movies does Morgan Freeman star in?"*).

In terms of automatic generation, though relations are often represented as
triples, Yin et al. (2015) create a large table containing 120M *n-tuple* relations using OpenIE (Etzioni et al., 2011), arguing that the extra expressivity afforded by these more detailed relations allows their system to answer more complex questions. Yin et al. use this to successfully reason over the WebQuestions dataset, as well as their own set of questions with more complex prepositional and adverbial constraints.

Elementary science exams contain a variety of complex and challenging inference problems (Clark et al., 2013; Jansen et al., 2016), with nearly 70% of questions requiring some form of causal, process, or model-based reasoning to solve and produce an explanation for. In spite of these exams being taken by millions of students each year, elementary students tend not to be fast or voluminous readers by adult standards, making this a surprisingly low-resource domain for grade-appropriate study guides and other materials. The questions also tend to require world knowledge expressed in grade-appropriate language (like that *bears have fur* and that *fur keeps animals warm*) to solve. Because of these requirements and limitations, table stores for elementary science QA tend to be manually or semi-automatically constructed, and comparatively small.

Khashabi et al. (2016) provide the largest elementary science table store to date, containing approximately 5,000 manually-authored rows across 65 tables based on science curriculum topics obtained from study guides and a small corpus of questions. Khashabi et al. also augment their tablestore with 4 tables containing 2,600 automatically generated table rows using OpenIE triples. Reasoning is accomplished using an integer-linear programming algorithm to chain table rows, with Khashabi et al. reporting that an average of 2 table rows are used to answer each question. Evaluation on a small set of 129 science questions achieved passing performance (61%), with an ablation study showing that the bulk of their model's performance was from the manually authored tables.

To help improve the quality of automatically generated tables, Dalvi et al. (2016) introduce an interactive tool for semi-automatic table generation that allows annotators to query patterns over large corpora. They demonstrate that this tool can improve the speed of knowledge generation by up to a factor of 4 over manual methods, while increasing the precision and utility of the tables up to seven fold compared to completely automatic methods.

All of the above systems share the commonality that they work to connect (or aggregate) multiple pieces of knowledge that, through a variety of inference methods, move towards the goal of answering questions. Fried et al. (2015) report that information aggregation for QA is currently very challenging, with few methods able to combine more than two pieces of knowledge before succumbing to *semantic drift*, or the phenomenon of two pieces of knowledge being erroneously connected due to shared lexical overlap, incomplete word-sense disambiguation, or other noisy signals (e.g. erroneously aggregating a sentence about *Apple computers* to an inference when working to determine whether *apples* are a kind of fruit). In a generating a corpus of natural-language explanations for 432 elementary science questions, Jansen et al. (2016) found that the average question requires aggregating 4 separate pieces of knowledge to explainably answer, with some questions requiring much longer explanations.

Though few QA solvers explicitly report the aggregation limits of their algorithms, Fried et al. (2015), Khabashi et al. (2016) and Jansen et al. (2017) appear to show limits or substantial decreases in performance after aggregating two pieces of knowledge. To the best of our knowledge, of systems that use information aggregation, only Jansen et al. (2017) explicitly rate the explanatory performance of the justifications from their model, with good explanations generated for only 60% of correctly answered questions. Taken together, all of this suggests that performance on information aggregation and explainable question answering is still far from human performance, and could substantially benefit from a large corpus of training data for these tasks.

3. Design Goals

We began with the following design goals:

Computable explanations: Explanations should be represented at different levels of structure (explanation, then sentences, then relations within sentences). The knowledge links between explanation sentences should be explicit through lexical overlap, which can be used to form an "explanation graph" that describes how each sentence is linked in an explanation.

Depth: Sufficient knowledge should be present in explanations such that that the answer could be arrived at with little extra domain or world knowledge – i.e. where possible, explanations should be targeted at the level of knowledge of a 5-year old child, or lower (see below for a more detailed discussion of explanatory depth).

