A Study of Automatically Acquiring Explanatory Inference Patterns from Corpora of Explanations: Lessons from Elementary Science Exams

Peter A. Jansen
School of Information, University of Arizona, Tucson, AZ
pajansen@email.arizona.edu

Abstract

Our long term interest is in building inference algorithms capable of answering questions and producing human-readable explanations by aggregating information from multiple sources and knowledge bases. Currently information aggregation (also referred to as "multi-hop inference") is challenging for more than two facts due to "semantic drift", or the tendency for natural language inference algorithms to quickly move off-topic when assembling long chains of knowledge. In this paper we explore the possibility of generating large explanations with an average of six facts by automatically extracting common explanatory patterns from a corpus of manually authored elementary science explanations represented as lexically-connected explanation graphs grounded in a semi-structured knowledge base of tables. We empirically demonstrate that there are sufficient common explanatory patterns in this corpus that it is possible in principle to reconstruct unseen explanation graphs by merging multiple explanatory patterns, then adapting and/or adding to their knowledge. This may ultimately provide a mechanism to allow inference algorithms to surpass the two-fact "aggregation horizon" in practice by using common explanatory patterns as constraints to limit the search space during information aggregation.

1 Introduction

Explainable methods of inference have emerged as an important area of study in natural language processing and machine learning [11], particularly for domains such as scientific discovery or automated medical diagnosis, where user trust is paramount, and the cost of making errors is high. Standardized science exams provide a challenge task for explainable question answering [1], as the questions contain a variety of complex and challenging inference problems [3, 5], and yet are expressed in simple-enough language that they provide a proving ground for developing algorithms specifically aimed at improving our capacity for automated inference and explanation generation while controlling for language complexity. Underscoring the difficulty of this task, to date the best ensemble solvers answer approximately 60% of elementary and middle-school science questions correctly [2, 12], and the best explanation-centered solver using information aggregation produces high-quality human-readable explanations for only about 25% of questions [6].

QA is often modeled as either a retrieval task, or an inference task. With respect to retrieval, models perform an answer sentence selection task where they learn to identify single sentences or short continuous passages of text in a corpus that answer the question. Many questions that require complex inference often can't be answered by a single continuous passage of text, and inference models work to solve this by identifying and aggregating information from multiple sources using a variety of techniques including logic [9], heuristics for matching rows of semi-structured tables [8], or learning to combine free-text sentences decomposed into graphs on clausal and prepositional boundaries [6].

