CIS-620

Spring 2021

Learning in Few-Labels Settings

Dan Roth

Computer and Information Science

University of Pennsylvania

This class

- Presentations
  - Key part of the class
  - Send me your presentation by Wednesday before you present.

- Discussion/Discussants
- Projects

Questions?

- Understand early and current work on Learning in Few-Labels Settings
  - (Learn to) read critically, present, and discuss papers
- Think about, and Understand realistic learning situations
  - Move away from the “easy cases” to the challenging ones
  - Conceptual and technical
- Try some new ideas
- How:
  - Presenting/discussing papers
    - Probably: 1-2 presentations each;
    - Each paper will have 2 discussants: pro/con
  - Writing 4 critical reviews
  - “Small” individual project (reproducing);
  - Large project (pairs)
  - Tentative details are on the web site.

All the material will be available on the class’ web site, open to all. Let me know if you don’t want your presentation to be availalble.

- Machine Learning
  - 519/419
  - 520
  - Other?
- NLP
  - Yoav Goldberg’s book
  - Jurafsky and Martin
  - Jacob Eisenstein
- Attendance is mandatory
- Participation is mandatory

- Time: Monday 3pm, break, 4:30 pm.
- Zoom Meeting https://upenn.zoom.us/j/95494190734?pwd=MzhMek83U0hCSVgrblZkenZjL1hIUT09
- **TA: Soham Dan**
  - **Office hours: 6-7pm Monday**

What Should We Address?

- Zero-shot (few shot) Learning
  - Label-Aware methods
  - Transfer learning methods
  - Representation driven methods
- Incidental Supervision Signals
  - Where can we get signals from?
  - How to use them?
  - Is it art?
- Low Resource Languages
  - New Signals & projection
  - Representations
- End-task Supervision
  - When and How?
  - How to use indirect supervision signals?
- Knowledge as supervision
  - Constraints driven paradigms
  - Partial supervision
- Transfer Learning & Adaptation
  - Domain shift
  - Label space shift
- Theory

Supervised Machine Learning

- **Goal:** Learning a function that maps an input to an output based on a given training set of input-output pairs.
- **Labeled Training Data:** $\boldsymbol{D}^{\,train}$: $(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_N, \boxed{y_N})$
  - $\boldsymbol{x}_i \in X$; $y_i \in \boxed{Y}$
- Learning algorithms produces a model: $g(\boldsymbol{x})$
- **Labeled Test Data:** $\boldsymbol{D}^{\,test}$: $(\boldsymbol{x}'_1, y'_1), (\boldsymbol{x}'_2, y'_2), \ldots, (\boldsymbol{x}'_M, y'_M)$
- We measure **performance** by comparing $\{g(\boldsymbol{x}'_i), y'_i\}_{i\,=\,1,\,M}$
- Why does it work?
  - Test Data is assumed to be sampled form the same distribution as the Training Data
  - TrueError < TrainError + F(Complexity(H), 1/N)

Does the label carry any meaning?

What if the set Y changes after training?

What if that doesn't hold?

What if N is small?

Zero-Shot

Zero-Shot Learning

- If Labels have meaning, we can imagine being “label-aware”.
  - Developing models that “understand” the label and classify an instance x to an appropriate label given this “understanding”.
  - In this case, the notion of a test-set may not be important anymore:
    - Given a single example, we should be able to classify it.
  - There are multiple ways to be “label-aware”
    - Prototypes; definitions, other forms of knowledge
- But the power of “label-aware” is not the only power that can drive zero-shot
  - Maybe you have seen training data for some labels (but not all) and you are aware of “relations” between labels.
    - Common in Computer Vision
- Transfer Learning:
  - Maybe learning model for task T_1, can be used to make predictions on task T_2.

Text Classification

- A lot of the work in NLP can still be viewed as **text classification**
  - ☐ Categorization into topics
  - ☐ Identify intention of text
  - ☐ Identify abusing text
  - ☐ Be admitted again soon/no
  - ☐ Classify fact/opinion
  - ☐ ……

The research community uses **data** embeddings broadly!
And, heavy reliance on task specific supervision
But there is almost no use of task/label understanding.
Promote **Label-Aware Models** as a form of **Incidental Supervision**

- While you understand the labels, models are not given information about the labels
- We simply view these tasks as multi-class classification
- The only thing that has changed since the 70-ies is slightly better learning algorithms, and slightly better word representations

Text Categorization

Is it possible to map a document to an entry in a taxonomy of semantic categories, without training with labeled data?

Second seed Rafael Nadal continues his quest for an unprecedented 12th title at this event against Grigor Dimitrov of Bulgaria. The Spaniard leads their FedEx ATP Head2Head 11-1 and has won their past four matches, in addition to all four of their previous battles on clay. Their most recent
semi-finals, which saw

It's about Sport
It's about Tennis

- Traditional text categorization requires training a classifier over a set of labeled documents (1,2,….k)
- Someone needs to label the data (costly)
- All your model knows is to classify into these given labels

Total costs for on-the-j
an average of 5% in 2018, surpassing $14,000 a year per employee, according to a National Business Group on Health survey of large employers. Specialty drugs continue to be the top driver of increasing costs. Companies will pick up nearly 70% of the tab, but employees must still bear about 30%, or roughly $4,400, on average.

It's about Money
It's about Health Care

You can classify these documents without task specific annotation, since you have an "understanding" of the labels

Categorization without Labeled Data [AAAI'08, AAAI'14, IJCAI'16]

This is **not an unsupervised learning** scenario. Unsupervised learning assumes a coherent collection of data points, where similar data points are assigned similar labels. It does not work on a single document. **0-shot learning**.

- Given:
  - A single document (or: a collection of documents)
  - A taxonomy of categories into which we want to classify the documents
- Dataless/Zero-Shot procedure:
  - Let $f(l_i)$ be the semantic representation of the labels (label descriptions)
  - Let $f(\mathbf{d})$ be the semantic representation of a document
  - Select the most appropriate category:

    $$l_i^* = \text{argmin}_i \text{ dist } (f(l_i) - f(\mathbf{d}))$$
  - Bootstrap
    - Label the most confident documents; use this to train a model.
- Key Question:
  - How to generate good Semantic Representations?
- Originally:
  - Task Independent Representations: best results with Wikipedia-based (ESA) [Gabrilovich & Markovitch AAAI'06]
  - Sparse representation: a TF/IDF weighted list of URL a concept appears in

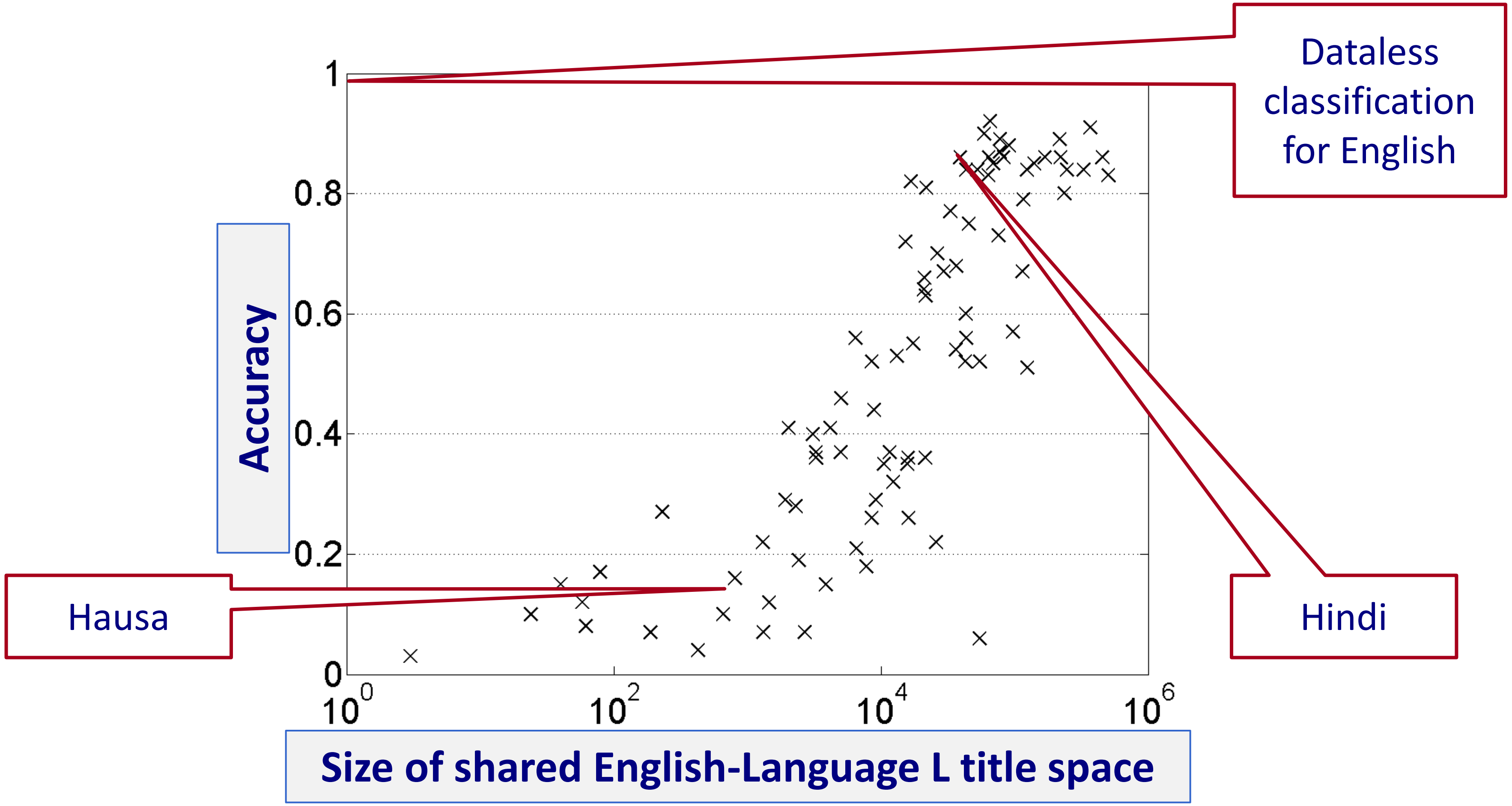
Hard to beat

Single Document Classification (88 Languages)