Reuse: Where possible, knowledge should be re-used across explanations to facilitate automated analysis of knowledge use, and identifying common explanation patterns across questions.

3.1. Explanation Depth

The level of knowledge required to convincingly explain why an answer to a question is correct depends upon one's familiarity with the domain of the question. For a domain expert (such as an elementary science teacher), a convincing explanation to why *thick bark* is the correct answer to *"Which characteristic could best help a tree survive the heat of a forest fire?"* might need only take the form of explaining that *one of bark's primary functions is to provide protection for the tree*. In contrast, for a domain novice, such as an elementary science student, this explanation

the web introduce more noise into the inference process than the high-quality tables from Wikipedia

Question	Which of the following characteristics would best help a tree survive the heat of a forest fire?
Answers	[A] large leaves [B] shallow roots *[*C] thick bark* [D] thin trunks

Levels of explanatory knowledge:

Domain Expert (e.g. teacher)
Bark is a protective covering around the trunk and branches of a tree.

Domain Novice (e.g. 4th grade student)
As an object's thickness increases, it's resistance to damage will also increase.

Young child (e.g. 5-year old)
Protecting something means preventing harm.
Fire causes harm to trees, forests, and other living things.
Thickness is a measure of how thick an object is.
A tree is a kind of living thing.

First Principles
Protecting a living thing has a positive impact on it's survival and health.

Table 1: Levels of explanatory knowledge depth in order of increasing specificity, and example explanatory sentences for each level. For a domain expert who is already fluent in the reasoning of a domain, brief explanations may be sufficient to completely understand why a given answer is correct. As the level of explanatory knowledge moves towards increasing specificity, less domain and world knowledge is assumed, and this knowledge must be explicitly included in the explanations. Explanatory levels are additive, i.e. an explanation targeted at the *young child* level would also include the knowledge at the *domain novice* and *domain expert* levels. In this work, we target authoring explanations at a level between *young child* and *first principles*.

might need to be elaborated to include more knowledge to make this inference, such as that *thicker things tend to provide more protection*. Here we identify four coarse levels of increasing explanatory knowledge depth, shown in Table 1.

For training explainable inference systems, a high level of explanatory depth is likely required. As such, in this work we target authoring explanations between the levels of *young child* and *first principles*. Pragmatically, in spite of their ultimate utility for training inference systems, building explanations too close to *first principles* becomes laborious and challenging for annotators given the level of abstraction and the large amount of implicit world knowledge that must be enumerated, and we leave developing protocols and methods for building such detailed explanations for future work.

4. Explanation Authoring

We describe our representations, tools, and annotation process below.

4.1. Questions

We author explanation graphs for a corpus of 2,201 elementary science questions (through grade) from the AI2 Science Questions V2 corpus, consisting of both standardized exam questions from 12 US states, as well as the separate AI2 Science Questions Mercury dataset, a set of questions licensed from a student assessment entity. Each question is a 4-way multiple choice question, and only those questions that do not involve diagram interpretation (a separate spatial task) are included. Approximately 20% of explanations required specialized domain knowledge (for example, spatial or mathematical knowledge) that did not easily lend itself to explanation using our formalism, resulting in a corpus of 1,680 questions and explanations.

4.2. Tables and Table Rows

Explanations for a given question consist of a set of sentences, each of which is on a single topic and centered around a particular kind of relation, such as *water is a kind of liquid (a taxonomic relation)*, or *melting means changing from a solid to a liquid through the addition of heat energy (a change relation)*.

Each explanation sentence is represented as a single row from a semi-structured table defined around a particular relation. Our tablestore includes 62 such tables, each centered around a particular relation such as *taxonomy, meronymy, causality, changes, actions, requirements,* or *affordances*, and a number of tables specified around specific properties, such as *average lifespans of living things, the magnetic properties of materials,* or the *nominal durations of certain processes (like the Earth orbiting the Sun)*. The initial selection of table relations was drawn from a list of 21 common relations required for science explanations identified by Jansen et al. (2016) on a smaller corpus, and expanded as new knowledge types were identified. Subsets of example tables are included in Figure 2. Each explanation in this corpus contains an average of 6.3 rows.

