Extracting Common Inference Patterns from Semi-Structured Explanations

Sebastian Thiem and **Peter Jansen**
School of Information, University of Arizona
{sthiem,pajansen}@email.arizona.edu

Abstract

Complex questions often require combining multiple facts to correctly answer, particularly when generating detailed explanations for why those answers are correct. Combining multiple facts to answer questions is often modeled as a "multi-hop" graph traversal problem, where a given solver must find a series of interconnected facts in a knowledge graph that, taken together, answer the question and explain the reasoning behind that answer. Multi-hop inference currently suffers from semantic drift, or the tendency for chains of reasoning to "drift" to unrelated topics, and this semantic drift greatly limits the number of facts that can be combined in both free text or knowledge base inference. In this work we present our effort to mitigate semantic drift by extracting large high-confidence multi-hop inference patterns, generated by abstracting large-scale explanatory structure from a corpus of detailed explanations. We represent these inference patterns as sets of generalized constraints over sentences represented as rows in a knowledge base of semi-structured tables. We present a prototype tool for identifying common inference patterns from corpora of semi-structured explanations, and use it to successfully extract 67 inference patterns from a "matter" subset of standardized elementary science exam questions that span scientific and world knowledge.

1 Introduction

Combining separate pieces of knowledge to answer complex natural language questions is a central contemporary challenge in natural language inference. For complex questions, a single passage in a corpus or single fact in a knowledge base is often insufficient to arrive at an answer, and multiple sentences or facts must be combined through some inference process. A benefit and goal of this "multi-hop" inference process is for the set of combined facts to form a human-readable explanation detailing why the inference and answer are correct.

Most recent approaches to combining knowledge to answer questions (e.g. Das et al., 2017; Jansen et al., 2017; Ding et al., 2019) model inference as a progressive construction, iteratively adding nodes (facts) one at a time to a graph that represents the inference (and explanation) required to answer a question. This approach suffers from the phenomenon of semantic drift (Fried et al., 2015), which is the observation that determining whether two facts can be meaningfully combined to answer a question is an extremely noisy process, and most often results in adding erroneous facts unrelated to answering a question that causes the inference to fail. A common signal to determine whether two facts might be combined is whether those facts have shared words or entities. For example, for a question asking about the possible *effects of sunlight on an ice cube*, a given solver might choose to meaningfully connect the facts *"melting means changing from a [solid] to a liquid by adding heat energy"* and *"water is a kind of [solid], called ice, at temperatures below 0°C"* on the shared word *solid*. Unfortunately, using shared words alone, either of these facts could also be connected to the fact *"sound travels fastest through a [solid]"*, which is irrelevant to answering this problem, and allows further traversals to unrelated facts about sound that can produce incorrect answers.

Jansen (2018) empirically demonstrated that combining facts based on lexical overlap has very low chance of success, which was measured at between 0.1% and 5% for elementary science questions, depending on the source corpus of the facts being retrieved. This is a significant limitation, as even elementary science questions require combining an average of 4 to 6 facts (and as many as 16 facts) that span scientific and common-sense or world knowledge in order to answer and provide

Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing, pages 53–65
Hongkong, China, November 3, 2019.

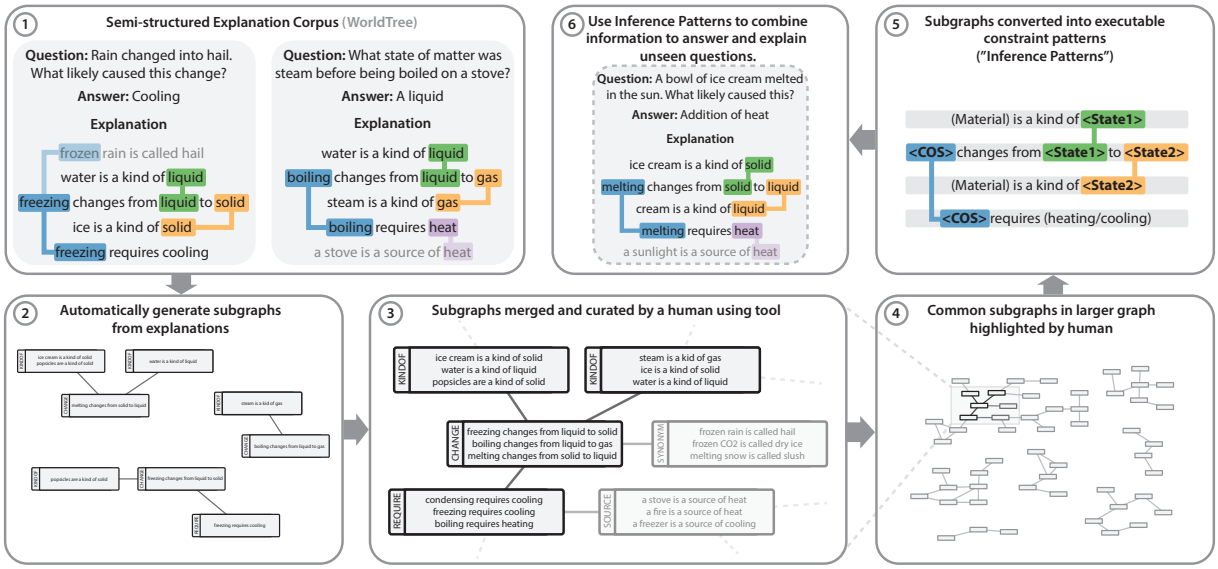

Figure 1: An overview of our inference pattern extraction approach. A corpus of semi-structured explanations (2) is preprocessed through a set of heuristics that generate a large number of small (often disconnected) subgraphs in a large graph (2). Those subgraphs are merged and curated (3). Inference patterns, or subgraphs of nodes can then be extracted from the curated graph, by the user (4). These patterns for executable constraint satisfaction patterns that can be executed over the knowledge base (5). In this work we address steps 2 through 5, whereas using these inference patterns to answer and explain unseen questions (6) is part of ongoing efforts.

a detailed explanation for their reasoning (Jansen et al., 2018, 2016), and such a low probability of successfully traversing the knowledge graph places strong limits on the length of inferences that can be made (Khashabi et al., 2019). In response to this challenge, a number of datasets such as HotpotQA (Yang et al., 2018) and WorldTree (Jansen et al., 2018) have emerged to provide explicit gold explanations that serve as training and evaluation instruments for multi-hop inference models.