[Song et. al. IJCAI'16; AIJ'19]

Can be done to all 293 languages represented in Wikipedia

Performance depends on the Wikipedia size (Cross-language links)



Understanding a Taxonomy

*A former Democrat, **Bloomberg**, switched his party registration in 2001.*

label aware models

- Key question: how do we "understand" the taxonomy?
- "Type" as a conceptual container binding entities together.
  - Defined extensionally as **a set of members** of the type
  - (**Not** examples annotated in context; Wikipedia pages, say)

politician

Richard Nixon	***Bill Clinton***
Elizabeth Warren	***Boris Johnson***

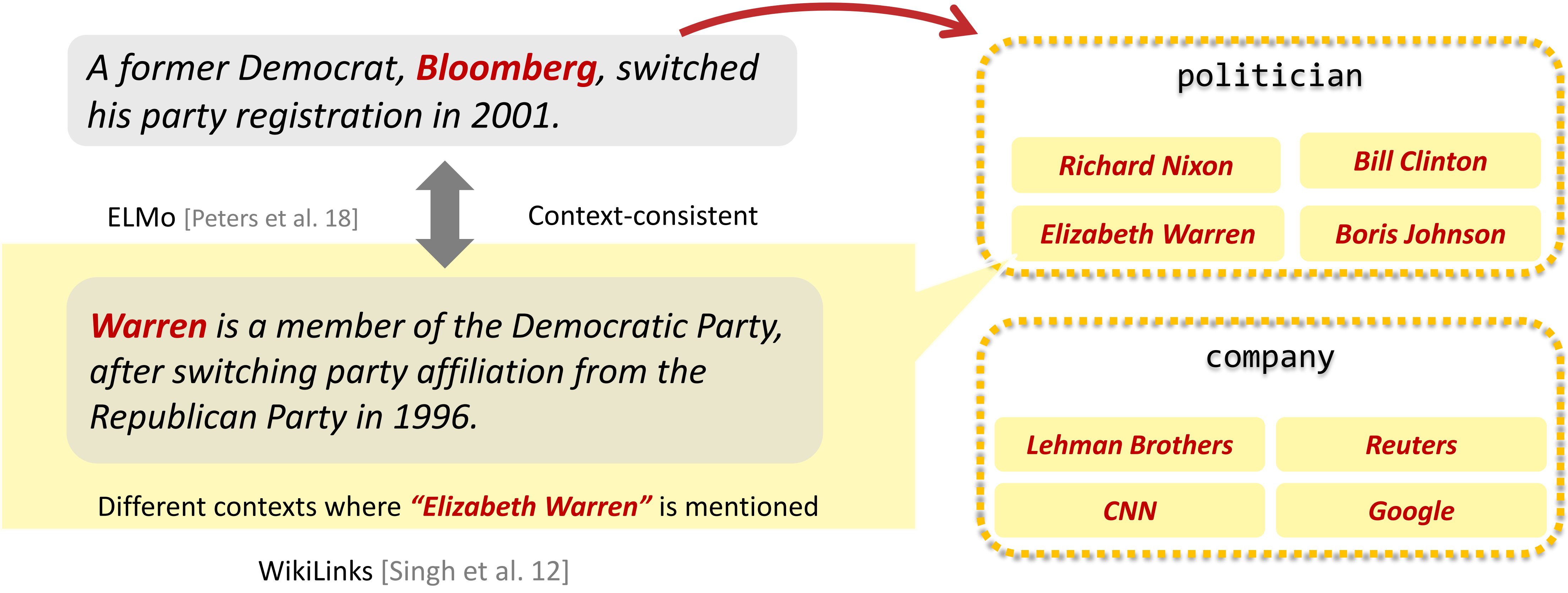
company

Lehman Brothers	***Reuters***
CNN	***Google***

Computational Approach: Determine the **type** of an **input mention** by **finding entities in the type-defining-set** that share a similar context
Of course, each entity will be in multiple such buckets

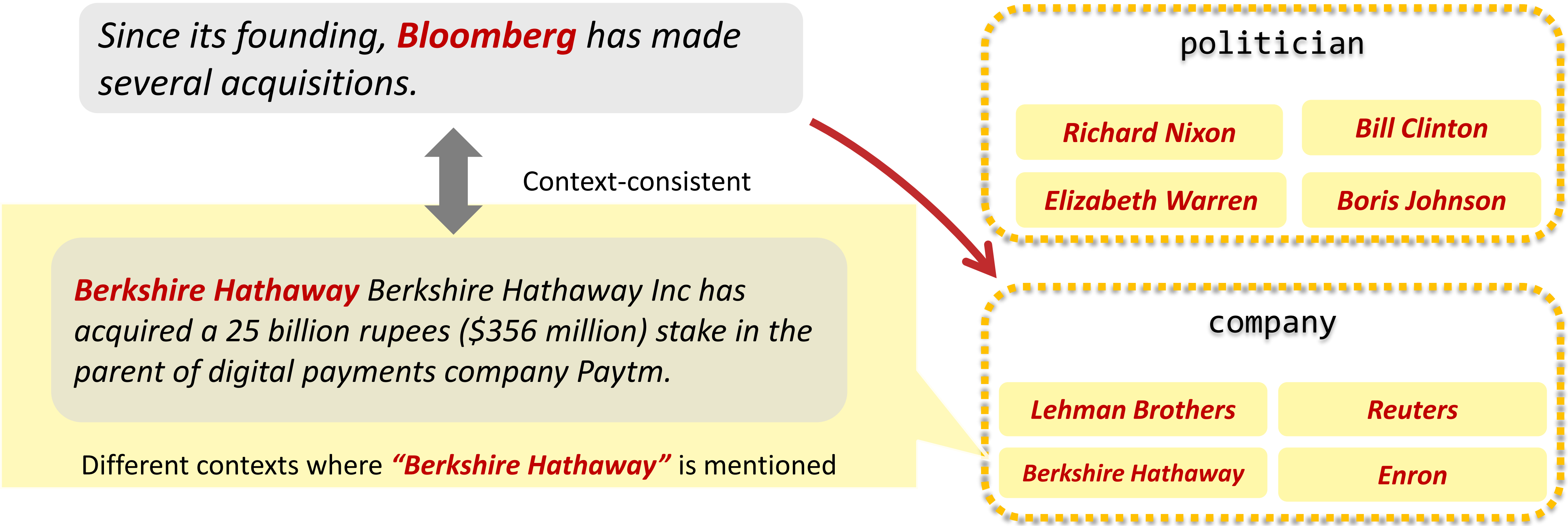
ZOE: Type-Compatible Grounding

- "Type" as conceptual container binding entities together.



