CIS-620

Spring 2021

Learning in Few-Labels Settings

Dan Roth

Computer and Information Science

University of Pennsylvania

Meeting # 3

2/8/21

Admin

- If you haven't selected a paper to present, please do so.
  - Papers for the 15th are on the spreadsheet. Please send me your draft presentation soon; no later than Friday night.
- Recall that you need to be a discussant on two papers.
  - Please send your questions/bullets by Sunday.
- Please follow the presentation guidelines
- Please follow the schedule on the website:
  - February 15:
  - Select a paper to reproduce
    - Reproduction papers will be released today.
  - First Critical Survey due
    - Guidelines will be released later this week
    - Do not survey papers that were already presented in class.

Today's Papers

- Zero-Shot Learning
  - [Zero-Shot Relation Extraction via Reading Comprehension](#) (Kevin Xie)
- Incidental Signals
  - Learning Dependency-Based Compositional Semantics (Krunal Shah)
- Knowledge as Supervision
  - A Logic-Driven Framework for Consistency of Neural Models (Jiayao Zhang)
- Zero-Shot + Knowledge
  - Zero-shot Learning of Classifiers from Natural Language Quantification (Young-Min Cho)

Zero-Shot

Zero-Shot Learning

- Protocols: Multiple protocols are referred to as Zero-Shot in the literature.
- Assume that we are talking about multi-class classification
  1. The decision model has not seen any task-specific training examples
  2. The decision model has been trained on *some* of the labels and needs to predict also on unseen labels.
  - [Yin et al. EMNLP’19] called these protocols: *label-fully-unseen* and *label-partially-unseen*
- Methods:
  1. Representation-based: examples & labels are mapped to a common semantic space
     - Sparse representations or Dense representations
  2. Transfer: a model that was trained on decision task T is being used (via some mapping) to support decisions on task T’.
     - Typically, transfer is done from Textual Entailment or Questions Answering
  3. Learning from definitions (or other external sources)

