USC Viterbi School of Engineering

Robust and Indirectly Supervised Information Extraction

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

Robust Information Extraction (Robust IE)

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

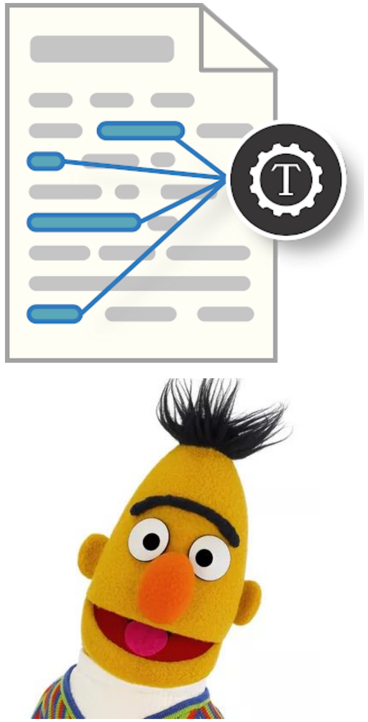
Honolulu

From Wikipedia, the free encyclopedia

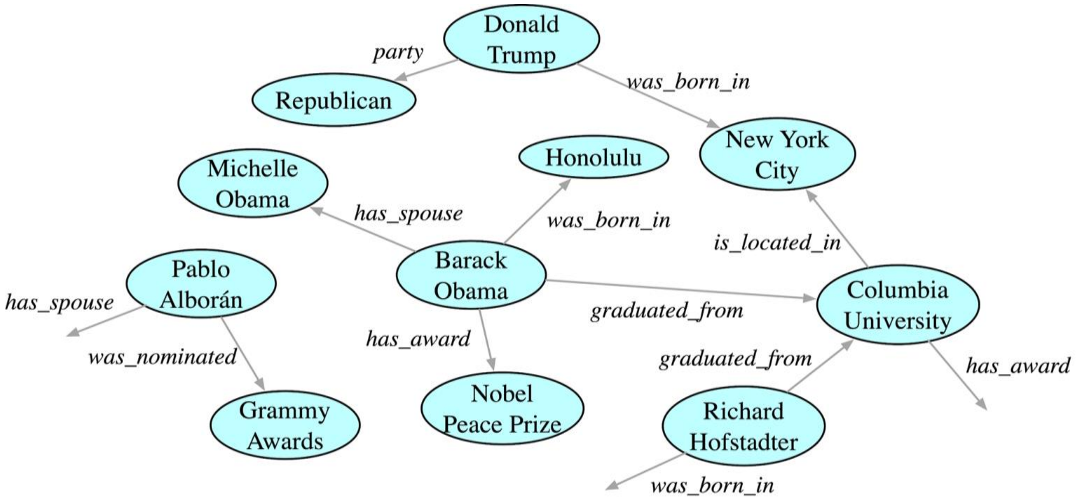
This article is about the largest city and state capital city of Hawai... Honolulu itself, see Honolulu County, Hawaii. For other uses, see

Honolulu (/ˌhɑːnəˈluːluː/;[7] Hawaiian: [honoˈlulu]) 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 Oʻahu,[a] 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 Nuʻuanu 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

Who scored the longest touchdown pass of the game?

In the fi... 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 F... fired a 6-yard TD pass to tight end Visanthe Shiancoe. The ...ings … 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-ya...D pass to wide receiver Devin Aromashodu.

What is her seizure frequency?

... reports worse ... w occurring ...p 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?

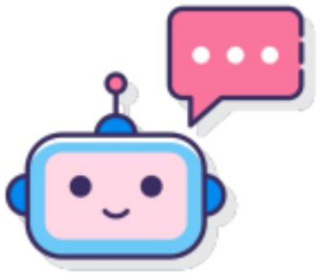
Mayor Rahm Ema... toward his bid for a third term – more than five times the total raised by his 10 challengers combined, campaign finance records show.

