



USCViterbi

School of Engineering

Robust and Indirectly Supervised Information Extraction

Muhao Chen

Department of Computer Science / Information Sciences Institute
University of Southern California



How do we make IE models *more reliable*?

Information Extraction (IE): A Fundamental Problem of NLP



The process of automatically inducing structural information (about concepts and their relations) from unstructured text

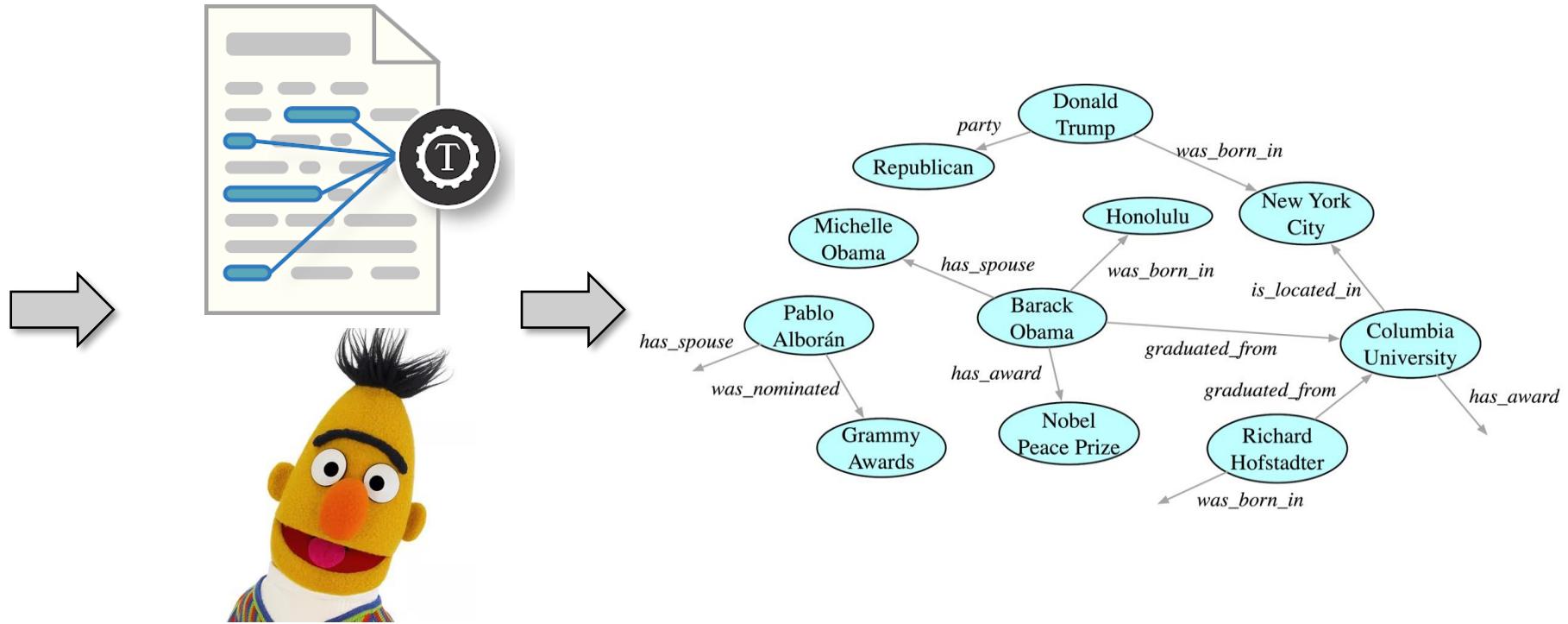
Honolulu

From Wikipedia, the free encyclopedia

This article is about the largest city and state capital city of Hawaii. Honolulu itself, see [Honolulu County, Hawaii](#). For other uses, see

Honolulu (/ha'ne lu:lu:/; [7] Hawaiian: [hono'lulu]) is the capital and largest city of the U.S. state of Hawaii, which is located in the Pacific Ocean. It is an unincorporated county seat of the consolidated City and County of Honolulu, situated along the southeast coast of the island of Oahu,[8] and is the westernmost and southernmost major U.S. city. Honolulu is Hawaii's main gateway to the world. It is also a major hub for international business, finance, hospitality, and military defense in both the state and Oceania. The city is characterized by a mix of various Asian, Western, and Pacific cultures, as reflected in its diverse demography, cuisine, and traditions.

Honolulu means "sheltered harbor"[9] or "calm port" in Hawaiian;[10] its old name, **Kou**, roughly encompasses the area from Nuuanu Avenue to Alakea Street and from Hotel Street to Queen Street, which is the heart of the present downtown district.[11] The city's desirability as a port accounts for its historical growth and importance in the Hawaiian archipelago and the broader Pacific region. Honolulu has been the capital of the Hawaiian Islands since 1845, first of the independent Hawaiian Kingdom, and after 1898 of the U.S. territory and state of Hawaii. The city gained worldwide recognition following Japan's attack on nearby Pearl Harbor on December 7, 1941, which prompted decisive entry of the U.S. into World War II; the harbor remains a major naval base, hosting the U.S. Pacific Fleet, the world's largest naval command.[12]



IE Model/System



IE is Integral to Natural language Understanding

Understanding text depends on the ability to extract information from it

- › Identifying and contextualizing
 - » entities,
 - » quantities (and their scope),
 - » events,
 - » relations, etc.
- › Inferring the identities of concepts



- › Answering questions about the text

In the first quarter, the Bears scored the longest touchdown pass of the game? Who scored the longest touchdown pass of the game?
Greg Olsen. ... In the third quarter, the ... back Adrian Peterson's 1-yard touchdown run. The **Bears** increased their lead over the **Vikings** with Cutler's 2-yard TD pass to tight end **Desmond Clark**. The gap was reduced when **Evan** fired a 6-yard TD pass to **tight end Visanthe Shiancoe**. The Vikings ... with Adrian Peterson's second 1-yard TD run. The Bears ... responded with Cutler firing a 20-yard TD pass to wide receiver ... Bennett. The Bears then won on Jay Cutler's game-winning 39-yard TD pass to wide receiver Devin Aromashodu.

She reports worse seizures, 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,...

Mayor Rahm Emanuel How much did his challengers raise? toward his bid for a third term – more than five times the total raised by his 10 challengers combined, campaign finance records show.

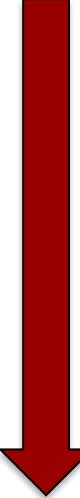
The COVID-19 pandemic in the United States is part of the worldwide pandemic of coronavirus disease 2019 (COVID-19). As of October 2020, there were more than **9,000,000 cases** and 230,000 COVID-19-related deaths in the U.S., representing **20%** of the world's known COVID-19 deaths, and the most deaths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

IE Benefits For Content Management (DARPA KMASS Project)

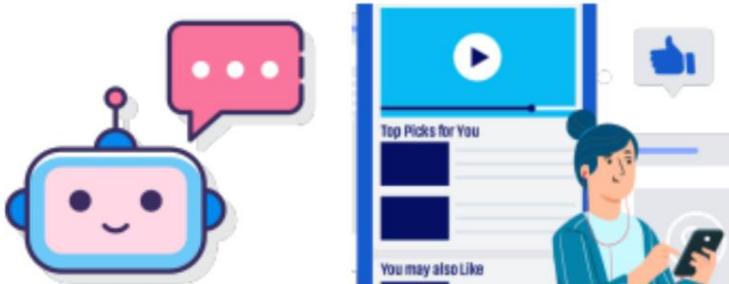
\$6.64M



Extracting structures about tasks, steps and concepts



A consolidated semantic index



Timely *in-context* content delivery in HCI

Deep Learning: Feedforward Neural Networks

The feedforward neural network is the simplest type of artificial neural network which has lots of applications in machine learning. It was the first type of neural network ever created, and a firm understanding of this network can help you understand the more complicated architectures like convolutional or recurrent neural nets. This article is inspired by the [Deep Learning Specialization course](#) of Andrew Ng in Coursera, and I have used a similar notation to describe the neural net architecture and the related mathematical equations. This course is a very good online resource to start learning about neural nets, but since it was created for a broad range of audiences, some of the mathematical details have been omitted. In this article, I will try to derive all the mathematical equations that describe the feedforward neural net.

The architecture of neural networks

The leftmost layer in this network is called the input layer, and the neurons within the layer are called input neur. The rightmost or output layer contains the output neurons, or, as in this case, a single output neuron. The middle layer is called a hidden layer, since the neurons in this layer are neither inputs nor outputs.

Regularization in Deep Learning

Regularization is a set of techniques that can prevent overfitting in neural networks and thus improve the accuracy of a Deep Learning model when facing completely new data from the problem domain. In this article, we will address the most popular regularization techniques which are called L1, L2, and dropout...

Procedural Index

Training a Feed-Forward Neural Network

Procedural Goal: Training a FFNN

Data Preprocessing

TempRel

Design Neural Architecture

Weight Initialization

Optimizing Parameters

Procedural Goal: Data Preprocessing

Data Collection

Data Cleaning

Data Augmentation

Procedural Goal: Optimizing Parameters

Selecting Optimizers

Tuning Learning Rates

Adding Regularizers

Factual Index



Consolidated Knowledge Nuggets

Skewed Initialization

Contradicts

Xavier/Glorot Initializer

HasProperty

Dimension Orthogonality

Causes

Regularizer

Factual

Preventing Overfitting

Conf: 0.90

Conf: 0.65

Conf: 0.45

Conf: 0.87

Conf: 0.40

Conf: 0.87

Conf: 0.65

Consolidated Knowledge Nuggets

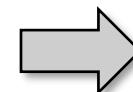


IE Is the Backbone of Any Knowledge-driven Tasks



WIKIPEDIA The New York Times

PubMed bioRxiv



Knowledge Bases



Bio/Med Databanks



Knowledge Representation



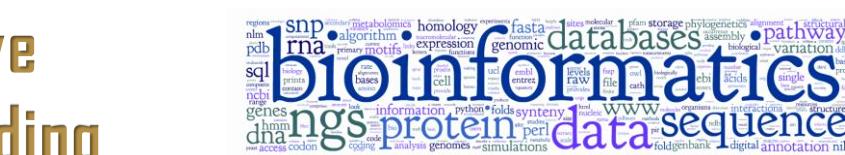
Commonsense QA

Event Prediction

Intent Prediction



Storytelling Content Selection Newsworthiness Detection



Proteomic Interaction Prediction Mutation Effect Estimation Genomic Function Prediction



Medical INFORMATICS

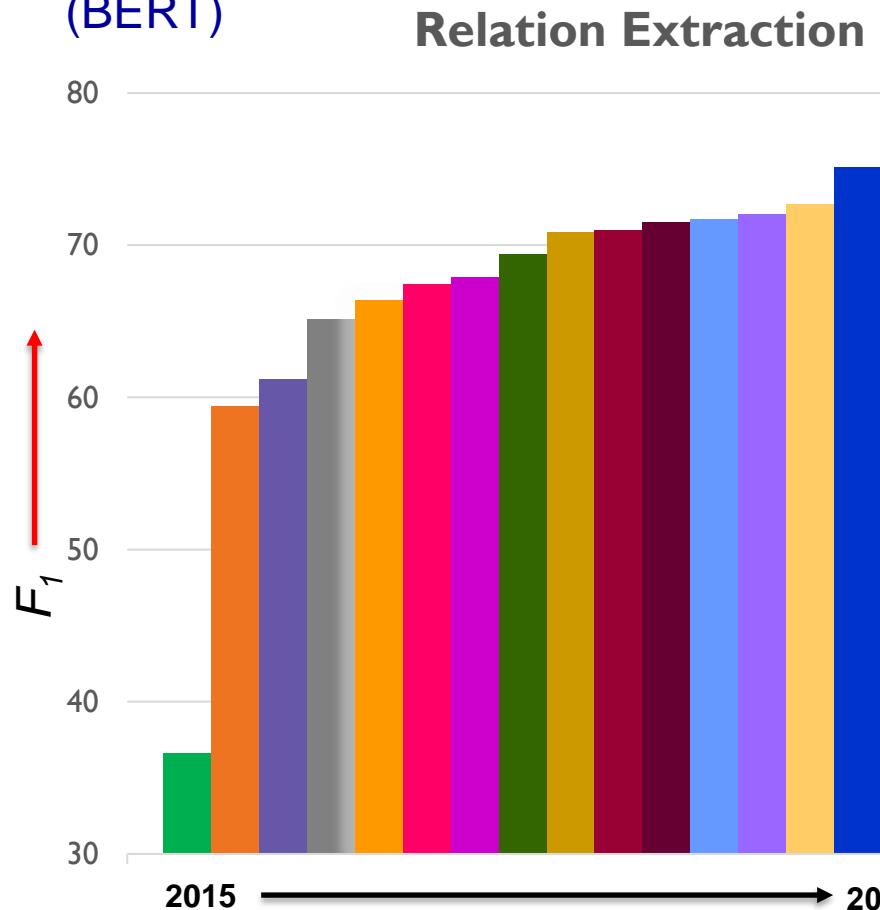


How IE Is Doing Today



On Benchmarks

Rise of the Pre-trained Models



In Reality
IE is still brittle.
for two nights when I was in Las Vegas.

I stayed in Treasure Island
LOC

Type?



Island X



limitedly generalizable,



and costly to develop.



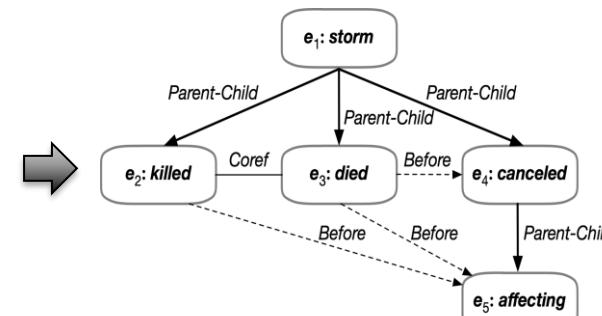
AIDA, KAIROS, BETTER, LwLL,
KMASS, GAIA, ECOLE...: all
costing *tens of millions \$*.



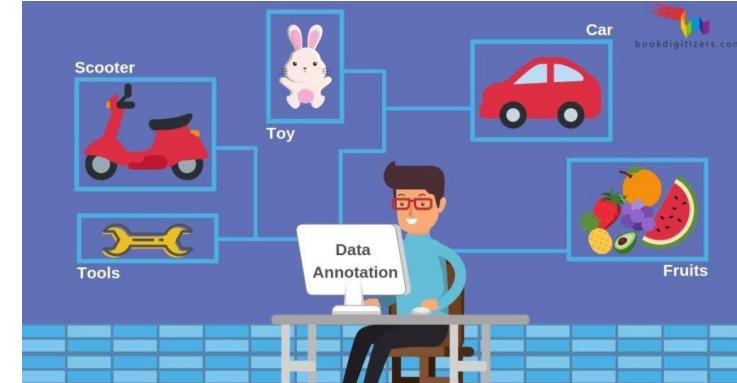
Challenge: Expensive Supervision

Obtaining direct supervision for IE is **difficult and expensive**

On Tuesday, there was a typhoon-strength ($e_1:\text{storm}$) in Japan. One man got ($e_2:\text{killed}$) and thousands of people were left stranded. Police said an 81-year-old man ($e_3:\text{died}$) 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 ($e_4:\text{canceled}$) 230 domestic flights, ($e_5:\text{affecting}$) 31,600 passengers.



Reading long documents, recognizing complex structures



Costs \$2-\$6 and >3 minutes for just 1 relation [Paulheim+ 2018]

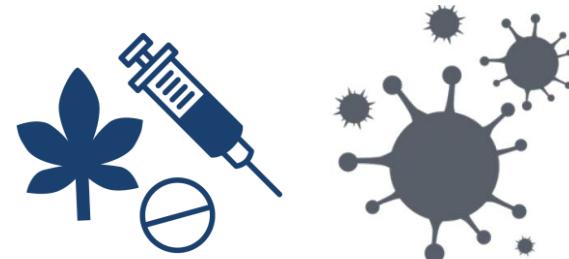
Insufficiency

- **General domain:** A few hundred documents or ten thousand scale sentences with annotation
- **Specific domain:** Up to several thousand sentences.

Noise

- **In-correct labels:** e.g. 5-9% errors in TACRED, CoNLL03, DocRED
- **Low agreement:** <70% IAA in HiEve, Intelligence Community, etc.

