Composing Process Knowledge using Semantic Roles

Niranjan Balasubramanian

September 29, 2015

Contents

1 Introduction

Knowledge extraction and reasoning are central for many Artificial Intelligence systems. Recently introduced tasks such as *MCTest* reading comprehension challenge [24], grade-level science exams [6], and process comprehension tasks [2] serve as excellent benchmarks for developing reasoning-based question answering systems. These tasks test the ability of the systems to interpret and reason about scenarios and situations. Knowledge requirements analysis for grade-level science exams shows the need for deep inference supporting knowledge [8, 7]. Also, even with advanced state-of-the-art reasoning techniques, shallow text-derived knowledge is ineffective due to lexical and syntactic variations in language [14]. Scalable acquisition of deeper semantic knowledge is essential for effective reasoning in these tasks.

Much progress has been made on question answering involving simple facts [1, 9, 4, 23]. This is in large part due to the availability of large scale curated relational knowledge bases such as Freebase, coupled with significant advances in automatic relation extraction [20, 5, 26]. Similar advances in large scale inference-supporting knowledge is vital for making significant progress in reasoning-based QA tasks.

In this work, we focus on knowledge acquisition methods with grade science exams as an end application. In particular we focus on a subset of knowledge about physical, chemical, and other natural phenomena. The goal is to derive knowledge that allows effective reasoning about scenarios involving these phenomena. We refer to this as process knowledge. [1]

1.1 Motivation

We illustrate the need for inference supporting knowledge using the following example from a 4^{th} grade science exam.

A pot is heated on a stove. What causes the metal handle of the pot to become hot?
(A) conduction (B) convection (C) radiation (D) combustion.

The question tests reasoning about heat transfer mechanisms. To establish that conduction is the cause, a reasoning system must establish that there is some heat transfer happening through *direct contact*. We envision a system that first interpret the scenario and apply its knowledge about the heating process and the conduction process. In this case, the system needs to first identify *what is being heated* (the pot), and *what is the purported result* of the heating (the pot handle becoming hot). Then, knowledge about *heating* should allow it to conclude that the thing being heated (the pot) will become hot. Knowledge about *conduction* should suggest that anything *in direct contact* with a hot object (the pot handle) should also become hot. Combining these bits the system can verify that the described scenario indeed matches the conduction process. Replicating this type of reasoning requires knowledge about conduction and heating in a suitable computational representation. TODO: [Reads very rough.]

At a high level the main bits of required knowledge are: what is undergoing the process, what is the result, what is the main action etc. These bits of information are naturally encoded as semantic roles. PropBank [] and FrameNet [] are two of the highly successful efforts which provide this kind of semantic role based knowledge. They have spurred tremendous advances in automatic semantic role labeling and their applications []. While these resources provide exhaustive coverage for modeling general open-domain actions, they

[1] Our representation and acquisition methods will be targeted towards the grade science benchmarks but the methodology is general and can be applied to other tasks and domains.

do not cover knowledge about scientific processes. FrameNet, for instance, does not have entries for nearly half of the processes described in 4th grade science exams. The coverage is likely worse for higher grade levels with deeper knowledge domains.

1.2 Proposal

In response, we propose to build a knowledge base about physical, chemical, and biological processes from their textual descriptions.

Representation

Our primary motivation is to design a representation that effectively supports reasoning while also being amenable for automatic extraction. We propose a representation that includes pre-specified semantic roles that apply to many processes, as well as a set of process specific roles that are induced automatically.

We propose to compose knowledge about each process using a set of pre-determined vocabulary of semantic roles and derive additional new roles automatically as needed. While many competing role sets exist, we propose a role vocabulary that covers most inferential needs in the target application [19]. By leveraging a common vocabulary we expect to gain from shared learning for the same roles across different processes.

Role	Types	Instances	Patterns
Undergoer	Physical Object	solid, pot, …	when <x> is heated, <x> conducts heat, …
Enabler/Enabling Event	Physical Object	stove, flame, …	heated on the <x>, heated by <x>, …
Theme	Energy	heat, radiation, …	conducts <x>, transfers <x>, …
Output	-	-	-
Purpose/Consequence	-	maintain temperature, …	helps to <x>, …
Benefactive	-	-	-
Source	Physical Object	solid, pot, …	when <x> is heated, <x> conducts heat, …
Target	Physical Object	handle, vessel, …	becomes <x> hot, …
Medium	solid, contact	solid, contact, …	via <x>, in touch with <x>, …

Figure 1: An example entry for the process *thermal conduction*.

The process knowledge we target is quite diverse, some of which is definitional in nature – describing the key classes of entities and types of actions of a process. Others are specific to certain occurrence of the processes in specific situations or contexts – providing details on the specific entities and actions. To capture these diverse bits of information, we propose to associate the following information in the knowledge:

- Type information which encodes class level information about the fillers
- Most frequent role fillers for each role, augmented with internal structure where possible (e.g., noun-compound relations, prepositional relations).
- The extraction patterns associated with these fillers.
- Source sentences from which the knowledge was derived.