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

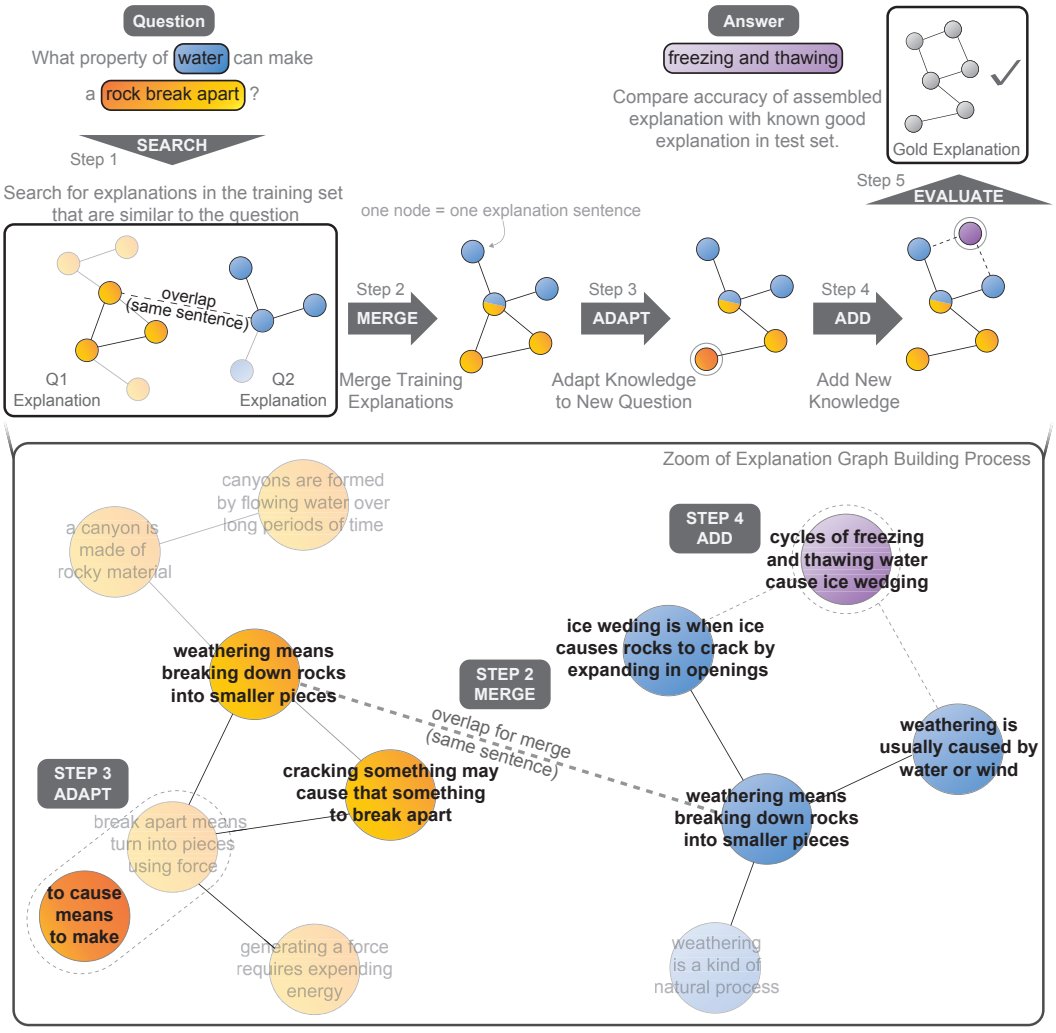

Figure 1: An overview of our proposed approach: inferences and explanations to novel questions are constructed by merging, adapting, and adding to subsets of gold explanations from known questions.

In the elementary science domain, building corpora of explanations has allowed us to estimate that each question requires an average of 4 to 6 separate facts to both correctly answer and to generate an explanation for [5]. Currently, nearly all inference models for QA that we are aware of have difficulty successfully aggregating more than two facts, due to the phenomenon of "semantic drift", or the tendency for inference to quickly drift off topic when creating long chains of facts connected by imperfect signals, like lexical overlap (i.e. shared words in two sentences), in a large search space [4]. To understand precisely how much semantic drift affects inference, we recently characterized how often lexical overlap leads to the meaningful aggregation of facts using 9,500 manual ratings of aggregation quality on both in-domain (elementary study guide) and open-domain (Wikipedia) corpora. We found the chance of meaningfully aggregating two sentences to be approximately 5% for in-domain resources, and 0.1% for the open-domain corpus, making chance performance for successfully constructing the average 6-fact explanation at less than one in a million. [1]. Overcoming the issue of semantic drift and the current "aggregation horizon" of 2 facts is one of the substantial barriers to successfully answering complex questions.

Characterizing explanations may provide a vehicle for reducing semantic drift, by identifying common explanatory patterns present in explanations that can provide a scaffold for constructing new

[1]The details of this characterization of lexical overlap as a vehicle for information aggregation and explanation generation are described in a paper currently under review