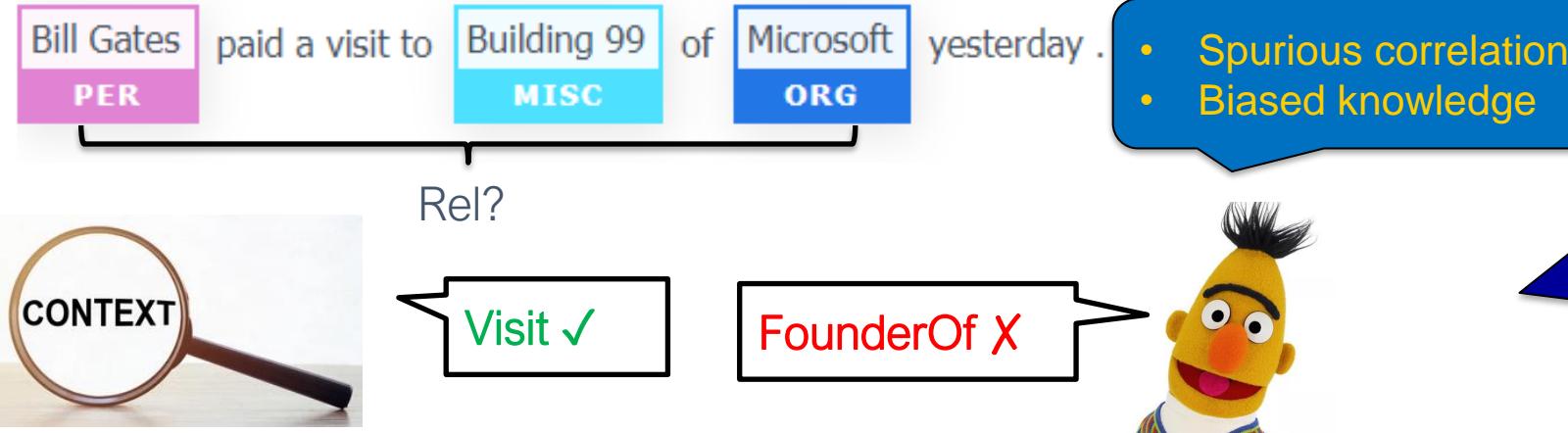
Low-resource Domains with Almost No Annotations





Challenge: Accountability

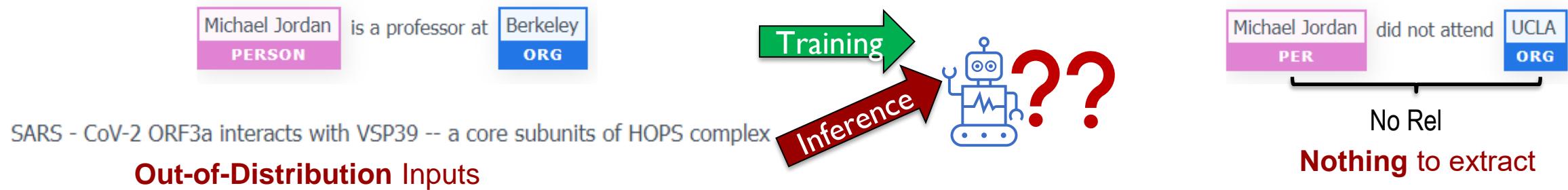
Making *Faithful* Extraction



- Harmful in many scenarios
- Extracting drug information
 - Extracting disease phenotypes
 - Extracting disaster events
 - Software version compatibility

Knowing the Decision Boundary

Real-world application often exposes much more diverse inputs, **with lots of exceptions**, to IE systems.



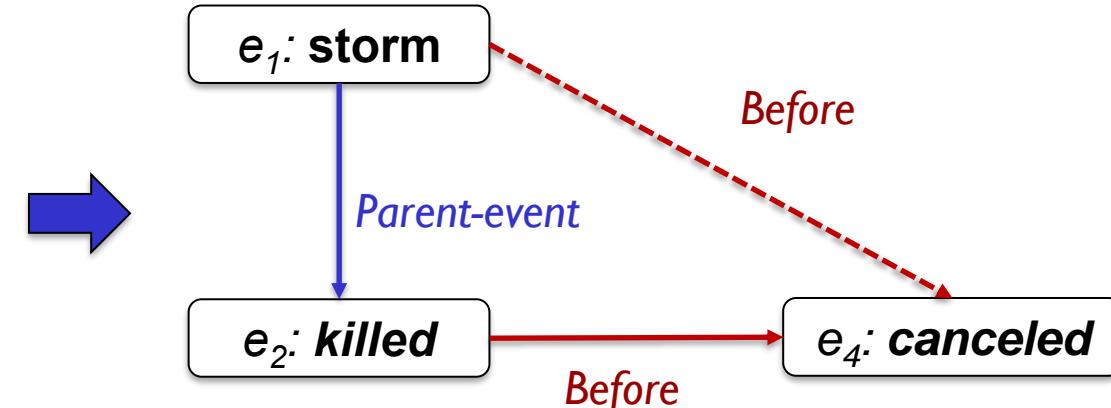


Challenge: Consistency

Extracts are interdependent decisions.

An article about a storm hazard in Japan

On Tuesday, there was a typhoon-strength ($e_1:\text{storm}$) in Japan. One man got ($e_2:\text{killed}$) and thousands of people were left stranded. Police said an 81-year-old man ($e_3:\text{died}$) 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 ($e_4:\text{canceled}$) 230 domestic flights, ($e_5:\text{affecting}$) 31,600 passengers.



Extraction Should be Globally Consistent

Symmetry: $e_3:\text{died}$ is BEFORE $e_4:\text{canceled} \Rightarrow e_4:\text{canceled}$ is AFTER $e_3:\text{died}$

Conjunction: $e_3:\text{died}$ is BEFORE $e_4:\text{canceled} \wedge e_4:\text{canceled}$ is a PARENT EVENT of $e_5:\text{affecting} \Rightarrow e_3:\text{died}$ BEFORE $e_5:\text{affecting}$

Implication, Negation ...

A BERT-based model getting 90% of correct pairwise decision still violates 46% of triplet constraints [Li et al. ACL-20]

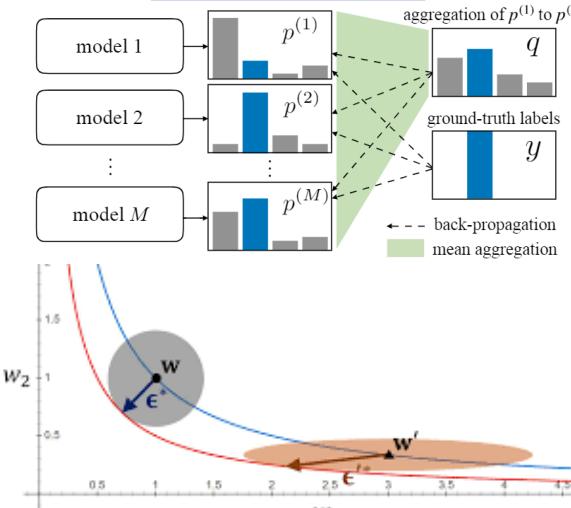
How do we **enforce logical constraints** for consistent/self-contained IE?
How do we **discover the constraints**?

Our Goal: More Reliable IE



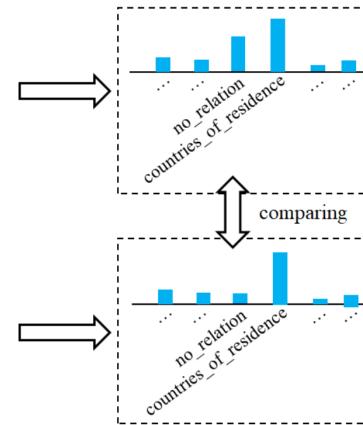
Training

Robustness



Proactively Mitigating Training Noise

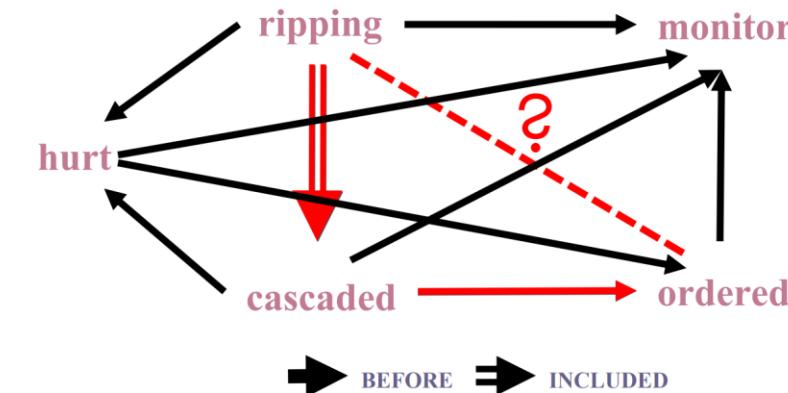
Eugenio Vagni, the Italian worker of the ICRC, Andreas Notter of Switzerland, and Mary Jean Lacaba of the Philippines were released by their Abu Sayyaf captors separately.



Inference

Mitigating Spurious Correlations

Indirect Supervision

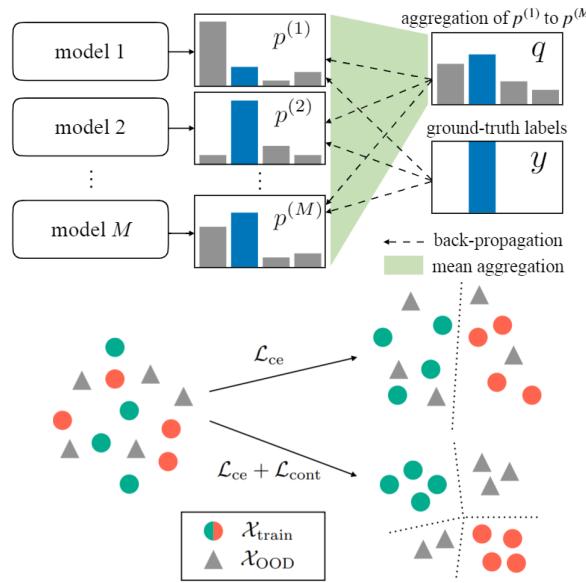


Logical Constraints and Cross-task Transfer

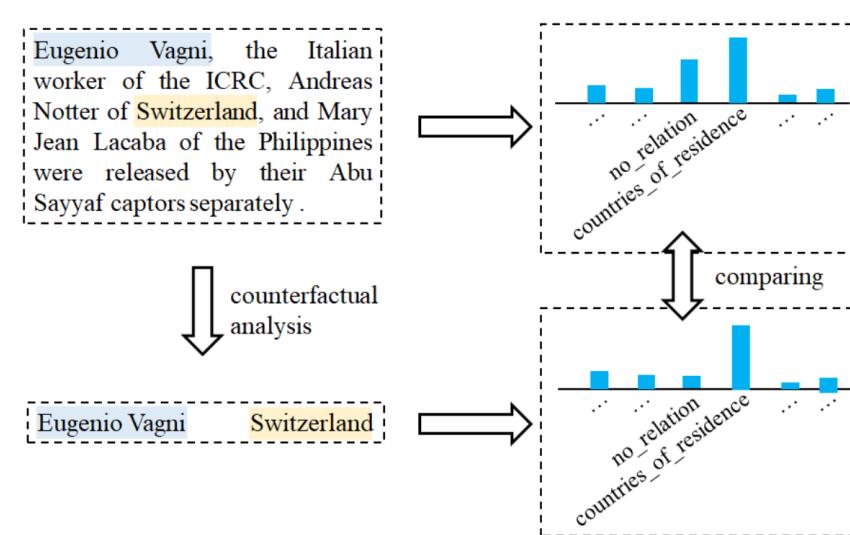
In This Talk



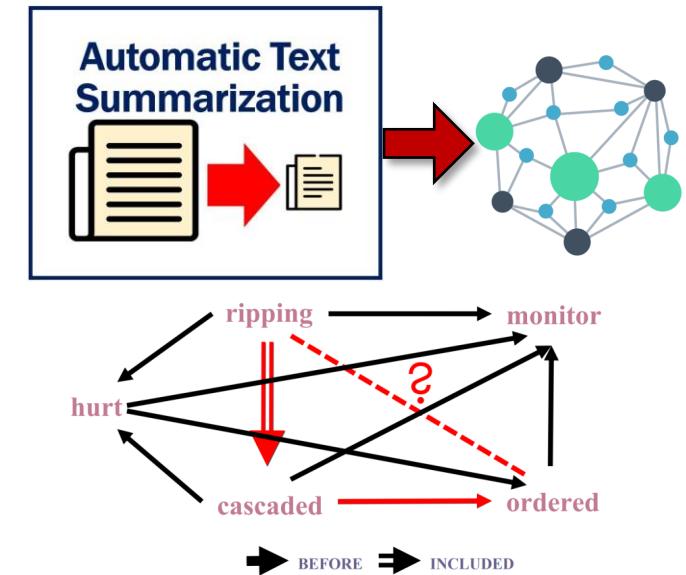
1. Noise-robust IE



2. Faithful IE



3. Indirectly Supervised IE



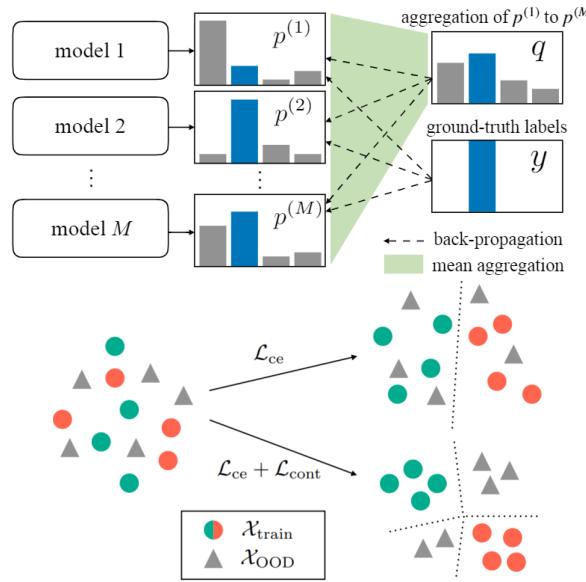
4. Future Directions



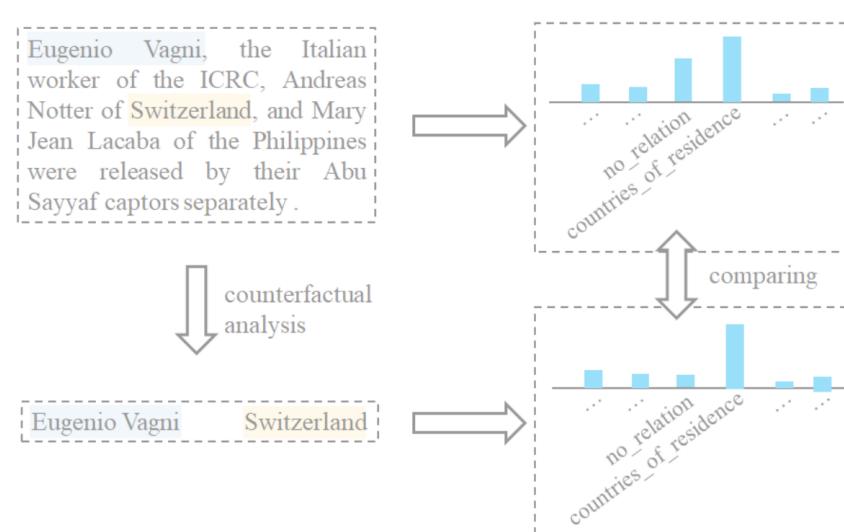


Agenda

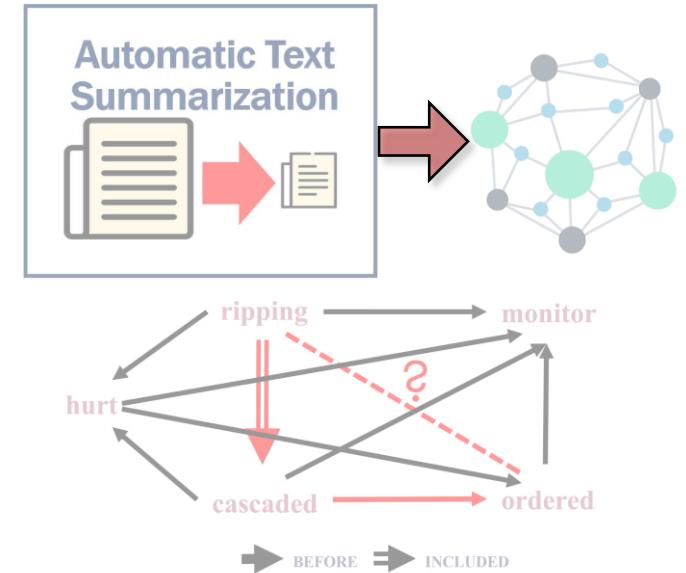
1. Noise-robust IE



2. Faithful IE



3. Indirectly Supervised IE



4. Future Directions

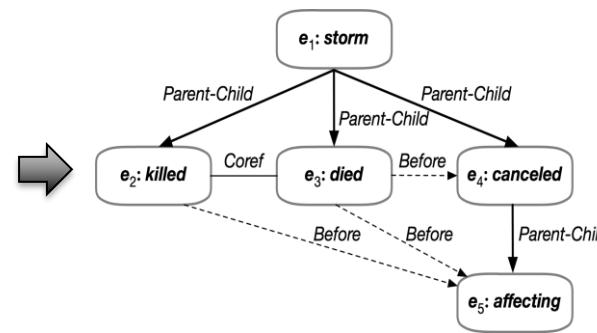




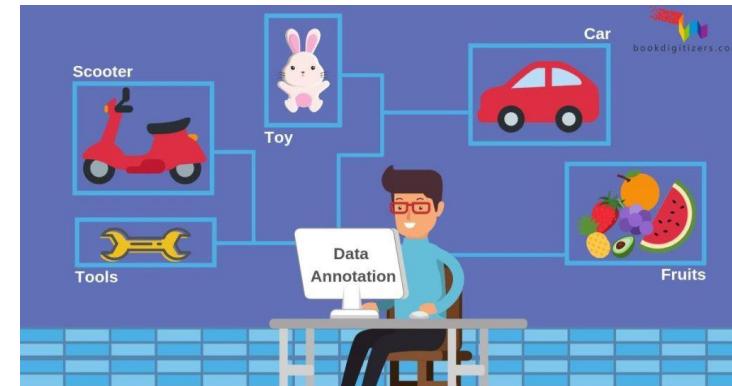
Imperfect Supervision

On Tuesday, there was a typhoon-strength ($e_1:\text{storm}$) in Japan. One man got ($e_2:\text{killed}$) and thousands of people were left stranded. Police said an 81-year-old man ($e_3:\text{died}$) 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 ($e_4:\text{canceled}$) 230 domestic flights, ($e_5:\text{affecting}$) 31,600 passengers.

