Automatic Construction of Inference-Supporting Knowledge Bases

Peter Clark, Niranjan Balasubramanian, Sumithra Bhakthavatsalam, Kevin Humphreys, Jesse Kinkead, Ashish Sabharwal, Oyvind Tafjord
Allen Institute for AI, 2157 North Northlake Way, Seattle, WA 98103
{firstname+initial-of-lastname}@allenai.org

Abstract

While there has been tremendous progress in automatic database population in recent years, most of human knowledge does not naturally fit into a database form. For example, knowledge that "metal objects can conduct electricity" or "animals grow fur to help them stay warm" requires a substantially different approach to both acquisition and representation. This kind of knowledge is important because it can support inference e.g., (with some associated confidence) if an object is made of metal then it can conduct electricity; if an animal grows fur then it will stay warm. If we want our AI systems to understand and reason about the world, then acquisition of this kind of inferential knowledge is essential.

In this paper, we describe our work on automatically constructing an inferential knowledge base, and applying it to a question-answering task. Rather than trying to induce rules from examples, or enter them by hand, our goal is to acquire much of this knowledge directly from text. Our premise is that much inferential knowledge *is* written down explicitly, in particular in textbooks, and can be extracted with reasonable reliability. We describe several challenges that this approach poses, and innovative, partial solutions that we have developed. Finally we speculate on the longer-term evolution of this work.

1. Introduction

Our long-term goal is the construction of "knowledgeable machines" - machines that contain large volumes of general and scientific knowledge, stored in a computable form, supporting reasoning and explanation. Our short-term instantiation of this goal is to build a machine that can pass elementary grade science exams, itself a hard task that has not yet been achieved in AI. The challenge is formidable because many of the exam questions require reasoning with both domain-specific and commonsense knowledge [1,12]. Given the cost of manually encoding that knowledge, our goal is to acquire much of it automatically.

This performance task helps inform both the AKBC and reasoning requirements. For example, consider the following multiple choice questions, taken from a 4th grade exam:

> What form of energy is being used when a person pushes a wooden block across the floor? (A) mechanical (B) magnetic (C) sound (D) electrical
>
> Fourth graders are planning a roller-skate race. Which surface would be the best for this race? (A) gravel (B) sand (C) blacktop (D) grass

This style of question is typical of 4th grade science, but is very different from the "factoid" questions that are often the focus of AKBC and question-answering [2]. There are no pre-constructed factoid databases (e.g., Freebase tables) where answers can be looked up. Similarly,

while Web search or word correlation methods can sometimes guess answers correctly, such methods are unreliable (e.g., in the second question, correlations between grass and race, and sand and surface, mislead such methods). Instead, they require knowledge about general truths and their application to specific (hypothetical) situations or scenarios. For the first question, we need access to knowledge that relates mechanical energy with movement of objects and an inference mechanism that draws conclusions about the "the block being pushed" using this knowledge. For the second question, we need to know that a roller-skate race involves roller skating, that roller skating is best on a smooth surface, and that blacktop is smooth. Obtaining these fragments of world knowledge and integrating them together is challenging.

Text does provide much of the required knowledge, but often expressed using different wordings and requiring interpretation. For example, for the first question, a study guide textbook [7] states:

> Mechanical energy is produced when two objects move together.
> Mechanical energy exerted by one object can push or pull another object.

and for the second we find in dictionaries:

> Roller skating: glide across a smooth surface wearing roller skates
> Blacktop: a bituminous material used for providing a smooth paving to a road.