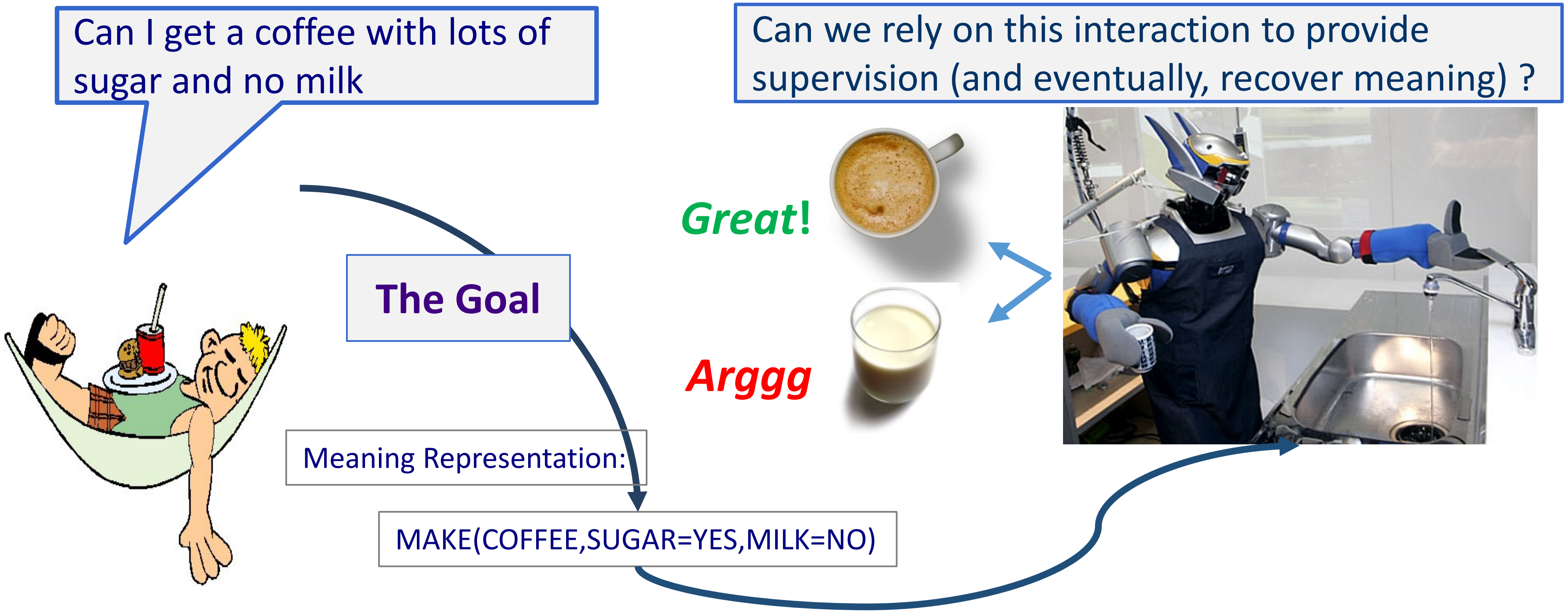
Zero-Shot Paper

- Zero-Shot Relation Extraction via Reading Comprehension (Kevin Xie)
  - Transfer learning for relation extraction.
  - Note that the standard relation extraction is defined as:
    - **Input:** Sentence, $mention_1$, $mention_2$, taxonomy of relations $\{R_1, R_2, \ldots R_k\}$ (includes a no-relation)
    - **Learn** a model that maps the mention pair into one of the relations R_i
  - **Example:** *Sanders' wife* is a native of *North Carolina* ➔ (born_in (sander's wide, NC)

Incidental Signals

Learning from Responses

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?

Incidental Supervision Paper

- [Learning Dependency-Based Compositional Semantics](#)(Krunal Shah)
  - Will present significant improvements over the original paper

Knowledge as Supervision

Joint Inference with General Constraint Structure [Roth&Yih'04,07,….]

Recognizing Entities and Relations

Joint inference gives good improvement

other	0.05
per	**0.85**
[illegible]	0.10

other	0.10
per	**0.60**
loc	0.30

other	0.05
per	0.50
loc	**0.45**

Brooklyn

Key Questions:

…rn the model(s)?

…source of the knowledge?

…de the global inference?

An Objective function that incorporates learned models with knowledge **(output constraints)**

A Constrained Conditional Model

irrelevant	0.05
spouse_of	**0.45**
born_in	0.50

[illegible]	[illegible]
spouse_of	0.05
born_in	**0.85**

Models could be learned separately/jointly; constraints may come up only at decision time.

Constrained Conditional Models [Abductive Reasoning; Chang et al.'12]

ILP Formulation

Variables are models

Penalty for violating the constraints.

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

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

Knowledge component: (Soft) constraints

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

- How to train models?
  1. Without the constraints; apply constraints only at decision time.
  2. With constraints
     - More costly
  3. What to learn during training? The objective function (**w, u**)? Learning all the intermediate functions **Ø(x, y)?**
- How to encode the constraints?
  1. Linear inequalities? Gives rise to LP/ILP
  2. Differentiable encoding of the linear constraints?

Knowledge as Supervision Paper

- A Logic-Driven Framework for Consistency of Neural Models (Jiayao Zhang)
  - Will present an interesting instance of this framework

Information extraction [Chang et al. ACL'07, MLJ'12]

Lars Ole Andersen . Program analysis and specialization for the C Programming language. PhD thesis. DIKU , University of Copenhagen, May 1994 .

$$\underset{y}{\operatorname{argmax}}\, \boldsymbol{\lambda} \cdot F(x, y)$$

Prediction result of a trained HMM

[AUTHOR]	Lars Ole Andersen . Program analysis and
[TITLE]	specialization for the
[EDITOR]	C
[BOOKTITLE]	Programming language
[TECH-REPORT]	. PhD thesis .
[INSTITUTION]	DIKU , University of Copenhagen , May
[DATE]	1994 .

Violates lots of natural constraints!

Strategies for Improving the Results

- (Pure) Machine Learning Approaches
  - Higher Order HMM/CRF?
  - Increasing the window size?
  - Use neural models
  - Adding a lot of new features
    - Requires a lot of labeled examples
  - What if we only have a few labeled examples?

Increasing the model complexity

Increase difficulty of Learning

Can we keep the learned model simple and still make expressive decisions?