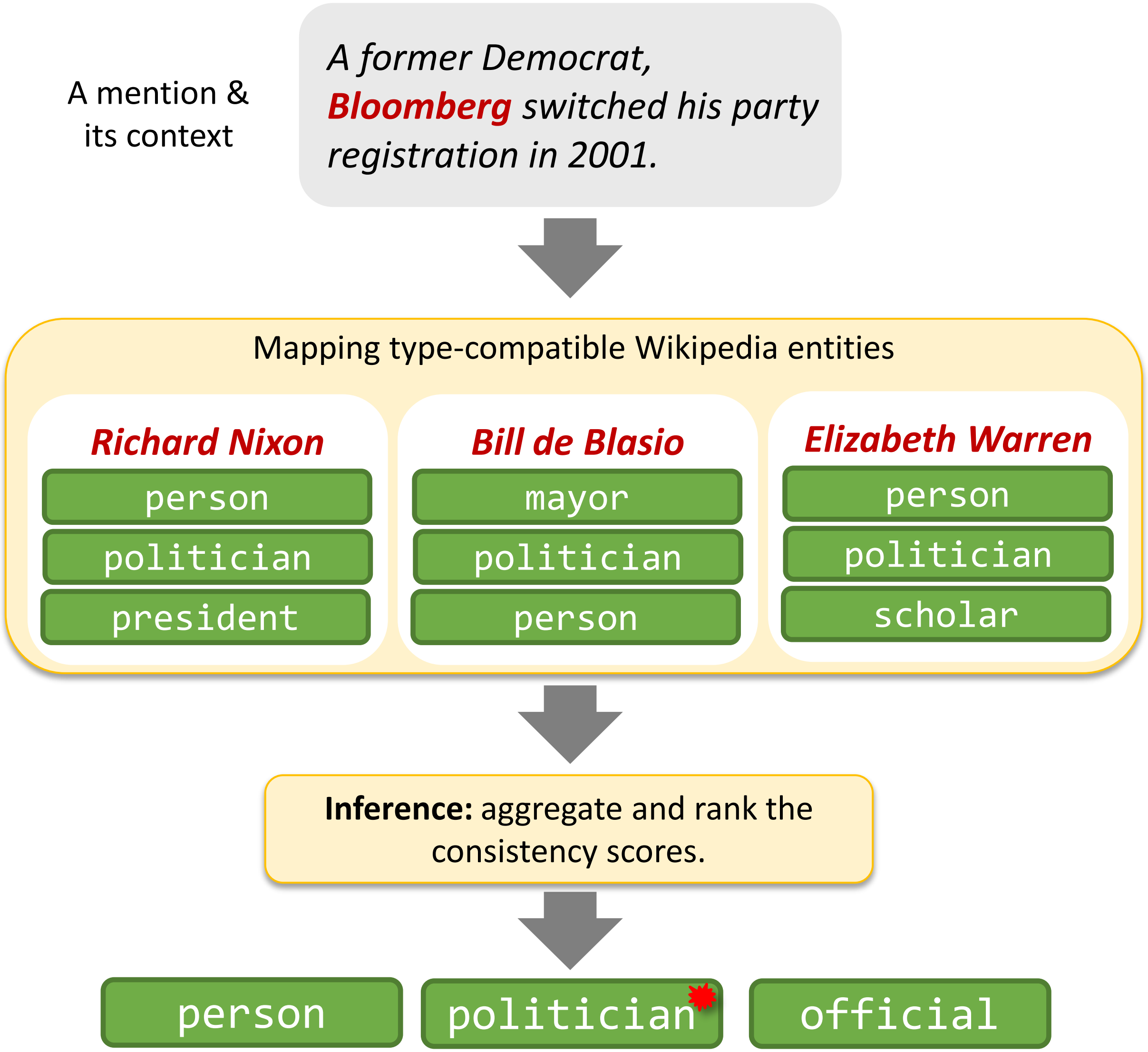
ZOE: Type-Compatible Grounding

- “Type” as conceptual container binding entities together.



Context consistency allows us to determine good candidates for **type compatibility.**

Zero-Shot Open Typing: Big Picture



High-level Algorithm:

1. Map the mention to **context-consistent** Wikipedia concepts (noting that each belong to multiple types)
   - Simple Entity Linking

2. Rank candidate titles by **context-consistency** and infer the types according to the type taxonomy.

Zero-Shot Via Transfer

- Assume that I already know how to solve some tasks
  - I have models that can support QA: (Context, Question) → Answer
  - I have models that can support Textual Entailment (Premise, Hypothesis) → [Entails, Contr, IDK]

- Can we use it to solve other tasks without tasks specific annotation?

Zero-Shot Event Extraction

Input: "*China purchased two nuclear submarines from Russia last month.*"

Output:

Event type: TRANSFER-OWNERSHIP

China has purchased two nuclear submarines from Russia last month.

Buyer-Arg Trigger Artifact-Arg Seller-Arg Time-Arg

- ❑ Annotation at this level is costly and requires expertise
- ❑ And, it needs to be done whenever we update the event anthology.

- ❑ On the other hand, it makes sense to assume that one can get a definition of each event type of interest.
- ❑ **Can this be used to classify events and their arguments?**

Events' Definitions

- Given Event Schema for each event type in the anthology
- Given a text snippet

- (1) Identify the event type
  - Zero-shot text classification
- (2) Choose the appropriate schema, and, with this guidance, generate questions that can determine the event's arguments.
- Importantly, there is a need to support also I-don't-know, since some of the questions will have know answer

Event type:

TRANSFER-OWNERSHIP

Argument slots:

- **Buyer-Arg**: The buying agent
- **Seller-Arg**: The selling agent
- **Beneficiary-Arg**: The agent that benefits from the transaction
- **Artifact-Arg**: The item or organization that was bought or sold
- **Price-Arg**: The sale price of the *ARTIFACT-ARG*
- **Time-Arg**: When the sale takes place
- **Place-Arg**: Where the sale takes place

Event Extraction as Question Answering

- **Input:** China purchased two nuclear submarines from Russia last month.

- **Trigger:** purchased
- **Event Type:**
  - Q0: Did someone transfer ownership? (multiple questions are being asked)
  - A0: Yes ⇒TRANSFER-OWNERSHIP (TC)
- **Arguments:** (now we know the event type)
  - **Q1:** What was purchased? (multiple questions for each arg type)
  - **A1:** Two nuclear submarines. ⇒Artifact-Arg
  - **Q2:** Who purchased two nuclear submarines?
  - **A2:** China. ⇒Buyer-Arg
  - **Q3:** Who did China purchase two nuclear submarines from?
  - **A3:** Russia. ⇒Seller-Arg
  - …..
  - **Q7:** Where was the transaction?
  - **A7:** I-don't-know. ⇒Place-Arg

Work in progress

We are relying on the ability to answer extractive questions and the fact that there are multiple large datasets for this task.
e.g., [He et al. 2019]
The ability to support **I-don't-know** is harder.

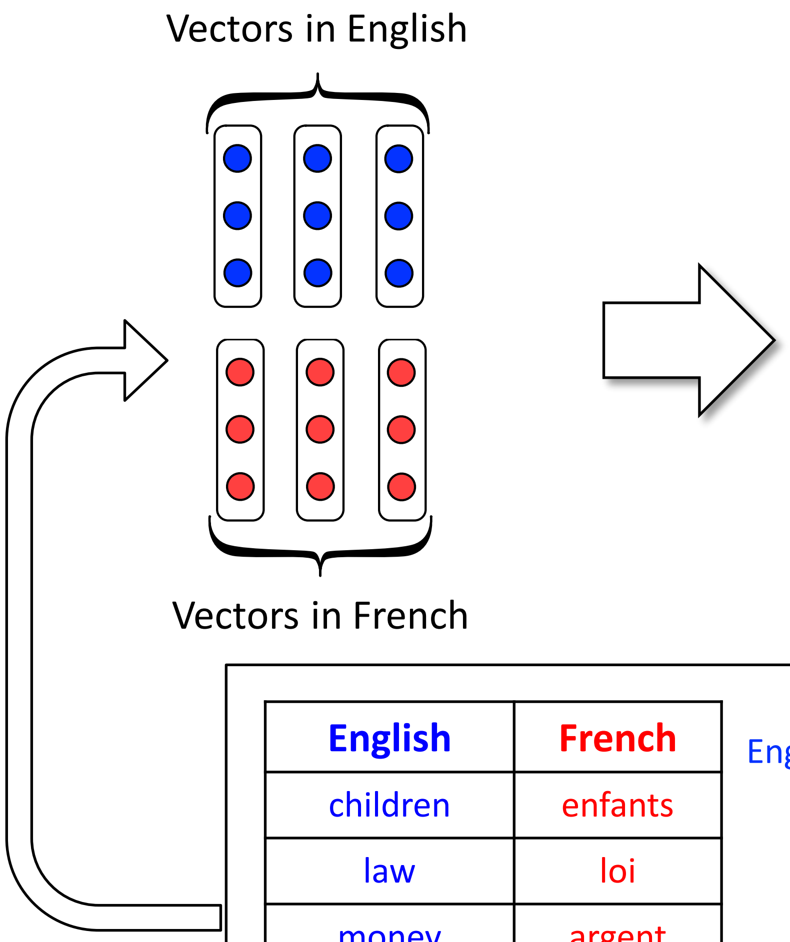
Low Resource Languages

Low Resource Languages

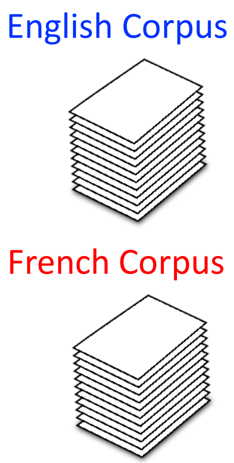
- We have annotations for many NLP tasks in English
- We have almost very little annotation for most NLP tasks in other languages
  - We have no annotation in low resource languages (some spoken by dozens of millions)