Annotation for IE is **difficult** and **expensive**



Reading long documents, annotating complex structures



Requiring time and effort of annotators with expert knowledge

Hence, annotations are **inevitably noisy** (even in most popular benchmarks)

- 5-8% errors in TACRED and CoNLL03
- 9% errors in DocRED
- <70% inter-annotator agreement in HiEve, Intelligence Community, etc.

...

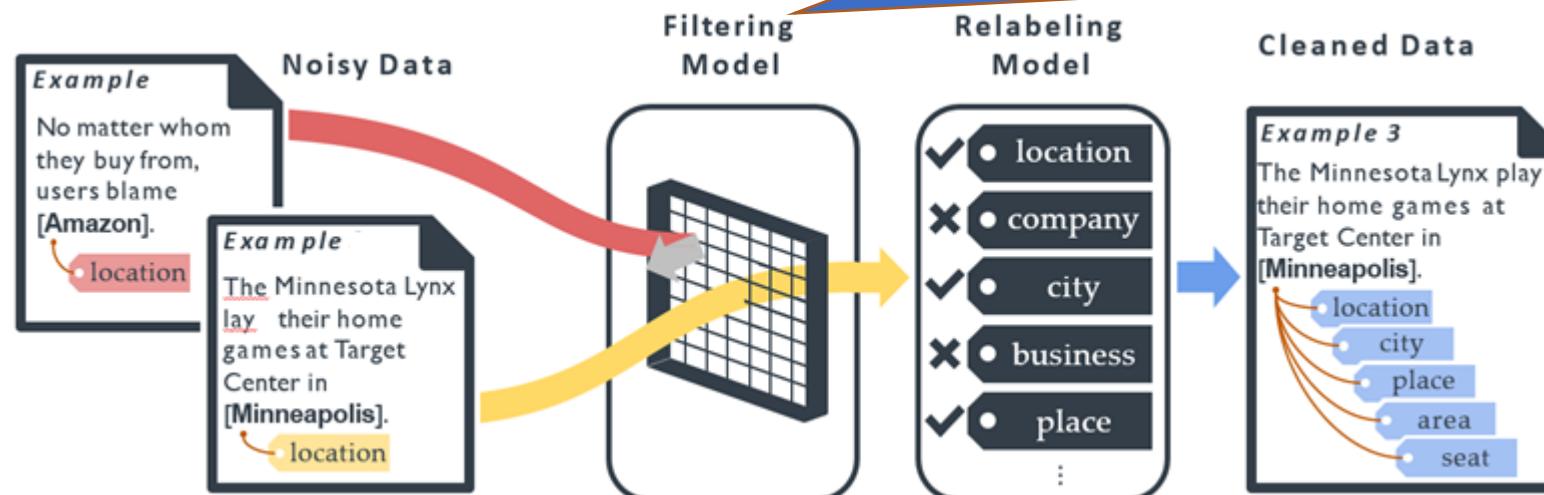
Even moderate training noise leads to significant flaws

High-performance IE models must be achievable under ***noisy supervision***



A Glance at Prior Solutions

Supervised Denoising

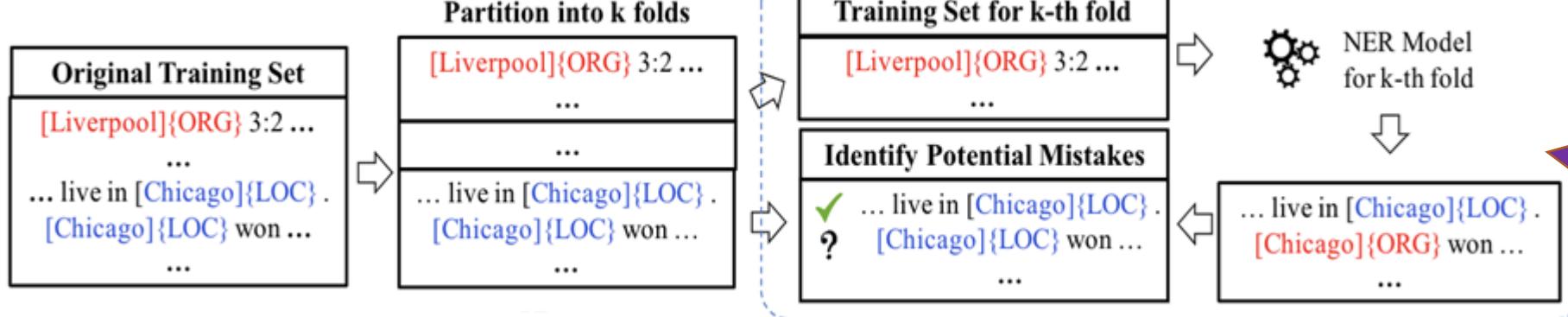


Noise filtering and relabeling models trained with **labeled clean data**.

Cost: manually labeling enough clean data is nowhere cheaper.

Ensemble-based Denoising

Reweighting k-folds of data based on **cross-validation**.

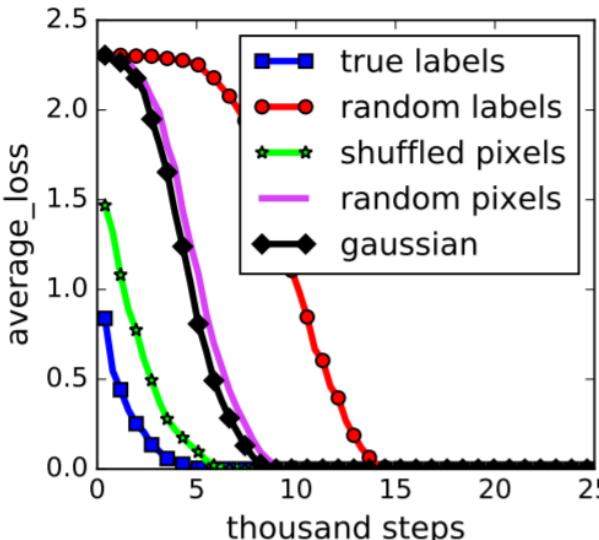


Coarse-grained noise estimation and **redundant** training and testing.

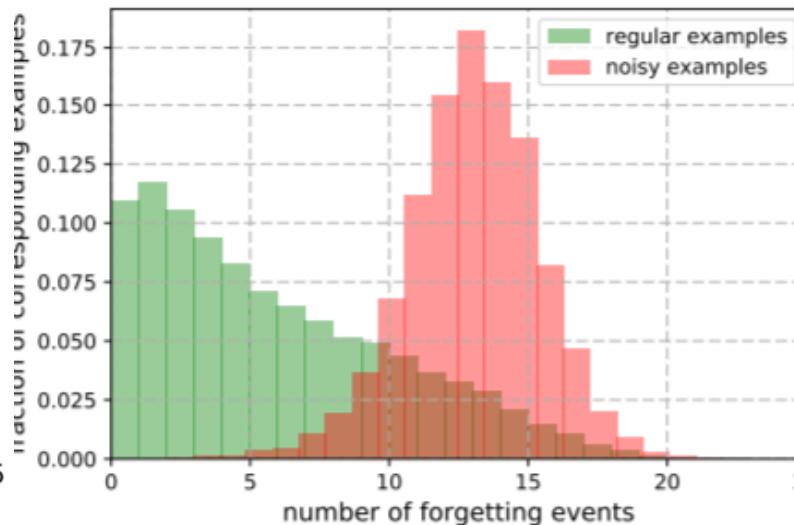
Wang et al. *CrossWeigh: Training named entity tagger from imperfect annotations*. **EMNLP 2019 (UIUC)**



Unsupervised Denoising: Co-regularized Knowledge Distillation

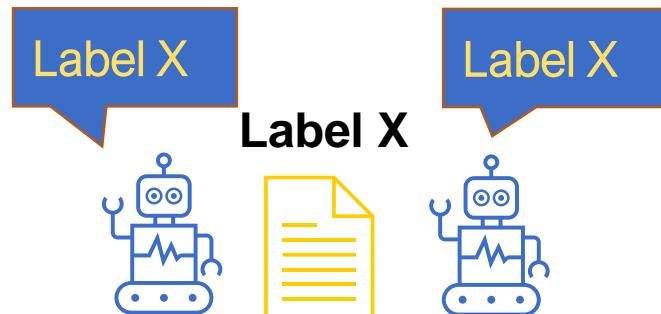


(1)



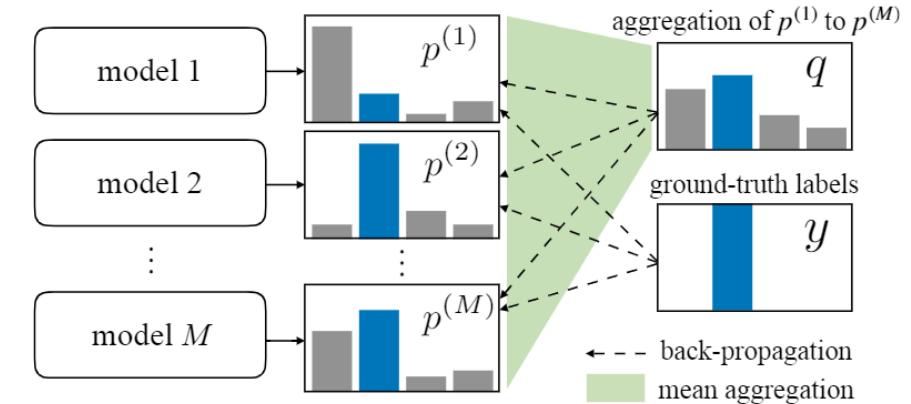
(2)

Noisy labels lead to delayed learning curves [Toneva+ ICLR-19]

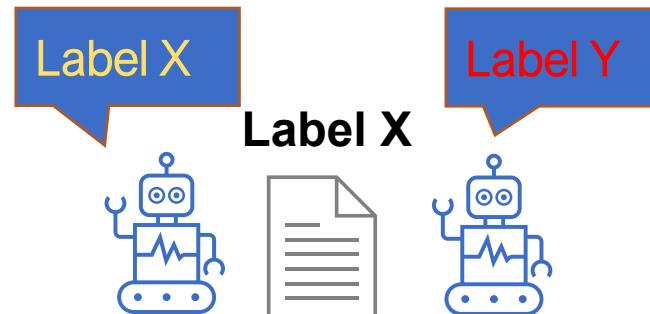


Mutual agreement by models indicates clean/noisy labels

Noisy instances are outliers to the task inductive bias.

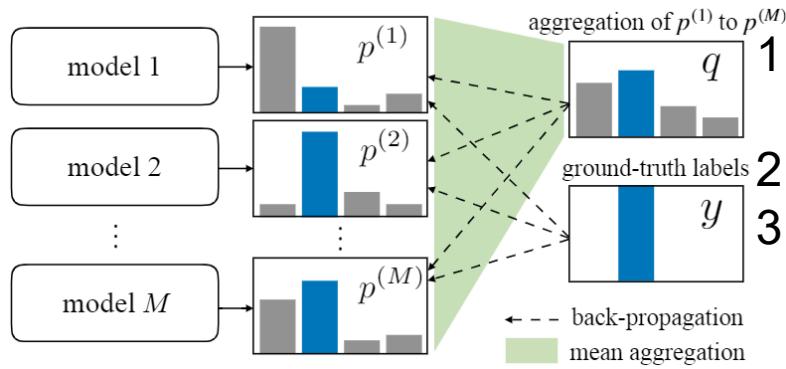


Co-regularization Framework





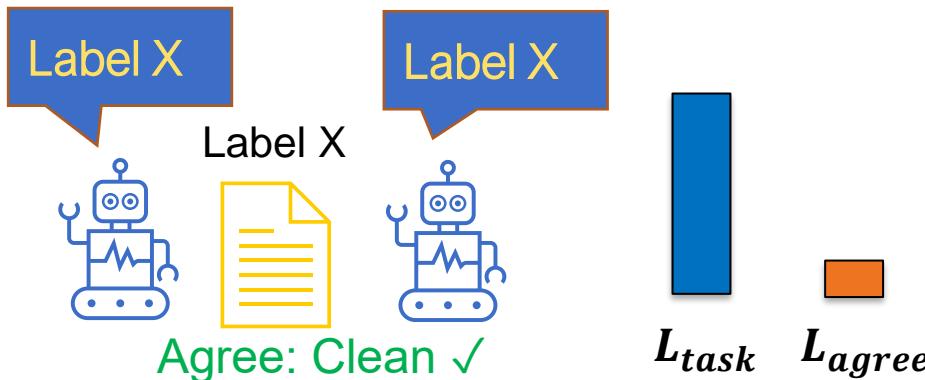
Unsupervised Denoising: Co-regularized Knowledge Distillation



1. Create $M (\geq 2)$; 2 is enough) identical models with **different initialization** (w/ **warm up**).
2. Train the models with both **the task loss** and an additional **agreement loss**.
3. Return one of the models.

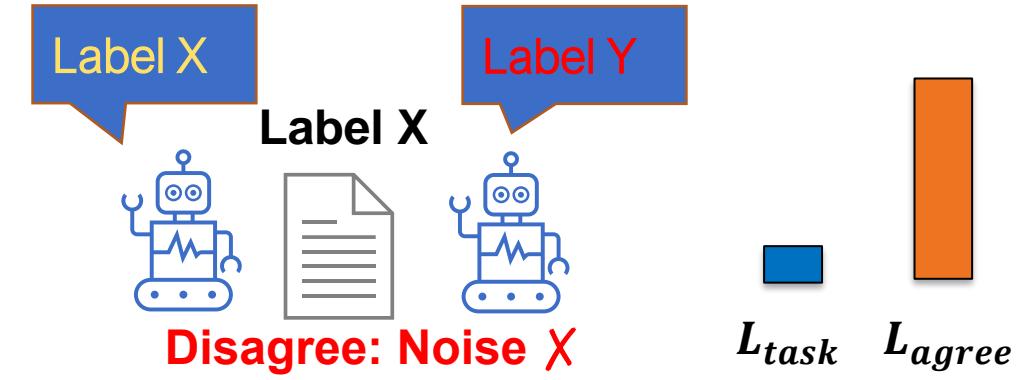
Cross-entropy L_{task}

K-L divergence between
model predictions L_{agree}



On clean data

- Lower agreement loss
- Focusing on task optimization



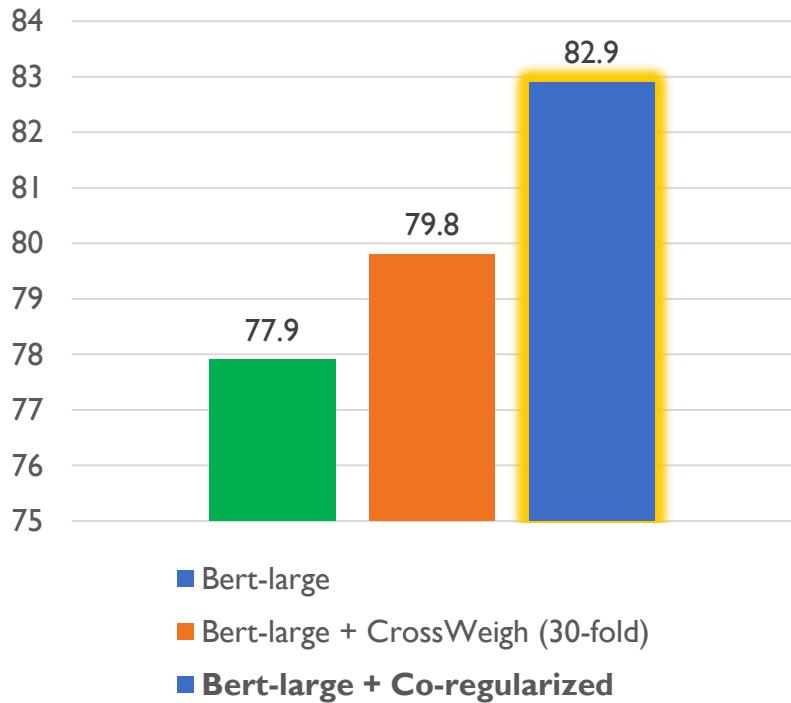
On noisy data

- Higher agreement loss
- Task optimization proactively prevents fitting those data

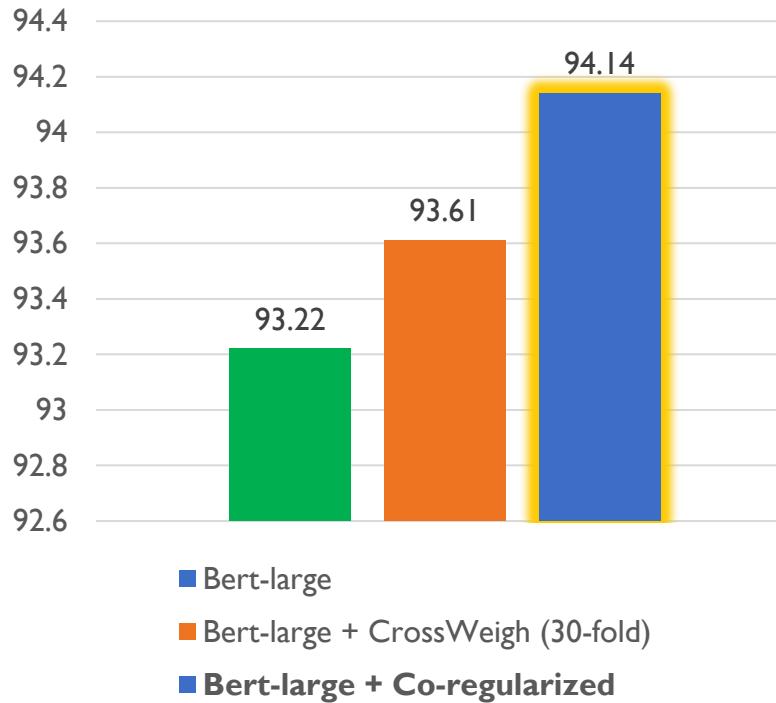
Unsupervised Denoising: Co-regularized Knowledge Distillation