These examples illustrate some obstacles for acquiring and using knowledge from them:

- Their wordings do not neatly align with the wording in the questions; rather, there are differences in both generality and phrasing that the system must bridge
- They are almost all generic sentences, a difficult type to process
- They are imprecise and/or have assumed (unstated) context
- They are linguistically complex, meaning interpretations are likely to contain errors

These lead to several requirements for acquisition, representation, and reasoning:

- the system must have knowledge of linguistic variability and be able to tolerate it
- the system must be able to relate specific cases to their generalizations (taxonomic reasoning)
- the system must have a means for interpreting and applying generic statements
- reasoning must be robust to an incomplete/noisy KB (some notion of probability/confidence is essential). In particular, the system needs to be able to make plausible leaps ("jump the gaps") between facts during inference if knowledge is missing.

We have made some initial progress towards a representation language (partially) meeting these requirements, and AKBC technology for extracting knowledge in that form. We give a short overview of our progress, and then describe our longer-term directions.

2. Approach

2.1 Representation and Inference

The input to our AKBC task is a collection of textbook-style texts and dictionary definitions, and the output is a knowledge base (KB), expressed in a probabilistic (subset of) first-order logic. Because the process is automated, the resulting logic expressions all have a similar structure, namely a "forall...exists..." format that we illustrate shortly. To tolerate a degree of linguistic variability, we include original text strings (words/phases) in the representation to denote concepts, and replace equality with a (phrasal) textual entailment service that provides a degree of confidence in equality. To tolerate incompleteness in the KB, we use an inference process that mixes logical reasoning with lexical matching, allowing the system to weakly reach conclusions even if some evidence is not known, thus "jump knowledge gaps" and avoid the brittleness of traditional deductive reasoning. A semantics for this is given in [6]. For other work on mixed inference, see [5,10,11].

2.2 AKBC Methodology

Our approach to AKBC occurs in three steps:

1. Specification: Identify the different classes of knowledge required for the performance task, and create a vocabulary for representing them (typically a small number of predicates).
2. Extraction: Build and apply extractors that identify expressions of those knowledge types in text.

3. Interpretation: Generate inference rules from the extractions. These rules denote the inference(s) that the extractions are deemed to sanction.

We now briefly describe these steps.

2.2.1 Specification

Although our performance task requires substantial knowledge, there are recurring *types* of knowledge that can be identified. For these types, we define a simple modeling vocabulary (typically one or two predicates), so that they can be represented semantically rather than lexically in the KB. These predicates are the targets for extraction. From an analysis of exam questions, our initial target predicates were: CAUSE/EFFECT, ENABLES, PURPOSE, REQUIREMENT, CONDITION, PART-OF, and EXAMPLE-OF relationships. We have recently added a more specialized vocabulary for representing statements about discrete and qualitative changes.

2.2.2 Extraction

Rule acquisition is performed in two steps. In the first step, we identify where the relations of interest are expressed in text. This is done using a (currently) hand-authored set of extraction rules that associates particular syntactic dependency patterns with a semantic relationship. For example, a rule states that the pattern "X results in Y" may denote (with some confidence) a causality relationship CAUSES(X,Y). Each identification produces a data structure containing the syntactic elements involved in the pattern. Where X and Y are events, we further decompose them into *subject+verb+object[+prepositional-phrase]** structures. For example, for one of the earlier example sentences, the extraction is as follows:

"Mechanical energy is produced when two objects move together."
("two objects"/?x "produce" "Mechanical energy") "when"/CONDITION
("two objects"/?x "move" "" ["together"])

2.2.3 Semantic Interpretation

Extractions are just a semi-formal data structure reflecting generic statements in text. To use these for reasoning we need to make explicit what inferences they sanction, i.e., make a semantic interpretation. While generic statements can be interpreted in many ways, e.g., [3,4], we adopt a simple, "forall...exists..." default reading where (A RELATION B) is interpreted as "for all instances of A, there will (with some confidence c1) exist a B that is in RELATION to A.". We also add a reverse reading, namely "for all instances of B, there will (with some confidence c2) exist an A that is in inverse(RELATION) to B". These interpretations are mechanically generated from the extractions by an algorithm that embodies this interpretation policy. We currently assume equal confidences c1 and c2, but are developing algorithms for better estimating confidences (using machine learning over hand-annotated examples). An example of a synthesized rule is:

```
// IF two objects move together THEN mechanical energy is produced.
forall m, t, o
isa(m,"move"), isa(t,"together"), isa(o,"two objects"), agent(m,o), arg(m,t)
  -> exists p, e
     isa(p,"produce"),isa(e,"Mechanical energy"), agent(p,o), object(p,e), condition(p,m).
```

We have used this approach to construct an initial KB containing approximately 30,000 rules from science texts and 15,000 rules from dictionary definitions.