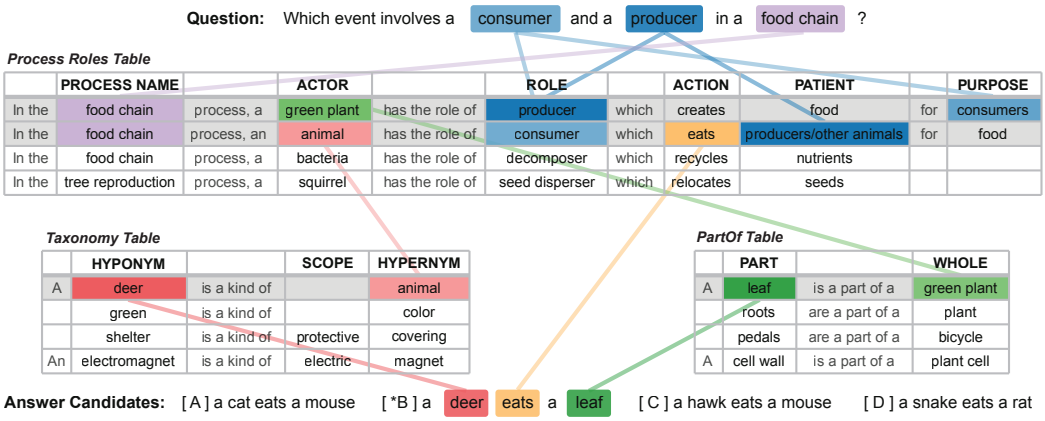

Figure 2: An example 4-sentence explanation, expressed as rows in semi-structured tables. These rows form a lexically-connected "explanation graph".

explanations. Here we approach reasoning not as a problem of building a "universal reasoning engine", but by building a large set of very specific competencies that together are able to answer a wide variety of questions, similar to Minsky's Society of Mind hypothesis [10]. We hypothesize that sufficiently detailed explanations can form explicit records of the reasoning required by a human to move from question to correct answer, and that large corpora of explanations will show common explanatory patterns (at various levels of abstraction) that can be reused to develop domain competencies and perform inference on new, unseen questions. We further hypothesize that many of these patterns will be highly domain-specific, applying only to a narrow subset of questions, while others may be general and be useful across a variety of question types.

This project is only at the preliminary stages, as we have only recently developed and demonstrated tools, representations, and training protocols that allow the construction of corpora of explanations in computable form at the scale of thousands of questions. This paper presents our first steps in identifying possible forms that common explanatory patterns might take, and we investigate representations of explanation patterns as graphs at different levels of abstraction. We also explore how explanatory patterns might be used in practice by being combined, adapted, or added to with additional facts that together provide both the inference required to answer science questions, as well as a compelling human-readable justification for why those answers are correct (this process is illustrated in Figure 1). We conclude with a study of the feasability of this approach on rebuilding gold explanations that require aggregating up to 6 facts to successfully answer and explain.

2 Representing explanations as lexically-connected graphs

In this study, we make use of an in-house[2] corpus of 1,680 computable explanations for elementary (3^{rd} to 5^{th} grade) science questions from standardized exams in 12 US states, drawn from the AI2 Mercury and Licensed datasets.[3] Because a repository of explanations does not exist for this domain, each explanation was manually authored by an annotator to be an explicit natural language record of the reasoning required for a given question and problem solving method. Explanations were targeted at the level of a 5-year old, and contain detailed world knowledge that might seem overly verbose to adults, but that is likely required to learn to perform computational inference.

To facilitate automated analysis, explanations are represented as a series of facts, with each fact expressed as a row in a knowledge base of semi-structured tables. Each table is centered around a particular relation, initially seeded from 20 common explanatory relations identified by Jansen et al. [5], then expanded to include additional subtypes as identified in the data. The knowledge base includes a total of 4,950 rows across 62 tables, and three broad types of knowledge: retrieval types (e.g. *taxonomic, part-of, properties*), inference-supporting types (e.g. *actions, uses, requirements,*

[2]We call this explanation graph corpus "WorldTree". The development of the tools, representations, and annotation protocols to construct this large-scale corpus of explanation graphs is described in a paper currently under review.