Jansen (2017) proposed combining "common explanatory patterns", or groups of frequently interconnected facts observed in explanations, as a possible means of mitigating the semantic drift associated with combining facts one at a time. Human-authored explanations contain meaningful connections between their component facts. Each edge in an explanatory pattern extracted from a human-authored explanation is a high-confidence edge that does not require a solver to use other more noisy signals (such as lexical overlap) to populate, reducing the opportunity for semantic drift. An empirical evaluation using the WorldTree explanation corpus demonstrated that this approach could in principle regenerate the majority of unseen gold explanation graphs by using only 2 or 3 hops between these "explanatory pattern" subgraphs, which is substantially fewer hops than the up to 16 hops required if aggregating single facts. The disadvantages of this technique are that (a) it requires the (currently manual) construction of a large corpus of detailed explanations to learn these common explanatory patterns from, which is an expensive process, and (b) it requires developing automatic or semi-automatic methods to abstract the structure of training explanations to mitigate sparsity and allow known explanations to generalize to unseen scenarios.

In this work, we explore a hybrid human-in-the-loop method and tool for abstracting the structure of common explanatory patterns found in the WorldTree corpus of structured explanations. We use this tool to extract 67 inference patterns, specified as constraint satisfaction patterns over a knowledge base of tables, from detailed explanations to standardized elementary science exam questions. Our long-term interests are in generating a corpus of common inference patterns at scale, and constructing an inference system that combines and uses those patterns to answer questions and produce detailed explanations for its answers. Conceptually, this is similar to Explanation-Based Learning (DeJong and Mooney, 1986; Baillargeon and DeJong, 2017), but using semi-structured text and constraint patterns instead of first-order logic. This approach is also similar to efforts at using scripts or semantic frames for inference (e.g. Wang et al., 2015; Ostermann et al., 2017), or automatically extracted

proxies (e.g. Khashabi et al., 2018), though confined to the subdomain of elementary science, and semi-automatically extracted from semi-structured explanation graphs.

2 Approach and Workflow

The workflow describing our process of taking a corpus of semi-structured explanations through the inference pattern discovery process is described in Figure 1, with further details below.

2.1 Semi-Structured Explanation Corpus

Our technique for discovering inference patterns requires extracting these patterns from a pre-existing corpus of semi-structured explanations. We make use of the WorldTree explanation corpus[1] (Jansen et al., 2018), a set of 1,680 detailed explanation graphs for standardized elementary science questions. These questions represent the elementary (3^{rd} through 5^{th} grade) subset of the Aristo Reasoning Challenge (ARC) corpus[2] (Clark, 2015), a set of 4-choice multiple choice elementary and middle-school science questions drawn from 12 US states.

Each question in Worldtree is paired with an explanation graph composed of a set of facts that, taken together, provide a detailed explanation for why the answer to a given question is correct. Each "fact" is a natural language sentence that takes the form of a row in a knowledge base of 62 semi-structured tables containing a total of 4,950 unique rows. Each table centers around encoding a particular type of knowledge, such as taxonomic relations (e.g. *a bird is a kind of animal*), part-of relations (*a wing is a part of a bird*), property knowledge (*metals are electrical conductors*), or other more complex relations, such as changes (*boiling means changing from a liquid to a gas by adding heat energy*), coupled relationships (*as altitude increases, air pressure deceases*), causality (*bacteria can cause diseases by infecting organisms*), and if-then relationships (*if an animal relies on plants for food, then it must store enough food for winter*).

Each semi-structured table contains between 2 and 16 content columns, which form an $n-ary$ relation between the columns in a given row, and are often used by inference frameworks (e.g. Pasupat and Liang, 2015; Sun et al., 2016; Khashabi et al., 2016) as they afford more fine-grained decomposition than triple representations (e.g. Etzioni et al., 2011; Schmitz et al., 2012) common in other inference methods (e.g. Das et al., 2017; Khot et al., 2017; Kwon et al., 2018). The knowledge base construction was data-driven, where each fact exists because it was authored to be used in at least one real explanation. As such, the knowledge base contains a mix of scientific and world knowledge, some of which is commonly found in other knowledge bases (e.g., taxonomic, part-of, used-for, Speer and Havasi, 2012; Tandon et al., 2017), while other kinds of knowledge (e.g. coupled relationships, how processes change actors, if-then relationships centered around elementary science concepts) are less common. When authoring explanations, the annotation protocol required annotators to attempt to reference existing rows (facts) first rather than create duplicate knowledge. The most highly reused row (*an animal is a kind of organism*) occurs in 89 different explanations, and 31% of rows in the knowledge base occur in more than one explanation. This suggests that a subset of core facts are frequently reused, but that some form of abstraction or generalization of explanations would be required for those core facts to connect to the 69% of facts used in only a single explanation, or to knowledge imported from other knowledge bases that is not currently used in any explanation.

2.2 Automatic Generation of Subgraphs

In this work we frame the process of discovering inference patterns as a process of clustering similar groups of facts together, and discovering meaningful connections between different groups of facts in the forms of constraints (see Figure 1, steps 2 to 5). These constraints take the form of edges between two tables, that can be satisfied by one row from each table having the same words in specific columns (see Figure 1, step 5, for an example).

Clustering Facts: Clustering similar facts requires recognizing that certain groups of facts tend to describe specific instances of a high-level process, even when those facts may have little or no lexical overlap with each other (as in grouping *"freezing means changing from a liquid to a solid"* and *"boiling means changing from a liquid to a gas"*, in the context of a *change of state of matter* process).

Discovering Connections: Discovering connections (i.e. edges) between two or more groups of facts that tend to occur together in gold explanation

[1] http://www.cognitiveai.org/explanationbank/

[2] http://www.allenai.org/data.html

Error Class	
Sparsity in Explanation Annotation	
Fact 1	**Friction** occurs when two **object's surfaces** move against each other
Fact 2	As an **object's** smoothness increases, it's **friction** will decrease when it's **surface** moves against another **surface**.
Issue	These facts are not observed together in a single question's explanation, so they are not connected.
Sparsity in Knowledge Base	
Fact 1	If food is **cooked** then **heat** energy is added to that food.
Fact 2	A stove generates **heat** for **cooking**.
Missing	A campfire generates **heat** for **cooking**.
Issue	Missing facts in the knowledge base limit the generalization of patterns to new scenarios (e.g. campfire).
Permissiveness in automatically populated edges	
Fact 1	Melting means changing from a solid to a liquid by adding heat **energy**
Fact 2	Wax is an electrical **energy** insulator
Issue	Creating edges based on shared words (here, **"energy"**) does not always generate meaningful connections.
Permissiveness in automatically populated column links	
Fact 2	A tape **measure*** is used to **measure distance**.
Fact 2	centimeters (cm) are a unit used for **measuring** is **distance**.
Issue	Ideally this edge should generalize to all kinds of measuring tools and units (e.g. **X** is used to measure **Y**, **Z** is a unit for measuring **Y**). The connection between **tape measure*** in Fact 1 and **measure** in Fact 2 makes generalization unlikely, and should be removed.