- Other options?
  - Constrain the output to make sense
  - Push the (simple) model in a direction that makes sense

Examples of Constraints

- Each field must be a consecutive list of words and can appear at most once in a citation.
- State transitions must occur on punctuation marks.
- The citation can only start with AUTHOR or EDITOR.
- The words pp., pages correspond to PAGE.
- Four digits starting with 20xx and 19xx are DATE.
- Quotations can appear only in TITLE
- …….

Easy to express pieces of “knowledge”

Non Propositional; May use Quantifiers

Information Extraction with Constraints

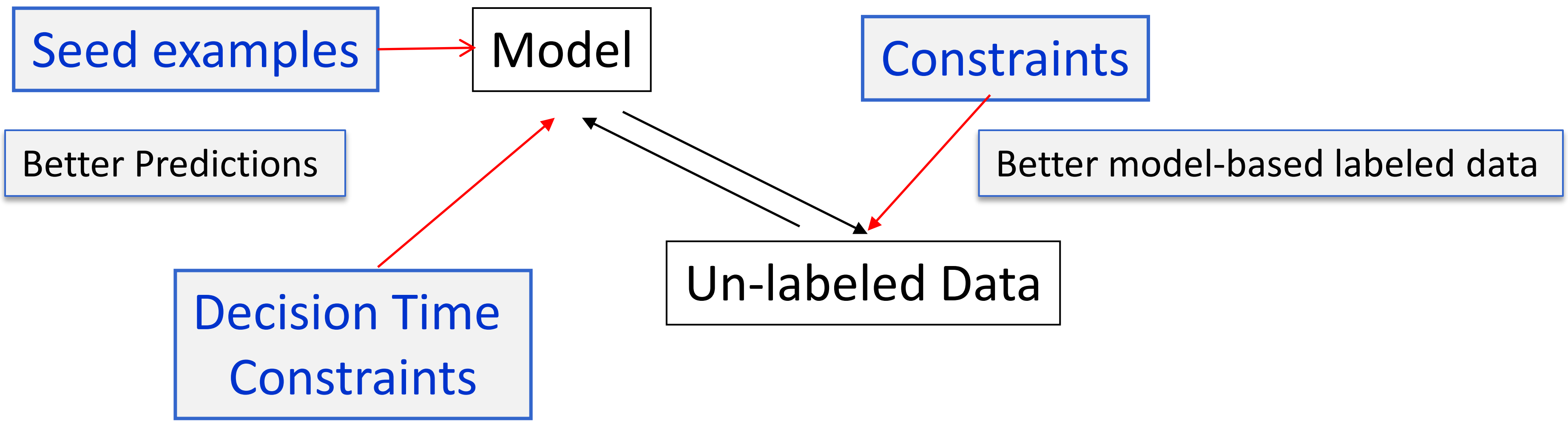
- Adding constraints, we get correct results!
  - **Without changing the model**

$$\underset{y}{\operatorname{argmax}}\, \boldsymbol{\lambda} \cdot F(x, y)$$

[AUTHOR]	Lars Ole Andersen .
[TITLE]	Program analysis and specialization for the C Programming language .
[TECH-REPORT]	PhD thesis .
[INSTITUTION]	DIKU , University of Copenhagen ,
[DATE]	May, 1994 .