Fine-grained column structure: In tabular representations, columns represent specific roles or arguments to a specific relation (such as *X is when Y changes from A to B using mechanism C*). In our tablestore we attempt to minimize the amount of information per cell, instead favouring tables with many columns that explicitly identify common roles, conditions, or other relations. This finer-grained structure eases the annotator's cognitive load when authoring new rows, while also better compartmentalizing the relational knowledge in each row for inference algorithms. The tables in our tablestore contain between *2 and 16* content columns, as compared to 2 to 5 columns for the Ariso tablestore (Khashabi et al., 2016).

Natural language sentences: QA models use a variety of different representations for inference, from semantic roles and syntactic dependencies to discourse and embeddings. Following Khashabi et al. (2016), we make use of a specific form of table representation that includes "filler" columns that allow each row to be directly read off as a stand-alone natural language sentence, and serve as input to any model. Examples of these filler columns can be seen in Figure 2.

4.3. Explanation Graphs and Sentence Roles

Explanations for a given question here take the form of a list of sentences, where each sentence is a reference to a specific table row in the table store. To increase their utility for knowledge and inference analyses, we require that

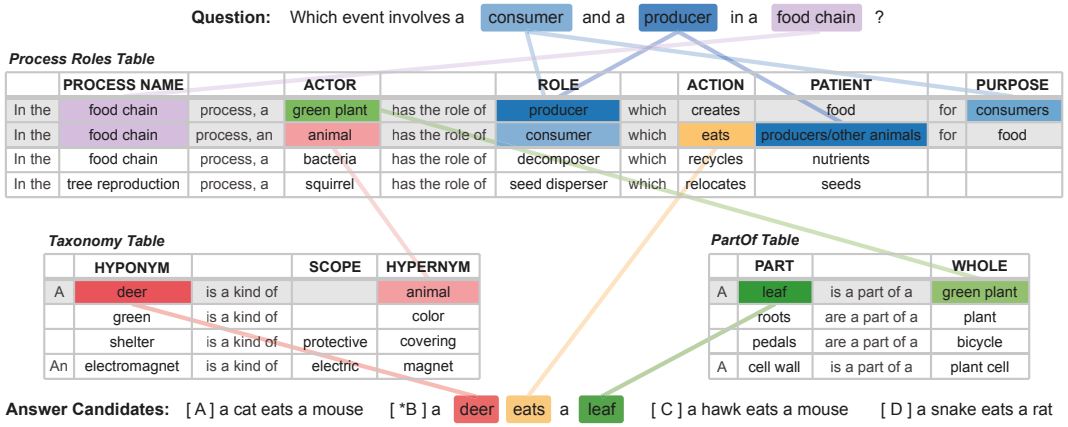

Figure 2: Examples of tables and table rows from the tablestore, grounded in an example question and explanation. Table columns define the primary roles or arguments for a given relation (e.g. *process name, actor, role, etc*). Unlabeled "filler" columns allow each row to be used as a stand-alone natural language sentence. Note that for clarity only 4 example rows per table are shown.[3]

Question	Which occurs as the kinetic energy of water molecules increases?
Answer	[*D] liquid water becomes water vapor
Central role	
	As a molecule's kinetic energy increases, temperature will increase.
	Boiling means changing from a liquid into a gas by adding heat energy.
Grounding role	
	Water is a kind of liquid.
	Water is in the gas state, called water vapor, for temperatures greater than 100 degrees celsius.
Background role	
	Matter is made of molecules.
Lexical glue role	
	To add means to increase.
	Temperature is a measure of heat energy.

Table 2: Examples of the four coarse classes of explanation sentence roles, *central, grounding, background, and lexical glue*.

each sentence in an explanation be explicitly lexically connected (i.e. share words) with either the question, answer, or other sentences in the explanation. We call this lexically-connected set of sentences an *explanation graph*.