Table 1: Example classes of errors when automatically generating inference pattern graphs. Fact 1 and Fact 2 represent facts (rows) drawn from the knowledge base of semi-structured tables. Boldface words represent lexical connections between those facts (edges between tables, on the specific columns those words occupy).

graphs. For example, facts about *change of state* processes *(freezing, boiling, melting, condensing)* may tend to connect to other groups of facts that discuss specific solids, liquids, or gasses that are undergoing the change of state (as in *"water is a kind of liquid"*, or *"ice is a kind of solid"*).

Our initial hypothesis was that it would be possible to extract a large corpus of inference patterns automatically from a sufficiently large and structured corpus of explanations. Instead, we discovered that both the clustering and connection processes are susceptible to a number of common opportunities for error (described in Table 1) that limit this process in practice. In addition to these error classes, we discovered challenges due to inference patterns existing at different levels of abstraction, with patterns at different levels of abstraction frequently overlapping. For example, a high-level domain-specific pattern might describe the process of *changing from one state of matter to another through the addition or subtraction of heat energy*, while describing *specific substances* and *sources of heat or cooling*. A substantially more low-level, domain-general, and common pattern in the corpus is taxonomic inheritance – the idea that if X is a kind of Y, and Y is a kind of Z, then X is a kind of Z (e.g. *a bird is a kind of animal, an animal is a kind of living thing, therefore a bird is a kind of living thing*). Similar low-level science-domain patterns are common (e.g. X is a kind of Y, Y is made of Z, as in *"an ice cube is a kind of object, and objects are made of matter")*. High-level and low-level patterns frequently overlap – that is, a high-level pattern may contain one or more low-level patterns. This caused challenges for the pilot experiments in entirely automatic extraction, either "over-grouping" facts into a single pattern that a human annotator would likely consider different patterns, or vice-versa.

Because of the high-precision requirements of multi-hop inference, our pragmatic solution to the above technical challenge is to build a hybrid system that combines automatic and manual methods. First, a preprocessing system assembles and connects groups of facts using a set of minimal high-precision low-recall heuristics. We then provide the user with a graphical tool to streamline the workflow for manually editing groupings, adding or removing edges between groups of facts, and speeding the inspection and repair of any errors made by the automated heuristics. Summary statistics on the proportion of these changes and errors on our analysis are included in Table 2.

2.3 Merging and Curating Subgraphs

To facilitate the assembly of subgraphs into large high-quality inference patterns, we developed and iterated the graphical authoring tool shown in Figure 2. The tool includes four main components:

Graph View: The graph view allows the annotator to inspect the entire graph in its current state, and to merge nodes (that represent groups of facts/table