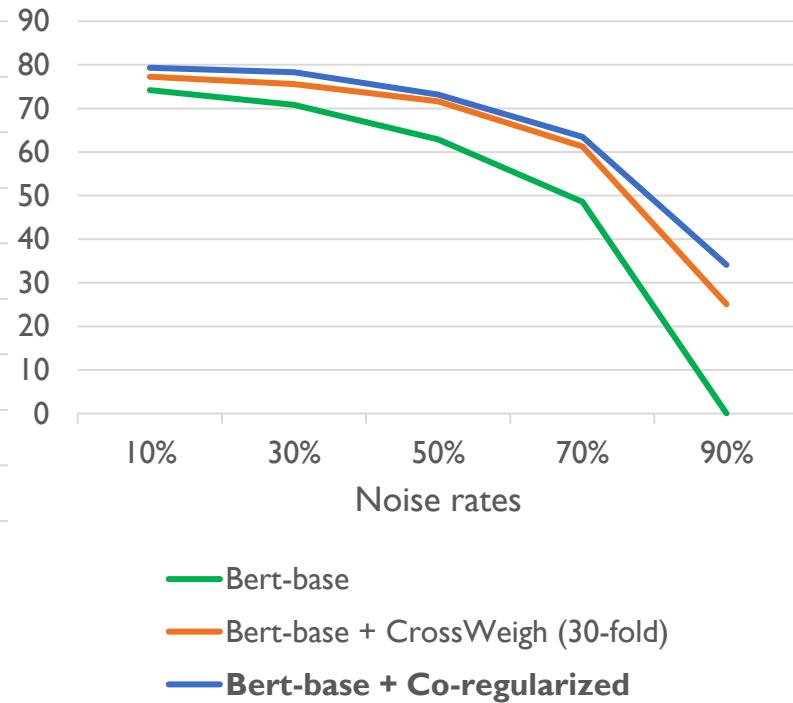
Relation Extraction (F1) on TACREV
(~8% training noise)



NER (F1) on Relabeled CoNLL-03
(~5.4% training noise)



Relation Extraction (F1) on TACREV
(varied noise rate via label flipping)

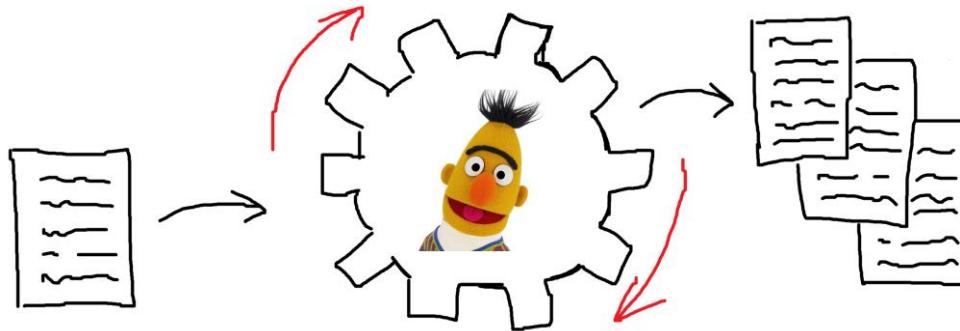


Merits of co-regularized knowledge distillation

- More robust than ensemble (e.g. CrossWeigh), especially when noise rates are higher
- More efficient (no redundant training/inference pass) and fine-grained denoising (instance-level)
- Applicable to any backbone IE models (see results w/ LUKE and C-GCN in the paper)



Robust Data Augmentation

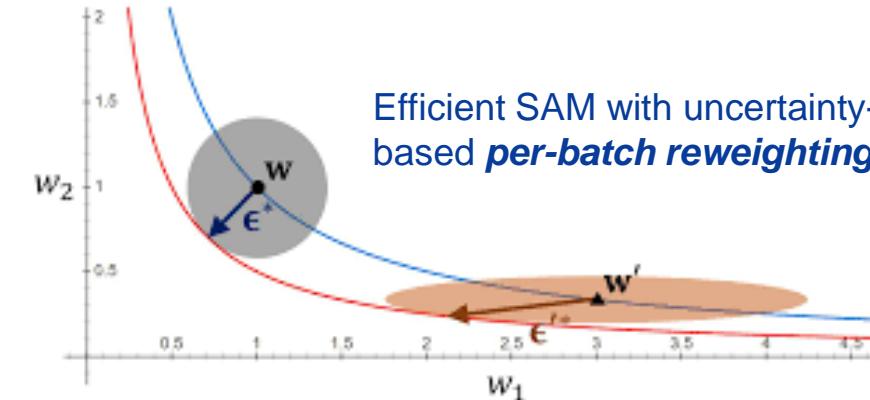


Denoising automatically augmented training data

- Long-term uncertainty measure
- Consistency training between original data and the augmentation

3.6-5.6% improvement on edit-based augmentation (EDA).
2.7-4.6% improvement on CSQA with generative data augmentation (G-DAUG).

Perturbation Robustness



δ -SAM: fast adversarial parameter perturbation [EMNLP-22]

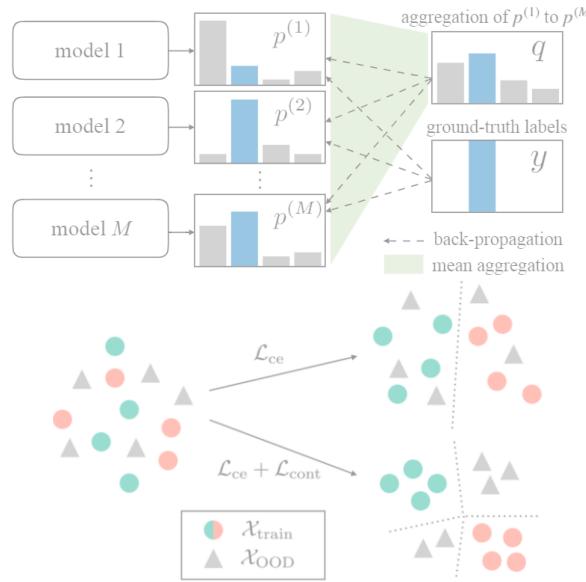
- Improving model robustness by finding flatter minima
- Consistent improvement on [IE](#), [textual retrieval](#), [summarization](#), and [NLU](#) tasks

Theoretically principled reweighting efficiently approximates *per-instance* adversarial perturbation.

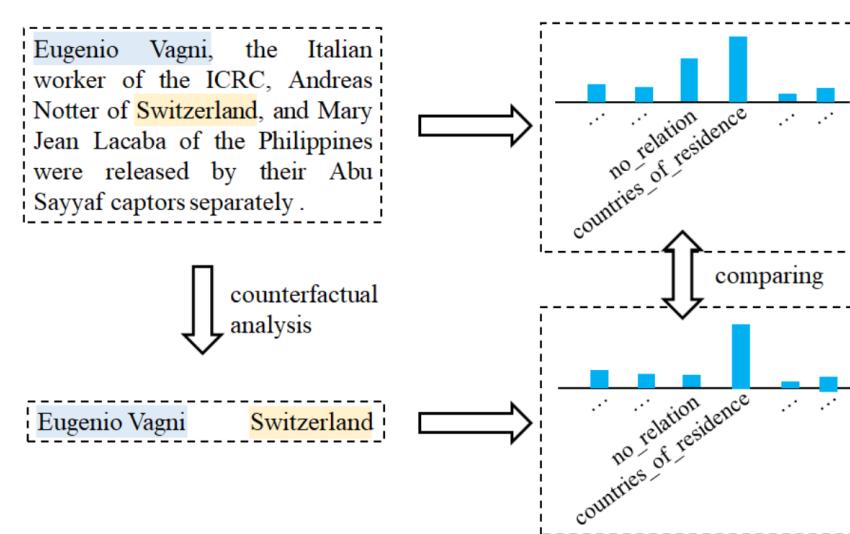


Agenda

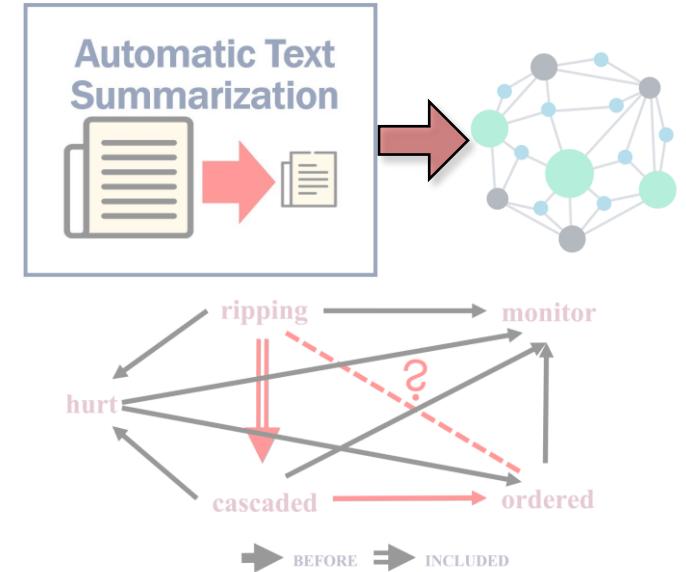
1. Noise-robust IE



2. Faithful IE



3. Indirectly Supervised IE



4. Future Directions





Faithfulness Issues

IE systems may not **faithfully** extract what is described in the **context**

Entity relation extraction:



Visit ✓

FounderOf X



According to prior knowledge

Event relation extraction:



Prior knowledge (in PLMs) may lead to unfaithful extraction

I went to see the doctor event1 However, I got more seriously sick. event2

Before? After?

```
graph LR; S1[see the doctor] --> S2[sick]; S1 --- RQ[Before? After?];
```

Before ✓

After X



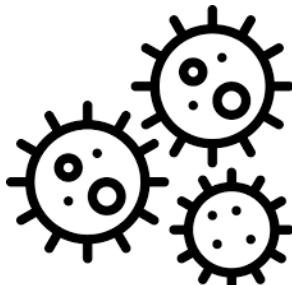
According to statistics

(Statistically) Biased training may lead to unfaithful extraction

Why Faithful IE Is So Important



Drug-drug Interaction



Disease-target detection



The amount of metformin absorbed while taking Acarbose was bioequivalent to the amount absorbed when taking placebo, as indicated by the plasma AUC values. However, the peak plasma level of metformin was reduced by approximately 20% when taking Acarbose due to a slight delay in the absorption of metformin.

Interaction type: mechanism

TOMM70, the most frequent binding partner of SARS-CoV-2 ORF9b, was identified in more than 1000 PSMs of the prey.

Interaction type: binding

More risky tasks where we couldn't afford any GUESSES from unfaithful IE

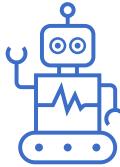
- **Disease phenotype extraction** from medical reports
- **Disaster event extraction** from social media
- **API version compatibility** detection from software documents
- **Travel event extraction** from emails and meeting logs
- ...



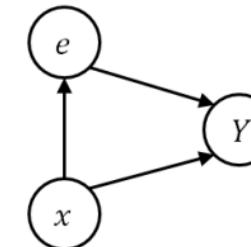
Spurious Correlation: Take Relation Extraction as An Example

What we hope the IE model to do

Bill Gates paid a visit to Building 99 of Microsoft yesterday.



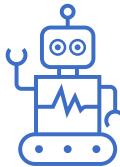
Comprehend the *context*, and induce the mentioned *relation* of *entities*.



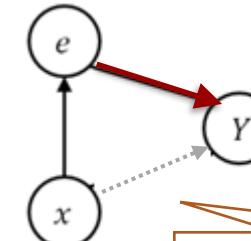
Relations should be inferred based on both mentions and the context

What it may actually do

Bill Gates paid a visit to Building 99 of Microsoft yesterday.



Read the *entities* and guess the *relation* without referring to the *context*.



Relation prediction is no longer attributed to the context.

Overly relying on entity mentions lead to a shortcut for RE

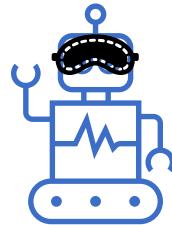
How do we mitigate the entity-relation spurious correlation?



Mention masks: mask out entity names with their types

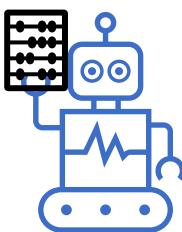
Person paid a visit to Building 99 of Org yesterday.

Similarly for event *RE*, we can mask using trigger types and tense



Mask mentions in both training and inference

- Pro: reduces mention biases
- Con: loses semantic information about entities \Rightarrow performance drop



Instance reweighting: FoCal loss, two-stage optimization, etc.

$$\text{FL}(p_t) = -(1 - p_t)^\gamma \log(p_t)$$

Upweight hard instances

- Pro: reduces training biases by (indirectly) upweighting some “underrepresented” instances
- Con: hard instances are not always “underrepresented” instances

Our Strategy: Counterfactual Inference



Distilling model biases with causal intervention [Pearl and Mackenzie, 2018]

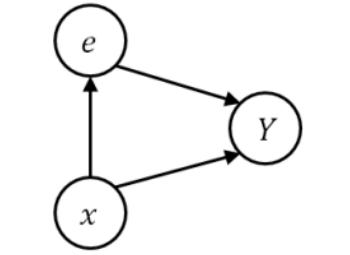
Instance: Bill Gates paid a visit to Building 99 of Microsoft yesterday.

We view the Inference as $f(X) = P(Y|X) = \sum_E P(Y|X, E)P(E|X)$

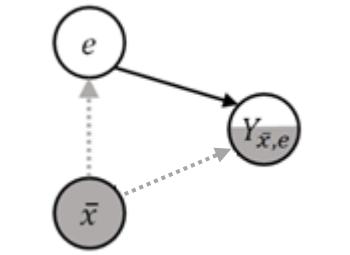
Counterfactual instance: Bill Gates Microsoft

Distilling the entity bias $f(e) = P(Y|do(X)) = P(Y|(X = e))$

- Fix the mediator value e (entity mentions)
- Intervention operation $do(X) = e$ creates a **counterfactual** that wipes out the entire context



Original Prediction



Entity-specific Bias

The difference: entity-debiased prediction $f(X) \setminus f(e)$

- Estimating the *Natural Direct Effect* [Pearl and Mackenzie, 2018] from X to Y



Our Strategy: Counterfactual Inference

Distilling model biases with causal intervention [Pearl and Mackenzie, 2018]

Instance: Bill Gates paid a visit to Building 99 of Microsoft yesterday.

We view the Inference as $f(X) = P(Y|X) = \sum_E P(Y|X, E)P(E|X)$

Counterfactual instance: \emptyset

Distilling the (global) label bias $f(\bar{x}) = P(Y|do(X)) = P(Y|(X = \bar{x}))$

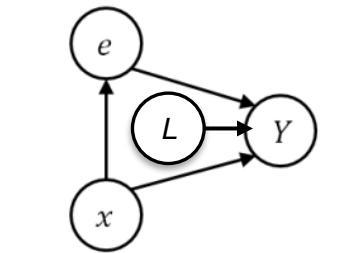
- Intervention operation $do(X) = \bar{x}$ encourages the model to make inference without seeing any input data

The difference: label-debiased prediction $f(X) \setminus f(\bar{x})$

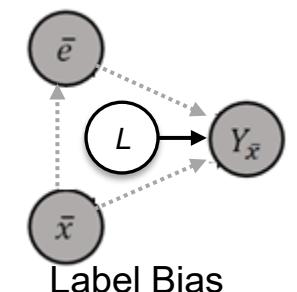
- Also estimates the Total Effect [Pearl and Mackenzie, 2018] of X

The final debiased prediction: $f(X) \setminus f(e) \setminus f(\bar{x})$

- Combining both effects
- Do not need to retrain the model
- Can easily adapt to different data distributions



Original Prediction



Label Bias



Our Strategy: Counterfactual Inference

Deducting the distilled biases from the original prediction

① Original Instance (x)

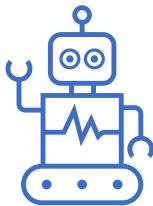
Bill Gates paid a visit to Building 99 of Microsoft yesterday.



