CIS-620

Spring 2021

Learning in Few-Labels Settings

Dan Roth

Computer and Information Science

University of Pennsylvania

Meeting # 6

2/29/21

Admin

Questions?

- Please follow the schedule on the web site
  - You should know when you are scheduled to present
  - And discuss.
  - Please send your presentation before Friday, and your questions/bullets by Sunday.
- Note that the presentations are not independent.
  - Things that we have mentioned in earlier meetings are relevant to later papers. It would be nice if you can make the connections.
- Please follow the presentation guidelines
- Late policy (for critical surveys and project repots; not for presentations)
  - 4 Days (96 hours).
- Please note the guidelines for Project #1.
  - Choose what you want to change and announce/discuss on Piazza. (By March 8)
  - Presentation and final report (March 15)

Presentations:

- Please read the **guidelines.**
- Do not **cut-and-paste** the paper to the slides.
  - Not everything should be presented.
  - The order of the paper may not be the right order for a presentation.
- When you read the paper:
  - You can **go back and forth** to check things (notation, details, math).
  - You can consult outside resources if needed.
- **Your audience cannot do it.**
  - Your job as the presented is to teach your students the paper despite this limitation.
- Think about what you need to do.
- **Experiments:** Just putting a table on the slide is not useful. Instead, discuss:
  - What is the goal of this experiment.
  - How do the results in the table achieve it (or not)
  - You don't need to show all the results

- So far, I've given very long list of comments to all of you.
- My goal is that you will learn from earlier presentations, so that I will not need to do it…

Today’s Papers

- Zero/Few-Shot Learning
  - Few-Shot Text Classification with Distributional Signatures (Chaitanya Malaviya)
- Semi-Supervised; Constrained Driven Learning
  - Constrained semi-supervised learning using attributes and comparative attributes (Rahul Shekhar)
- Weak Supervision
  - A Discrete Hard EM Approach for Weakly Supervised Question Answering (Venkata Sai Nikhil Thodupunuri)
- Partial Supervision
  - Sentiment Tagging with Partial Labels using Modular Architectures (Lishuo Pan)

Constraints Driven Learning

Constrained Conditional Models [Cheng et al.'07; Chang et al.'12]

ILP Formulation

$$y = \text{argmax}_y \sum \mathbf{1}_{\phi(x,\,y)}\, \mathbf{w}_{x,y} \quad \text{subject to Constraints } C(x,y)$$

Penalty for violating the constraints.

Formulation goes back to (Roth & Yih 2004). Also related to PR (Ganchev et al. 2010)

Knowledge component: (Soft) constraints
E.g., given x, y has even # of 1's

A linear function over models – can be used to model any logical function

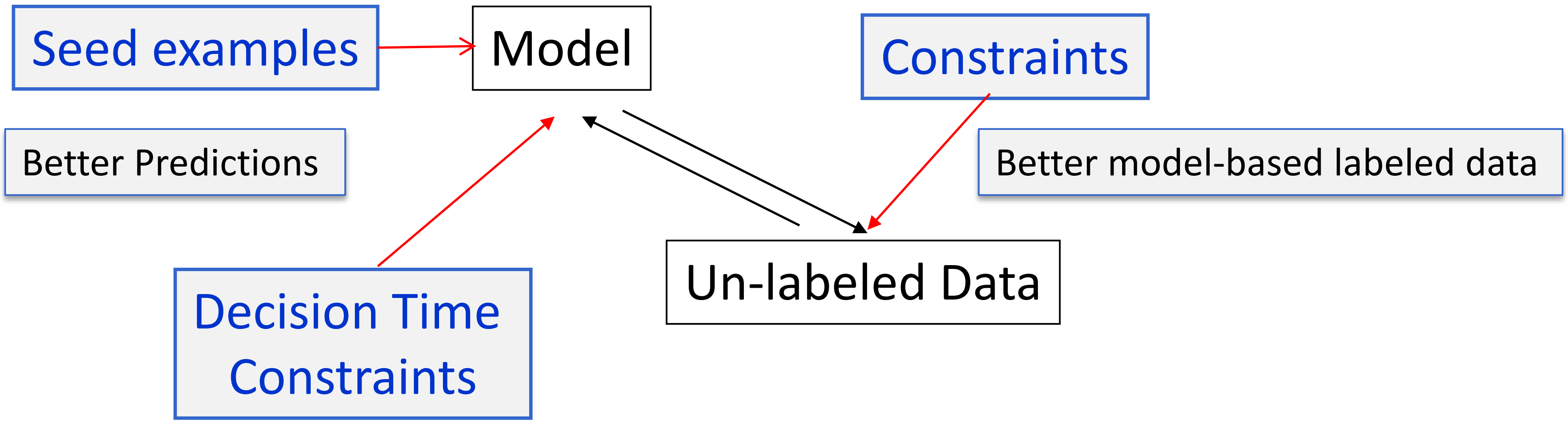
Features, Models, NN (non-linearity comes here)

