

Composing Representations for Processes using Semantic Roles

Niranjan Balasubramanian, Stony Brook University

September 10, 2015

Knowledge-intensive benchmarks including standardized grade-level science exams [1], textual entailment [4], and reading comprehension tasks [7] have renewed interests in automatic knowledge extraction and reasoning. These tasks motivate reasoning-based systems that go beyond simple retrieval and syntactic structure matching [2, 3, 5].

Effective semantic representations of relevant information is critical for making significant advances in building such systems. Semantic role-based representations have shown promise for open-domain factoid question answering [8, 6]. Some of the knowledge required for grade-level science exams are naturally expressed via semantic roles. Consider the following example from an actual 4th grade science exam.

*When plants use stored sugar for energy, they go through a process called
(A) photosynthesis (B) transpiration (C) respiration (D) perspiration.*

Photosynthesis and *respiration* both involve sugar and energy. Photosynthesis converts light energy to sugar, whereas (cellular) respiration releases energy in the sugar by breaking it down. Not surprisingly these processes are described using similar words, which makes bag-of-words style reasoning unreliable. Understanding the different roles that energy and sugar play in these processes is key to effective reasoning. Such semantic roles are critical for performing chained automated reasoning.

Several existing resources provide semantic representations of general open-domain actions e.g., FrameNet, PropBank, and VerbNet. However, these resources do not adequately cover many of the 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 to be worse for higher grade levels.

We propose to build a knowledge base about physical, chemical, and biological processes from their textual descriptions. The central idea is to automatically compose a semantic representation using a pre-determined vocabulary of semantic roles. Having identified the roles of interest, we will seek out sentences that express these roles and build extractors for these roles via bootstrapping. We will conduct intrinsic and external evaluations. We will create a manual target representation for intrinsic evaluation, and use the 4th grade science questions as an external application.

Intellectual Merit

This work investigates a new direction in acquiring semantic roles by targeting sentences that express information in expected ways. The primary contribution of this work will include development of joint alignment and collective labeling mechanisms for semantic role labeling. A secondary contribution is the development of methods for automatically identifying roles that apply for a given process. The advances will result in new ways to leverage unlabeled data and balance representational needs and extraction capabilities.

Broader Impact

Automatic knowledge extraction is fundamental to advancing Artificial Intelligence. This work contributes towards building systems capable of understanding knowledge in texts and reason with them. Better reasoning systems help information access and reduce information overload thereby accelerating research and discovery, as well as serve the information needs of the population at large.

Keywords Knowledge extraction, Information extraction, Event extraction, Summarization

References

- [1] Peter Clark. Elementary school science and math tests as a driver for ai: Take the aristo challenge. *to appear*, 2015.
- [2] Peter Clark, Niranjana Balasubramanian, Sumithra Bhakthavatsalam, Kevin Humphreys, Jesse Kinkead, Ashish Sabharwal, and Oyvind Tafjord. Automatic construction of inference-supporting knowledge bases. Montreal, Canada, December 2014.
- [3] Peter Clark, Phil Harrison, and Niranjana Balasubramanian. A study of the AKBC/requirements for passing an elementary science test. In *Proc. of the AKBC-WEKEX workshop at CIKM*, 2013.
- [4] Ido Dagan, Bill Dolan, Bernardo Magnini, and Dan Roth. Recognizing textual entailment: Rational, evaluation and approaches. *Natural Language Engineering*, 16(01):105–105, 2010.
- [5] Tushar Khot, Niranjana 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.
- [6] 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.
- [7] 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.
- [8] Dan Shen and Mirella Lapata. Using semantic roles to improve question answering. In *EMNLP-CoNLL*, pages 12–21, 2007.