2.4 Question-Answering

To answer a 4-way multiple choice question, the question is first converted to four true/false questions (one for each option) of the form "Is it true that X_i?". Each X_i is then converted into an inference rule using similar NLP machinery, with the important difference that the rule is to be proved, rather than asserted. To prove the rule, the LHS variables are existentially rather than universally quantified, then the result considered to hold for all variable values by the principle of Universal Generalization. For the earlier example ("What form of energy is being used when a person pushes a wooden block across the floor? (A) mechanical…"), the question interpretation for the correct option (A) looks:

```
// IF a person pushes a wooden block across the floor using X THEN X is mechanical energy
exists p, u, pu, b
  isa(p,"a person"), isa(u,"used"), isa(p,"pushes"), isa(b,"a wooden block across the floor"),
  agent(pu,p), object(pu,b), condition(u,pu), agent(u,p), object(u,x)
    -> isa(x,"mechanical energy").
```

To "prove" this (or more precisely, derive a confidence for this), the LHS is asserted as the setup (known facts), the RHS is the goal to prove, and a (currently best-first) inference procedure searches for chains of reasoning from setup to goal.

As mentioned earlier, soft matching (combining predicate and lexical methods) is used to trigger a rule with a degree of confidence, and an entailment service replaces equality. In this case, the rule shown earlier applies, the entailment service returning (among other things) a confidence that "pushes" entails "moves". An overall confidence in the goal is computed, and the answer option with the highest confidence returned. Note that soft matching may result in inference chains to incorrect answers also, but typically with lower overall confidence.

In preliminary experiments, the inference system currently scores 60% on the target task (non-diagram multiple choice questions on the NY Regents 4th grade science exam), using a small (68 question) dataset of unseen test data, 5% above the best retrieval-based bag-of-word method we have developed. Equally importantly, the inference system returns an explanation for its conclusion, based on the inference chain that was used. The retrieval-based method submits each X_i as a search query to our corpus of text (using Lucene as the search engine), collects the Lucene score of the best-matching sentence retrieved for each X_i, and picks the answer option i with the highest score. This provides an initial indication that the approach is viable, but further experiments with significantly larger datasets are needed and planned.

3. The Longer-Term Vision

This work is significant because we are acquiring knowledge in the form of inference-supporting rules, allowing a previously inaccessible class of questions to be attempted. However, the work is also preliminary, with clear limitations in the current treatment of generics, confidences, and inference, that we continue to investigate. For our purposes here, we now describe three longer-term themes we see as important to the longer-term vision.

3.1 Richer Models

Mapping natural language to semantic predicates canonicalizes the expressed knowledge, and allows general axioms (e.g., CAUSES is transitive) to be exploited for question-answering. More generally, we can identify *families* of semantic predicates that together can be used to model a phenomenon (e.g., next-event(), subevent() for processes; Q+, Q-, I+, I- for qualitative influences [8]) and that can be interrelated via axioms. A family of such predicates constitutes a modeling language, and their axioms can be seen as a piece of "computational clockwork" with which to model the world. These are useful because, if we map text onto one of these modeling languages, the system can then draw the conclusions sanctioned by the model. For example, if an extractor converts a paragraph about photosynthesis into an instance of a qualitative model (Q+, Q-, etc.), then axioms from the model can be used to reason about the process (e.g., less sunlight implies less O2 produced). A good example of this for processes is [9].