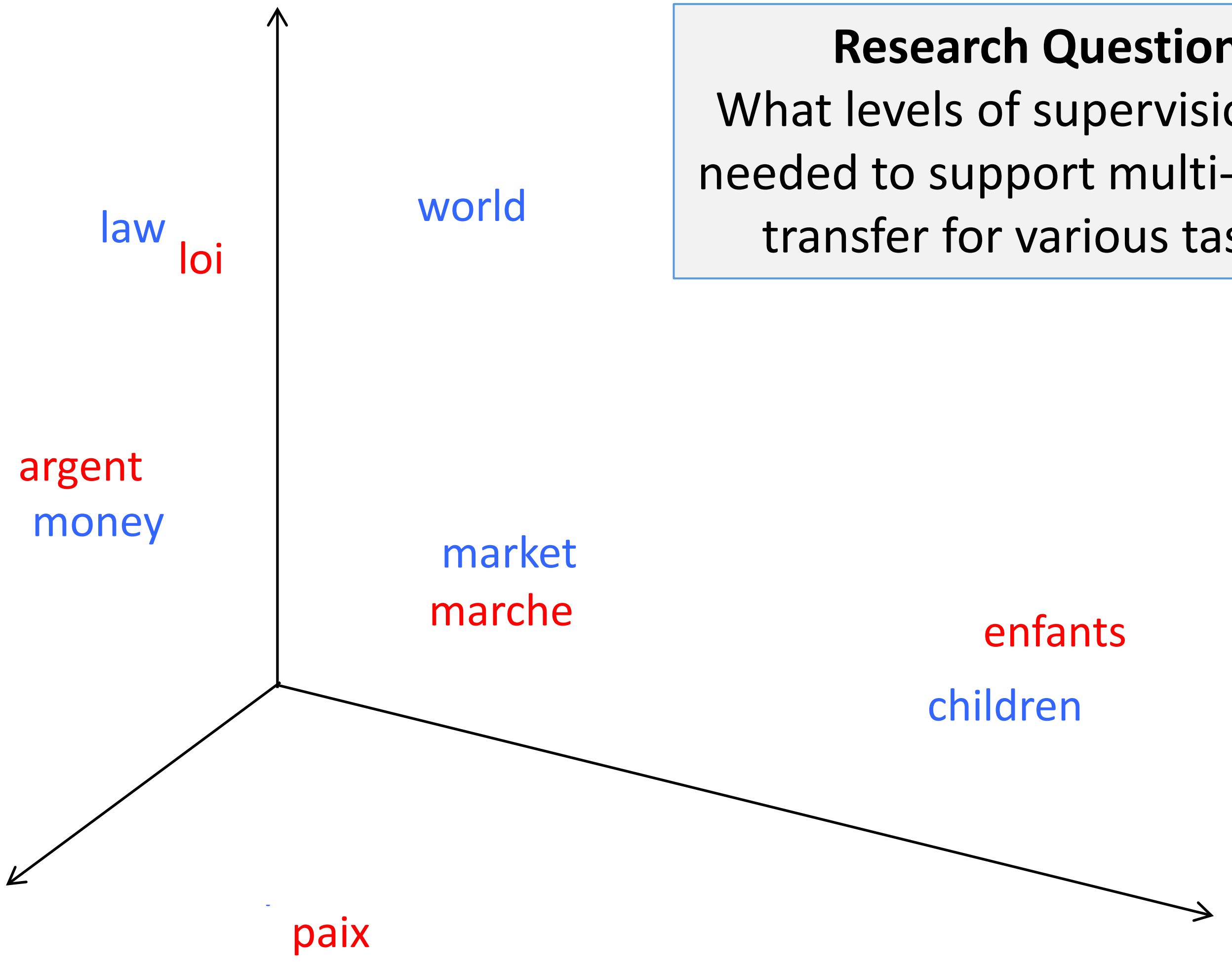
- What options do we have?
  - Translation
    - Not feasible in most cases. Requires a lot of annotation
  - Clever use of existing resources (dictionaries, similar “rich” languages)
  - Projection of annotation
  - Multilingual representations of language

Cross-lingual Representations [Upadhyay et al. (ACL’16)]



English	French
children	enfants
law	loi
money	argent
world	??
??	paix
??	marche





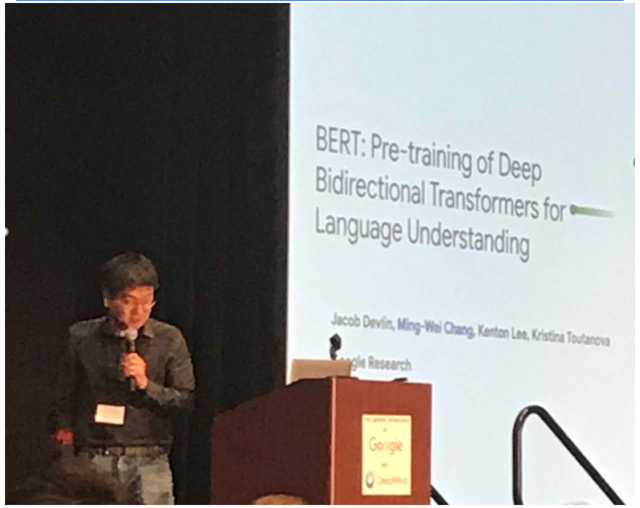
Research Question
What levels of supervision are needed to support multi-lingual transfer for various tasks?

Continuous approximation of discrete translation dictionaries in a vector space.

What Happened?

- It has been established that *multilingual embeddings* are essential
  - NER, EDL, SF all rely on these representations.
- However, it's also clear that in order to develop tools that span many languages we may need *many models*, and some (minimal) level of supervision.
- BERT: powerful **contextual** language model
  - mBERT: a multilingual version – multilingual embeddings
  - A single multilingual embedding for many languages.
  - No direct supervision – only needs sufficient data in each language.
- Needs to be extended to make use in low-resource languages
- Many questions remain
  - Why and when does it work?

Ming-Wei Chang
NAACL'19 Best Paper Award



Contextual embeddings are behind many of the advances in NLP in the last couple of years. But, on their own, they are not sufficient to address low-resource languages

Massively Multilingual Analysis of NER

- Low-resource NER:
  - different methods, parameters, languages

- Evaluation in 22 languages (LORELEI)
  - 10 different scripts
  - 10 language families (Niger-Congo most popular)

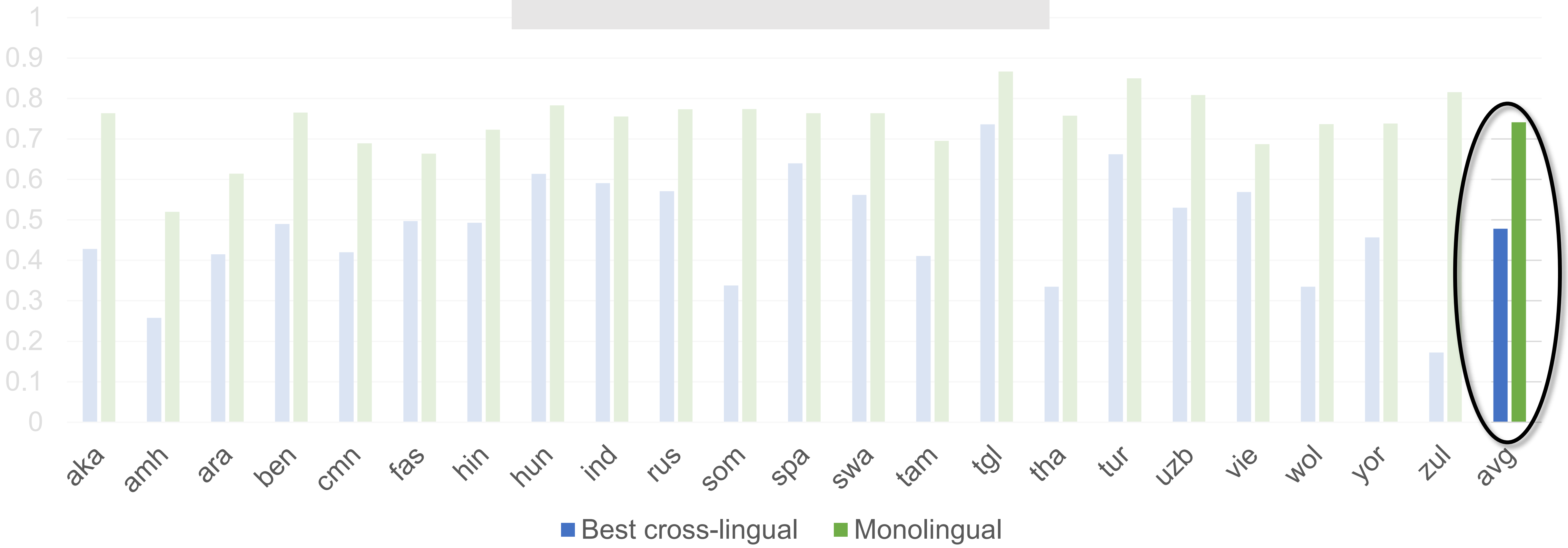
Language	3 letter code
Akan (Twi)	aka
Amharic	amh
Arabic	ara
Bengali	ben
Farsi	fas
Hindi	hin
Hungarian	hun
Indonesian	ind
Chinese	cmn
Russian	rus
Somali	som
Spanish	spa
Swahili	swa
Tagalog	tgl
Tamil	tam
Thai	tha
Turkish	tur
Uzbek	uzb
Vietnamese	vie
Wolof	wol
Yoruba	yor
Zulu	zul

Overall: Still Ways to Go

- Average of best cross-lingual (47 F1) is still less than monolingual (74 F1)

- **Amazing: this can be done almost "for free"**
- Cross-lingual transfer by itself isn't sufficient

The gap can be narrowed with additional methods



What Else can be Done? Case of NER

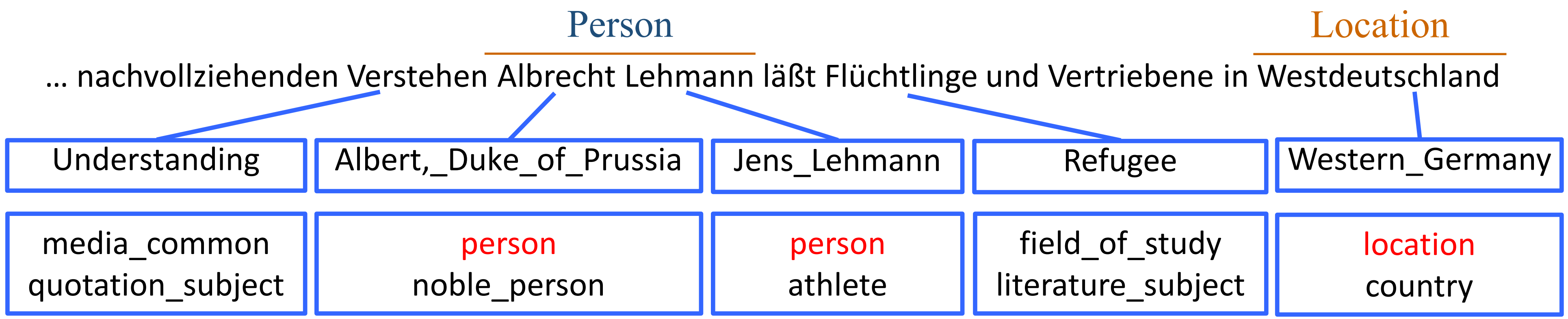
- Multilingual NER with no Target Language Training Data [Tsai et al. ‘16]
  - Using Wikipedia-based language-independent features for NER