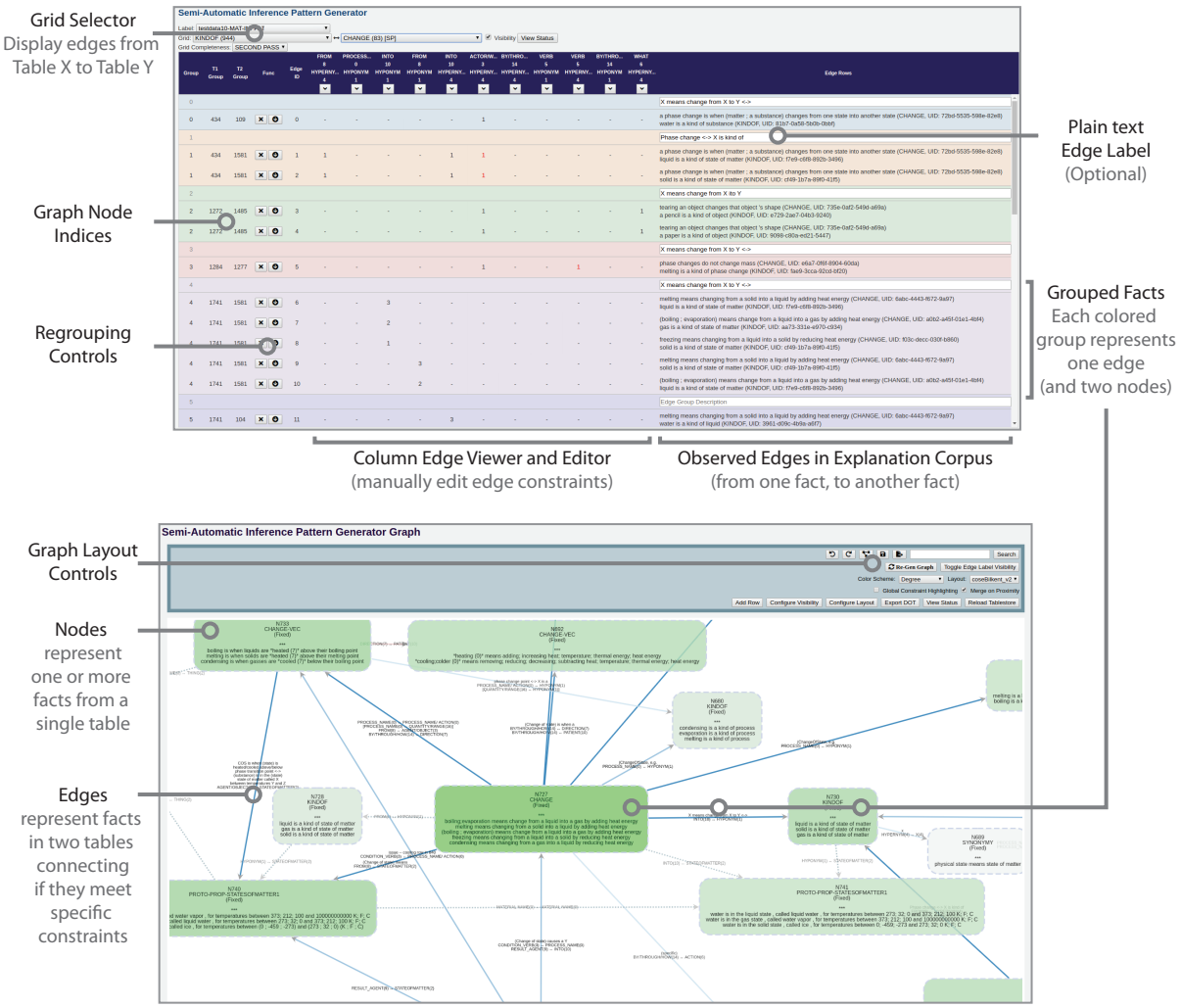

Figure 2: A screenshot of the grid view (top) and graph view (bottom) of our inference pattern extraction tool. The constraint view and the tablestore spreadsheet integration are not shown for space.

rows) together to perform the fact clustering procedure. The graph view also allows the annotator to highlight specific subgraphs to mark as inference patterns, which enables further functionality in the constraint view.

Grid View: The grid view enables the curation of the edges between nodes by visually displaying them in an interface that allows the user to (a) remove automatically populated edges that are not meaningful, (b) remove only part of edges (i.e. specific links between columns between two tables), and (c) manually edit the automatic clustering by dragging and dropping specific rows in one edge group either into another existing group, or into a new group.

Constraint View: Once a user has identified and marked a subgraph to extract as an inference pattern, the constraint view allows "running" that inference pattern to generate all possible sets of rows in the tablestore that satisfy that pattern's constraints. As subgraphs extracted directly from the large curated graph built from the explanation corpus tend to require edits to their nodes and constraints before they are generic and runnable inference patterns, the constraint view also includes a number of debugging tools to facilitate diagnosing constraints that are unable to be satisfied.[3]

Table View: The tool also includes an interface to a Google Sheet[4] storing a live copy of the Tablestore that the annotator can edit to refine existing knowledge, or incorporate additional knowledge, while curating and debugging inference patterns.

The tool runs in a Chrome browser window, and is implemented as a Javascript application with a node.js backend server. We make use of Cy-

[3]We include exports from the constraint view tool for all extracted patterns in our supplementary material.

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

Measure	Count
Graph Nodes:	
Nodes before merging	700
Nodes after merging	540 (77%)
Graph Edges:	
Edges before curation	637
Edges after curation	771 (21%)
Grid Row-to-Row Connections:	
Row-to-row connections before curation	1384
Row-to-row connections modified	631 (46%)
Row-to-row connections removed	224 (16%)
Grid Edge Constraints:	
Edge constraints before curation	2101
Edge constraints removed	133 (6%)
Edge constraints marked optional	27 (1%)

Table 2: Manual edits done to the automatically generated graph and grid during the merging and curation steps. Values in parentheses represent percent change.

toscape.js (Franz et al., 2015) as a graph visualization plugin, while primarily using the CoSE-Bilkent graph layout algorithm (Dogrusoz et al., 2009) modified to allow variable edge lengths based on the maximum degree of connected nodes to make the graph easier to visualize when assembling densely-connected patterns. The tool was iterated for usability to maximize throughput for the merging and curation steps, and includes functionality for quickly finding knowledge in the graph while seamlessly moving between graph (graphical) and grid (tabular) views, filtering subsets of nodes and edges by various metrics (completeness, table connection, user-selected utility rating), and keeping track of where the annotator is in the curation workflow. A Scala preprocessing tool reads in gold explanations (which can be filtered to include only subsets of questions by a question classification label, such as only *matter*, *energy*, or *life science* questions), applies the initial clustering heuristics, and outputs tab delimited files that are then read in by the tool. Edges between rows in WorldTree are determined by rows have overlapping content lemmas (defined as nouns, verbs, adjectives, or adverbs), with Stanford CoreNLP (Manning et al., 2014) used for lemmatization and POS tagging.

3 Preliminary Evaluation

To evaluate the utility of our approach, we made use of the tool to extract inference patterns present in all questions in the training subset of the WorldTree corpus categorized as belonging to the *Matter* topic, one of the 9 broad science curriculum categories of question topics, using the ARC question classification labels of Xu et al. (2019). This represents 43 of 902 (5%) of questions and explanations in the training corpus, covering topics such as Changes of State of Matter *(e.g. melting, boiling)*, Measuring Properties of Matter *(e.g. temperature, mass)*, Physical vs Chemical Changes *(e.g. length vs composition)*, Properties of Materials *(e.g. electrical or thermal conductivity, taste)*, Properties of Objects *(e.g. shape or volume)*, and Mixtures *(e.g. alloys)*.

Inference Pattern	Nodes	Edges
Alloys	5	4
*Altitude**	8	10
Building requires measuring	11	13
Burning-Preventing Harm	12	15
Change of State	68	128
Chemical Changes	11	12
Containers contain things	6	6
Cooking Food	9	11
Electrical Conductivity	27	52
Friction	15	24
*General Motion**	3	3
*Ice Wedging**	4	4
*Long lasting vs replacing**	5	4
Magnetism	14	20
Manufacturers use mats. for products	5	5
Measurements	22	34
Navigation lost at sea	6	7
Physical Changes	13	14
Seeing	19	29
*Soil erosion**	6	6
*Solutions - Dissolving substances**	4	5
*Sources of Heat**	3	2
*Sunlight as a source of energy**	14	30
*Sunlight location and shadow size**	7	7
*Taste**	9	11
Taxonomic Inheritance	2	1
*Texture**	4	3
Thermal Conductivity	27	34
*Touch-Hardness**	4	3

Table 3: A list of high-level inference patterns discovered in the corpus of explanations for *Matter* science exam questions using this tool. A full list of patterns is provided in Table 5 (see `Appendix`). An asterisk (*) signifies patterns that are partial or otherwise limited in size because they overlap with other topics (e.g. from Earth or Life Science) not examined in this preliminary study.

3.1 Initial merging and curation

The preprocessing procedure generated 273 grids for this subset of the explanation corpus, representing the specific pairs of tables (e.g. KINDOF $\leftrightarrow$ CHANGE) that have direct connections in the explanations for these questions. A total of 1,384 unique row-row connections populated these grids, and required manual verification. Summary statistics for the edits to these grids is shown in Table 2.