Figure 1 shows an example for the envisioned knowledge. The table includes a set of pre-specified general purpose roles such as *Undergoer*, *Enabler* and also a process specific role *Medium* that is automatically derived from inspecting sentences that describe the process.

Similar to FrameNet we propose to determine roles with respect to the semantics of the process rather than with respect to the specific verb (or predicate) that is used to describe the process. This canonicalization is desirable as it reduces reasoning time burden by eliminating need for steps that only establish equivalence of information realized via different predicates.

Iterative Role Acquisition and Refinement

General semantic role labeling task is challenging because of the lexical and syntactic variations in role realizations. Handling the variations requires learning from large amounts of training data, which is laborious and requires expert labor. Also as discussed earlier, existing semantic resources such as FrameNet or PropBank cannot be directly used for training as they do not cover the target concepts.

Our key premise is that to gather role knowledge we don't need a semantic role labeler that works well on all sentences. We are interested in role acquisition and not role interpretation. With the abundance and variety of information available on the web, we can target extraction from sentences that convey the same information in expected (easy to extract) constructs.

We propose an iterative role acquisition and expansion pipeline that includes the following:

- **Targeted Pattern-based Extraction** – We build a simple pattern-based local role extractor augmented with a classifier. Beginning with a set of manually constructed query patterns we search the web to find sentences that match these patterns. For instance "<process name> causes <x>" is a simple pattern that can be used to find the *result* role for a process. A simple classifier then assesses if the extraction is valid. The strong expectation of the type of role and where the filler is likely to be located allows us to design features that generalize better across different roles and processes thereby reducing need for large amounts of training data.

- **Joint Inference Across Sentences** – We propose a joint inference model that operates over multiple sentences to avoid errors in the local extraction. Despite the targeted acquisition, local sentence-level extraction alone is not adequate, because not all patterns are unambiguous. For example, "evaporates into <x>" extracts steam, a *result*, as well as atmosphere, a *location*. Because our goal is to acquire knowledge about a process, we turn the variety of expression on the web to our advantage. If the *atmosphere* were in fact a *result* then we might expect to see other sentences where it is expressed as a result with unambiguous patterns. Prior work proposed Integer Linear Program (ILP) based approach for joint inference for roles *within* sentences. We build upon this work to find an assignment of roles that maximizes the sum of extraction confidences, while also minimizing disagreements in labels for similar text spans in similar sentences.

- **Iterative Expansion and Refinement of Knowledge** – We propose a continuous iterative expansion and refinement of the knowledge. Determining the best set of knowledge bearing sentences *a priori* is extremely difficult. The quality and scope of the extracted knowledge is effectively determined by the set of sentences from which it were derived: Query patterns constrain the coverage for roles and pre-specified role vocabulary may miss critical process specific roles.

  An inspection of the gathered knowledge can provide valuable guidance in expanding and refining the knowledge. We propose to investigate methods to 1) assess the coverage of the roles, 2) induce new roles by identifying consistently repeated arguments that don't fit existing roles, 3) bootstrap extractors and query patterns using relevance feedback techniques, 4) refine the knowledge by organizing the knowledge in clusters of scenarios/instances.

Evaluation

The generated knowledge base will be evaluated for intrinsic quality and external utility. For internal evaluations we will perform manual evaluation of the resulting KB. As part of the evaluation, **we will also create a large scale curated KB, which relies on correcting the outputs from the system, rather than careful annotation of each sentence**. We will also generate a gold standard of the desired roles and role fillers for evaluating coverage. As an external evaluation we will use the knowledge base for answering process recognition questions. All the resources and evaluation test beds will be shared with the research community for further research.

1.3 Prior Work and Contributions

The motivation and direction for this proposal stems directly from our prior work on grade science exams. Our earlier work studied knowledge requirements [8], developed inference-supporting rule knowledge bases [7], and investigated sophisticated state-of-the-art probabilistic reasoning methods for QA [14]. Our ongoing preliminary work investigated the value of a handful of semantic roles in answering process recognition questions and identified the knowledge, representation, and reasoning gaps [19]. This proposal aims to address the central challenges that we've identified through these prior works.

Upon successful completion this project would have made the following contributions:

- **Semantic Resource for Grade-level Science** – The knowledge base we build will cover process in grade-level science and will be made available to the public and the academic community for fostering research in automatic reasoning systems. To foster further easy exploration, all the code will be open sourced and we will also host a web service that will dynamically compose the knowledge for new unseen processes.
- **Targeted Acquisition and Local Extraction** – A targeted acquisition method that gathers sentences expressing roles, and a role classification formulation that allows use of features with better generalization.
- **Joint Inference Across Sentences** – A joint inference method that reconciles roles across different sentences.
- **Continuous Expansion and Refinement** – Methods for assessing expanding and refining knowledge based on the current state of the knowledge.

References

[1] Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. Semantic parsing on freebase from question-answer pairs. In *EMNLP*, pages 1533–1544, 2013.