Biased prediction Y_x

② Counterfactual instance w/o context (\bar{x}, e)

Bill Gates Microsoft

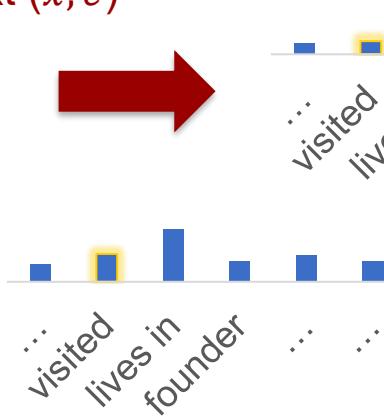


deduct

Entity bias $Y_{\bar{x},e}$

③ Empty counterfactual instance (\bar{x})

\emptyset



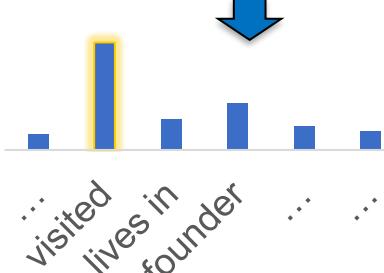
deduct

(Global) label bias $Y_{\bar{x}}$

$$Y_{\text{final}} = Y_x - \lambda_1 Y_{\bar{x},e} - \lambda_2 Y_{\bar{x}}$$

$$\lambda_1^*, \lambda_2^* = \arg \max_{\lambda_1, \lambda_2} \psi(\lambda_1, \lambda_2) \quad \lambda_1, \lambda_2 \in [a, b]$$

Obtained
on dev set

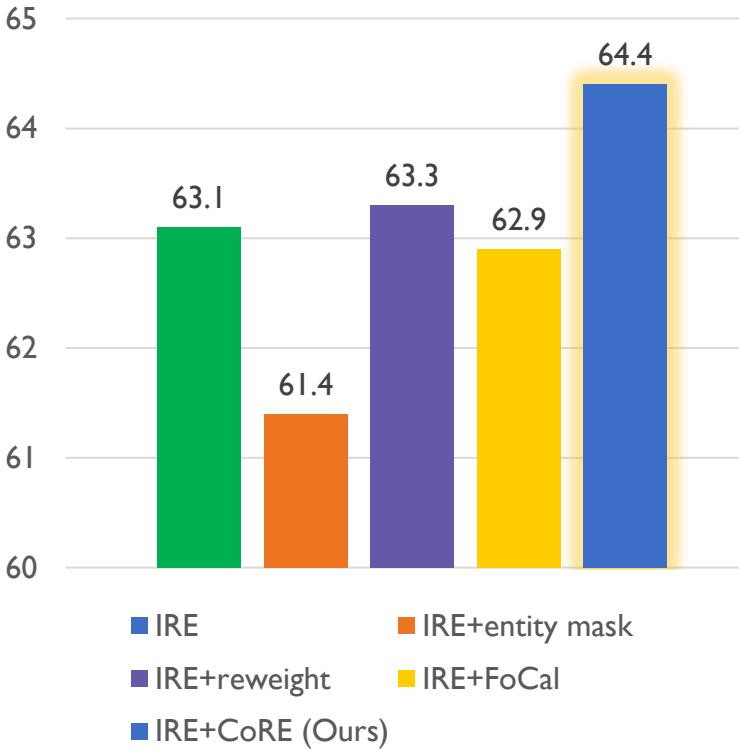


Debiased prediction Y_{final}

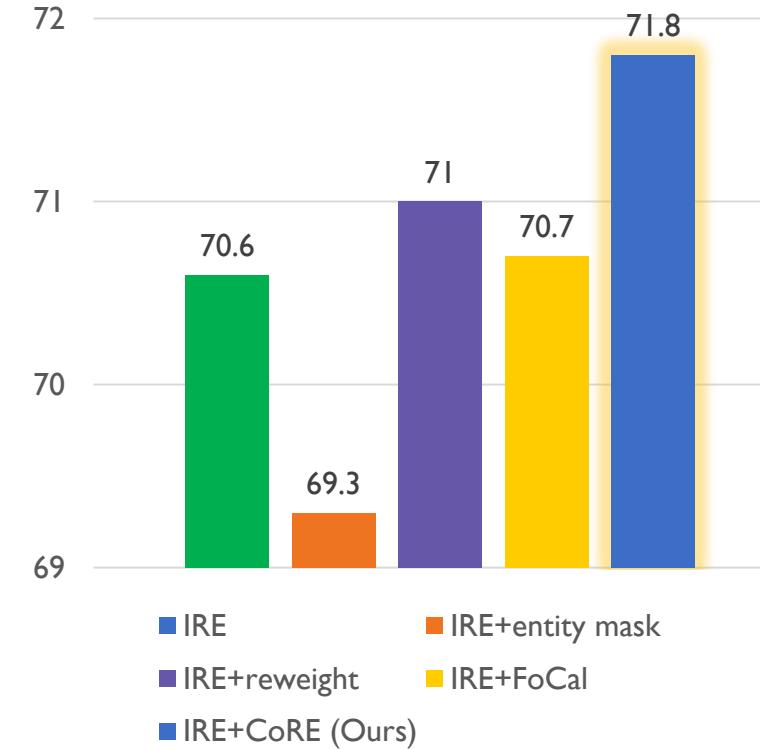
Counterfactual Inference



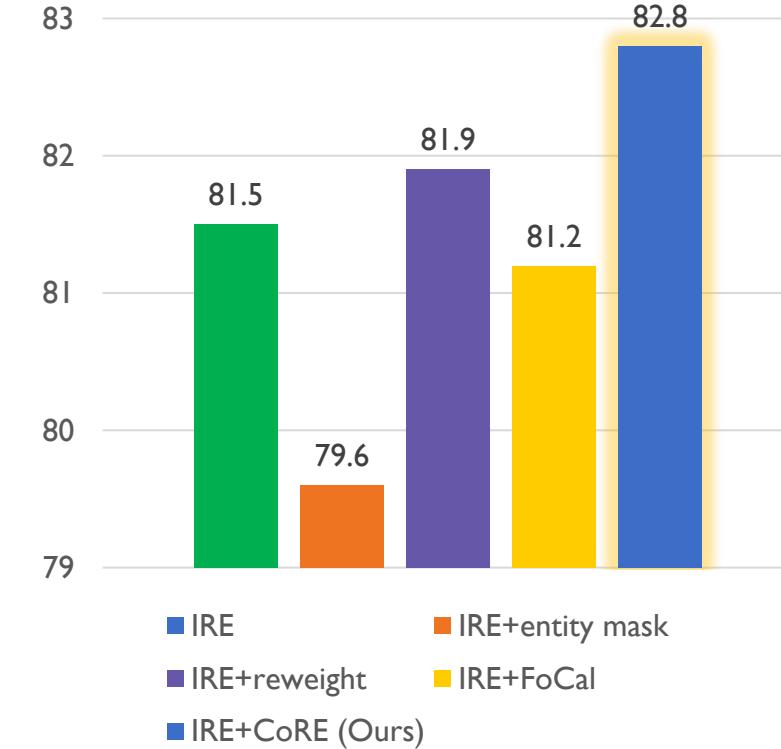
F1-macro on TACRED



F1-macro on TACREV



F1-macro on Re-TACRED



Counterfactual inference can lead to more precise and fairer relation extraction.

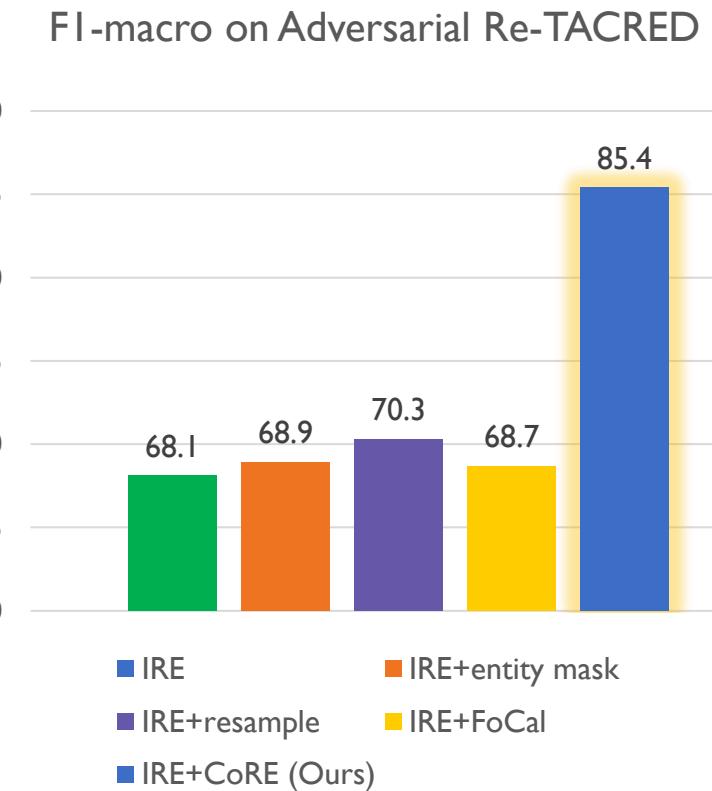
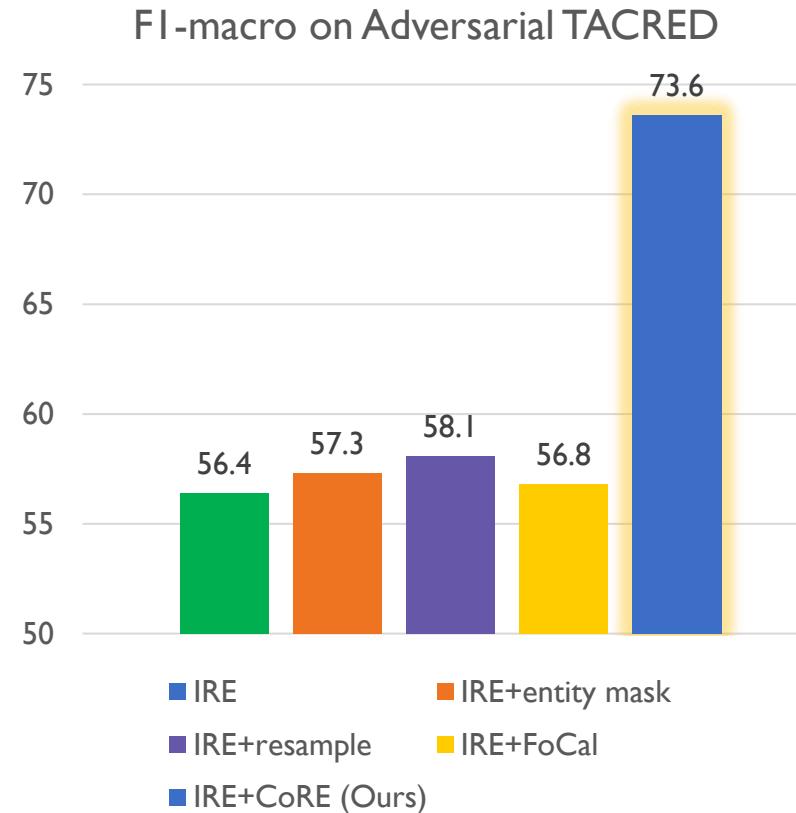
*IRE_{RoBERTa} is one of the best-performing sentence-level RE models ([Z&C ACL 2022](#)). Results are also available for LUKE.

Counterfactual Inference



Evaluation on adversarial TACRED and Re-TACRED.

- Filtered test sets where combinations of entities and relations have not appeared in training sets.
- Models **cannot guess the relations trivially based on entity mentions.**



Significantly more faithful relation extraction shown on OOD examples.



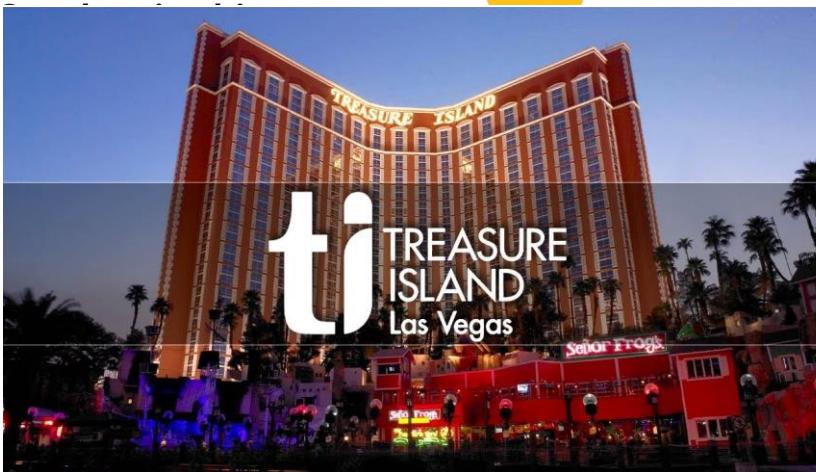
Our Continuing Studies in This New Direction

Faithfulness in IE is still an underexplored research direction.

More Complex Artifacts in Entity Typing
Mention-Context bias

Input: Last week I stayed in *Treasure Island* for two nights when visiting Las Vegas.

Gold labels: hotel, resort, location, place
Pred labels: island, land, location, place



Dependency bias

Input: Most car *spoilers* are made from polyurethane, while some are made from lightweight steel or fiberglass.

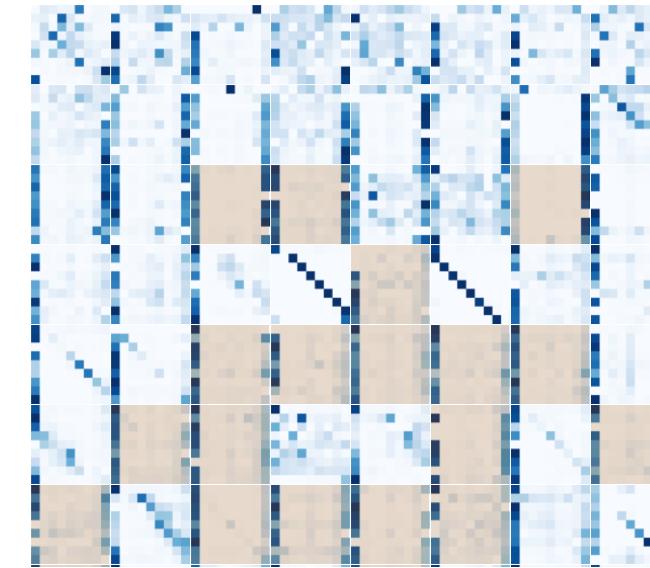
Gold labels: part, object

Pred labels: object, car, vehicle



- + pronoun, lexical overlapping, name frequency, overgeneralization
- Counterfactual data augmentation to address them all

General-purpose Feature Debiasing



Original Attention Flow

Year	Title	Role
1990	Sun-sai	Sun-sai
	Wai	Kit
	Tramp	Wai
		Siu-bo

LATTICE

Year	Title	Role
1990	Justice, my foot!	Sun-sai
	Royal Tramp	Kit
		Wai
		Siu-bo

$P_{ij}=l_j l_i$ ● ● $P_{ij}=P_{\max}$ ● ●

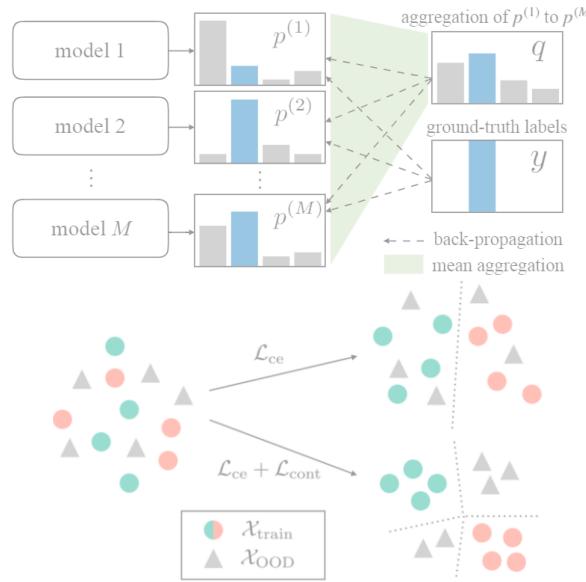
Year	Title	Role
1990	Justice, my foot!	Sun-sai
	Royal Tramp	Kit
		Wai
		Siu-bo

- Attention smoothing, perturbation, constrained PoE
- Feature-equivariance learning