[3]`http://www.allenai.org/data.html`

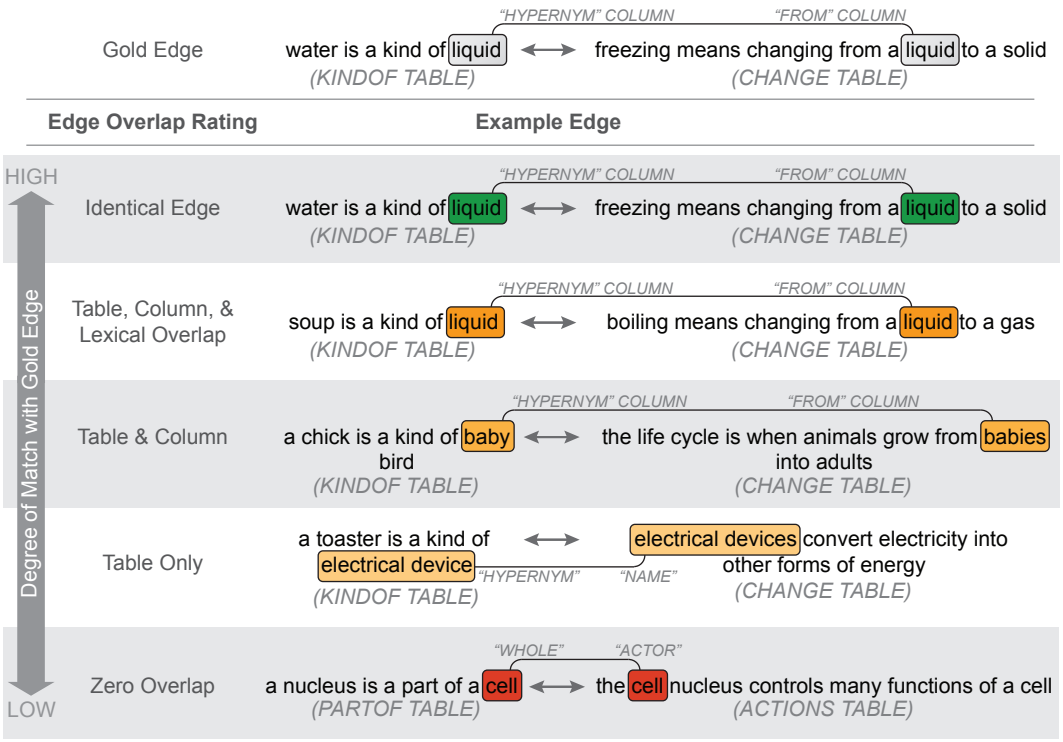

Figure 3: Patterns of edge overlap in explanation graphs at four levels of abstraction, grounded in an example gold edge, and five edges that have a given level of abstracted overlap with the gold edge.

sources), and complex inference types (e.g. *causality, coupled relationships, if/then, changes*). Similar to [8, 7], each table includes "filler" columns, allowing rows to concurrently function both as computable semi-structured relational knowledge, as well as plain-text natural language sentences in human-readable form. An example question and explanation constructed from four rows in three separate tables is shown in Figure 2. In this corpus, each of the 1,680 questions requires an average of 6 facts (table rows) to answer and explain the reasoning behind that answer.

We include the additional requirement that each sentence (or table row) in an explanation must be lexically connected (i.e. have shared words with) at least one other sentence in the explanation. In this way the explanations form lexically-connected "explanation graphs" that explicitly represent both the knowledge required to answer and explain a question, as well as *how* the knowledge in an explanation is interconnected. As shown in Figures 1 and 2, the explanation graph can be represented both at the level of the table cell (low-level) and the level of the table row/sentence (high-level).

3 Identifying common explanatory patterns

What is an explanatory pattern? At a high-level, these might take the form of broad recipes for answering a specific kind of question. At a low level, these might be thought of as n-grams (or, subgraphs) of gold explanation graphs that commonly reoccur across different questions in the explanation corpus, and might ultimately be used like puzzle pieces to construct new explanations.

Here we characterize explanation patterns by representing edges in gold explanation graphs at 4 levels of abstraction, shown in Figure 3. The edges can be abstracted in several ways, where a specific table row is replaced with progressively less specific references to that row. For example, the two rows on an edge might be replaced with only a reference to the table that they come from (e.g. *KINDOF* and *CHANGES*). Progressively more specific references might include both the table and columns that overlap on a given edge, and this could be further specified to also include the specific words in those two columns that overlap.

Table 1: Example explanation patterns involving the KINDOF relation for each level of abstraction. These explanatory patterns are small, involving only a single edge in an explanation graph. Example linked sentences provided satisfy at least the minimum constraints of the pattern. Frequency represents the number of questions whose explanations contain one of these patterns. Bold words highlight the lexical overlap (shared words) on a given edge.