In our preliminary analysis, we observed that the sentences in our explanations can take on very different roles, and we hypothesize that differentiating these roles is likely important for inference algorithms. We identified four coarse roles, listed in Table 2, and described below:

- **Central:** The central concept(s) that a question is testing, such as *changes of state* or the *coupled relationship between kinetic energy and temperature*.

- **Grounding:** Sentences linking generic or abstract terms in a *central* sentence with specific instances of those terms in the question or answer. For example, for questions about changes of state, grounding sentences might identify specific instances of *liquids* (such as water) or *gasses* (such as water vapor).

- **Background:** Extra information elaborating on the topic, but that (strictly speaking) isn't required to arrive at the correct inference.

- **Lexical glue:** Sentences that lexically link two concepts, such as *"to add means to increase"*, or *"heating means adding heat"*. This is an artificial category in our corpus, brought about by the need for explanation graphs to be explicitly lexically linked.

For each sentence in each authored explanation, we provide annotation indicating which of these four roles the sentence serves in that explanation.

4.4. Annotation Tool

To facilitate explanation authoring, we developed and iterated the web-based collaborative authoring tool shown in Figure 3. The tool displays a given question to the explanation author, and allows the author to progressively build an explanation graph for that question by querying the tablestore for relevant rows based on keyword searches, as well as past explanations that are likely to contain similar content or structure (increasing consistency across explanations, while reducing annotation time). A graphical visualization of the explanation graph helps the author quickly assess gaps in the explanation content to address by highlighting lexical overlap between sentences with coloured edges and labels. The tablestore takes the form of a shared Google Sheet[4] that the annotators populate, with each table represented as a separate tab on the sheet.

4.5. Procedure and Explanation Review

For a given question, annotators identified the central concept the question was testing, as well as the inference required to correctly answer the question, then began progressively constructing the explanation graph. Sentences in the graph were added by querying the tablestore based on keywords, which retrieved both single sentences/table rows, as well as entire explanations that had been previously annotated. If any knowledge required to build an explanation did

[3] Note that this figure also appears in an earlier workshop submission on identifying explanatory patterns (Jansen, 2017)

[4] http://sheets.google.com

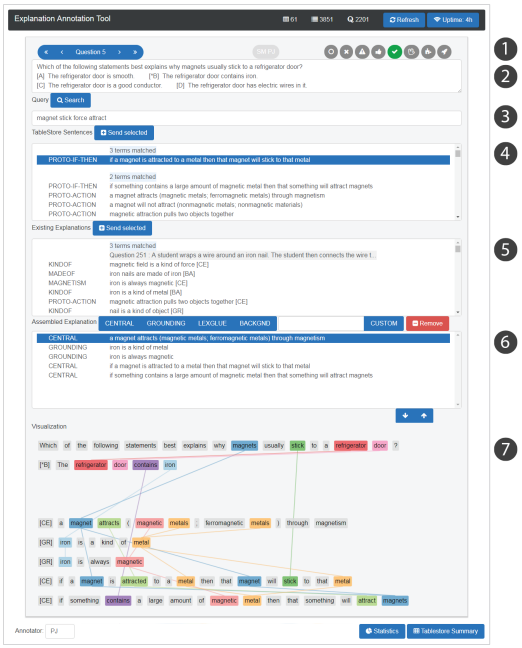

Figure 3: The explanation authoring web tool. Interface components include: (1) A list of user-settable flags to assist in the annotation and quality review process; (2) Question and answer candidates; (3) Query terms for search; (4) Query results (tablestore); (5) Query results (complete explanations); (6) Current explanation being assembled; (7) Explanation graph visualization of lexical overlap within the explanation.

not exist in the tablestore, this was added to an appropriate table, then added to the explanation.