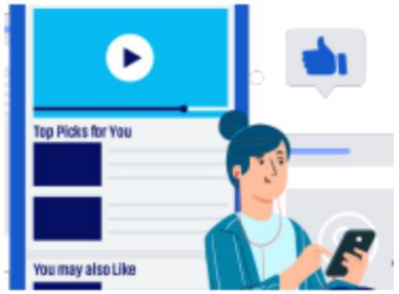
The COVID-19 pandemic in the United States is part of the worldwide pandemic of coronavirus disease 2019 (COVID-19). As of October 2020, there were more than 9,000,000 cases and 230,000 COVID-19-related deaths in the U.S., representing 20% of the world's known COVID-19 deaths, and the most deaths of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

IE Benefits For Content Management

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

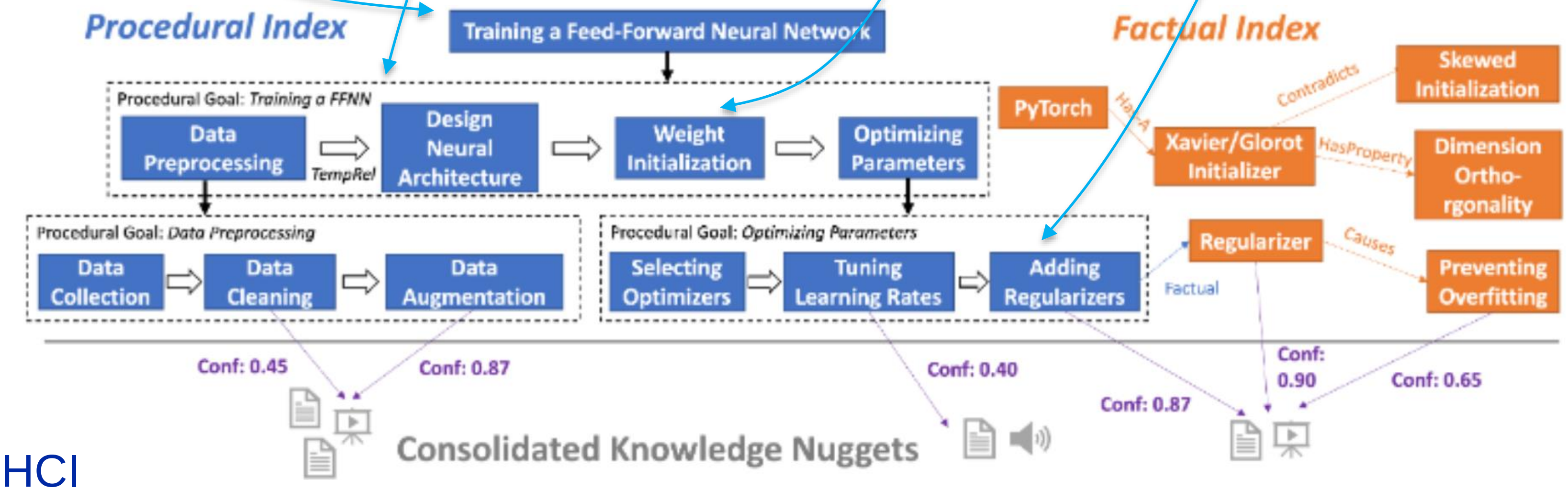
The leftmost layer in this network is called the input layer, and the neurons within the layer are called input neur. The rightmost or output layer contains the output neurons, or, as in this case, a single output neuron. The middle layer is called a hidden layer, since the neurons in this layer are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