Freq.	Pattern	Abs.
11/800	*water is a kind of **liquid** (KINDOF) ↔ boiling means changing from a **liquid** to a gas by adding heat energy (CHANGE).*	ROW
5/800	(KINDOF, HYPONYM COL) ↔ (PARTOF, WHOLE COL) ["plant"] *e.g. a tree is a kind of **plant** ↔ a leaf is a part of most **plants***	T/C/L
119/800	(KINDOF, HYPONYM COL) ↔ (KINDOF, HYPERNYM COL) *e.g. a dog is a kind of **animal** ↔ an **animal** is a kind of organism*	T/C
80/800	(KINDOF) ↔ (USEDFOR) *e.g. a **magnifying glass** is a kind of tool ↔ a **magnifying glass** is used to see small objects by making them appear bigger*	T

3.1 Common Explanatory Patterns

We examined explanation graphs for approximately half the corpus (800 questions), representing their edges at each of the 4 levels of abstraction, from identical edge to only references to tables. Example patterns centered around the *KINDOF* table are shown in Table 1, and a complete list is available in the supplementary data. The most common identical edge is *water is a kind of liquid (KINDOF) ↔ boiling means changing from a solid to a liquid by adding heat energy (CHANGE)*, an edge common for change-of-state questions that discuss boiling, and which occurs in 11 separate questions (1.4% of all questions). In total, 514 totally lexicalized edges (i.e. same rows) repeat more than once in the corpus, with other frequently recurring edges discussing genetically inherited versus learned characteristics, photosynthesis, electricity, astronomy, and other common test topics.

Progressively more abstracted versions of the edges occur more commonly, at the risk of reduced specificity. The most common rule at the table/column level of abstraction is taxonomic traversal – i.e., knowing that *a dog is a kind of animal, and an animal is a kind of organism*, which occurs at least once in 15% of all explanations (119 of 800, see Table 1). Taxonomic traversal is often manually implemented as a high-confidence rule in inference solvers, but here emerges naturally as a frequently used explanatory pattern in the corpus, highlighting the potential utility of this approach to ultimately aid automatic methods of inference.

3.2 Reusing common explanatory patterns through merges, adaptations, and additions

While specific edges at different levels of abstraction are commonly reused in explanations, is it possible that larger subgraphs also commonly reoccur, potentially dramatically reducing both the search space of knowledge, and need for novel information aggregation when making an inference for novel questions?

To answer this question, we found all questions in the corpus whose gold explanation graphs contain up to 9 edges, and then exhaustively attempted to reconstruct those gold explanations using subgraphs of explanations taken from other questions, also exhaustively represented at all possible combinations of abstraction. We allow up to 3 subgraphs to merge together based on having common edges, and rank the resulting merged subgraphs based on the number of missing edges when compared to the gold graph, the level of abstraction of common edges that subgraphs were merged on (favoring identical edges before abstracted matches), as well as the least overall combination of abstractness across all edges in the subgraph. In this way, the combined graph that emerges as the top ranked reconstruction attempts to prioritize having some representation for all the edges in the gold graph, while making high-confidence (i.e. less abstract) merges between subgraphs, and having the least abstract representation of edges compared to the gold graph.

An example pattern is included in Table 2, and the average of the top-ranked reconstructed graphs is shown in Figure 4, averaged across gold graphs with between 1 and 9 edges, and broken down by the number of subgraphs merged. Grounding this in a hypothetical gold explanation graph with 9 edges, this analysis shows that there are at least two subgraphs in the explanation corpus that, when merged, will contain an average of 3 identical edges, 1.5 missing edges that require information