While we have used just a handful of semantic predicates so far, our longer-term picture of the KB includes a (small) collection of modeling languages, hand-built, and a (large) library of model instances expressed in those languages, extracted automatically. These will provide additional inferential power for question answering, exploiting the axioms of the modeling languages.

3.2. Confidence Measures

We currently do not have a good measure of confidence/reliability of extracted rules, and sometimes "bad" rules cause bad results. To reduce this problem, confidences could be associated with rules and used to bias reasoning (higher weight to apparently "good" rules). Signals for confidences can come from several places, e.g., repetition, use in (successful) question-answering tasks, degree to which other rules conflict with a given rule.

3.3. Bootstrapping

Our current pipeline (parse→extraction→axiom synthesis) is noisy, with errors introduced at every stage. In many cases, e.g., bad PP attachment, the errors are "obvious" to a human reader because they conflict with the reader's knowledge of the world. Given that the system is acquiring world knowledge itself, an obvious and desirable approach is be to similarly exploit that acquired knowledge in future reading: Knowledge creates expectations to guide text interpretation, and text interpretation creates new knowledge.

4. Summary

Our goal is "knowledgeable machines", and our approach is the automatic acquisition of inference-supporting knowledge from text. We have identified some key requirements for the approach, and summarized a preliminary way of meeting them that, although involving substantial simplifications, has competitive performance on (unseen) test data. Further development of these techniques is essential if we are to build machines that have a richer understanding of the world.

Acknowledgements

We also thank Isaac Cowhey, Oren Etzioni, Dirk Groeneveld, Sunil Mishra, Mark Schaake, and Michael Schmitz for their critical contributions to this work.

References

[1] Clark P.,Harrison P., Balasubramanian N. A Study of the AKBC/Requirements for Passing an Elementary Science Test. In Proceedings of the AKBC-WEKEX workshop at CIKM 2013.

[2] Fader, A., Zettlemoyer, L., and Etzioni, O., Open question answering over curated and extracted knowledge bases. In Proceedings of the KDD 2014 conference.

[3] Schubert, Lenhart K., and Francis Jeffry Pelletier. "Generically speaking, or, using discourse representation theory to interpret generics." Properties, types and meaning. Springer Netherlands, 1989. 193-268.

[4] Liebesman, David. "Simple generics." Noûs 45.3 (2011): 409-442.

[5] Islam Beltagy and Stephen Roller and Gemma Boleda and and Katrin Erk and Raymond J. Mooney. Natural Language Semantics using Distributional Semantics and Probabilistic Logic. SemEval-2014.

[6] A. Sabharwal, N. Balasubramanian, E. Gribkoff. Markov Logic Networks for Natural Language Question Answering. AI2 Technical Report, 2014.

[7] J. Barry, K. Cahill. Barrons Fourth Grade Science Study guide. 2007.

[8] B. J. Kuipers. Qualitative Reasoning: Modeling and Simulation with Incomplete Knowledge. Cambridge, MA: MIT Press. 1994.

[9] Berant, J., Srikumar, V., Chen, P., Huang, B., Manning, C., Vander Linden, A., Harding, B., Clark, P., Modeling Biological Processes for Reading Comprehension. Proc. EMNLP 2014.

[10] Angelika Kimmig, Stephen H. Bach, Matthias Broecheler, Bert Huang, and Lise Getoor. 2012. A short introduction to Probabilistic Soft Logic. In Proceedings of NIPS Workshop on Probabilistic Programming: Foundations and Applications (NIPS Workshop-12).

[11] A. McCallum, E. Gabrilovich, R. Guha, K. Murphy. Proc. AAAI Symposium on Integrating Symbolic and Distributional Approaches, 2015 (to appear).

[12] Clark, P. Elementary School Science and Math Tests as a Driver for AI. Proc. IAAI'15, 2015. (To appear).