Weight initialization for neural networks

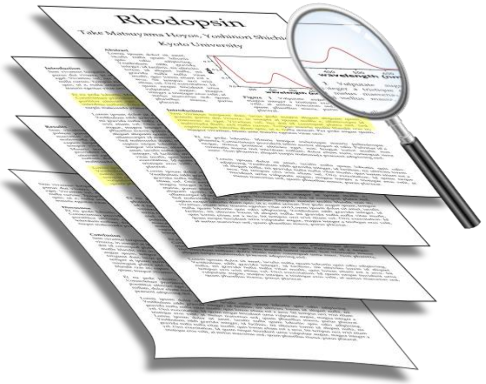
Historically, weight initialization involved using small random numbers, although over the last decade, more specific heuristics have been developed that use information, such as the type of activation function that is being used and the number of inputs to the node. These more tailored heuristics can result in more effective training of neural network models using the stochastic gradient descent optimization algorithm. In this tutorial …

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…



IE Is the Backbone of Any Knowledge-driven Tasks



Knowledge Bases

Bio/Med Databanks

Knowledge Representation

Common Sense

Commonsense QA
Event Prediction
Intent Prediction

Narrative Understanding

Storytelling
Content Selection
Newsworthiness Detection

Proteomic Interaction Prediction
Mutation Effect Estimation
Genomic Function Prediction