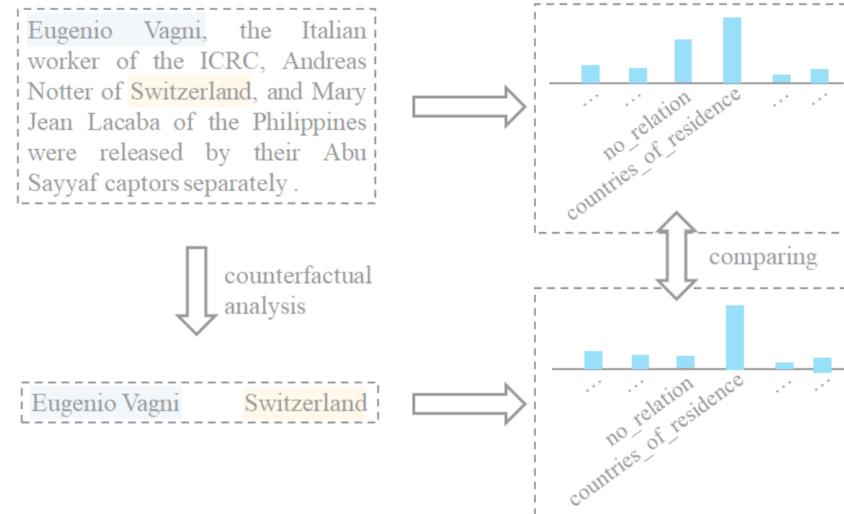


Agenda

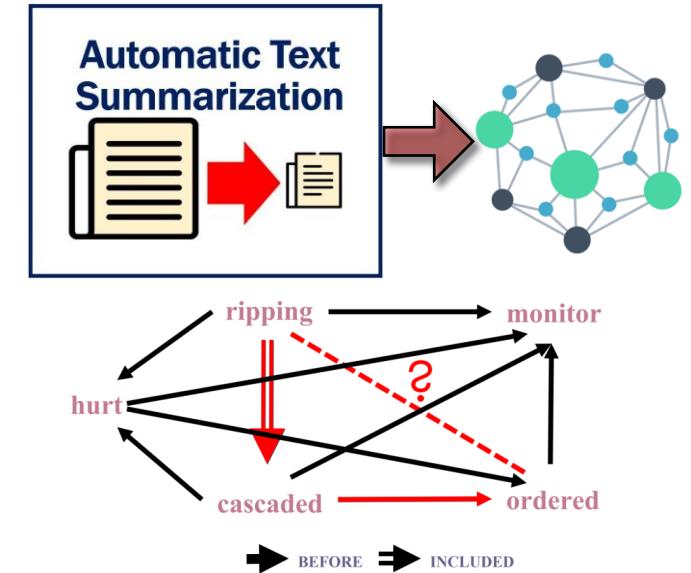
1. Noise-robust IE



2. Faithful IE



3. Indirectly Supervised IE



4. Future Directions



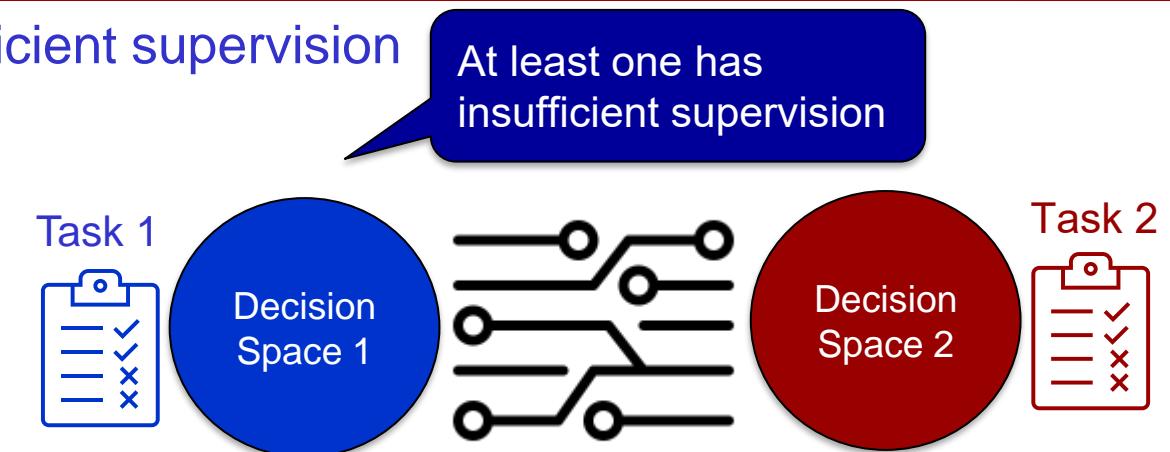


Two Forms of Indirect Supervision



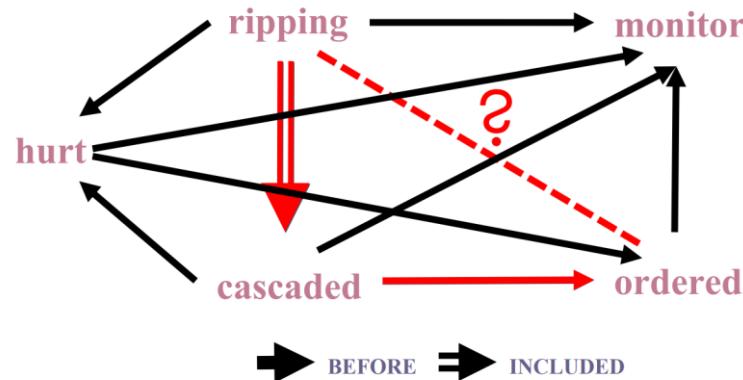
Direct annotation is difficult and expensive

IE suffers from insufficient supervision

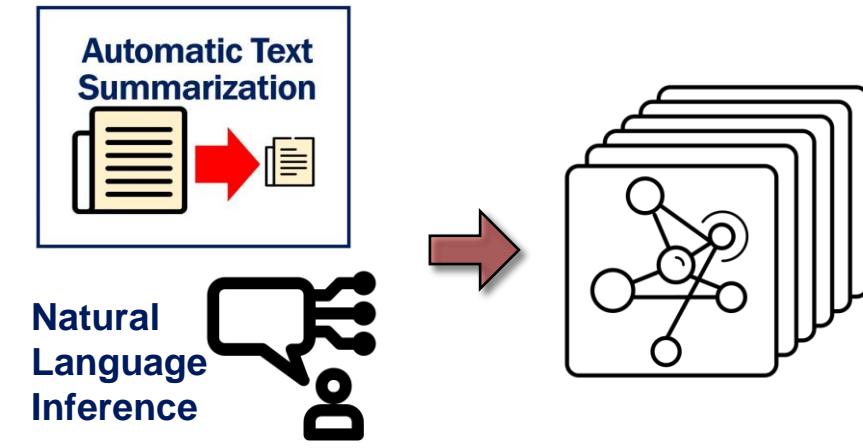


Can we bridge the supervision of different tasks?

Two Forms of Indirect supervision



(Logically) Constrained Learning



Cross-task Transfer

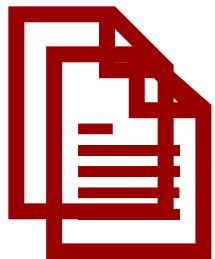
Constrained Learning: Bridging Learning Resources with Logical Constraints



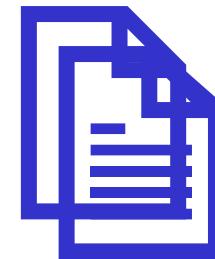
Take Event Relation Extraction as an Example

- Temporal relation extraction (Before, After, ...)
- Membership detection (Subevents, Coreference)

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.

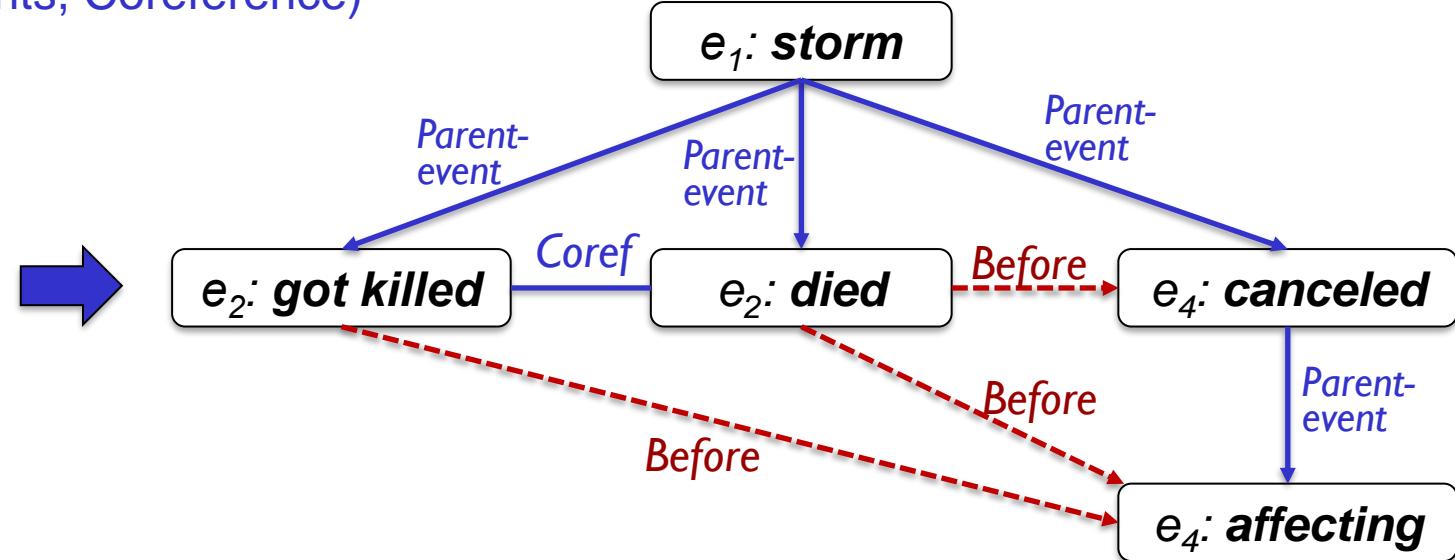


TempRel Corpora
(MATRES, TB-Dense, etc.)



Membership Corpora
(HiEve, ECB+, etc.)

Could we connect these supervision data?



Implication

$e_1: storm$ is PARENT of $e_4: canceled \Rightarrow e_1: storm$ is BEFORE $e_4: canceled$

Conjunction

$e_3: died$ is BEFORE $e_4: canceled \wedge e_4: canceled$ is a PARENT of $e_5: affecting \Rightarrow$

$e_3: died$ is BEFORE $e_5: affecting$

Implication, Negation, ...

Use logical constraints!



Logical Constraints Of Relations

Dependency of Decisions

Symmetry

$e3:\text{died}$ is BEFORE $e4:\text{canceled}$
 $\Rightarrow e4:\text{canceled}$ is AFTER $e3:\text{died}$

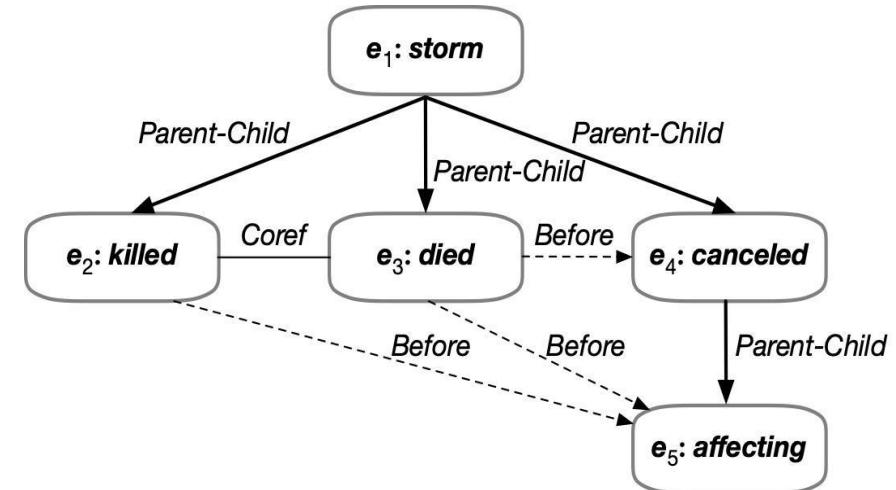
Conjunction

$e3:\text{died}$ is BEFORE $e4:\text{canceled}$
 $\wedge e4:\text{canceled}$ is a PARENT of $e5:\text{affecting}$
 $\Rightarrow e3:\text{died}$ BEFORE $e5:\text{affecting}$

(we also consider *Implication* and *Negation*)

Transitivity

$e1:\text{storm}$ is PARENT of $e4:\text{canceled}$
 $\wedge e4:\text{canceled}$ is a PARENT of $e5:\text{affecting}$
 $\Rightarrow e1:\text{storm}$ is a PARENT of $e5:\text{affecting}$



- Goal: incorporating logical constraints into neural model training.
- Learning to provide **globally consistent** predictions
 - Providing indirect supervision across tasks/decision spaces



Incorporating Logical Constraints in A Neural Architecture

Using product t -norm model constraints as differentiable functions

Symmetry and negation are subsumed by implication loss; Transitivity is also captured by conjunction loss.

- L_A Task Loss: $\top \rightarrow r(e_1, e_2) \rightarrow -w_r \log r_{(e_1, e_2)}$
- L_S Implication Loss: $\alpha(e_1, e_2) \rightarrow \bar{\alpha}(e_2, e_1) \rightarrow \log \alpha_{(e_1, e_2)} - \log \bar{\alpha}_{(e_2, e_1)}$
- L_C Conjunction Loss: $\alpha(e_1, e_2) \wedge \beta(e_2, e_3) \rightarrow \gamma(e_1, e_3) \rightarrow \log \alpha_{(e_1, e_2)} + \log \beta_{(e_2, e_3)} - \log \gamma_{(e_1, e_3)}$
 $\alpha(e_1, e_2) \wedge \beta(e_2, e_3) \rightarrow \neg\delta(e_1, e_3) \rightarrow \log \alpha_{(e_1, e_2)} + \log \beta_{(e_2, e_3)} - \log(1 - \delta_{(e_1, e_3)})$
- Training Objective: $L = L_A + \lambda_S L_S + \lambda_C L_C$

Constraints become entropy regularizers

$\begin{array}{c} \beta \\ \diagdown \\ \alpha \end{array}$	PC	CP	CR	NR	BF	AF	EQ	VG
PC	PC, \neg AF	-	PC, \neg AF	\neg CP, \neg CR	BF, \neg CP, \neg CR	-	BF, \neg CP, \neg CR	-
CP	-	CP, \neg BF	CP, \neg BF	\neg PC, \neg CR	-	AF, \neg PC, \neg CR	AF, \neg PC, \neg CR	-
CR	PC, \neg AF	CP, \neg BF	CR, EQ	NR	BF, \neg CP, \neg CR	AF, \neg PC, \neg CR	EQ	VG
NR	\neg CP, \neg CR	\neg PC, \neg CR	NR	-	-	-	-	-
BF	BF, \neg CP, \neg CR	-	BF, \neg CP, \neg CR	-	BF, \neg CP, \neg CR	-	BF, \neg CP, \neg CR	\neg AF, \neg EQ
AF	-	AF, \neg PC, \neg CR	AF, \neg PC, \neg CR	-	-	AF, \neg PC, \neg CR	AF, \neg PC, \neg CR	\neg BF, \neg EQ
EQ	\neg AF	\neg BF	EQ	-	BF, \neg CP, \neg CR	AF, \neg PC, \neg CR	EQ	VG, \neg CR
VG	-	-	VG, \neg CR	-	\neg AF, \neg EQ	\neg BF, \neg EQ	VG	-