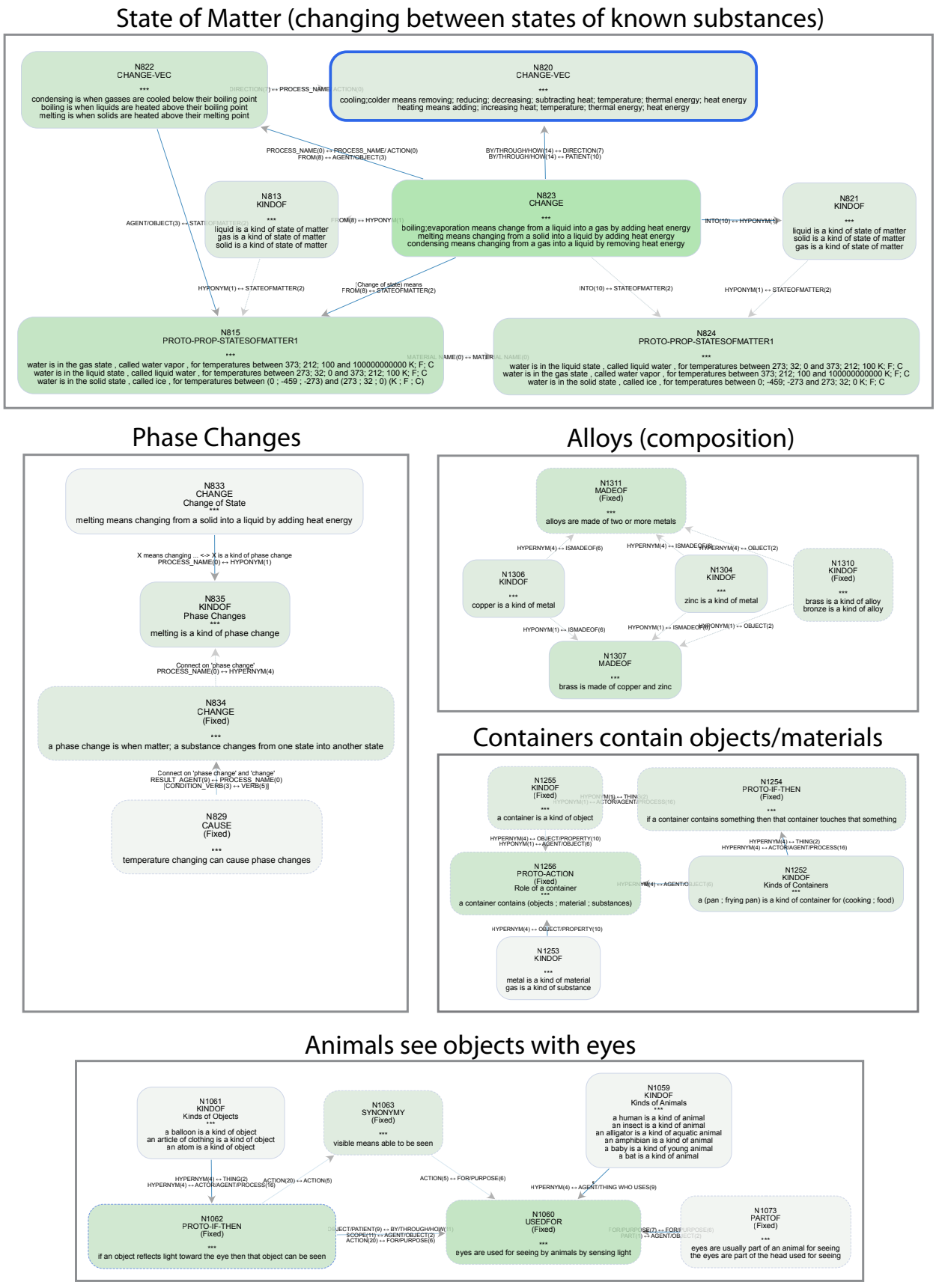

Figure 3: Example inference patterns extracted from the WorldTree explanation corpus using this tool. Nodes represent one or more example facts (table rows) from a specific table (e.g. CHANGE), and edges represent constraints between table rows that must be satisfied. Facts shown in nodes are examples, and not an exhaustive list of all rows that meet the constraints for a given inference pattern.

On average, each grid generally required minimal to moderate editing. Figure 4 (see `Appendix`) shows the full graph before and after the initial merging and curation process.

3.2 Extracting Inference Patterns

Due to it's size, the graph after merging and curation is included in the supplementary material. Manual inspection of the curated graph using the Graph View revealed 29 high-level inference patterns shown in Table 3, each containing between 3 and 66 nodes, and up to 107 edges.[5] These represent the high-level inferences being described in the *Matter* subset of the explanation corpus, and include scientific reasoning processes for topics such as *Measuring Properties with Instruments* and *Thermal Conductivity*, while also describing common world knowledge such as *Seeing*, *Tasting*, and *Cooking Food*. These world-knowledge-centered explanation patterns tend to be either directly required to answer questions (for example, about observing *material properties*), or to process the examples the questions are grounded in (such as *temperature* or *state changes* caused by cooking food). While high-level patterns can be classified as belonging more to scientific or world knowledge, the individual knowledge present in each pattern is generally a mix of both, including nodes that match either scientific knowledge (e.g. *"Matter in the gas phase has variable volume"*) or world knowledge at either a high-level (e.g., *"a balloon is a flexible container"*) or low-level (e.g., *"if a container contains something, then that container touches that something"*).

Examining the 29 high-level inference patterns, we further subdivided them into 38 smaller, more reusable component inference patterns that describe narrower inferences for a given problem domain. For example, the high-level *Change of State* inference pattern was subdivided into 3 smaller and more specialized patterns such as *Changing between states of known substances, Phase Changes,* and *Evaporating Liquids*, each containing between 4 and 9 nodes. Examples of these inference patterns are shown in Figure 3, while the full corpus of patterns generated is included in the supplementary material.

[5]These large inference patterns (up to 66 nodes and 107 edges) represent large topical patterns generated from analyzing many questions on similar topics, and were not derived from any one question. In these cases, it is likely that only a small subset of these larger inference patterns would be used to answer a given question. We describe further subdividing these larger patterns into smaller reusable pieces further in Section 3.2.

