Automatically Extracting Action Graphs from Materials Science Synthesis Procedures

Sheshera Mysore[1], Edward Kim[2], **Emma Strubell**[1], Ao Liu[1], Haw-Shiuan Chang[1], Srikrishna Kompella[1], Kevin Huang[2], Andrew McCallum[1], Elsa Olivetti[2]

UMassAmherst College of Information and Computer Sciences

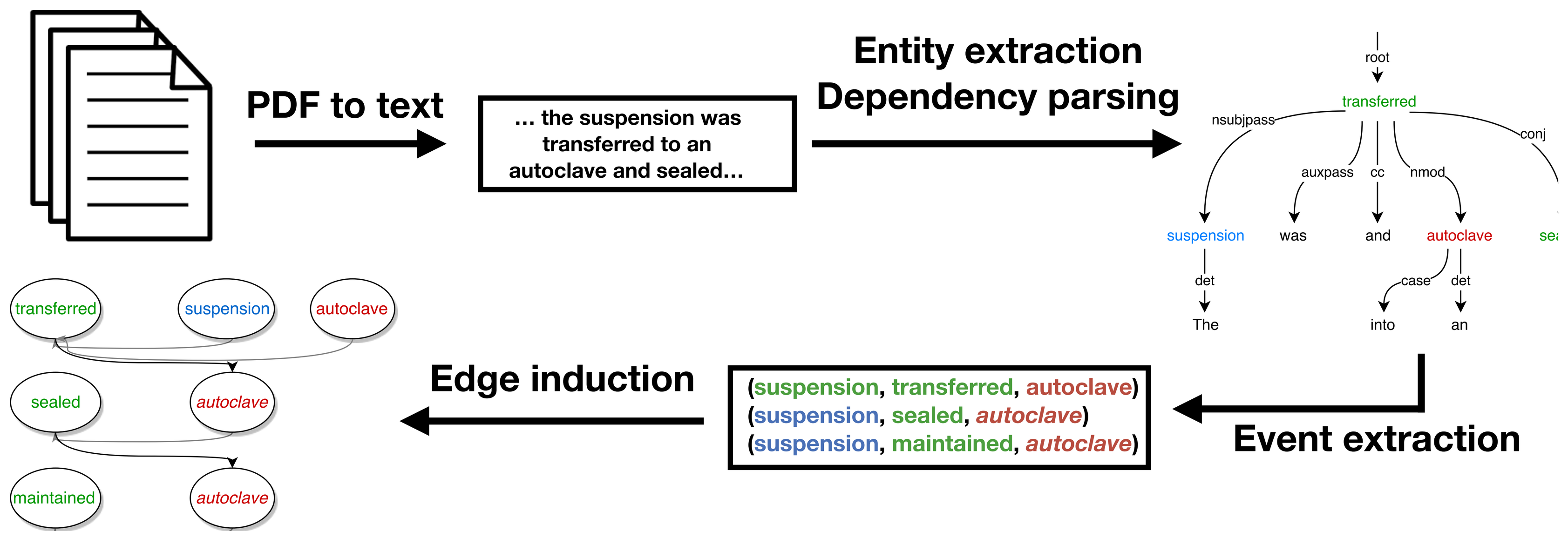
[1]College of Information and Computer Sciences, University of Massachusetts Amherst
[2]Department of Materials Science and Engineering, Massachusetts Institute of Technology

Massachusetts Institute of Technology

Summary

- Want to analyze, predict inorganic materials synthesis procedures at large scale.
- Unlike organic synthesis, inorganic procedures exist only as natural language narratives in scientific journal articles.
- We automatically extract structured representations of synthesis procedures from materials science article text using a pipeline of supervised and unsupervised NLP methods.

Overall extraction pipeline architecture



Entity extraction

Let $x = [x_1, \ldots, x_T]$ be a sentence of input text and $y = [y_1, \ldots, y_T]$ be per-token output tags. We predict the most likely y, given a conditional model $P(y \mid x)$. We experiment with two factorizations of $P(y \mid x)$. First:

$$P(y|x) = \prod_{t=1}^{T} P(y_t|F(x)), \quad (1)$$

tags are conditionally independent given features $F(x)$, encoded by a deep neural network (DCNN, Bi-LSTM). Second:

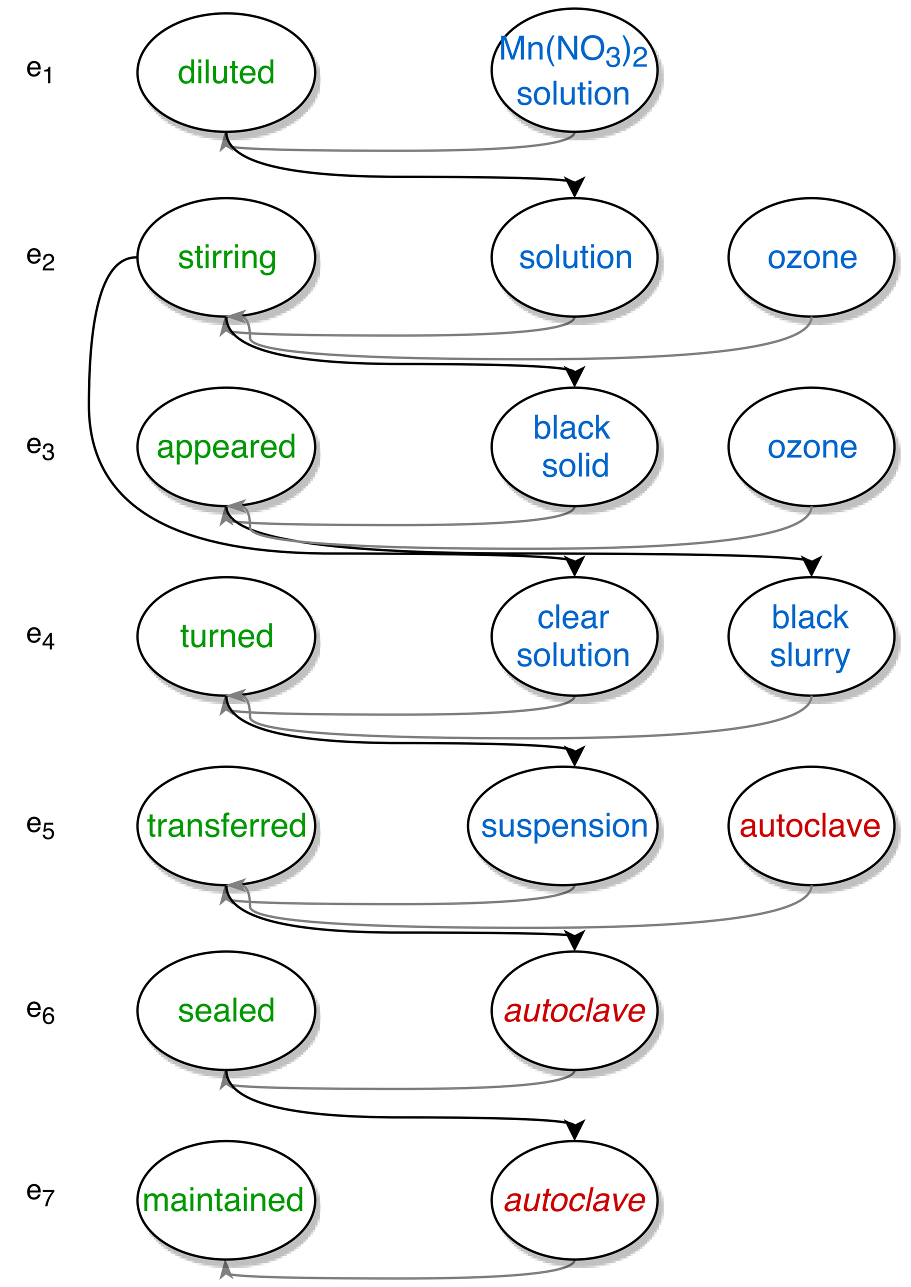
$$P(y|x) = \frac{1}{Z_x} \prod_{t=1}^{T} \psi_t(y_t|F(x))\psi_p(y_t, y_{t-1}), \quad (2)$$

a linear-chain CRF that couples all of y together, enforcing constraints between labels during prediction w/ local factor ψ_t, pairwise factor fψ_p, partition function Z_x (Lafferty et al., 2001). $F(x)$ is parameterized by a neural network (Bi-LSTM-CRF) or a log-linear model over sparse binary features (CRF-ling, CRF-hand, CRF-both) indicating e.g. lexicon membership, part-of-speech.