[2] Jonathan Berant, Vivek Srikumar, Pei-Chun Chen, Abby Vander Linden, Brittany Harding, Brad Huang, Peter Clark, and Christopher D. Manning. Modeling biological processes for reading comprehension. In *Proceedings of EMNLP*, 2014.

[3] Anders Björkelund, Love Hafdell, and Pierre Nugues. Multilingual semantic role labeling. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task*, pages 43–48. Association for Computational Linguistics, 2009.

[4] Antoine Bordes, Jason Weston, and Nicolas Usunier. Open question answering with weakly supervised embedding models. In *Machine Learning and Knowledge Discovery in Databases*, pages 165–180. Springer, 2014.

[5] Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R Hruschka Jr, and Tom M Mitchell. Toward an architecture for never-ending language learning. In *AAAI*, volume 5, page 3, 2010.

[6] Peter Clark. Elementary school science and math tests as a driver for ai: Take the aristo challenge. *to appear*, 2015.

[7] Peter Clark, Niranjan Balasubramanian, Sumithra Bhakthavatsalam, Kevin Humphreys, Jesse Kinkead, Ashish Sabharwal, and Oyvind Tafjord. Automatic construction of inference-supporting knowledge bases. In *AKBC2014*, Montreal, Canada, December 2014.

[8] Peter Clark, Phil Harrison, and Niranjan Balasubramanian. A study of the AKBC/requirements for passing an elementary science test. In *Proc. of the AKBC-WEKEX workshop at CIKM*, 2013.

[9] Anthony Fader, Luke Zettlemoyer, and Oren Etzioni. Open question answering over curated and extracted knowledge bases. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1156–1165. ACM, 2014.

[10] William R Foland Jr and James H Martin. Dependency-based semantic role labeling using convolutional neural networks. In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*, pages 279–288, 2015.

[11] Hagen Fürstenau and Mirella Lapata. Graph alignment for semi-supervised semantic role labeling. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 11–20, Singapore, 2009.

[12] Hagen Fürstenau and Mirella Lapata. Semi-supervised semantic role labeling via structural alignment. *Computational Linguistics*, 38(1):135–171, 2012.

[13] Daniel Gildea and Daniel Jurafsky. Automatic labeling of semantic roles. *Computational linguistics*, 28(3):245–288, 2002.

[14] Tushar Khot, Niranjan Balasubramanian, Eric Gribkoff, Ashish Sabharwal, Peter Clark, and Oren Etzioni. Exploring markov logic networks for question answering. In *Empirical Methods in Natural Language Processing*, 2015.

[15] Ivan Titov Alexandre Klementiev. Semi-supervised semantic role labeling: Approaching from an unsupervised perspective. In *Proceedings of the COLING Conference.*, 2012.

[16] Peter Koomen, Vasin Punyakanok, Dan Roth, and Wen-tau Yih. Generalized inference with multiple semantic role labeling systems. In *Proceedings of the Ninth Conference on Computational Natural Language Learning*, pages 181–184. Association for Computational Linguistics, 2005.

[17] Joel Lang and Mirella Lapata. Unsupervised induction of semantic roles. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 939–947, Los Angeles, California, June 2010. Association for Computational Linguistics.

[18] Joel Lang and Mirella Lapata. Unsupervised semantic role induction with graph partitioning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1320–1331, Edinburgh, Scotland, UK., July 2011. Association for Computational Linguistics.

[19] Samuel Louvan, Chetan Naik, Veronica Lynn, Ankit Arun, Niranjan Balasubramanian, and Peter Clark. Semantic role labeling for process recognition questions. In *K-CAP Scientific Knowledge Workshop*, 2015.

[20] Mausam, Michael Schmitz, Robert Bart, Stephen Soderland, and Oren Etzioni. 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, 2012.

[21] Luiz Augusto Pizzato and Diego Mollá. Indexing on semantic roles for question answering. In *Coling 2008: Proceedings of the 2nd workshop on Information Retrieval for Question Answering*, pages 74–81. Association for Computational Linguistics, 2008.

[22] Vasin Punyakanok, Dan Roth, Wen-tau Yih, and Dav Zimak. Semantic role labeling via integer linear programming inference. In *Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics.

[23] Siva Reddy, Mirella Lapata, and Mark Steedman. Large-scale semantic parsing without question-answer pairs. *Transactions of the Association for Computational Linguistics*, 2:377–392, 2014.

[24] Matthew Richardson, Christopher JC Burges, and Erin Renshaw. Mctest: A challenge dataset for the open-domain machine comprehension of text. In *EMNLP*, volume 1, page 2, 2013.

[25] Dan Shen and Mirella Lapata. Using semantic roles to improve question answering. In *EMNLP-CoNLL*, pages 12–21, 2007.

[26] Fabian M Suchanek, Gjergji Kasneci, and Gerhard Weikum. Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web*, pages 697–706. ACM, 2007.

[27] Oscar Täckström, Kuzman Ganchev, and Dipanjan Das. Efficient inference and structured learning for semantic role labeling. *Transactions of the Association for Computational Linguistics*, 3:29–41, 2015.