Change of State
Freezing means changing from a liquid to a solid by reducing heat energy
A liquid is a kind of state of matter
Water is in the liquid state, called liquid water, for temperatures between 0 C and 100 C
A solid is a kind of state of matter
Water is in the solid state, called liquid water, for temperatures between -273 C and 0 C
Cooling means reducing heat energy
Freezing is when liquids are cooled below freezing point
Phase Changes
Boiling means changing from a liquid to a gas by adding heat energy
Boiling is a kind of phase change
A phase change is when a substance changes from one state to another state
Temperature changes can cause phase changes
Alloys
Alloys are made of two or more metals
Bronze is a kind of alloy
Bronze is made of copper and tin
Tin is a kind of metal
Copper is a kind of metal
Containers contain objects
A container is a kind of object
If a container contains something, then that container touches that something
A bowl is a kind of container
A container contains objects
A rock is a kind of object

Table 4: A small subset of example combinations of knowledge base facts that satisfy the constraints of inference patterns extracted from the explanation corpus. Each example was generated from the inference pattern, and is not found in the training corpus.

3.3 Executing constraint patterns

Our long-term goal is to use the extracted inference patterns to answer unseen questions, and enable generating detailed coherent multi-fact explanations for the reasoning behind those answers. We are currently building a scripting language and development environment for easily authoring and evaluating constraint-based inference patterns.

In the near-term, to evaluate the executability of each pattern, we incorporate a constraint satisfaction framework into the extraction tool allowing the user to test each extracted pattern by querying the tablestore knowledge base and enumerating valid combinations of table rows that satisfy the constraints of a given inference pattern. Our Javascript table constraint solver is able to process approximately 2 million constraint evaluations per second,

which generally satisfies exhaustively testing small patterns in under one minute.[6] The graphical interface allows disabling subsections of larger inference patterns for speed to exhaustively test larger inference patterns piece-wise, or limiting specific nodes to only a small subset of possible facts to speed evaluation.

Examples of valid combinations of facts satisfying the extracted inference patterns in Figure 3 are shown in Table 4. Each of these short explanations was not observed in the training corpus, but rather was generated by satisfying the constraints of an inference pattern by querying the knowledge base, and could form explanations for unseen questions – either in whole, or as part of a combination of several patterns together (such as combining *Changes of State* and *Phase Changes*). At our current state of development, each inference pattern generally matches between one and several thousand unique patterns in the knowledge base, but precise counts are limited by the speed of our current constraint satisfaction solver.

4 Conclusion and Future Work

We present a method and tool for extracting inference patterns from corpora of explanations, where these inference patterns provide a mechanism to combine large amounts of knowledge with high-confidence. While this ability to combine facts into meaningful multi-fact patterns exceeds what is currently possible using contemporary algorithms for multi-hop reasoning, several challenges remain.

First, while significantly faster and more data-driven than our manual attempts at constructing inference patterns, the end-to-end process of constructing an explanation for a question, authoring knowledge base facts, merging and curating a central graph, extracting patterns from that graph, and debugging generic patterns currently comes at a significant labour cost – an average of approximately 2 hours per question[7] – that we are working to further reduce to allow the technique to scale. We hypothesize that a number of the time costs associated with this process scale sublinearly, and are currently working on demonstrating this by refining the protocol and evaluating on an order-of-magnitude more explanations.

Second, while these inference patterns have utility for answering and explaining science exam questions, this needs to be empirically demonstrated by incorporating the patterns into a question answering system to measure the overall recall of this technique. We are actively pursuing both the construction of a corpus of science-domain explanation patterns, at scale, while concurrently developing methods of using these inference patterns to answer questions and provide compelling multi-fact explanations for their answers.

[6] We are currently developing a high-performance standalone constraint satisfaction solver for these types of lexicalized table-based constraint satisfaction problems.

[7] Approximate durations of the most time consuming steps (average per question): explanation construction: 15 minutes; merging/graph curation/high-level pattern identification: 45 minutes; subpattern identification/debugging: 45 minutes.

5 Supplementary Material

This work contains supplementary material, including additional tables and figures in the `Appendix` below, and a corpus of 67 extracted inference patterns available at `http://www.cognitiveai.org/explanationbank/`.

6 Acknowledgements

We thank Elizabeth Wainwright, Steven Marmorstein, Zhengnan Xie, and Jaycie Martin for contributions to the WorldTree explanation corpus, who were funded by the Allen Institute for Artificial Intelligence (AI2). This work is supported by the National Science Foundation (NSF Award #1815948, "Explainable Natural Language Inference", to PJ).

References

Renée Baillargeon and Gerald F DeJong. 2017. Explanation-based learning in infancy. *Psychonomic bulletin & review*, 24(5):1511–1526.

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

Rajarshi Das, Shehzaad Dhuliawala, Manzil Zaheer, Luke Vilnis, Ishan Durugkar, Akshay Krishnamurthy, Alex Smola, and Andrew McCallum. 2017. Go for a walk and arrive at the answer: Reasoning over paths in knowledge bases using reinforcement learning. *AKBC*.

Gerald DeJong and Raymond Mooney. 1986. Explanation-based learning: An alternative view. *Machine learning*, 1(2):145–176.

Ming Ding, Chang Zhou, Qibin Chen, Hongxia Yang, and Jie Tang. 2019. Cognitive graph for multi-hop reading comprehension at scale. *ACL*.

Ugur Dogrusoz, Erhan Giral, Ahmet Cetintas, Ali Civril, and Emek Demir. 2009. A layout algorithm for undirected compound graphs. *Information Sciences*, 179(7):980–994.

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

Max Franz, Christian T Lopes, Gerardo Huck, Yue Dong, Onur Sumer, and Gary D Bader. 2015. Cytoscape. js: a graph theory library for visualisation and analysis. *Bioinformatics*, 32(2):309–311.

Daniel Fried, Peter Jansen, Gustave Hahn-Powell, Mihai Surdeanu, and Peter Clark. 2015. Higher-order lexical semantic models for non-factoid answer reranking. *Transactions of the Association for Computational Linguistics*, 3:197–210.

Peter Jansen. 2018. Multi-hop inference for sentence-level textgraphs: How challenging is meaningfully combining information for science question answering? *TextGraphs*.

Peter Jansen, Niranjan Balasubramanian, Mihai Surdeanu, and Peter Clark. 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.

Peter Jansen, Rebecca Sharp, Mihai Surdeanu, and Peter Clark. 2017. Framing qa as building and ranking intersentence answer justifications. *Computational Linguistics*.