New tables were regularly added, most commonly for property knowledge surrounding a particular topic (e.g. whether a particular material is recyclable). Because explanations are stored as lists of unique identifiers to table rows, tables and table rows could regularly be refactored, elaborated, or entirely reorganized without requiring existing explanations to be rewritten. We found this was critical for consistency and ensuring good organization throughout corpus construction.

One of the central difficulties with evaluating explanation authoring is determining metrics for interannotator agreement, as many correct explanations are possible for a given question, and there are many different wordings that an annotator might choose to express a given piece of knowledge in the tablestore. Similarly, the borders between different levels of explanatory depth are fuzzy, suggesting that one annotator may express their explanation with more or less specificity than another.

To address these difficulties we included two methods to increase consistency. First, as a passive intervention during the explanation generation process, annotators are presented with existing explanations that can be drawn from to compose a new explanation, where these existing explanations share many of the same query terms being used to construct the new explanation. Second, as an active intervention, each explanation goes through four review passes to ensure consistency. The first two passes are completed by the original annotator, before checking a flag on the annotation tool signifying that the question is ready for external review. A second annotator then checks the question for completeness and consistency with existing explanations, and composes a list of suggested edits and revisions. The fourth and final pass is completed by the original annotator, who implements these suggested revisions. This review process is expensive, taking approximately one third of the total time required to annotate each question.

Each annotator required approximately 60 hours of initial training for this explanation authoring task. We found that most explanations could be constructed within 5-10 minutes, with the review process taking approximately 5 more minutes per question.

5. Explanation Corpus Properties

Here we characterize three properties of the explanation corpus as they relate to developing methods of explainable inference: *knowledge frequency, explanation overlap,* and *tablestore growth.*

5.1. Knowledge Use and Row Frequency

The tables most frequently used to author explanations are shown in Table 3, broken down into three broad categories identified by Jansen et al. (2016): *retrieval types, inference-supporting types, and complex inference types*. Because the design of this corpus is data driven – i.e., knowledge is generally added to a table because it is required in one or more explanations[5] – we can calculate how frequently the rows in a given table are reused to obtain an approximate measure of the generality of that knowledge. On average, a given table row is used in 2.9 different explanations, with 1,535 rows used more than once, and 531 rows used 5 or more times. The most frequently reused row (*"an animal is a kind of organism"*) is used in 89 different explanations. Generic "change of state" knowledge (e.g. solids, liquids, and gasses) is also frequently reused, with each row in the StatesOfMatter table used in an average of 15.7 explanations. Usage statistics for other common tables are also provided in Table 3.

5.2. Explanation Overlap

One might hypothesize that questions that require similar inferences to correctly answer may also contain some of the same knowledge in their explanations, with the amount of knowledge overlap dependent upon the similarity of the questions. We plan to explore using this overlap as a method of inference that can generate new explanations by editing, merging, or expanding known explanations from similar, known questions (see Jansen (2017) for an initial study). For this to be possible, an explanation corpus must reach a sufficient size that a large majority of questions have substantial overlap in their explanations.

Figure 5 shows the proportion of questions in the corpus that have *1 or more*, *2 or more*, *3 or more*, *etc.*, overlapping rows in their explanations with *at least one other*

[5]For compatibility, we do include several property tables from the Aristo tablestore, though a large proportion of rows from these tables are not actively used. Our tablestore includes 4,950 rows, 3,686 of which are actively used in at least one explanation.