How far are the decisions (y) is from a "legal/expected" assignment

- This objective can give rise to multiple learning paradigms.
  - It is mostly used in the structured learning paradigm, since the constraints relate the assignments (predictions) of groups of output variables.
- One common usage is in the semi-supervised learning setting
  - Start with a small set of labeled data
  - Incorporate un-annotated data that is annotated by the model
    - But this could be very noisy, and cause a model to drift away
  - "Fix" the examples annotated by the model using constraints.

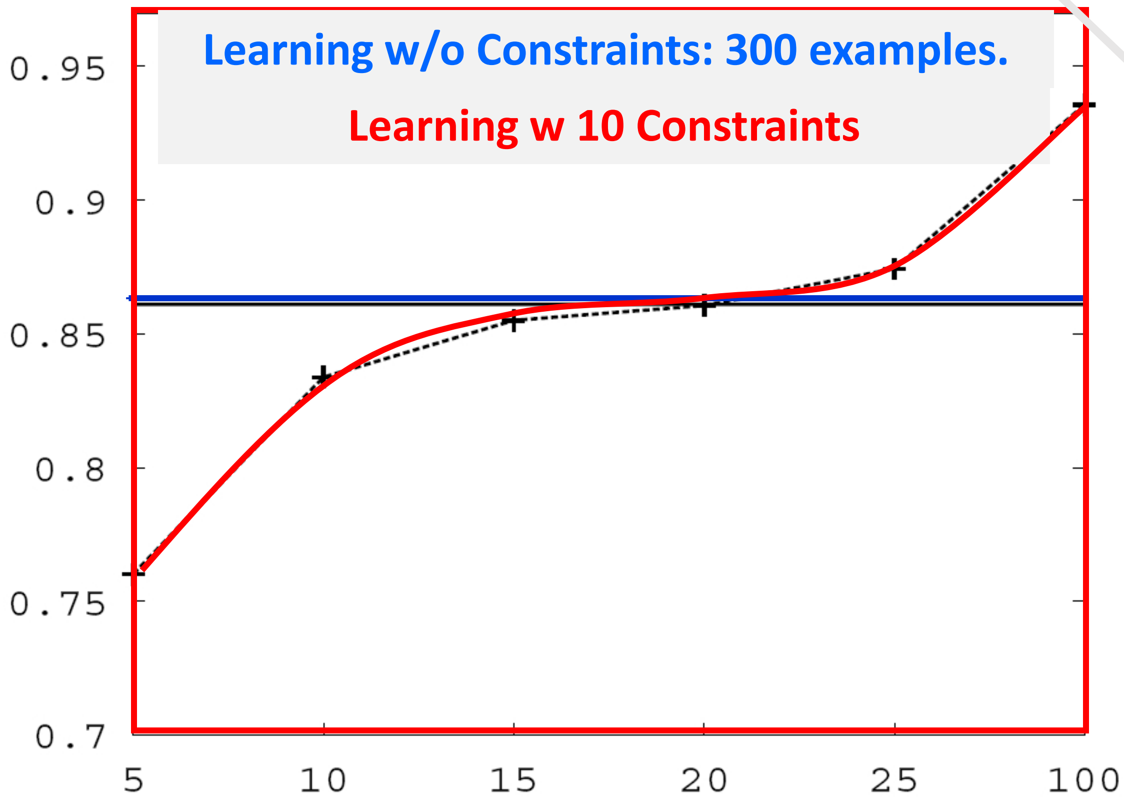
Guiding (Semi-Supervised) Learning with Constraints

- In traditional Semi-Supervised learning the model can drift away from the correct one.
- Constraints can be used to generate better training data
  - At training to improve labeling of un-labeled data (and thus improve the model)
  - At decision time, to bias the objective function towards favoring constraint satisfaction.



Value of Constraints in Semi-Supervised Learning

Objective function: $f_{\Phi,C}(\mathbf{x},\mathbf{y}) = \sum w_i \phi_i(\mathbf{x},\mathbf{y}) - \sum \rho_i d_{C_i}(\mathbf{x},\mathbf{y}).$



# of available labeled examples

Constraints are used to Bootstrap a semi-supervised learner

Poor model + constraints used to annotate unlabeled data, which in turn is used to keep training the model.

