Extraction of causal structure from procedural text for discourse representations

Michael Regan[1,2], James Pustejovsky[3], and William Croft[1]

[1]Department of Linguistics, University of New Mexico
[2]Department of Computer Science, University of Colorado Boulder
[3]Department of Computer Science, Brandeis University

The extraction of causal structure from scientific text is an important step in automating deep semantic analyses of synthesis procedures. Our objectives are (1) to leverage a finite, expressive set of semantic labels as a high-level representation of event decompositional causal structure, (2) to extract subevent semantics with a reliable signal from surface features of text using limited annotated data, and (3) to model discourse structure as subevent representation of participant interactions for analysis and inference.

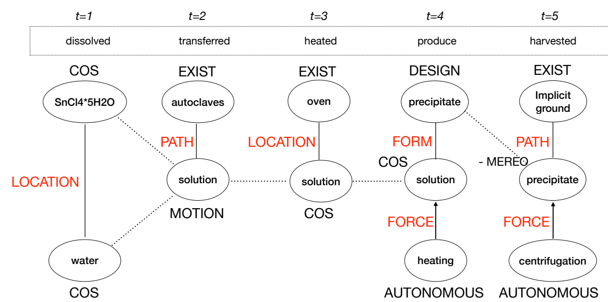

Figure 1: Storyline for: "2.0 g **SnCl4*5H2O** was first dissolved in 100 mL deionized **water**. The resulting **solution** was then transferred to two 100 mL stainless steel **autoclaves** and heated in an **oven** at 120 degC for 28 h to produce a **precipitate**, which was harvested by **centrifugation**." (Adapted from [Zhou et al., 2013]). Black CAPS indicate states or external agent (e.g., COS=change-of-state) with RED causal/non-causal relations left of links (e.g., FORCE). Dotted lines=coreference.

Our model of event decompositional structure is based on the theory of *force dynamics* [Croft, 2012, Talmy, 1988] and the claim that the meaning of syntactic form is causal in nature. Events are decomposed into aspectual, qualitative state, and causal dimensions to model change over time, representing directly participant interactions. From analyses of subevent structure [Croft et al., 2016], we have extended our work to entity-centered discourse representations [Croft et al., 2017, 2020] based on metro map models [van Erp et al., 2014] as in Fig. 1.

The extraction mechanism consists minimally of these steps: identify participants of each event (e.g., predicate-argument structure), classify causal and non-causal relations between participants, classify entity qualitative state changes (or no change), and infer entity coreference links (incl. set/member and part/whole relations). The semantic classification tasks depend largely on a survey of English language data, cross-linguistic analyses, and recent experiments using transfer learning that provide evidence of the highly predictive mapping between surface syntax and causal, force-dynamic meaning.

We examine the utility of this representation to support AI reasoning and causal inference [Peters et al., 2017]. A richer representation will incorporate temporal relations [Pustejovsky et al., 2003], event modality (e.g. actual, hypothetical) [O'Gorman et al., 2016], and implicit arguments [O'Gorman et al., 2018]. In future work, we will test our representation as a bias to build process graphs for material synthesis procedures [Mysore et al., 2019] and extend construction mappings for greater domain generalizability.

References

William Croft. *Verbs: Aspect and causal structure.* Oxford University Press, 2012.

William Croft, Pavlina Peskova, and Michael Regan. Annotation of causal and aspectual structure of events in RED: a preliminary report. *4th Workshop on Events, ACL*, 2016.

William Croft, Pavlina Peskova, and Michael Regan. Integrating Decompositional Event Structures into Storylines. *Events and Stories in the News Workshop, ACL*, 2017.

William Croft, Pavlina Kalm, and Michael Regan. Decomposing events and storylines. In Tommaso Caselli, Martha Palmer, Eduard Hovy, and Piek Vossen, editors, *Events and Stories.* Cambridge University Press, 2020.

Sheshera Mysore, Zach Jensen, Edward Kim, Kevin Huang, Haw-Shiuan Chang, Emma Strubell, Jeffrey Flanigan, Andrew McCallum, and Elsa Olivetti. The Materials Science Procedural Text Corpus: Annotating Materials Synthesis Procedures with Shallow Semantic Structures. *Linguistic Annotation Workshop, ACL*, 2019.

Tim O'Gorman, Kristin Wright-Bettner, and Martha Palmer. Richer Event Description: Integrating event coreference with temporal, causal and bridging annotation. *2nd Workshop on Computing News Storylines, ACL*, 2016.

Tim O'Gorman, Michael Regan, Kira Griffit, Ulf Hermjakob, Kevin Knight, and Martha Palmer. AMR Beyond the Sentence: the Multi-sentence AMR corpus. *COLING*, 2018.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms.* The MIT Press, 2017. ISBN 0262037319.

James Pustejovsky, Jose Castano, Robert Ingria, Roser Sauri, Robert Gaizauskas, Andrea Setzer, and Graham Katz. TimeML: Robust Specification of Event and Temporal Expressions in Text. *New directions in question answering*, 2003.

Leonard Talmy. Force dynamics in language and cognition. *Cognitive Science*, 2, 1988.

Marieke van Erp, Gleb Satyukov, Piek Vossen, and Marit Nijssen. Discovering and visualizing stories in the news. *LREC*, 2014.

Xiaosi Zhou, Li-Jun Wan, and Yu-Guo Guo. Binding sno 2 nanocrystals in nitrogen-doped graphene sheets as anode materials for lithium-ion batteries. *Advanced Materials*, 2013.