Knowledge Type	Prevalence (% of expl.)	Rows in Table	Avg. Row Freq.
Retrieval Types			
Taxonomic	78%	1,119	1.2
Synonymy	61%	639	1.6
PartOf	14%	148	1.6
Properties (Generic)	11%	173	1.1
MadeOf	7%	72	1.7
Contains	6%	75	1.4
Examples	5%	58	1.4
Measurements (P)	4%	23	3.0
Locations (P)	3%	47	1.1
InheritedTraits (P)	3%	22	2.3
StatesOfMatter (P)	3%	3	15.7
Conductivity (P)	3%	9	4.9
Resources (P)	3%	16	2.7
Inference Supporting Types			
Actions	25%	259	1.6
UsedFor	19%	191	1.7
Requires	15%	121	2.1
SourceOf	14%	81	2.8
Affect	12%	77	2.6
Opposites	8%	35	3.8
FormedBy	4%	40	1.9
Affordances	4%	48	1.3
Complex Inference Types			
If/Then	21%	229	1.6
Cause	17%	183	1.6
Changes (discrete)	14%	62	3.8
Transfer	9%	46	3.3
Changes (vector)	9%	62	2.4
CoupledRelationships	7%	126	0.9
ProcessRoles	3%	12	3.8

Table 3: The proportion of explanations that contain knowledge from a given table, sorted by most frequent knowledge, and broken down by the knowledge type of a given table. Tables not used in at least 3% of explanations are not shown. (P) indicates a given table describes properties, e.g. whether a given material is conductive. Average Row Frequency refers to the average number of explanations a given row from that table is used in.

question in the corpus.[6] Similarly, to ground this, Figure 4 shows a visualization of questions whose explanations have *2 or more* overlapping rows. For a given level of overlapping explanation sentences, Figure 5 shows that the proportion of questions with that level of overlap increases logarithmically with the number of questions.

This has two consequences. First, it allows us to estimate the size of corpus required to train hypothetical inference methods for the science exam domain capable of producing explanations. If a given inference method can work successfully with only minimal overlap (for example, 1 shared table row), then a training corpus of 500 explanations in this domain should be sufficient to answer 80% of questions. If an inference method requires 2 shared rows, the corpus requirements would increase to approximately 2,500 questions to answer 80% of questions. However, if an inference method requires 3 or more rows, this likely would not be possible without a corpus of at least 20,000 questions and explanations – a substantial undertaking. Second, because this relationship is strongly logarithmic, if it transfers to domains outside elementary science, it should be possible to estimate the corpus size requirements for those domains after authoring explanations for only a few hundred questions.

[6]Though not included for space, the number of questions with *N or more* rows in common in their explanations increases linearly with the number of questions. For this corpus, for a given question, on average there are 17 questions that have 1 or more overlapping rows in their explanation, 9 questions with 2 or more shared rows in their explanation, and 5 questions with 3 or more shared rows in their explanation.

5.3. Explanation Tablestore Growth

Finally, we examine the growth of the tablestore as it relates to the number of questions in the corpus. Figure 6 shows a monte-carlo simulation of the number of unique tablestore rows required to author explanations for specific corpus sizes. This relationship is strongly correlated *(R=0.99)* with an exponential proportional decrease.[7] For this elementary science corpus, this asymptotes at approximately 6,000 unique table rows, and 10,000 questions, providing an estimate of the upper-bound of knowledge required in this domain, and the number of unique questions that can be generated within the scope of the elementary science curriculum.

The caveat to this estimate is that it estimates the knowledge required for elementary science exams as they currently exist, with the natural level of variation introduced by the test designers. Questions are naturally grounded in examples, such as *"Which part of an oak tree is responsible for undertaking photosynthesis?" (Answer: the leaves)*. While the corpus often contains a number of variations of a given question that test the same curriculum topic and have similar explanations, many more variations on these questions are possible that ground the question in different examples, like *orchids*, *peach trees*, or other plants. As such, while we believe that these estimates likely cover the core knowledge of the domain, many times that knowledge would be required to make the explanation tablestore robust to small variations in the presentation of those existing exam questions, or to novel unseen questions.

6. Conclusion

We provide a corpus of explanation graphs for elementary science questions suitable for work in developing explainable methods of inference, and show that the knowledge frequency, explanation overlap, and tablestore growth properties of the corpus follow predictable relationships. This work is open source, with the corpus and generation tools available at `http://www.cognitiveai.org/explanationbank`.