Medical Informatics

Diagnostic Prediction
Disease Phenotyping
Drug Repurposing

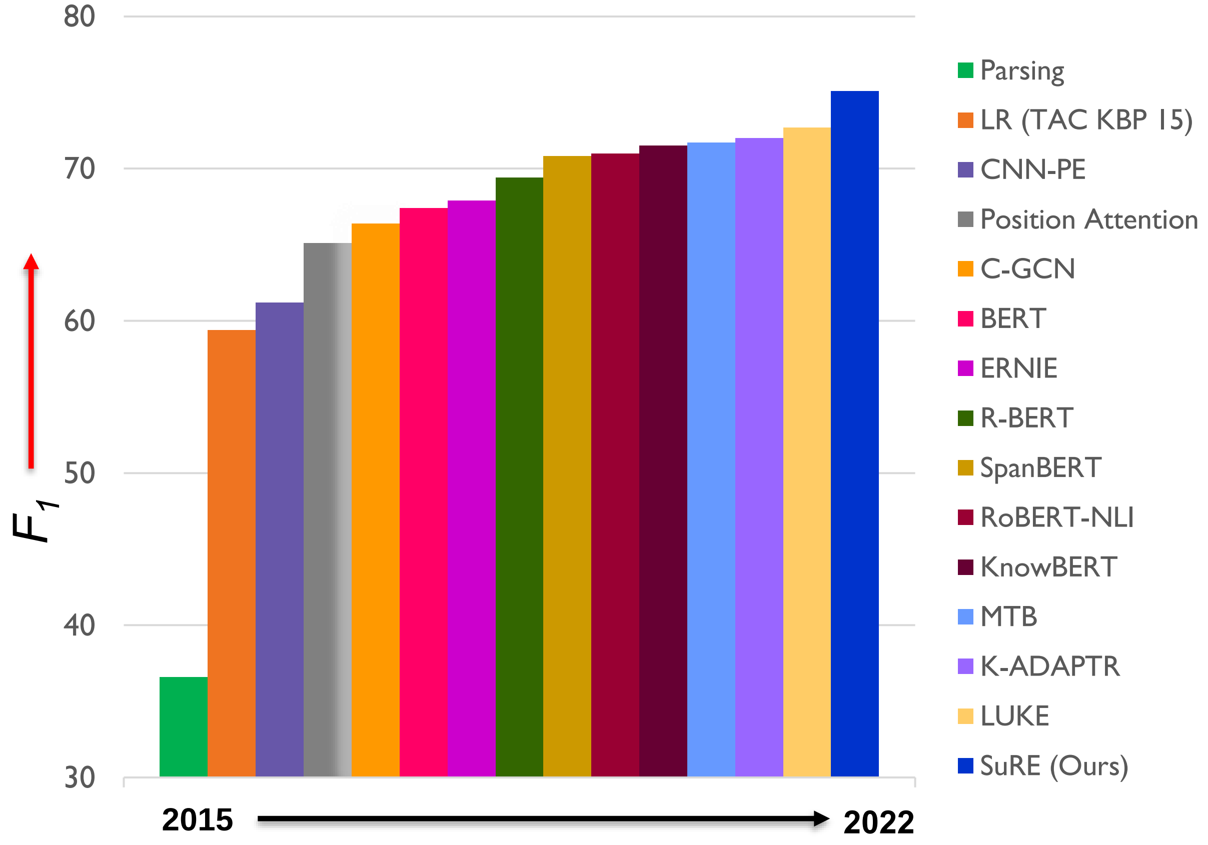
How IE Is Doing Today

On Benchmarks

Rise of the Pre-trained Models

(BERT)

Relation Extraction



In Reality

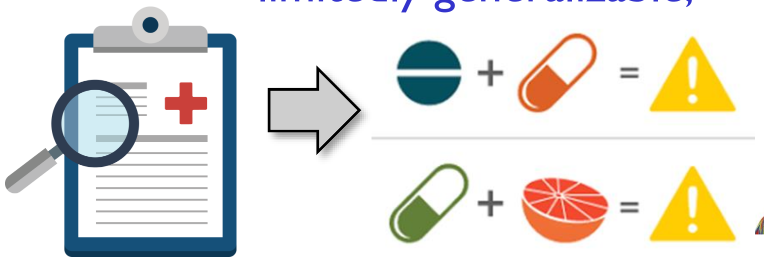
IE is still brittle,

I stayed in Treasure Island [LOC] for two nights when I was in Las Vegas [LOC] .

Type?

Island ✗

limitedly generalizable,



and costly to develop.

IARPA

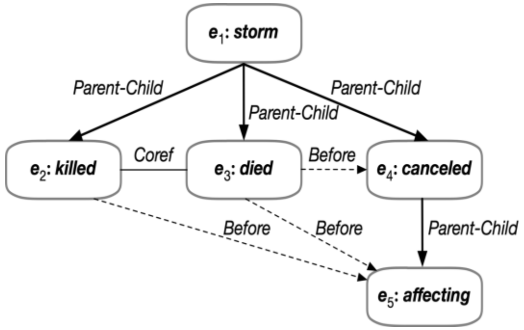
DARPA

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

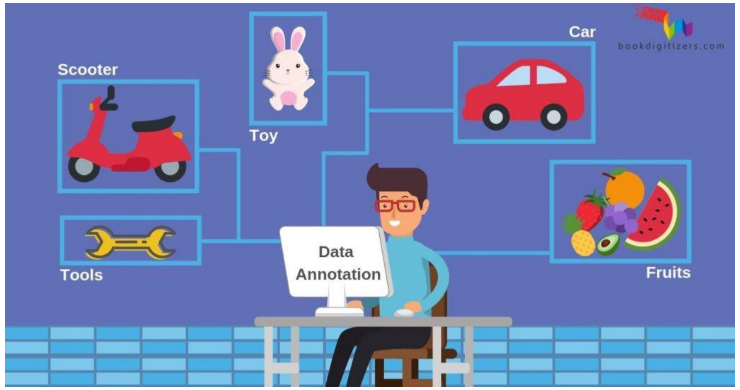
Challenge: Expensive Supervision

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

On Tuesday, there was a typhoon-strength (e_1:***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.



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

Bill Gates (PER) paid a visit to Building 99 (MISC) of Microsoft (ORG) yesterday .

Rel?

Visit ✓

FounderOf ✗

- Spurious correlation
- Biased knowledge

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.