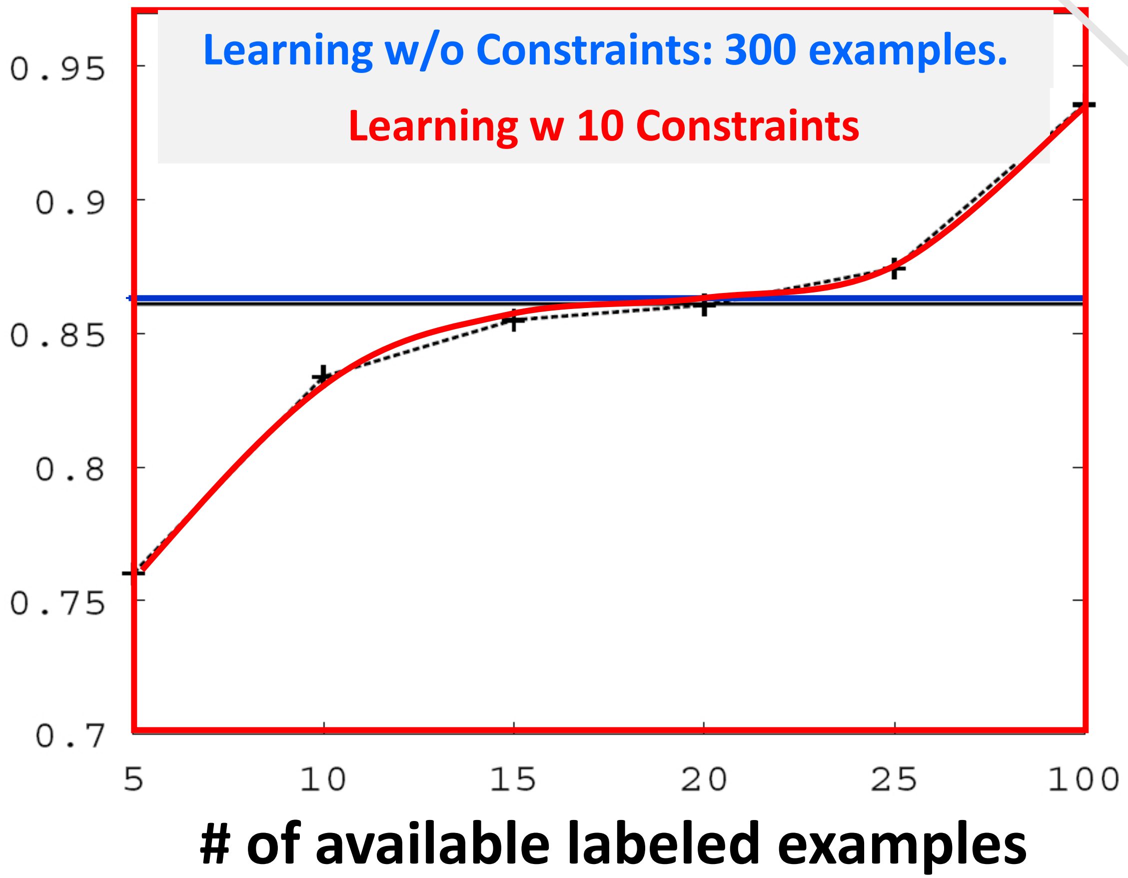
Guiding (Semi-Supervised) Learning with Constraints

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



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=ϕ

For each x in unlabeled dataset

h ← argmax$_{y \in Y}$ $w^T\emptyset(x, y) + u^TC(x, y)$

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

(w,u) = γ (w,u) + (1- γ) 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.

Constrained EM: Two Versions

- While Constrained-Driven Learning [CODL; Chang et al, 07,12]

  is a constrained version of hard EM:
- $$y^{*} = \mathrm{argmax}_{[y:Uy<b]} P_w(y|x)$$

Constraining the y feasible set

- … It is possible to derive a constrained version of EM:
- To do that, constraints are relaxed into expectation constraints on the posterior probability q:

$$E_q[Uy] <= b$$

Constraining a distribution over y

- The E-step now becomes: [Neal & Hinton '99 view of EM]

$$q' = \arg\min_{q:q(\mathbf{y})\geq 0, E_q[\mathbf{U}\mathbf{y}]\leq \mathbf{b}, \sum_{\mathbf{y}} q(\mathbf{y})=1} KL(q(\mathbf{y})||P(\mathbf{y}|\mathbf{x},\mathbf{w}))$$

- This is Posterior Regularization [PR] [Ganchev et al, 10]

The CoDL paper and the PR papers are doing a good job comparing these frameworks; also see Samdani & Roth, NAACL-12 for a unifying framework.

Zero-Shot + Knowledge Paper

- Zero-shot Learning of Classifiers from Natural Language Quantification (Young-Min Cho)
  - Using definitions to understand the target labels
  - Standard text classification problem.
  - **Input:** Text Snippet, taxonomy of labels $\{l_1, l_2, \ldots l_k\}$ (includes a none)
  - **Learn** a model that maps the text snippet into one of the labels.

- Key technical question is how to use the knowledge given by the "definitions"
  - Use of Posterior Regularization