7. Acknowledgements

We thank the Allen Institute of Artificial Intelligence for funding this work, Peter Clark at AI2 for thoughtful discussions, and Paul Hein for assistance constructing the annotation tool.

[7]Here, this exponential proportional decrease takes the form of $R = 434 - (-2.93/0.00054) \cdot (1 - e^{-0.00054 \cdot Q})$, where R is the size of the tablestore in rows, to explainably answer Q questions.

Figure 4: Questions in this explanation corpus connected by explanation overlap. Here, nodes represent questions and their explanations, and edges between nodes represent two questions having *at least 2 or more (i.e. 2+)* shared rows (i.e. sentences) in their explanations, with at least one of these shared rows being labelled as having a CENTRAL role to the explanation. Topic clusters (labels) naturally emerge for questions requiring similar methods of inference, based on the shared content of their explanations.

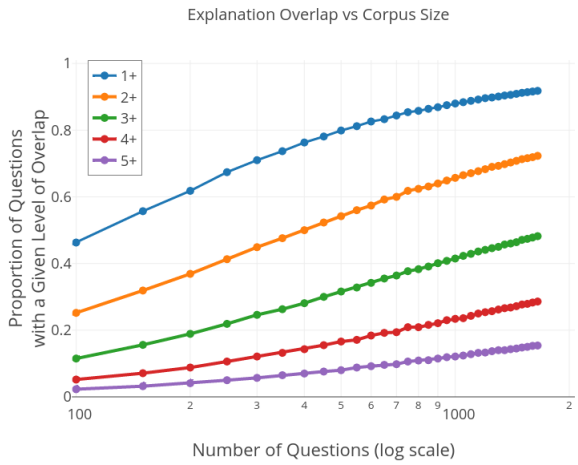

Figure 5: Monte-carlo simulation showing the proportion of questions whose explanations overlap by *1 or more, 2 or more, 3 or more, ...,* explanation sentences. The proportion increases logarithmically with the number of questions in the corpus. Each point represents the average of 100 simulations.

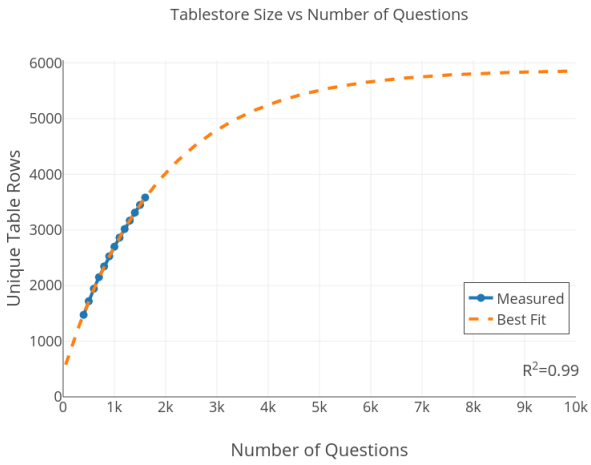

Figure 6: Monte-carlo simulation showing the number of unique table rows required to explainably answer a given number of questions. The line of best fit (dashed) suggests that this is a proportional decay relationship ($R^2 = 0.99$), asymptoting at approximately 6,000 table rows and 10,000 questions. Each point represents the average of 10,000 simulations.

8. Bibliographical References

Berant, J., Chou, A., Frostig, R., and Liang, P. (2013). Semantic parsing on freebase from question-answer pairs. In *EMNLP*.

Clark, P., Harrison, P., and Balasubramanian, N. (2013). A study of the knowledge base requirements for passing an elementary science test. In *Proceedings of the 2013 Workshop on Automated Knowledge Base Construction*, AKBC'13, pages 37–42.

Clark, P., Etzioni, O., Khot, T., Sabharwal, A., Tafjord, O., Turney, P. D., and Khashabi, D. (2016). Combining retrieval, statistics, and inference to answer elementary science questions. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA.*, pages 2580–2586.