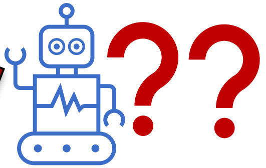
Michael Jordan (PERSON) is a professor at Berkeley (ORG)

SARS - CoV-2 ORF3a interacts with VSP39 -- a core subunits of HOPS complex

Out-of-Distribution Inputs

Training

Inference



Michael Jordan (PER) did not attend UCLA (ORG)

No Rel

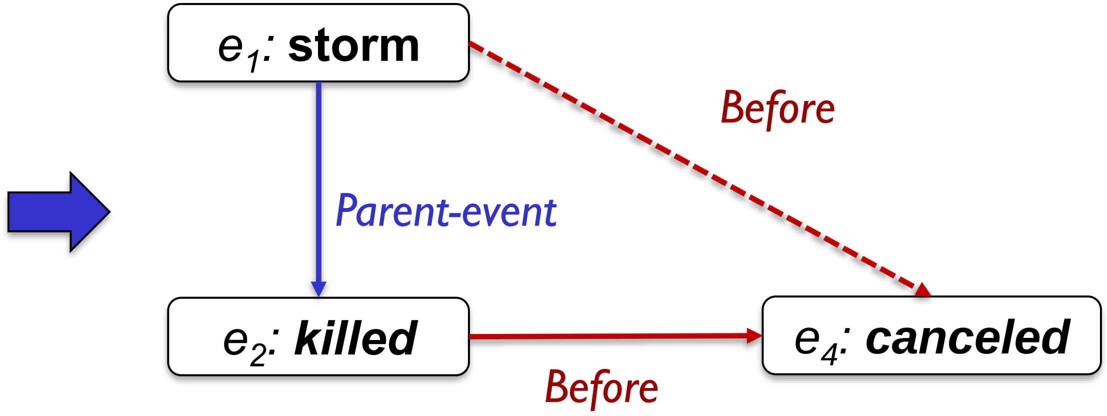
Nothing to extract

Challenge: Consistency

Extracts are interdependent decisions.

An article about a storm hazard in Japan

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.



*Extraction Should be **Globally Consistent***

Symmetry: *e3:died* is BEFORE *e4:canceled* => *e4:canceled* is AFTER *e3:died*

Conjunction: *e3:died* is BEFORE *e4:canceled*∧*e4:canceled* is a PARENT EVENT of *e5:affecting*=> *e3:died* BEFORE *e5:affecting*

Implication, *Negation* …

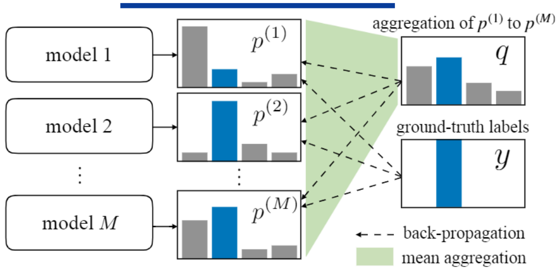
A BERT-based model getting **90%** of correct pairwise decision still **violates 46%** of triplet constraints [Li et al. ACL-20]

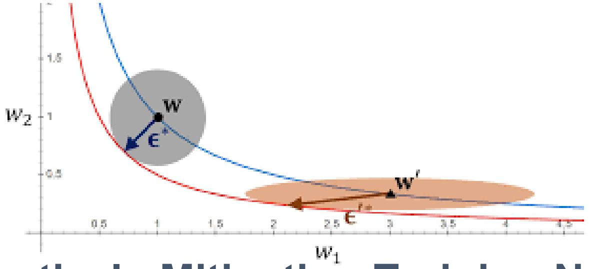
How do we *enforce logical constraints* for consistent/self-contained IE?
How do we *discover the constraints*?