Constraints Driven Learning (CoDL)

Archetypical Semi/un-supervised learning: **A constrained EM**

[Chang, Ratinov, Roth, ACL'07;ICML'08,MLJ'12]
See also: Ganchev et. al. 10 (PR)

(w,u)=learn(L)

For N iterations do

$T=\phi$

For each x in unlabeled dataset

$\mathbf{h \leftarrow argmax_{y \in Y}\ w^T\emptyset(x, y) + u^T C(x, y)}$

$T=T \cup \{(x, h)\}$

$(w,u) = \gamma\ (w,u) + (1-\gamma)\ \text{learn}(T)$

Supervised learning algorithm parameterized by (w,u).
(w,u) are latent variables

Inference with constraints:
(use the constraints to "correct" predictions)
Then augment the training set

Learn from new training data
Weigh supervised & unsupervised models.

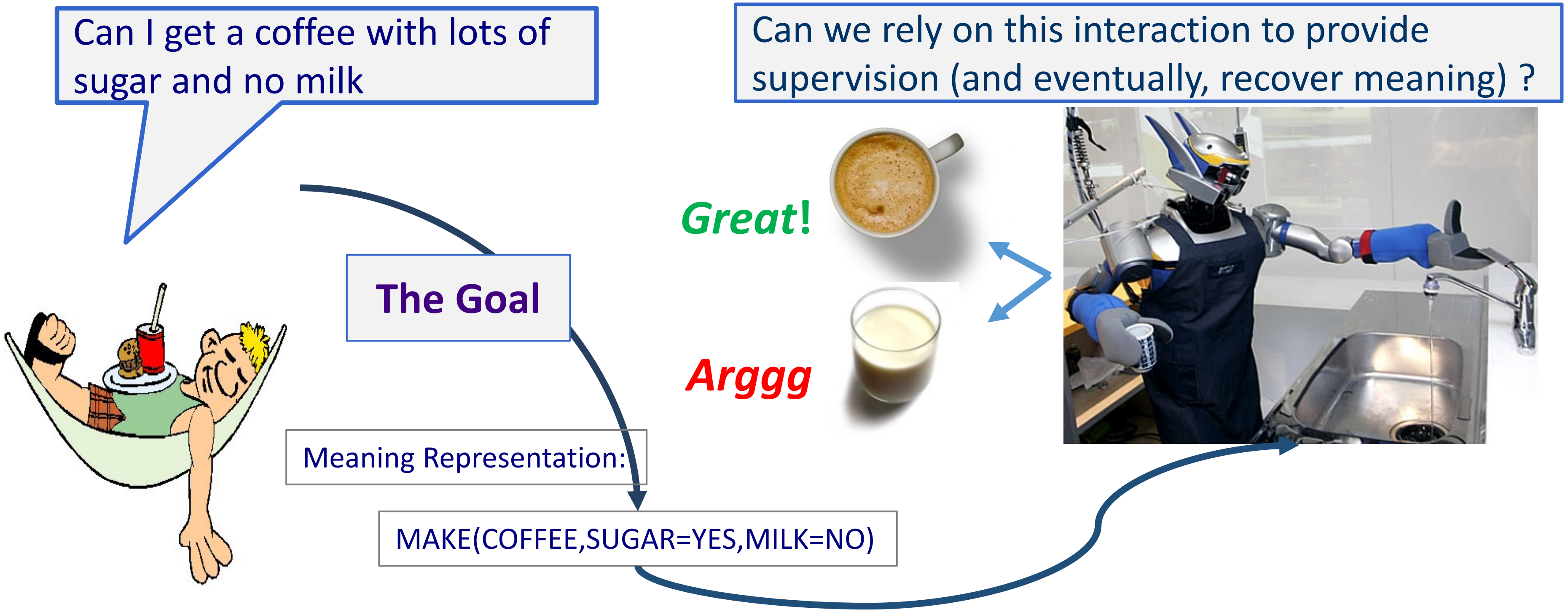
Notice that this is a (constrained) EM algorithm .