- Transfer of annotation via cheap translation [Mayhew et al. ‘17]

- Use of Weak signals:
  - Character-Based Language Models for Entities [Yu et al.’ 18]
    - Identifying Entities across languages with minimal supervision
  - Partial annotation [Mayhew et al.’19]
    - Forms of annotation that are typical to those provided by non-native speakers
    - High precision, low recall

Multilingual NER with no Target Language Training Data

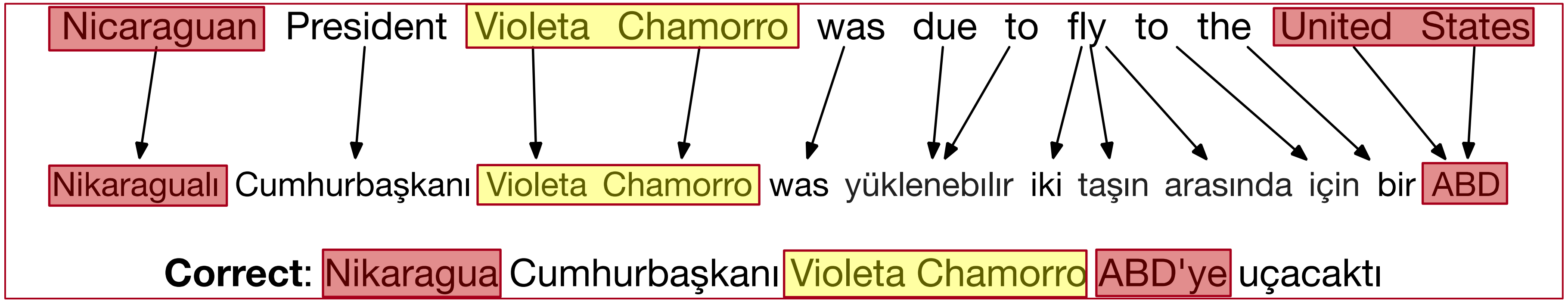
- Cross-lingual EDL generates good language-independent features for NER by grounding n-grams



- Words in any language are grounded to the English Wikipedia
  - Features extracted based on the titles can be used across languages
- Instead of the traditional pipeline: NER → Wikification/EDL
  - Simple-minded EDL of n-grams is sufficient to provide features for the NER model

Transfer via Cheap Translation [Mayhew, Tsai, Roth EMNLP’17]

- Start with English Gold NER data
- Translate word-by-word into Turkish
- Result is “Turkish-flavored English”
- Train with a standard, state-of-the-art NER
- Translation is bad:
  - Ignorance of morphology
  - Wrong word order
  - Missing vocabulary



NER with Partial Annotation

Romanized Bengali

ebisi’ra giliyyaana phinnddale aaja pyaalestaaina adhiinastha gaajaa theke aaja raate ekhabara jaaniyyechhena .

ABC’s Gillian Findale has reported from Gaza under Palestine today.

- Low recall, high precision annotation
- Can be solicited even from non-native speakers
- Challenge:

An algorithmic approach that allows training high quality NER from partial annotation given by non-native speaker.

Learning from Responses

Understanding Language Requires (some) Feedback

Can I get a coffee with lots of sugar and no milk

Can we rely on this interaction to provide supervision (and eventually, recover meaning) ?

The Goal

***Great*!**

Arggg

Meaning Representation:

MAKE(COFFEE,SUGAR=YES,MILK=NO)

- 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. 2010: Driving Semantic Parsing from the World's Response** – a lot more to do]

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

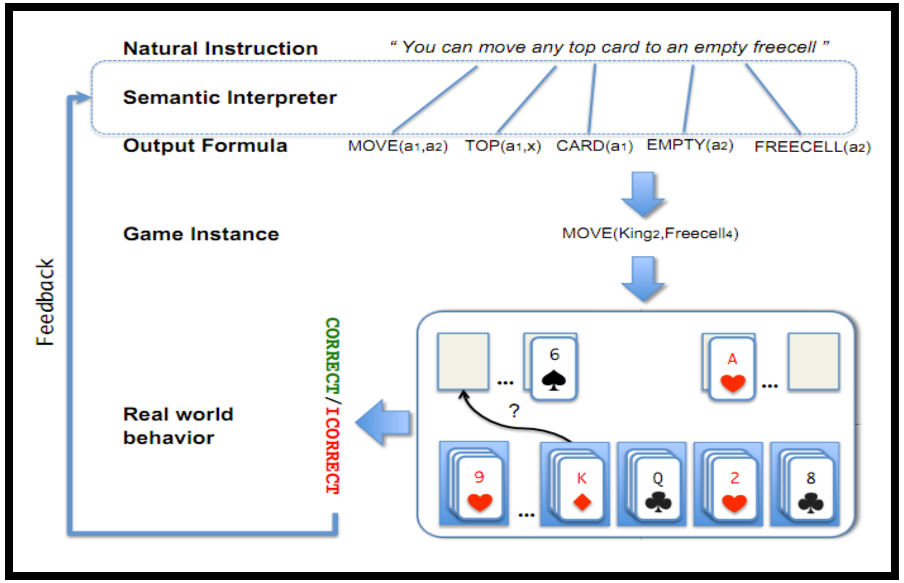
A Response based Learning Scenario

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

English Sentence ⇒ **Model** ⇒ Meaning Representation

A top card can be moved to the tableau if it has a different color than the color of the top tableau card, and the cards have successive values.

Move (a1,a2) top(a1,x1) card(a1) tableau(a2) top(x2,a2) color(a1,x3) color(x2,x4) not-equal(x3,x4) value(a1,x5) value(x2,x6) successor(x5,x6)



⇐ Play Freecell (solitaire)

- Simple derivatives of the models outputs: game API
  - Supervise the derivative and Propagate it to learn the transformation model

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 simple derivatives of the models outputs,
  - Supervise the derivative (easy!) and
  - Propagate it to learn the complex, structured, transformation model

LEARNING:

- Train a structured predictor (semantic parse) with this binary supervision
  - Many challenges: e.g., how to make a better use of a negative response?
- Learning with a constrained latent representation, use inference to expectations to exploit knowledge (e.g., on the structure of the meaning representation).
- [Clarke, Goldwasser, Chang, Roth CoNLL'10; Goldwasser, Roth IJCAI'11, MLJ'14]

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

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

De-compositions for Reasoning about Text [Gupta et al. 18', 19' 20']

- Text Comprehension challenges:
  - Understand the text
    - Identify and contextualize events, entities, quantities (and their scope), relations, etc.
  - Understand questions about the text
    - Often, requires decomposing the question in a way that depends on the text
  - Reason about the text
    - Combine and manipulate the identified information to accomplish information needs (e.g., answering questions)
- How can we supervise to support this level of understanding?
  - Too many (ill-defined) latent decisions
    - Annotating text for all is not scalable
  - End-task supervision is the only realistic approach [Clarke et. al.'10] but it is too loose – how can we learn all the latent decisions from end-to-end supervision?

Who scored the longest touchdown pass of the game?

In the ... Greg ... touchdown run. The Bears increased their lead over the Vikings with Cutler's 2-yard TD pass to tight end Desmond Clark. The Vikings … with Favre firing a 6-yard TD pass to tight end Visanthe Shiancoe. The Vikings … with Adrian Peterson's second 1-yard TD run. The Bears then responded with Cutler firing a 20-yard TD pass to wide receiver Earl Bennett. The Bears then won on Jay Cutler's game-winning 39-yard TD pass to wide receiver Devin Aromashodu.

What is her seizure frequency?

She reports wors... now occurring up to 10/week, in clusters about 2-3 day/week. Previously reported seizures occurring about 2-3 times per month, often around the time of menses,…

How much did his challengers raise?

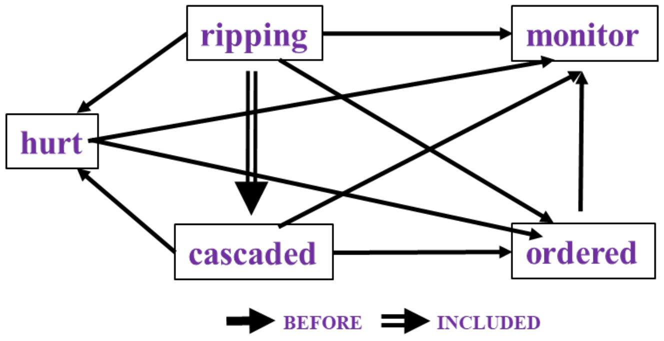
Mayor Rahm E... n toward his bid for a thi... l raised by his 10 challengers combined, campaign finance records show.