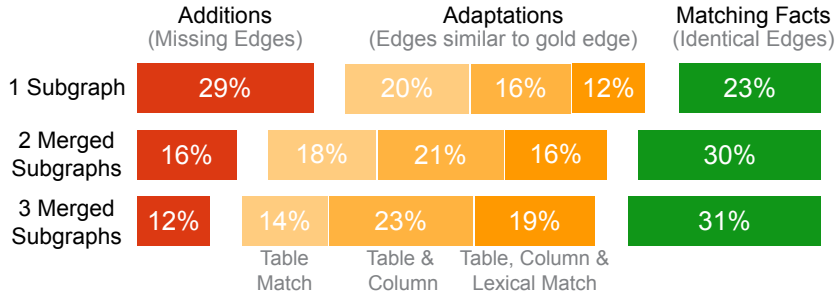

Figure 4: Average reconstruction quality of gold explanations of between 1 and 9 edges, for both single explanation subgraphs, and merges of up to 3 subgraphs.

Table 2: Example question, its gold explanation, and an explanation for a different question that contains zero of the same sentences/rows, and few of the same content (i.e. "non-fill") words, but that provides a recipe for answering the question and generating an explanation. *(Note, edges are not drawn for space, but are at the T/C, and T/C/L ["measure"] adaptation level of abstraction).*

Question **Answers**	Walter wanted to find out if faster wind speeds increased the amount of wind erosion. Which instrument should he use to measure wind speed? *(A*) anemometer* (B) barometer (C) rain gauge (D) thermometer
Gold Explanation	wind speed is a property of weather (PROPERTIES). an anemometer is a kind of instrument (KINDOF). an anemometer is used to measure wind speed (USEDFOR).
Similar Explanatory Pattern	air pressure is a property of air/the atmosphere (PROPERTIES). an barometer is a kind of instrument (KINDOF). a barometer is used to measure air pressure (USEDFOR).

aggregation, and 4.5 edges represented at various levels of abstraction away from the gold edges that require adaptation to reconstruct the gold explanation.

4 Conclusion

In this work we explore automatically extracting common inference patterns from corpora of explanations represented as graphs. Our analysis suggests that not only are common explanatory patterns present at various levels of abstraction, but that it may be possible to combine large subgraphs of explanations like puzzle pieces to generate novel explanations to unseen questions, reducing the search space for inference solvers, and increasing the number of facts that can be meaningfully aggregated. The supplementary data for this paper, including common explanatory patterns in this corpus, and a fine-grained characterization of reconstruction quality (Figure 4), are included in the supplementary data at `http://www.cognitiveai.org/explanations` .

Acknowledgements

Thanks to Elizabeth Wainwright, Steven Marmorstein, and Clayton T. Morrison for assistance in constructing the corpus of explanations analyzed in this paper, Peter Clark at the Allen Institute for Artificial Intelligence (AI2) for thoughtful discussions, and AI2 for providing financial support for this project.

References

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

[2] Peter Clark, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter D. Turney, and Daniel Khashabi. 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, 2016.

[3] Peter Clark, Philip Harrison, and Niranjan Balasubramanian. 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, 2013.

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

[5] Peter Jansen, Niranjan Balasubramanian, Mihai Surdeanu, and Peter Clark. 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 2016. The COLING 2016 Organizing Committee.

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

[7] Sujay Kumar Jauhar, Peter D Turney, and Eduard H Hovy. Tables as semi-structured knowledge for question answering. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2016.

[8] Daniel Khashabi, Tushar Khot, Ashish Sabharwal, Peter Clark, Oren Etzioni, and Dan Roth. Question answering via integer programming over semi-structured knowledge. In *Proceedings of the International Joint Conference on Artificial Intelligence*, IJCAI'16, pages 1145–1152, 2016.

[9] Tushar Khot, Niranjan Balasubramanian, Eric Gribkoff, Ashish Sabharwal, Peter Clark, and Oren Etzioni. Exploring markov logic networks for question answering. In *EMNLP*, 2015.

[10] Marvin Minsky. *The Society of Mind*. Simon & Schuster, Inc., New York, NY, USA, 1986.

[11] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "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, 2016. ACM.

[12] Carissa Schoenick, Peter Clark, Oyvind Tafjord, Peter Turney, and Oren Etzioni. Moving beyond the turing test with the allen ai science challenge. *Communications of the ACM*, 60(9):60–64, 2017.