Capturing 80 constraints in total



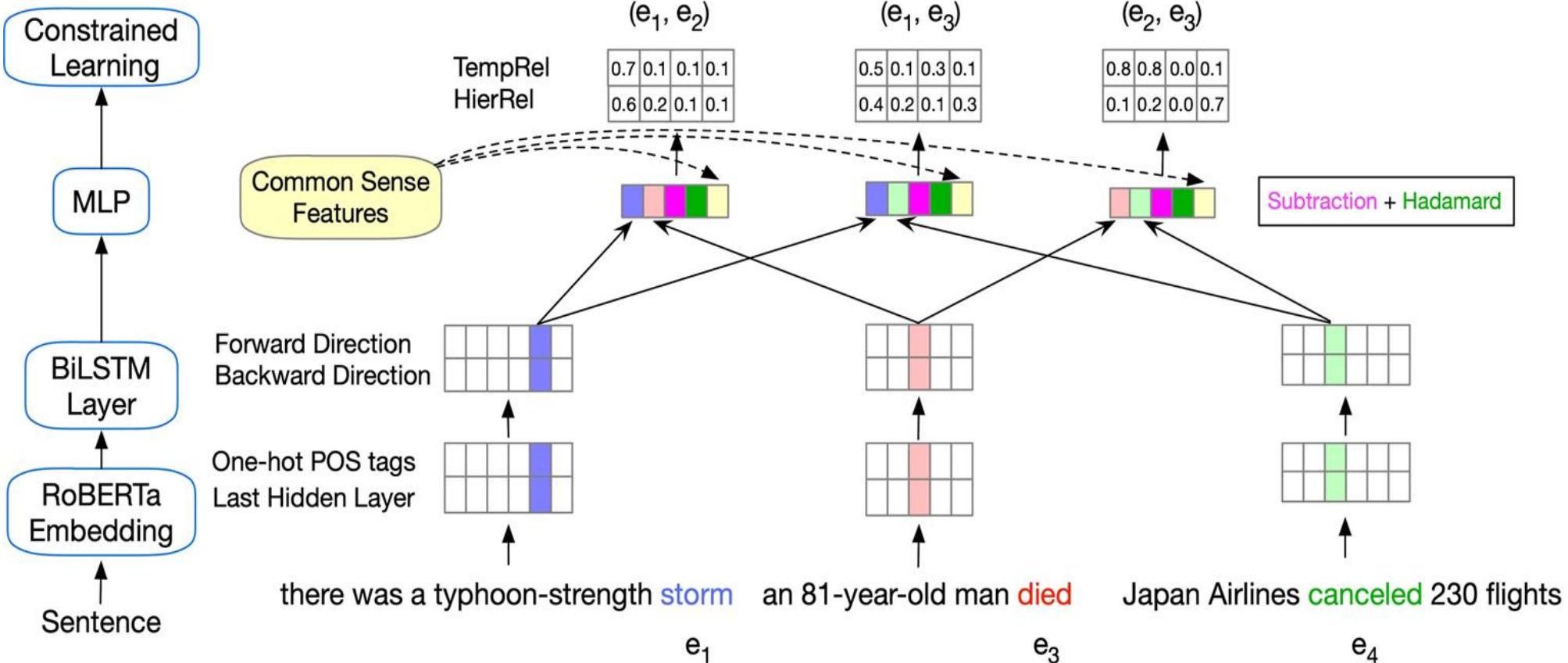
Joint Constrained Learning

- Temporal Relations
- Subevent Relations (Memberships)
- Event Coreference

Task loss

Implication and conjunction constraint losses

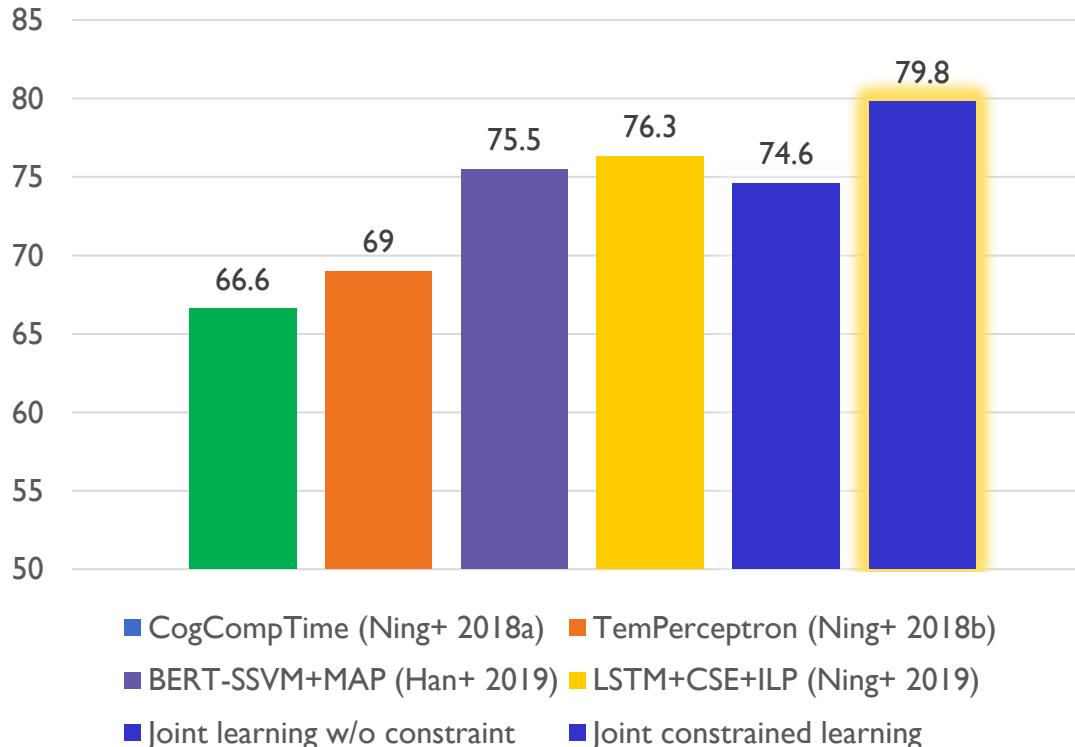
$$\text{Loss Function: } L = L_A + \lambda_S L_S + \lambda_C L_C$$



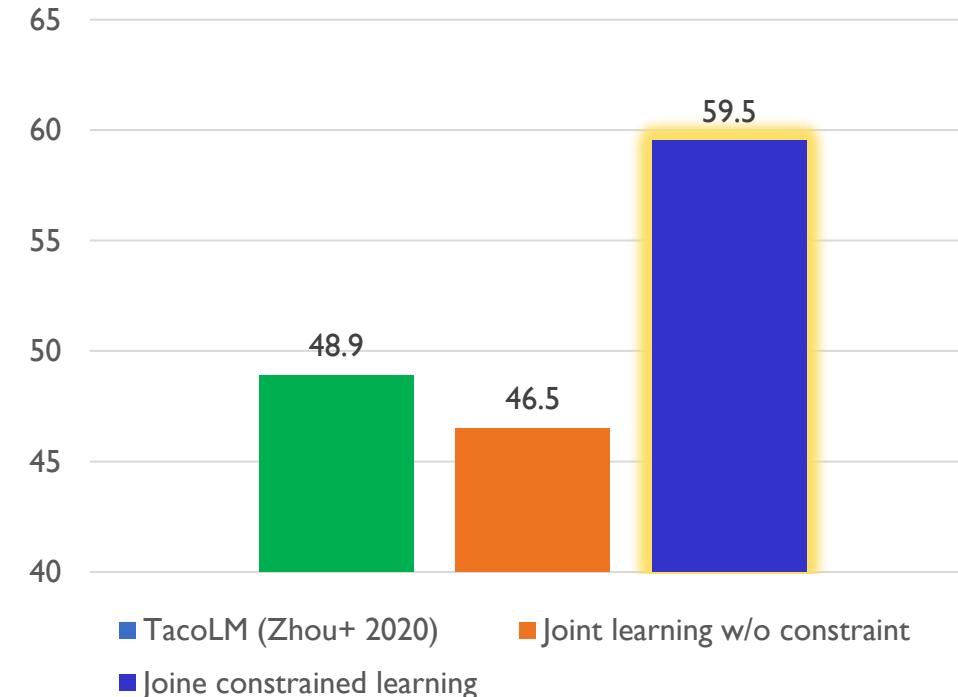
The Joint Constrained Learning Architecture



F1 on MATRES for TempRel Extraction



F1 on HiEve for Membership (Subevents and coref) Extraction



Key Observations

- Constraints are a natural bridge for learning resources with different sets of relations
- Adding constraints sufficiently enforces logical consistency of extraction, surpassing ILP in inference (w/o constrained learning) by 2.6-12.3% in ACC

Automatically Learning Constraints



Some logical constraints can be hard to articulate. We should automatically capture them!

Event-event relations are related to narrative segmentation

- Subevent relations happen much more often within the same narrative segment
[Lukasik+ EMNLP-20]

An (implicit) soft logic constraint. How do we capture it?

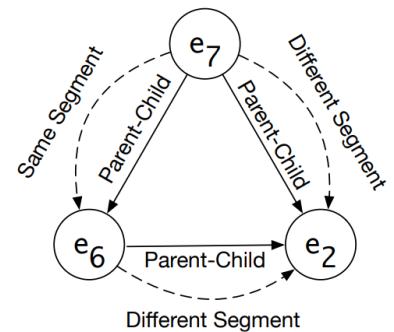
Former Penn State football coach Jerry Sandusky posted (e1) bail Thursday after spending a night in jail following a new round of sex-abuse charges (e2) filed against him. Sandusky secured his release using (e3) \$200,000 in real estate holdings and a \$50,000 certified check provided (e4) by his wife, Dorothy, according to online court record ... He was also charged (e5) last month with abusing eight boys, some on campus, over 15 years, allegations that were not immediately brought to the attention of authorities even though high-level people at Penn State apparently knew about them. In all, he faces more than 50 charges (e6). The scandal (e7) has resulted in the ousting (e8) of school President Graham Spanier and longtime coach Joe Paterno.

Constraint Learning

Training a single-layer rectifier network on all ``triangles'' to identify legitimate structures

$$\mathbf{w}_k \cdot \mathbf{X} + b_k \geq 0 \longrightarrow p = \sigma\left(1 - \sum_{k=1}^K \text{ReLU}(\mathbf{w}_k \cdot \mathbf{X} + b_k)\right)$$

Estimates probability of a legitimate triangle



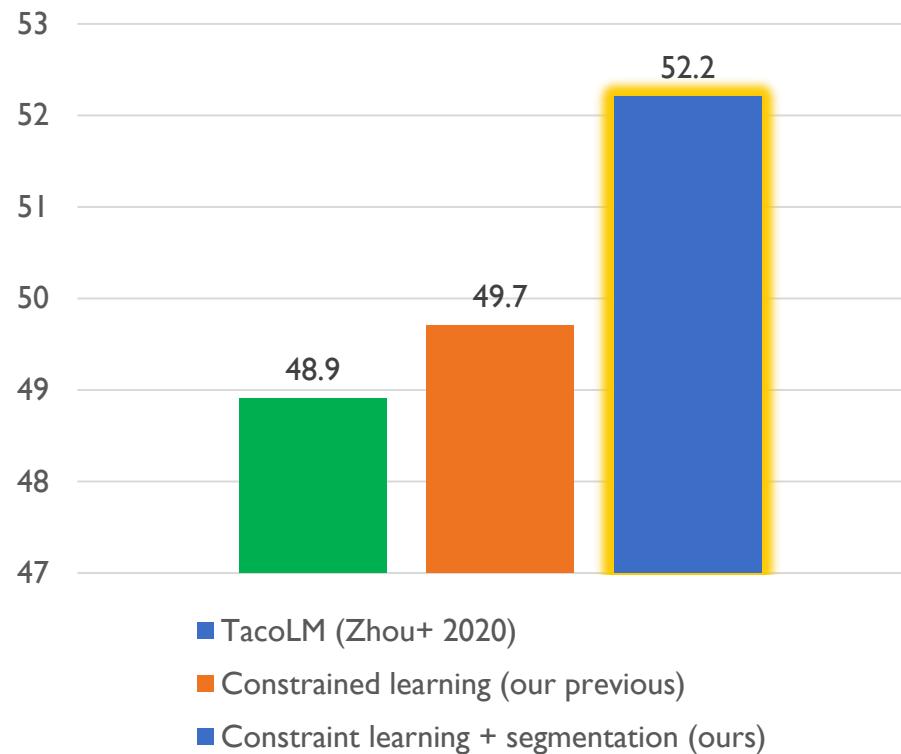
Adding constraint losses according to the the rectifier estimated “truth values” of constraints

$$L_{cons} = -\log\left(\sigma\left(1 - \sum_{k=1}^K \text{ReLU}(\mathbf{w}_k \cdot \mathbf{X} + b_k)\right)\right)$$

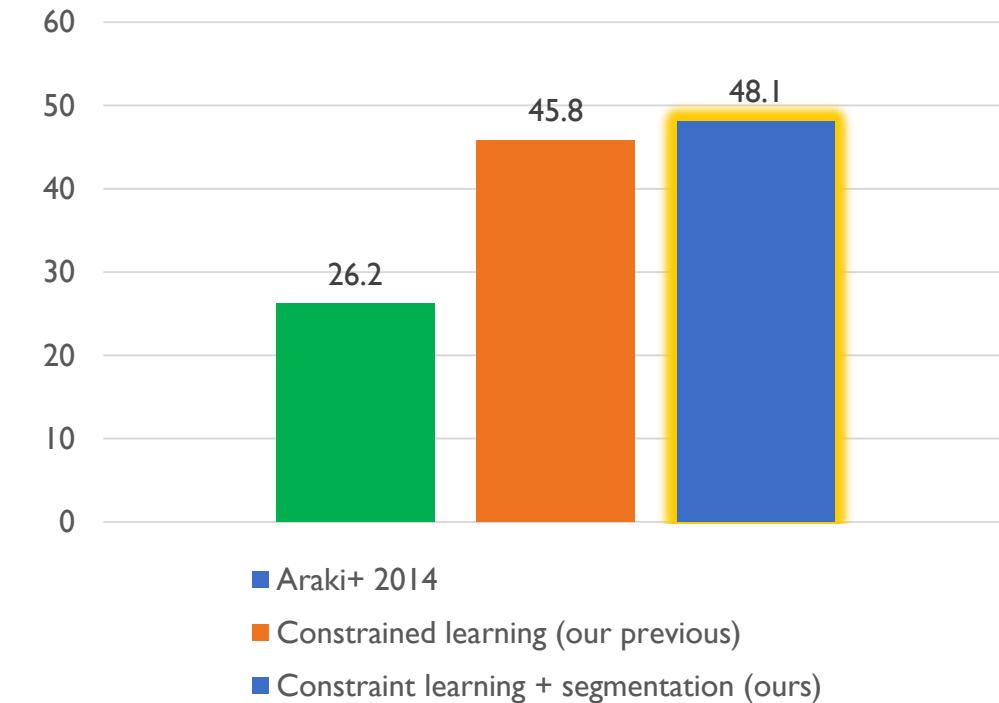


Automatically Learning Constraints

Subevent detection (F1) on HiEve



Subevent detection (F1) on Intelligence Community



- Constraint learning automatically captures soft constraints
- Allowing more indirect supervision signals to be introduced (from narrative segmentation).



Indirect supervision via Cross-task Transfer

Ultra-fine Entity Typing

- >10K free-form types
- Very few clean training cases (~2k)

Once Upon Andalasia is a video game based on the film of the same name.



film, art, movie, show, entertainment, creation

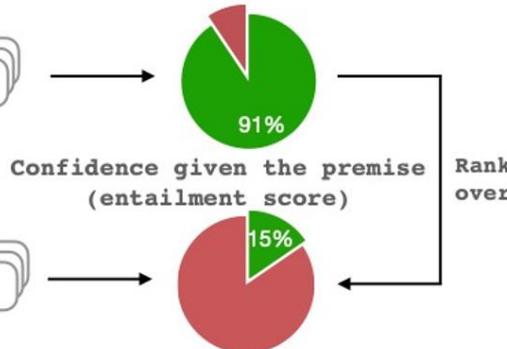
Indirect Supervision from Natural language Inference

In fact, Chrysler needs to convince investors it is on the right track if it wants to pay back billions in loans from the U.S. and Canadian governments.

Premise

Chrysler is a company.

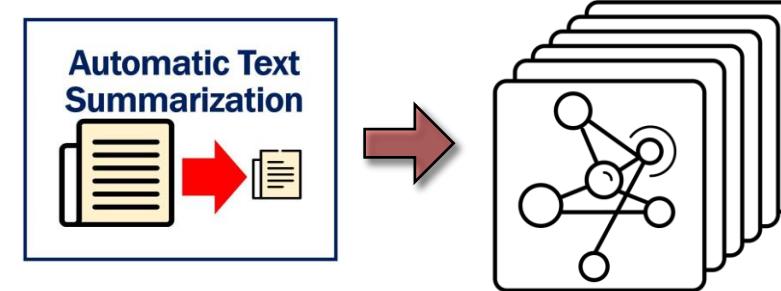
Hypotheses by TRUE labels



The first to reach >50% F1 (for >10k types) on UFET

Excellent generalization to unseen types

Relation Extraction



Abstractive summarization as indirect supervision

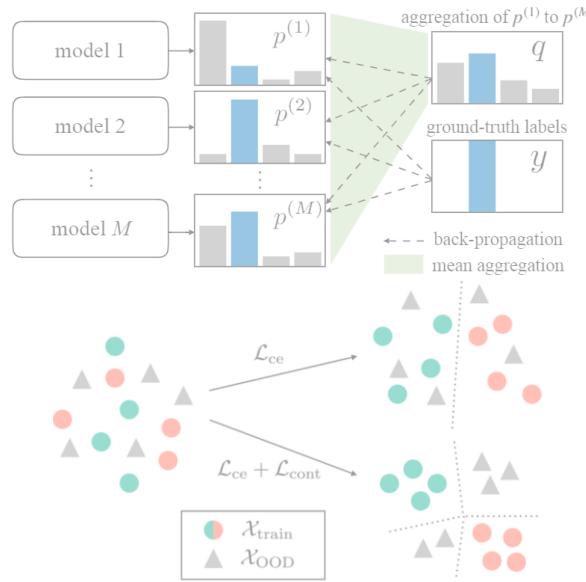
- Viewing relations as **one kind of salient information** to be summarized
- Transfer-tuning a summarization model for **constrained decoding** of verbalized relations

Close to SOTA performance using only 5% of training data on TACRED

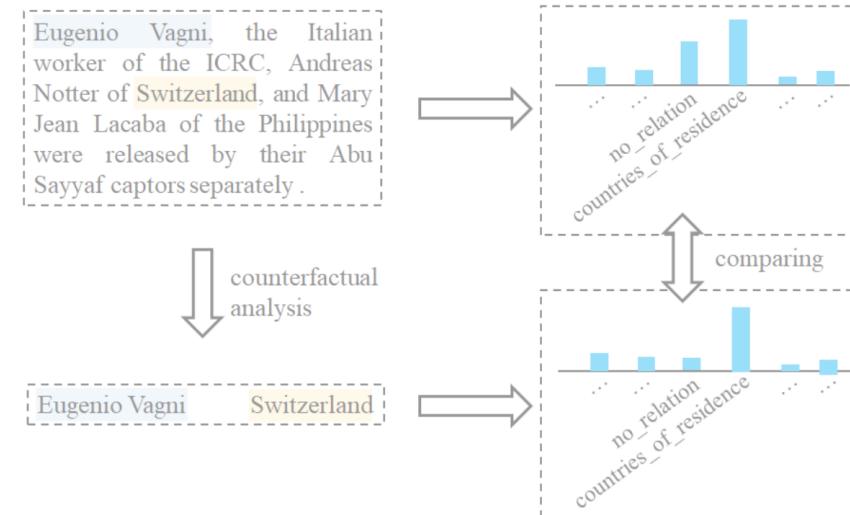


Agenda

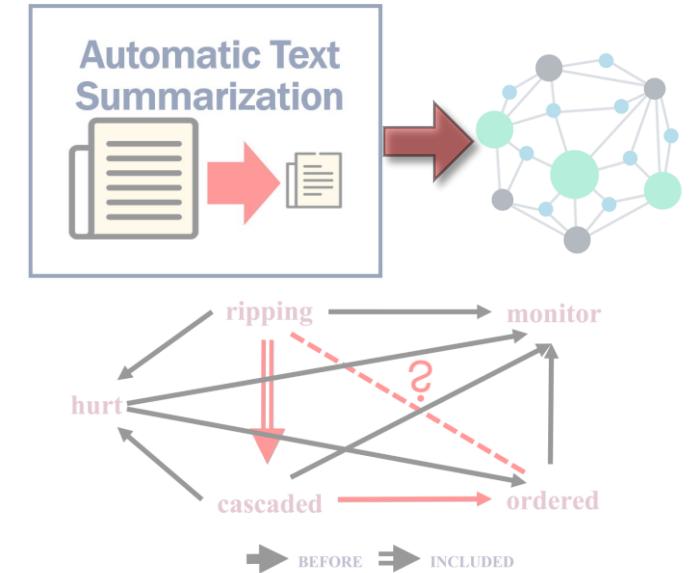
1. Noise-robust IE



2. Faithful IE



3. Indirectly Supervised IE



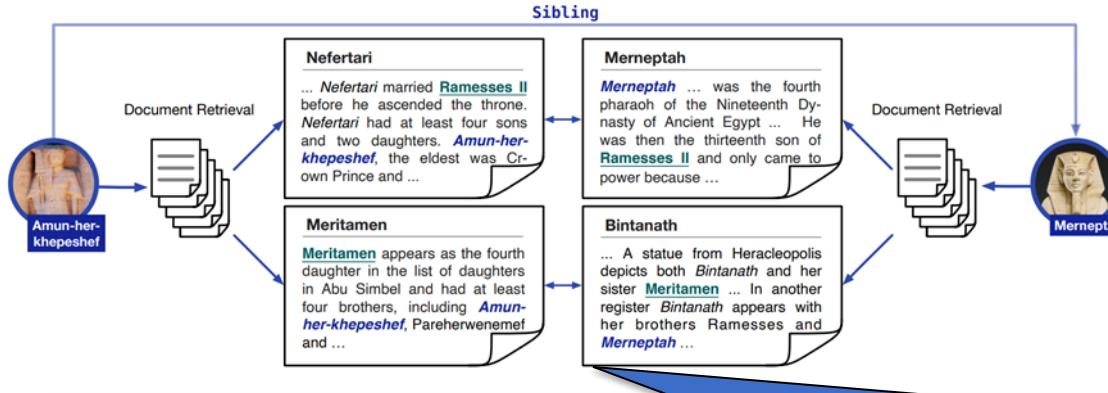
4. Future Directions





Thinking Across Documents

① Inducing relations across documents



~57.6% of Wikidata (En) facts do not find mentions in the same Wikipedia article [Yao+ 2021]

③ From understanding “what the text says” to “what is happening”



② Consolidating unevenly distributed knowledge



DBpedia

Novel



DBpedia

Monogatari (story)
Love story
Royal family story
Realistic novel
Ancient literature

Many more challenges to IE

- Multi-hop reasoning
- Consolidation
- Tracking information pollution
- Long-form document modeling
- Mitigating frequency biases
- ...