Today’s Papers

- Semi-Supervised; Constrained Driven Learning
  - Constrained semi-supervised learning using attributes and comparative attributes (Rahul Shekhar)

Response-Driven Learning / End-Supervision

Understanding Language Requires (some) Feedback



- How to recover meaning from text?
- Standard "example based" ML: annotate text with meaning representation
  - The teacher needs deep understanding of the **agent** ; not scalable.
- Response Driven Learning (current name: learning from denotation): Exploit indirect signals in the interaction between the learner and the teacher/environment
- [A lot of work in this direction, following **Clarke et al. CoNLL'10: Driving Semantic Parsing from the World's Response]**

Response Based Learning

- We want to learn a model that transforms a natural language sentence to some meaning representation.

English Sentence ⟹ **Model** ⟹ Meaning Representation

- Instead of training with (Sentence, Meaning Representation) pairs
- Think about/invent **behavioral derivative(s)** of the models outputs
  - Supervise the derivatives (easy!) and
  - Propagate it to learn the complex, structured, transformation model

Geoquery with Response based Learning

- We want to learn a model to transform a natural language sentence to some formal representation.

English Sentence → **Model** → Meaning Representation

What is the largest state that borders NY?

largest(state(next_to(const(NY))))

Query a GeoQuery Database. ← Simple derivatives of the model's outputs

The key challenge is computational. The space of possible semantic parses is huge. Approaches focused on trying to constrain this space.

- "Guess" a semantic parse. **Is [DB response == Expected response] ?**
  - ☐ **Expected**: Pennsylvania **DB Returns**: Pennsylvania→ **Positive Response**
  - ☐ **Expected**: Pennsylvania **DB Returns**: NYC, or ???? → **Negative Response**

If the response is "yes", it could still be so for the wrong reason, despite the semantic parse being wrong.

If the response is "no", the semantic parse must be wrong; how to supervise?

Key Challenges: Summary

- The **response** may not completely define the **intermediate representation**
  - But the supervision is really dictated by the intermediate representation
  - Consequently, the supervision is not completely defined
    - If the response is correct – it may not completely define the supervison
    - If the response is incorrect – credit (blame) assignment is still a problem
- The space of intermediate representation is very large
  - Computational issues
  - Constraints on intermediate representations should be used
- These issues are only beginning to be addressed now in the literature of End-Supervision, but were studied a lot earlier under **learning with latent representations**
  - Structured Output Learning with Indirect Supervision
    Ming-Wei Chang and Vivek Srikumar and Dan Goldwasser and Dan Roth, *ICML – 2010*
  - Discriminative Learning over Constrained Latent Representations
    Ming-Wei Chang and Dan Goldwasser and Dan Roth and Vivek Srikumar, *NAACL - 2010*

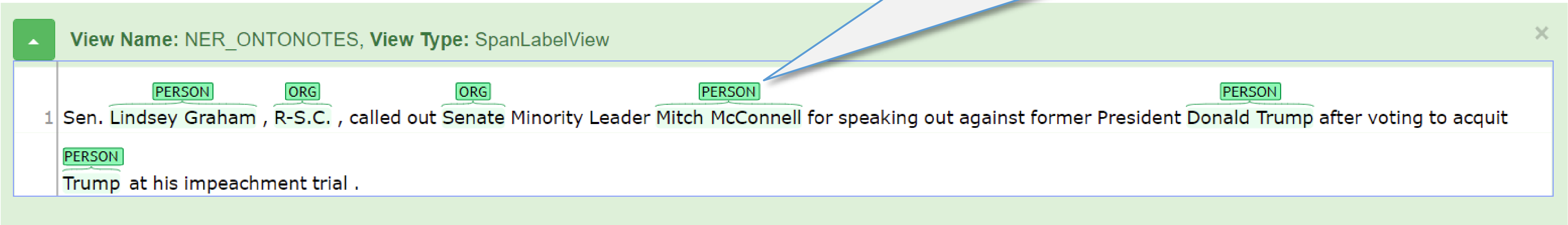
Today's Papers

- Weak Supervision
  - A Discrete Hard EM Approach for Weakly Supervised Question Answering (Venkata Sai Nikhil Thodupunuri)

Partial Supervision

- Exhaustive Annotation is often unrealistic Labels parts of sentences
  - Do we need it?

In most cases the annotation is at the token level, using the BIO convention, which has 1-1 mapping to the annotation below.
....[Leader O] [Mitch B-PER] [McConnell I-PER] [for-O]...



- E.g., what happens if the annotation only consists of:
  - [Mitch McConnell == PER; Lindsey Graham == PER]
- CoDL has been used in this context too:
  - Mayhew et al. Named Entity Recognition with Partially Annotated Training Data *CoNLL* (2019)

Today's Papers

- Partial Supervision
  - Sentiment Tagging with Partial Labels using Modular Architectures (Lishuo Pan)