Example action graph

Typical synthesis procedure text

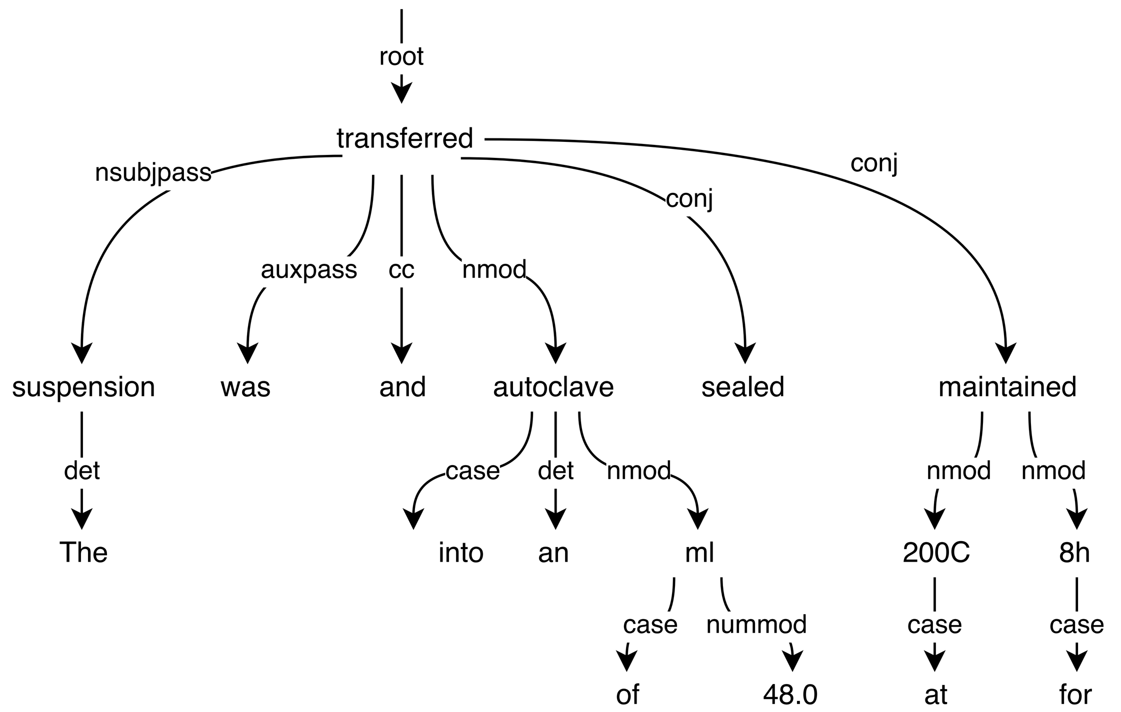
In a typical procedure for the synthesis of β-MnO_2 nanowires, 2.5 mL of 50 wt.% $Mn(NO_3)_2$ solution was diluted to 25.0 mL, and ozone was fed into the bottom of the solution for 30 min under vigorous stirring. With the indraught of ozone, black solid appeared gradually and the clear solution turned into black slurry finally. Then the suspension was transferred into an autoclave of 48.0 mL, sealed and maintained at 200 °C for 8 h. After this, the autoclave was cooled to room temperature naturally. The resulting solid products were washed with water, and dried at 120 °C for 8 h.



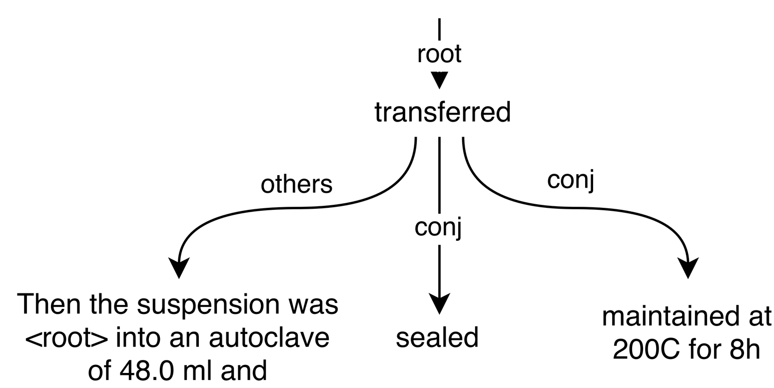
Event extraction

We extract events from sentences by applying heuristic rules on the syntactic dependency parse of the sentence. The most important of these splits every phrase whose head links to the root with a `conj` relation. All other tokens are associated with the root and constitute the main phrase. Each split phrase and the main phrase is considered an event.

Example dependency parse:



Post-processed parse representing 3 events:



Edge induction

We consider two methods for edge induction: Our baseline (Seq) simply attaches events in the order they occur in the text; Our generative probabilistic model (Prob) is based on that of Kiddon et al. (2015), which uses hard EM to learn an attachment model given strong priors based on parse tree structure and typical attachment in procedural text.

Experimental results

Entity extraction:

Model	Precision	Recall	F1
CRF-ling	76.98	67.41	71.88
CRF-hand	75.59	69.32	72.48
CRF-both	74.97	72.12	73.52
DCNN	**77.85**	77.16	77.50
Bi-LSTM	74.25	77.83	76.00
Bi-LSTM-CRF	74.64	**80.74**	**77.57**

Action graph extraction:

Setting 1: Ignore edges between unaligned events.

Model	Aligned	Unaligned	Precision	Recall	F1
End-to-end evaluation					
Seq	39.85%	30.95%	**73.04**	**94.38**	**82.35**
Prob	39.85%	30.95%	68.38	89.89	77.67
Perfect node segmentation					
Seq	63.80%	0%	**99.29**	**99.29**	**99.29**
Prob	63.80%	0%	95.36	95.36	95.36

Setting 2: Penalize edges between unaligned events.

Model	Aligned	Unaligned	Precision	Recall	F1
End-to-end evaluation					
Seq	39.85%	30.95%	**27.10**	**27.91**	**27.50**
Prob	39.85%	30.95%	25.81	26.58	26.19
Perfect node segmentation					
Seq	63.80%	0%	**99.29**	**92.36**	**95.70**
Prob	63.80%	0%	95.36	88.70	91.91