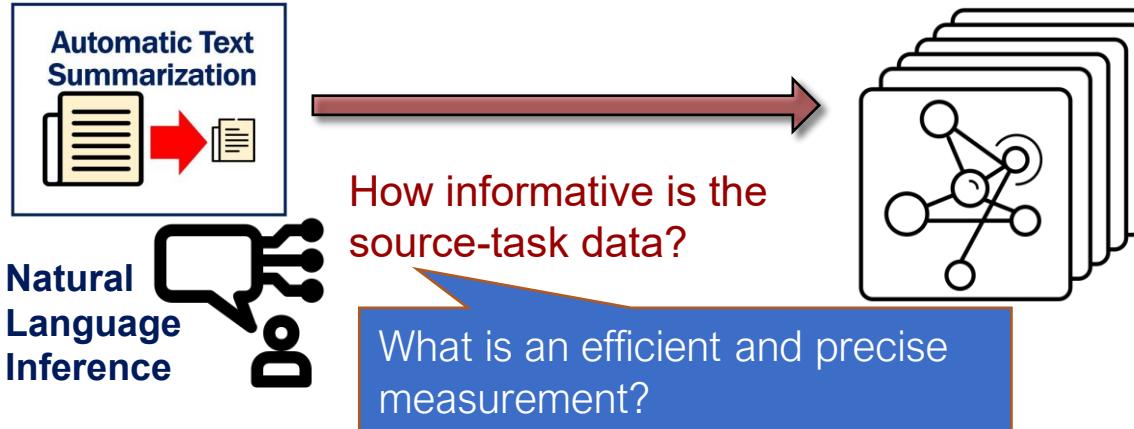
👉 A long way to go



Accountable Indirect Supervision

Decision choices for indirectly supervised learning needs to be accountable

Measuring the Affinity of Indirect Supervision Signals

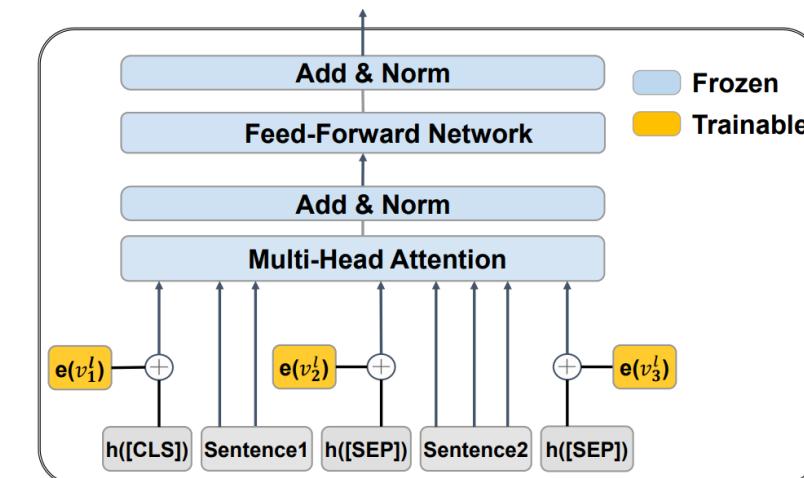
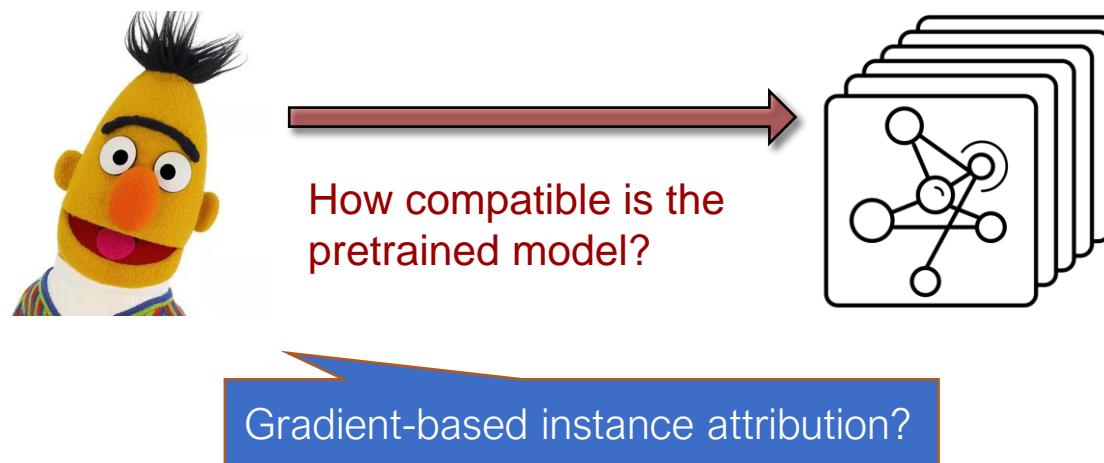


Finding Indirect Supervision from the Model Hub



A screenshot of the Hugging Face Model Hub interface. It shows a search bar with "Models 79,308", a "Filter by name" button, and an "Add filters" button. Below the search bar are two model cards: "xlm-roberta-base" and "bert-base-uncased". Each card includes the model name, a brief description, the last update date, the number of parameters, and the number of stars.

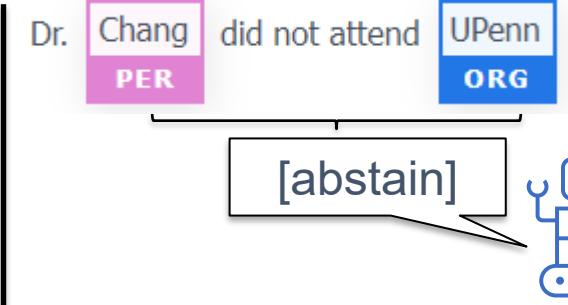
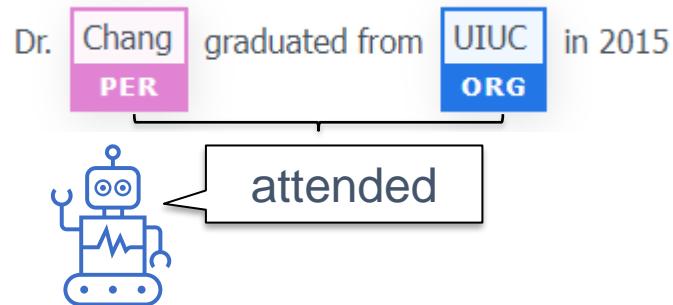
Efficient Adaptation of Large Models to Indirect Supervision





Selective Extraction

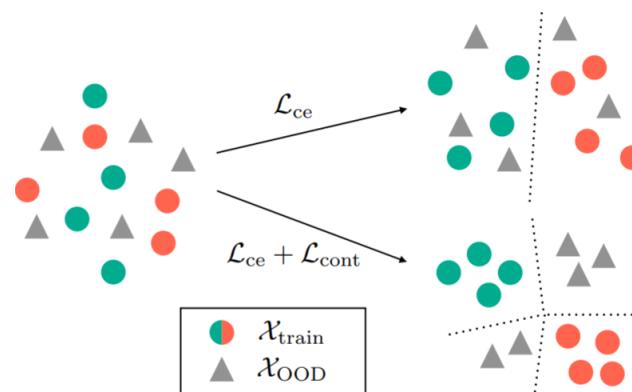
In inference, IE models need to know when to not extract



IE models can be exposed to many exception cases in real-world application.

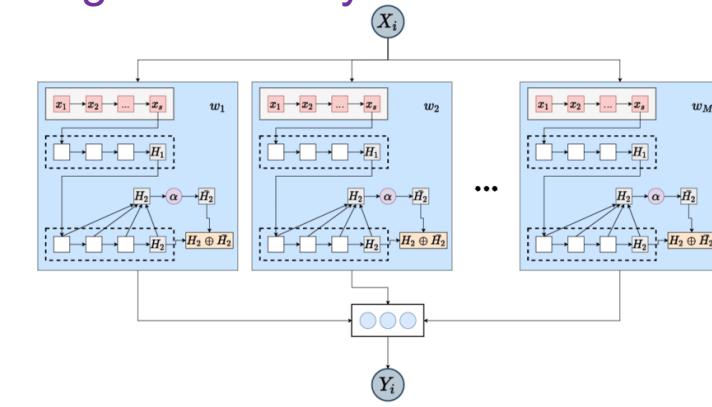
How to know its **decision boundary**.

Unsupervised out-of-distribution (OOD) detection



Increase inter-class discrepancy \Rightarrow Better OOD detection

Estimating Uncertainty for Prediction



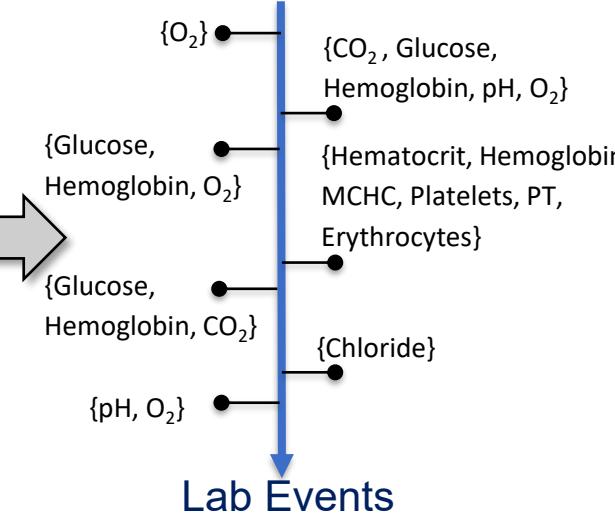
Softmax Response, Monte-Carlo Dropout, etc.



Medicine and Healthcare



Drug-drug Interaction



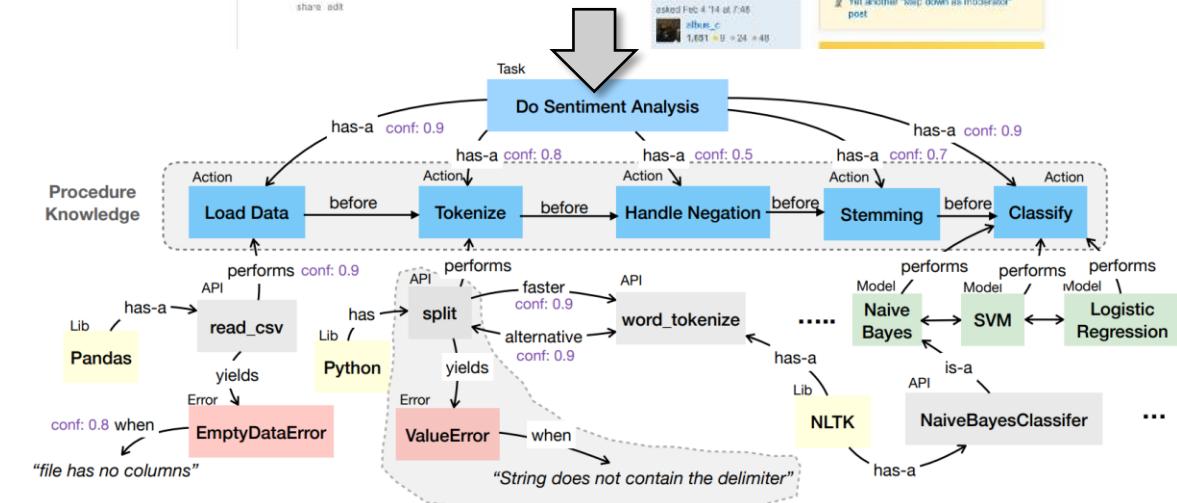
Programming Education

Load data from txt with pandas

```
import pandas as pd
data = pd.read_csv('output_list.txt', header = None)
print data
```

This is the structure of the input file: 1 E 2888.8 78.2836942112 1347.2836942112 /title,address,txt

Now the data are imported as a unique column. How can I divide it, so to store different elements separately (so I can call `data[1,i]`)? And how can I define a header?



- Low-resource domains that particularly
- need **indirect supervision** and **constrained learning**;
 - suffer from **noise** and **faithfulness issues**.

Acknowledgement



Student Researchers

Language Understanding and Knowledge Acquisition Lab



Wenxuan Zhou
(PhD Student)

Bangzheng Li
(Undergrad →
PhD Student)

Nancy Xu
(PhD Student)

Eric Qasemi
(PhD Student)

Fei Wang
(MS →
PhD Student)

James Y. Huang
(PhD Student)

Keming Lu
(MS Student)

Collaborating Institutes



Carnegie
Mellon
University



NUS
National University
of Singapore



UNIVERSITY OF
CAMBRIDGE



Microsoft
Research



David Geffen
School of Medicine



amazon

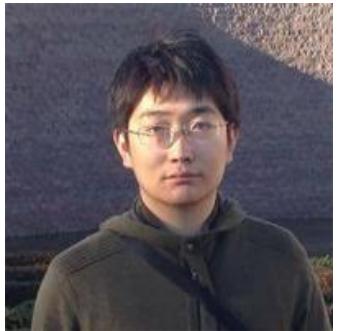


APPLIED RESEARCH LABORATORY FOR
INTELLIGENCE
AND SECURITY

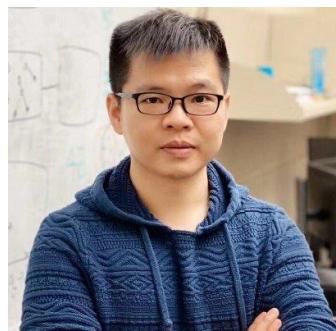
cisco

Sponsors

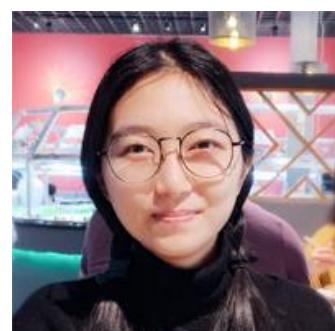
New Frontiers of Information Extraction



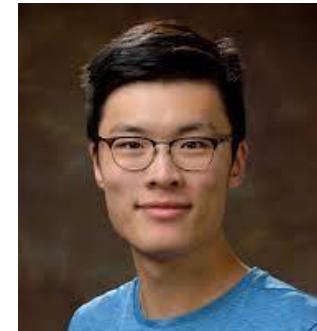
Muhao Chen



Lifu Huang



Manling Li



Ben Zhou



Heng Ji



Dan Roth

Contents

- Robustness of IE (Muhao@USC)
- Indirectly and Minimally Supervised IE (Ben@UPenn)
- Knowledge-guided IE (Heng@UIUC/Amazon)
- Transferability of IE (Lifu@VT)
- Multimodal IE (Manling@UIUC)
- Emerging Challenges of IE (Dan@UPenn/Amazon)

<https://cogcomp.seas.upenn.edu/page/tutorial.202207>

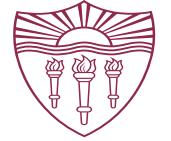


NAACL 2022

July 2022

NAACL Tutorials

New Frontiers of Information Extraction



Thank You