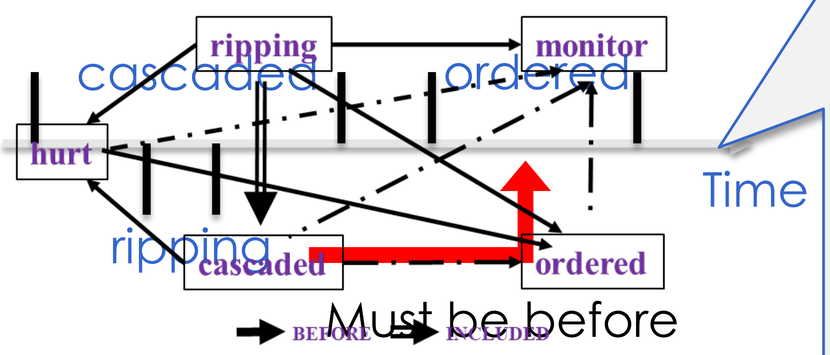
The COVID-19 pandemic in the United States is part of the worldwide pandemic of coronavirus disease 2019 (COVID-19). As of October 2020, there were more than 9,000,000 cases and 230,000 COVID-19-related deaths in the U.S., representing 20% of the world's known COVID-19 deaths, and the most deaths of any country.

Knowledge as Supervision

Time and Events [Ning et al. *SEM'2018; ACL'18, EMNLP'18, EMNLP'19; Wang et al. EMNLP'20]

- In Los Angeles that lesson was brought home today when tons of earth **cascaded** down a hillside, **ripping** two houses from their foundations. No one was **hurt**, but firefighters **ordered** the evacuation of nearby homes and said they'll **monitor** the shifting ground until March 23rd.



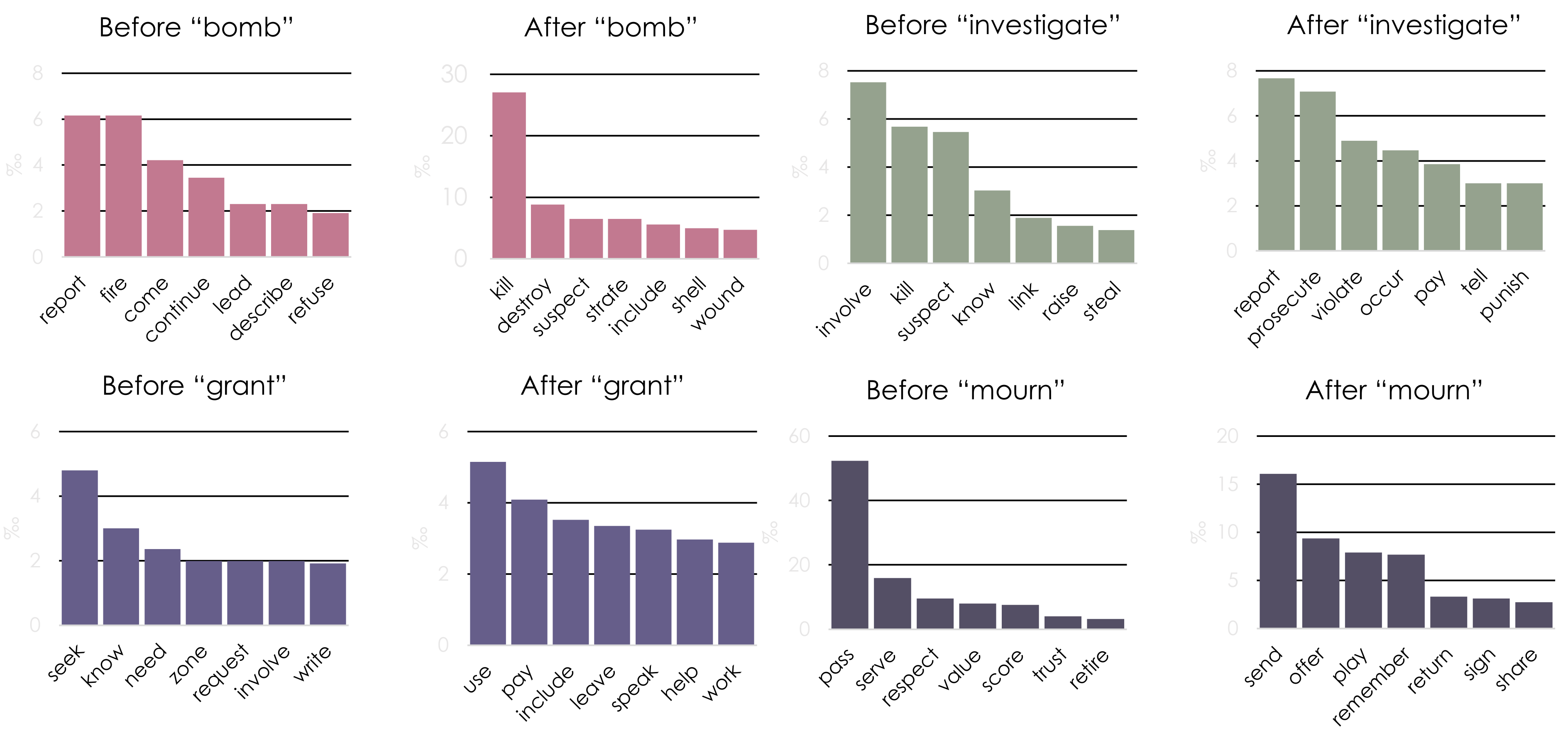


1. Reasoning: How to exploit these [declarative & statistical] "expectations"?
2. How/why does it impact generalization & supervision?

- Very difficult task— hinders exhaustive annotation (O(N^2) edges)
- But, it's rather easy to get partial annotation – some relations.
- And, we have **strong expectations** from the output
  - Transitivity
  - Some events tend to precede others, or follow others

More than 10 people have (**event1**), police said.
A car (**event2**) on Friday in a group of men.