Clark, P. (2015). Elementary school science and math tests as a driver for AI: take the aristo challenge! In Blai Bonet et al., editors, *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA.*, pages 4019–4021. AAAI Press.

Dalvi, B., Bhakthavatsalam, S., Clark, C., Clark, P., Etzioni, O., Fader, A., and Groeneveld, D. (2016). IKE - an interactive tool for knowledge extraction. In *Proceedings of the 5th Workshop on Automated Knowledge Base Construction, AKBC@NAACL-HLT 2016, San Diego, CA, USA, June 17, 2016*, pages 12–17.

Etzioni, O., Fader, A., Christensen, J., Soderland, S., and Mausam, M. (2011). Open information extraction: The second generation. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence (ICJAI)*, pages 3–10.

Fried, D., Jansen, P., Hahn-Powell, G., Surdeanu, M., and Clark, P. (2015). Higher-order lexical semantic models for non-factoid answer reranking. *Transactions of the Association for Computational Linguistics*, 3:197–210.

Jansen, P., Balasubramanian, N., Surdeanu, M., and Clark, P. (2016). What's in an explanation? characterizing knowledge and inference requirements for elementary science exams. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2956–2965, Osaka, Japan, December. The COLING 2016 Organizing Committee.

Jansen, P., Sharp, R., Surdeanu, M., and Clark, P. (2017). Framing qa as building and ranking intersentence answer justifications. *Computational Linguistics*.

Jansen, P. (2017). A study of automatically acquiring explanatory inference patterns from corpora of explanations: Lessons from elementary science exams. In *Proceedings of the 2017 Workshop on Automated Knowledge Base Construction*, AKBC'17.

Jauhar, S. K., Turney, P. D., and Hovy, E. H. (2016). Tables as semi-structured knowledge for question answering. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Khashabi, D., Khot, T., Sabharwal, A., Clark, P., Etzioni, O., and Roth, D. (2016). Question answering via integer programming over semi-structured knowledge. In *Proceedings of the International Joint Conference on Artificial Intelligence*, IJCAI'16, pages 1145–1152.

Khot, T., Balasubramanian, N., Gribkoff, E., Sabharwal, A., Clark, P., and Etzioni, O. (2015). Exploring markov logic networks for question answering. In *EMNLP*.

Khot, T., Sabharwal, A., and Clark, P. (2017). Answering complex questions using open information extraction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 2: Short Papers*, pages 311–316.

Legare, C. H. (2014). The contributions of explanation and exploration to children's scientific reasoning. *Child Development Perspectives*, 8(2):101–106.

Pasupat, P. and Liang, P. (2015). Compositional semantic parsing on semi-structured tables. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 1135–1144, New York, NY, USA. ACM.

Rittle-Johnson, B. and Loehr, A. M. (2016). Eliciting explanations: Constraints on when self-explanation aids learning. *Psychonomic bulletin & review*, pages 1–10.

Roscoe, R. D. and Chi, M. T. (2007). Understanding tutor learning: Knowledge-building and knowledge-telling in peer tutors' explanations and questions. *Review of Educational Research*, 77(4):534–574.

Schoenick, C., Clark, P., Tafjord, O., Turney, P., and Etzioni, O. (2017). Moving beyond the turing test with the allen ai science challenge. *Communications of the ACM*, 60(9):60–64.

Sharp, R., Surdeanu, M., Jansen, P., Valenzuela-EscÃ¡rcega, M. A., Clark, P., and Hammond, M. (2017). Tell me why: Using question answering as distant supervision for answer justification. In *Proceedings of the Conference on Computational Natural Language Learning (CoNLL)*.

Sun, H., Ma, H., He, X., Yih, W.-t., Su, Y., and Yan, X. (2016). Table cell search for question answering. In *Proceedings of the 25th International Conference on World Wide Web (WWW)*, pages 771–782.

Yin, P., Duan, N., Kao, B., Bao, J., and Zhou, M. (2015). Answering questions with complex semantic constraints on open knowledge bases. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 1301–1310. ACM.