Peter Jansen, Elizabeth Wainwright, Steven Marmorstein, and Clayton Morrison. 2018. Worldtree: A corpus of explanation graphs for elementary science questions supporting multi-hop inference. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.

Peter A Jansen. 2017. A study of automatically acquiring explanatory inference patterns from corpora of explanations: Lessons from elementary science exams. In *6th Workshop on Automated Knowledge Base Construction (AKBC 2017)*.

Daniel Khashabi, Erfan Sadeqi Azer, Tushar Khot, Ashish Sabharwal, and Dan Roth. 2019. On the capabilities and limitations of reasoning for natural language understanding. *arXiv preprint arXiv:1901.02522*.

Daniel Khashabi, Tushar Khot, Ashish Sabharwal, Peter Clark, Oren Etzioni, and Dan Roth. 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.

Daniel Khashabi, Tushar Khot, Ashish Sabharwal, and Dan Roth. 2018. Question answering as global reasoning over semantic abstractions. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Tushar Khot, Ashish Sabharwal, and Peter Clark. 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.

Heeyoung Kwon, Harsh Trivedi, Peter Jansen, Mihai Surdeanu, and Niranjan Balasubramanian. 2018. Controlling information aggregation for complex question answering. In *European Conference on Information Retrieval*, pages 750–757. Springer.

Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60.

Simon Ostermann, Michael Roth, Stefan Thater, and Manfred Pinkal. 2017. Aligning script events with narrative texts. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, pages 128–134, Vancouver, Canada. Association for Computational Linguistics.

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

Michael Schmitz, Robert Bart, Stephen Soderland, Oren Etzioni, et al. 2012. Open language learning for information extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 523–534. Association for Computational Linguistics.

Robert Speer and Catherine Havasi. 2012. Representing general relational knowledge in conceptnet 5. In *LREC*, pages 3679–3686.

Huan Sun, Hao Ma, Xiaodong He, Wen-tau Yih, Yu Su, and Xifeng Yan. 2016. Table cell search for question answering. In *Proceedings of the 25th International Conference on World Wide Web (WWW)*, pages 771–782.

Niket Tandon, Gerard De Melo, and Gerhard Weikum. 2017. Webchild 2.0: Fine-grained commonsense knowledge distillation. In *Proceedings of ACL 2017, System Demonstrations*, pages 115–120.

Hai Wang, Mohit Bansal, Kevin Gimpel, and David McAllester. 2015. Machine comprehension with syntax, frames, and semantics. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint*

Conference on Natural Language Processing (Volume 2: Short Papers), pages 700–706.

Dongfang Xu, Peter Jansen, Jaycie Martin, Zhengnan Xie, Vikas Yadav, Harish Tayyar Madabushi, Oyvind Tafjord, and Peter Clark. 2019. Multi-class hierarchical question classification for multiple choice science exams.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*.

7 Appendix

Annotation Workflow: The annotation workflow is as follows: The user selects a subset of questions to process (in this preliminary work, we select all *MATTER* questions in the WorldTree corpus). The user then switches to the Grid View, which displays one "grid" at a time, where each grid represents all the connections from a given table to another table in the tablestore (for example, all the connections from the *KINDOF* table to the *IF-THEN* table). The user then uses the Grid View to quickly verify that the automatic groupings are correct, and make adjustments or edits to these groupings. Here the user can also remove bad edges (two table rows that were automatically connected, but whose connection isn't meaningful), or remove subsets of the column links on edges that are partially correct (see Table 1). Once this is completed, the user then switches to the Graph View, where they click on each node group from the recently curated grid, highlight other nodes that contain similar rows, and make manual node merging decisions (by dragging and dropping nodes on top of each other). Notes can also be left on specific nodes or edges, to help describe what underlying concepts the nodes represent, and how they interconnect. Once this is completed, the user marks that grid completed, and moves on to the next grid. User-selectable filtering allows only nodes and edges from grids that have been completed to be displayed, greatly reducing clutter and visual search time.

Once the user has completed all grids, the graph is completed, and represents the interconnected knowledge of all of the explanations in the questions, typically itself clustered into a number of disconnected graphs that represent large high-level inference patterns (such as magnetic attraction, thermal transfer, or changes of state of matter). The user then manually inspects these, and highlights subgraphs of nodes to form a candidate inference pattern. These candidate patterns form a series of knowledge constraints for a series of tablestore rows that must be met in each node in order to satisfy the constraints. These constraints can then be run, debugged (as a whole, or as subsets of nodes or edges), and saved. During this process, missing knowledge or edits to existing knowledge in the tablestore that prevent generalization are often discovered – these edits can be immediately made to the Tablestore Google Sheet and the constraints rerun in seconds, to form a fast iteration cycle for debugging knowledge base and inference pattern constraint interactions.

7.1 Additional Resources

A full export of the inference patterns generated in this work, as well as example patterns from the knowledge base that satisfy their patterns of constraints, is available at `http://www.cognitiveai.org/explanationbank/`.

7.2 Additional Tables and Figures

Additional tables and figures are provided below.