Our Goal: More Reliable IE

Robustness

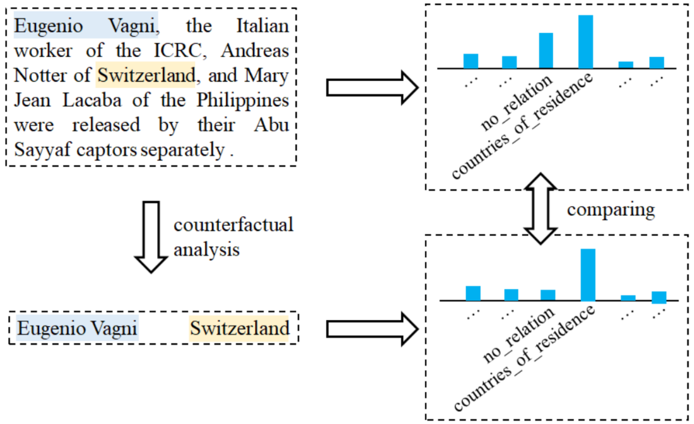
Training





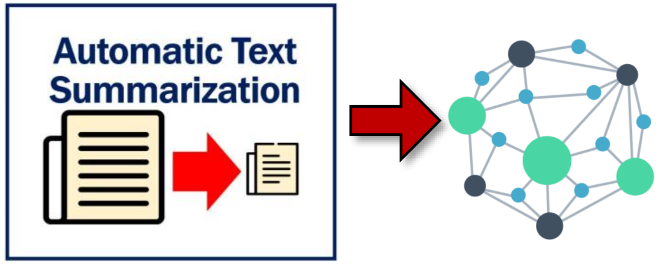
Proactively Mitigating Training Noise

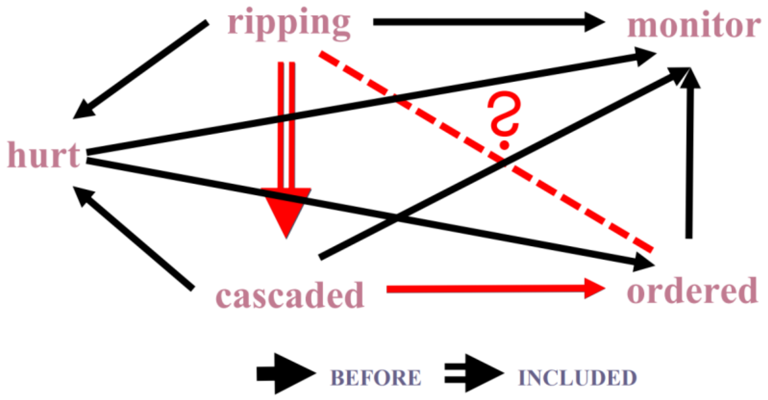
Inference



Mitigating Spurious Correlations

Indirect Supervision