Event distributions Support Temporal Relations

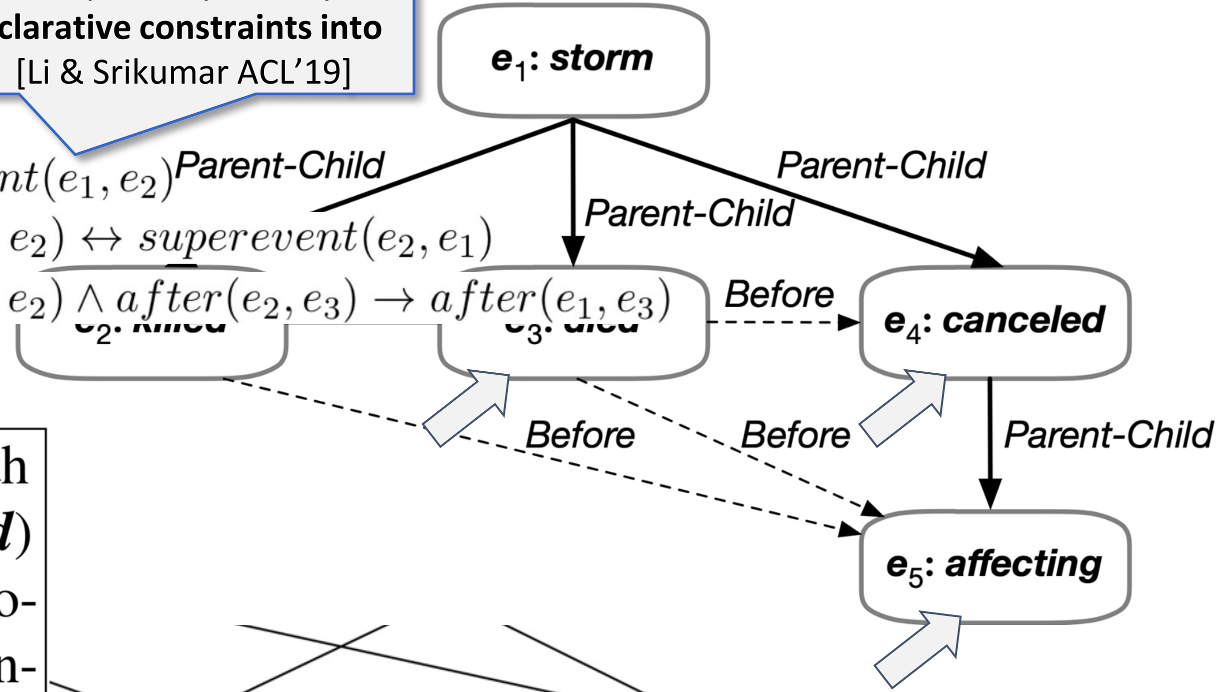


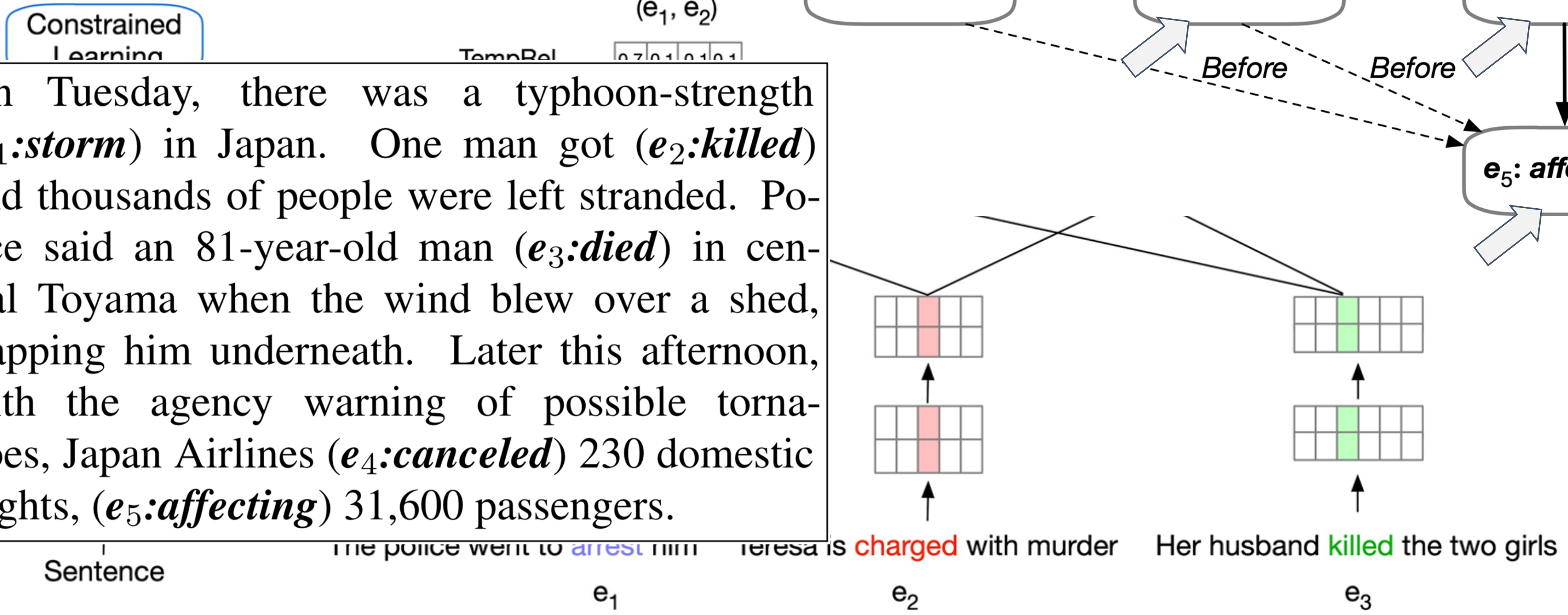
More Reasoning: Event Complexes [Wang et al. EMNLP'20]

- Temporal Relations
- Subevent Relations
- Event Coreference

Enforcing more logical constraints: Temporal, Symmetry, and Conjunctive by **converting declarative constraints into differentiable learning objectives** [Li & Srikumar ACL'19]

$$\top \rightarrow subevent(e_1, e_2)$$
$$subevent(e_1, e_2) \leftrightarrow superevent(e_2, e_1)$$
$$subevent(e_1, e_2) \wedge after(e_2, e_3) \rightarrow after(e_1, e_3)$$





On Tuesday, there was a typhoon-strength (***e_1:storm***) in Japan. One man got (***e_2:killed***) and thousands of people were left stranded. Police said an 81-year-old man (***e_3:died***) in central Toyama when the wind blew over a shed, trapping him underneath. Later this afternoon, with the agency warning of possible tornadoes, Japan Airlines (***e_4:canceled***) 230 domestic flights, (***e_5:affecting***) 31,600 passengers.

Information extraction

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?
  - 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

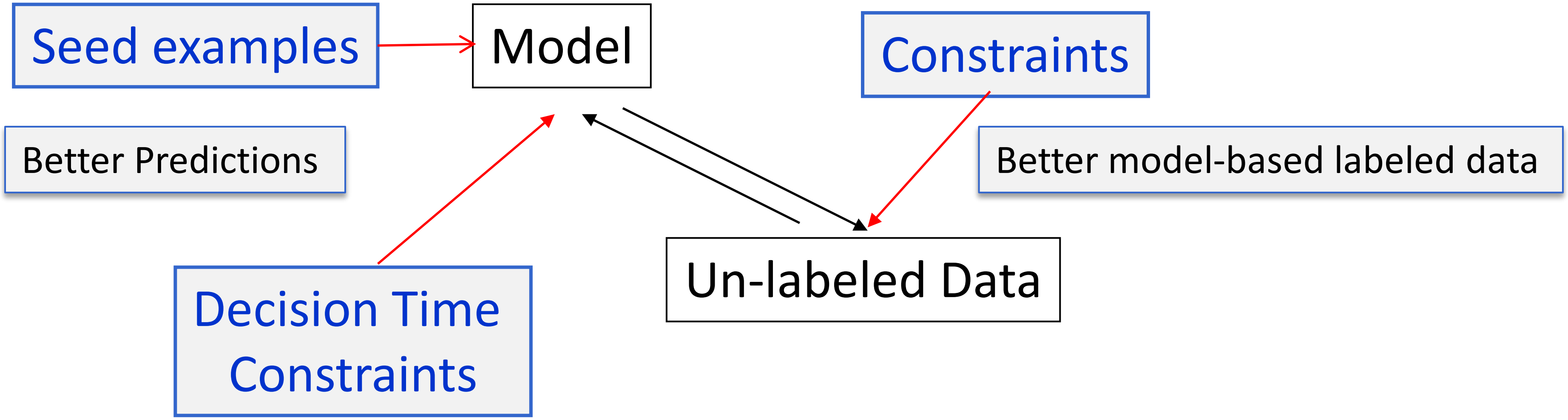
- Adding constraints, we get correct results!
  - **Without changing the model**

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

[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 .

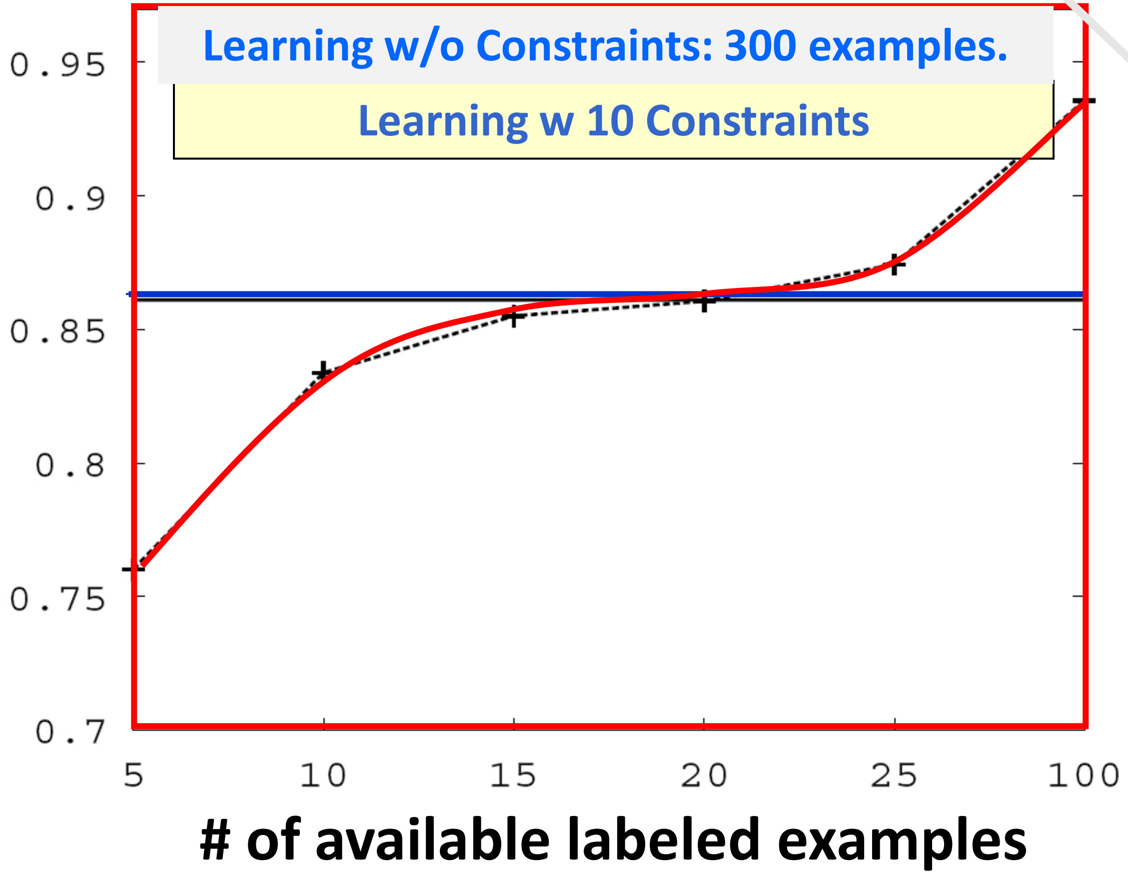
Guiding (Semi-Supervised) Learning with Constraints

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

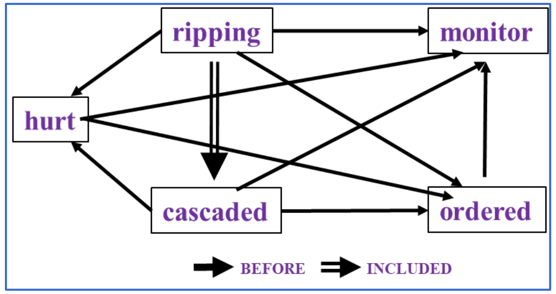


Constraints are used to Bootstrap a semi-supervised learner

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

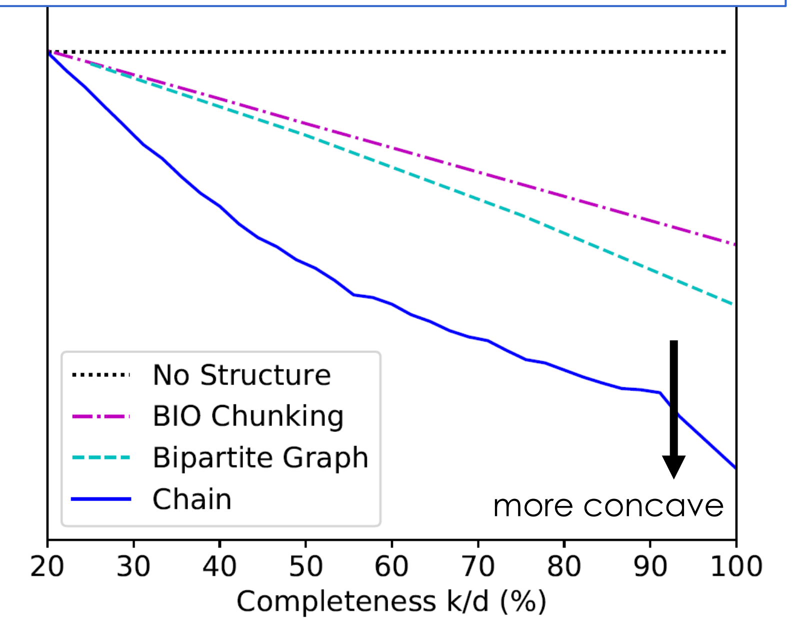
Why Does Structure Help? [Ning et al. NAACL'19]

- Information-theoretic considerations support quantifying the benefit of additional labels
- The tighter the structure is, the smaller the benefit of an additional label is.