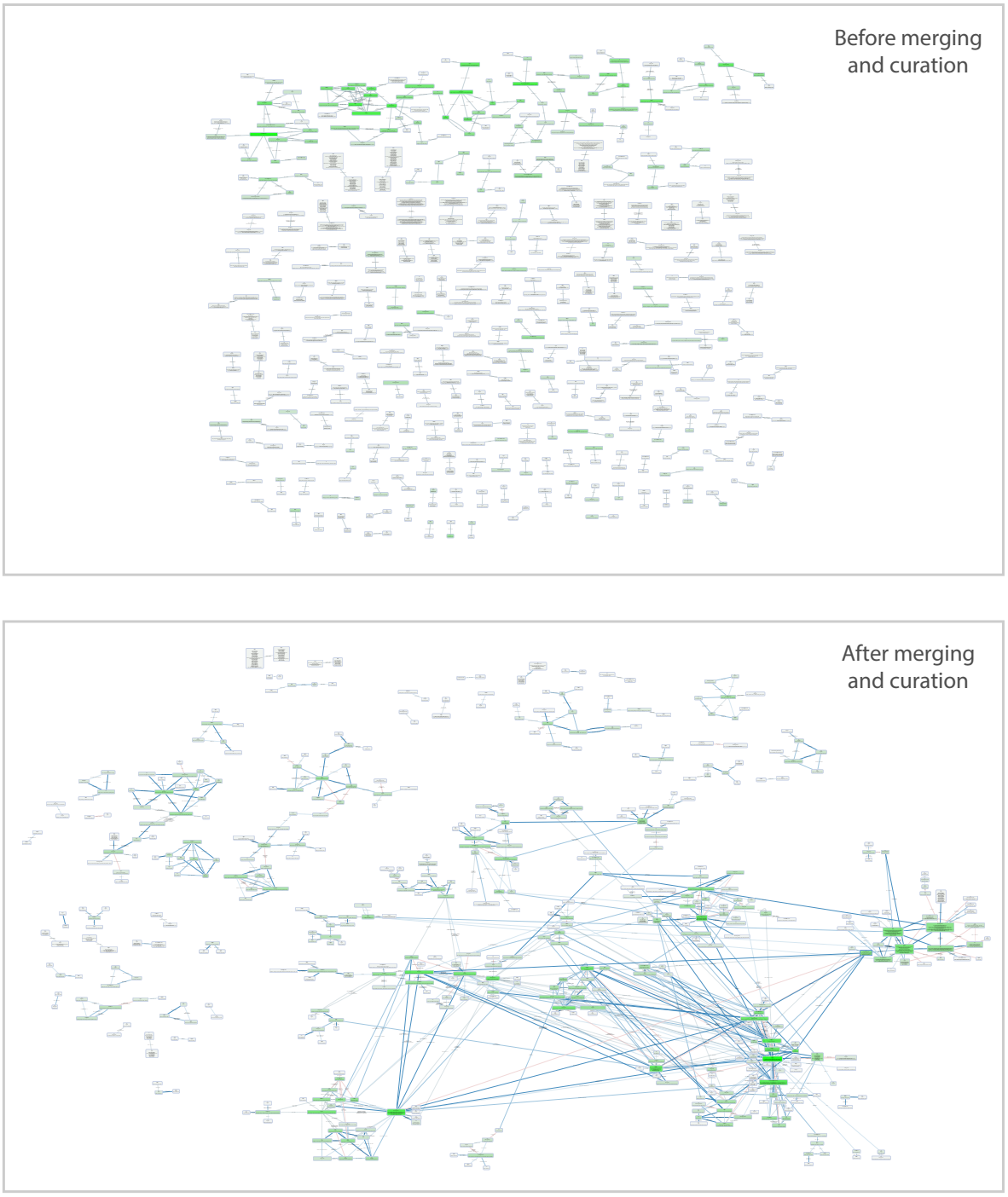

Figure 4: (top) the graph generated by the preprocessing tool, before manual curation and editing by the tool (step 2 in Figure 1). (bottom) the graph after manual curation and editing, and before inference patterns have been generated (step 3 in Figure 1). Clusters in the bottom graph approximately correspond to high-level inference patterns. The set of inference patterns is not shown for space, but each extracted pattern and it's enumerations are included as separate files in our supplementary material.

Inference Pattern	Nodes	Edges	Enumerated Instances in KB
Alloys	-	-	-
Alloy (Core)	3	2	27
Alloy (Composition)	5	6	8
Alloys (Single Elem Not Alloy)	3	2	6
*Altitude **	8	10	6
Benefits of long lasting vs replacement	5	4	2
Building requires measuring - Study materials	-	-	-
Building requires measuring	6	7	19
Sturdy materials for building	6	7	3
Burning-Electrocution-Preventing Harm	-	-	-
Harm caused by burning	9	10	546
Harm caused by electrocution	7	7	324
Change of State	-	-	-
Change of State (Evaporating Liquids)	9	9	11
Phase Changes	4	3	7
State of Matter (changing between states of known substances)	8	13	3
Chemical Changes	-	-	-
Chemical Changes (Core+Grounding Specific Chemical Change)	4	3	2
Chemical Reactions (Core)	4	5	15
Chemical Reactions (Core + Substance Grounding)	6	7	363
Chemical Reactions (e.g., acids)	4	3	3
Containers contain things	-	-	-
Containers (Abstracted)	5	5	1000
Containers (Application)	6	6	15
Cooking Food	-	-	-
Cooking (Core)	6	6	26
Cooking a particular food	7	7	338
Cooking (Containers for cooking)	6	6	1
Electrical Conductivity	-	-	-
Dangers of Electric Shock	4	3	414
Electrical Insulation	15	23	46
Electrical Circuits in Devices	7	11	18
Friction	-	-	-
Friction (core)	16	31	3
*General Motion **	3	3	6
*Ice Wedging **	4	4	2
Magnetism	-	-	-
Magnetic Objects	5	4	10
Manufacturers use material for products	-	-	-
Manufacturers use materials for products (core)	4	3	19
Measurements	-	-	-
Measurement Tools	4	4	130
Observations (Celestial Bodies)	5	6	6
Observations (Distant Objects)	5	6	208
Observations (Microscopic Things)	6	6	4
Observations (Small Things)	6	6	94
Navigation-Direction-Being lost at sea	-	-	-
Navigation (core)	3	2	1
Navigation (being lost/boat)	6	7	2
Physical Changes	-	-	-
Physical Changes (Changing Shape)	9	10	832
Seeing	-	-	-
Things that can see and what they can see	6	6	1000
*Soil erosion **	6	6	28
*Solutions - Dissolving substances **	4	5	1
*Sources of Heat **	3	2	6
*Sunlight as a source of energy **	14	30	80
*Sunlight location and shadow size **	7	7	312
*Taste **	9	11	26
Taxonomic Inheritance	2	1	1000
*Texture **	4	3	2
Thermal Conductivity	-	-	-
Thermal Conductivity (Core)	21	26	1000
Thermal Conductors 0	5	4	9
Thermal Conductors 1	5	4	8
Thermal Insulators	5	5	5
*Touch-Hardness **	4	3	8

Table 5: An extended list of inference patterns discovered in the corpus of explanations for *Matter* science exam questions using this tool. Indented inference patterns represent a subset of smaller, more generic sub-patterns extracted from the larger pattern. "Enumerated instances in KB" represents the number of unique combinations of facts the pattern generates in our current KB (note that for speed, this currently has a hard upper limit of 1,000 patterns). An asterisk *(*)* represents patterns that are partial or otherwise limited in size because they overlap with questions (e.g. from Earth or Life Science) not examined in this preliminary study.