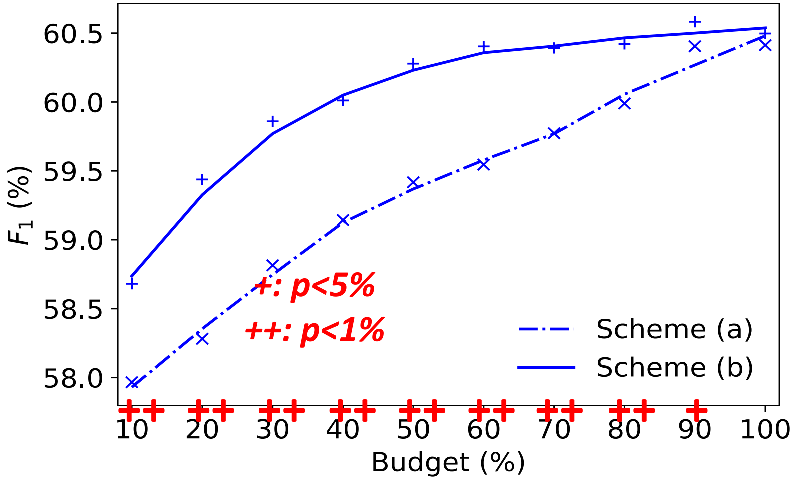


- Consequently, better generalization even with a partial set of supervision signals.

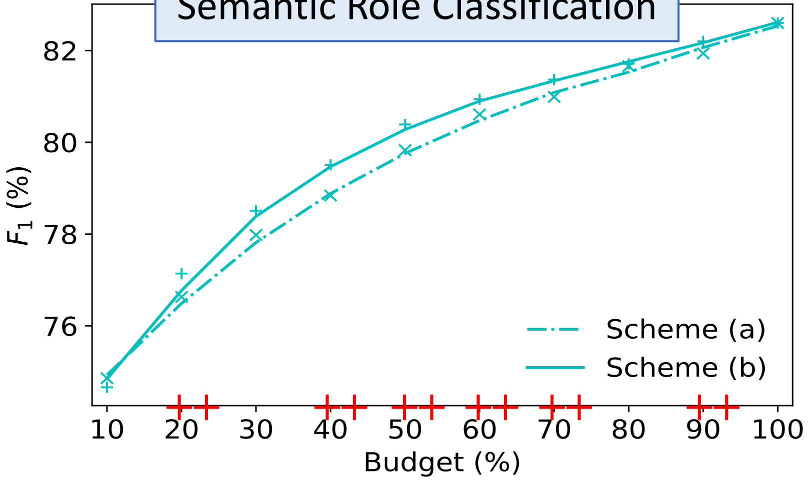
$\boldsymbol{\Delta_k = I_k - I_{k-1}}$: **The benefit of a new label**



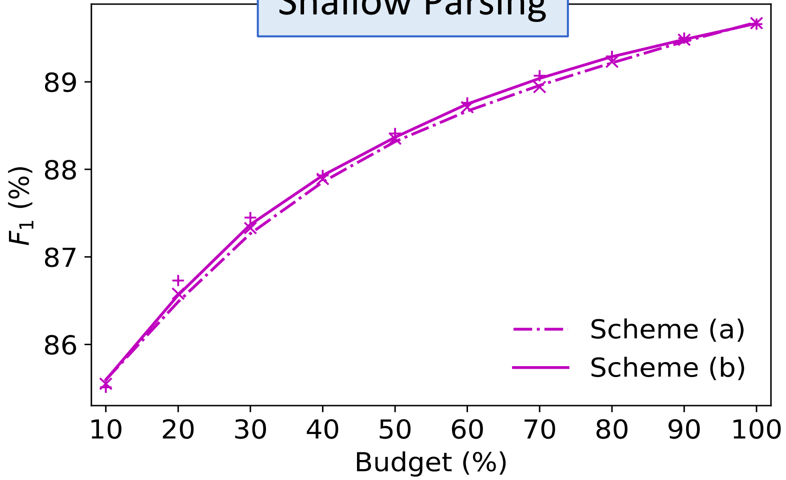
Chain:
Temporal Relation Extraction



Bipartite Graph:
Semantic Role Classification



Seq Tagging
Shallow Parsing



Theory

What can we Say about our ability to learn?

- Assume that instead of the “perfect” supervision signal you get:
  - Noisy signals
  - Partial Signals
  - Constraints
  - Supervision from other (related tasks)
- Is it possible to learn the target task from these signals?

Learnability with Indirect Supervision Signals (Wang, Ning, Roth, NeurIPS 2020)

Framework

- We studied the problem of learning with **general forms of supervision signals** (O), any signals that contain (partial) information about the true label.
  - Examples: noisy/partial labels, constraints, or indirect feedback from the environment.
- We proposed a general learning scheme, which uses the predicted label to induce predictions about O. The prediction is evaluated by the observed signals (fig. 1).

Question: Under what conditions on the indirect supervision can we guarantee learnability of the original classification problem?

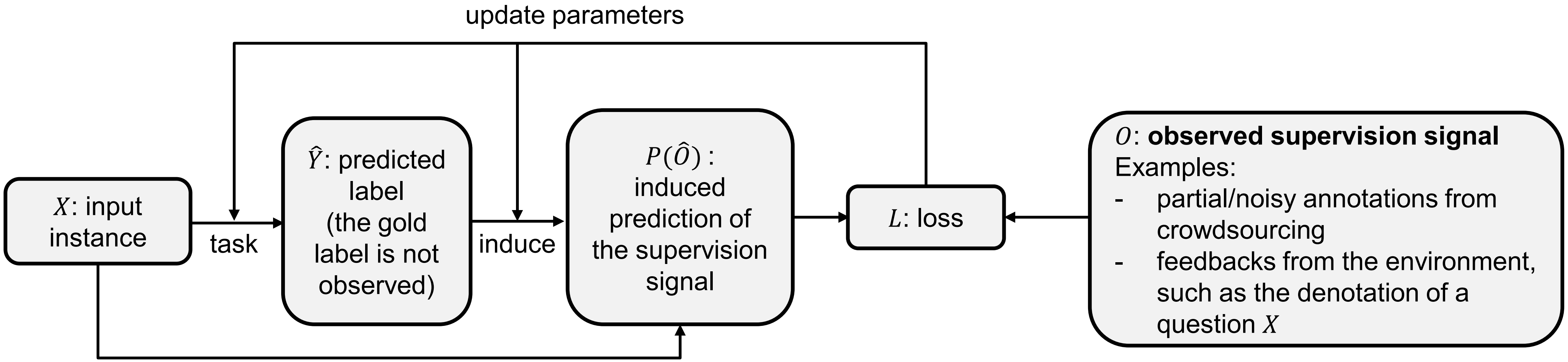
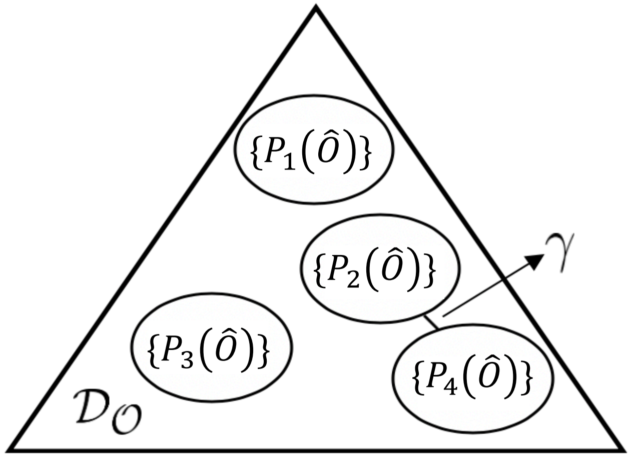


Figure 1: Learning Scheme

Learnability with Indirect Supervision Signals

(Wang, Ning, Roth, To appear in NeurIPS 2020)

$\{P_i(\hat{O})\}$ is the set of possible induced predictions given the label Y is classified as i



Separation condition: different $\{P_i(\hat{O})\}$ are separated by a minimum distance γ

Figure 2: Separation (4-class Classification Example)

Theory

- We proved: if the predictions induced by different labels are **separated**, then the original classification problem will be learnable (fig. 2).
- We also derived a unified generalization bound for learning with indirect supervisions using the separation degree γ.

Applications

- Our result can be applied to recover the previous results about learnability with noisy and partial labels.
- It will also help us to find and combine supervision signals that is easier to collect and ensures learnability.