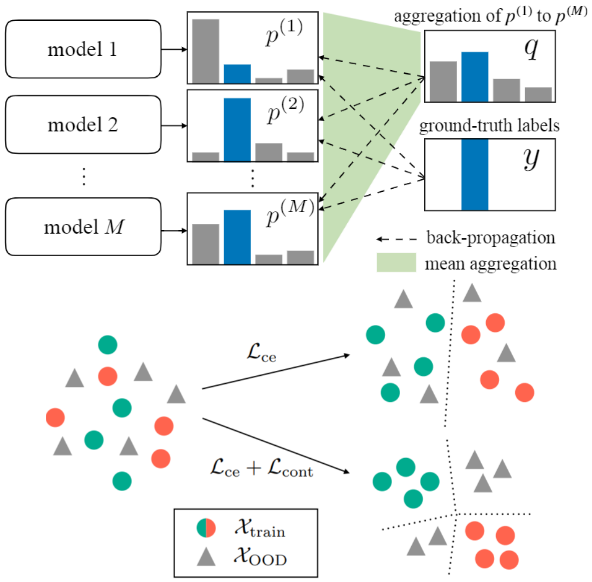




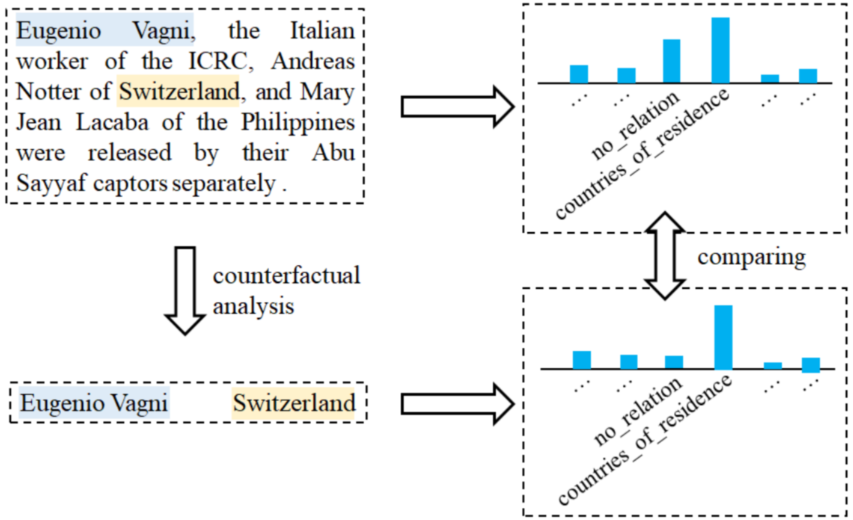
Logical Constraints and Cross-task Transfer

In This Talk

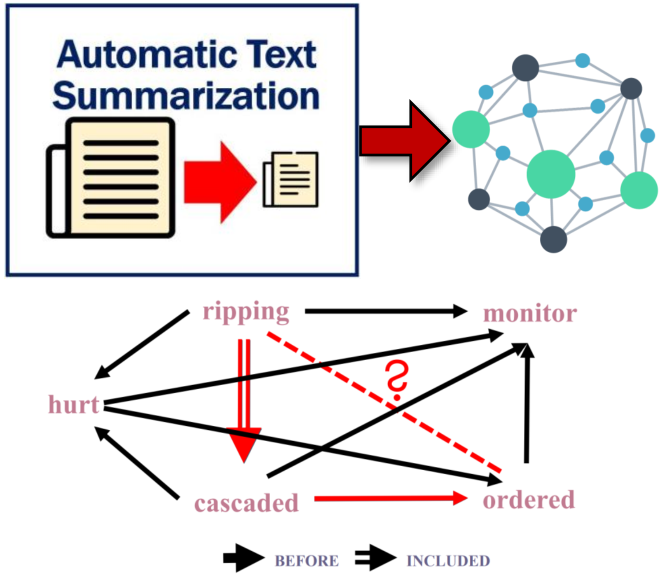
1. Noise-robust IE



2. Faithful IE



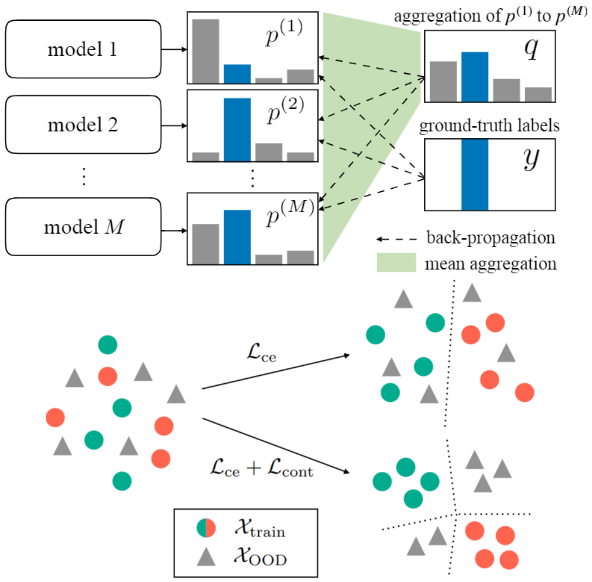
3. Indirectly Supervised IE



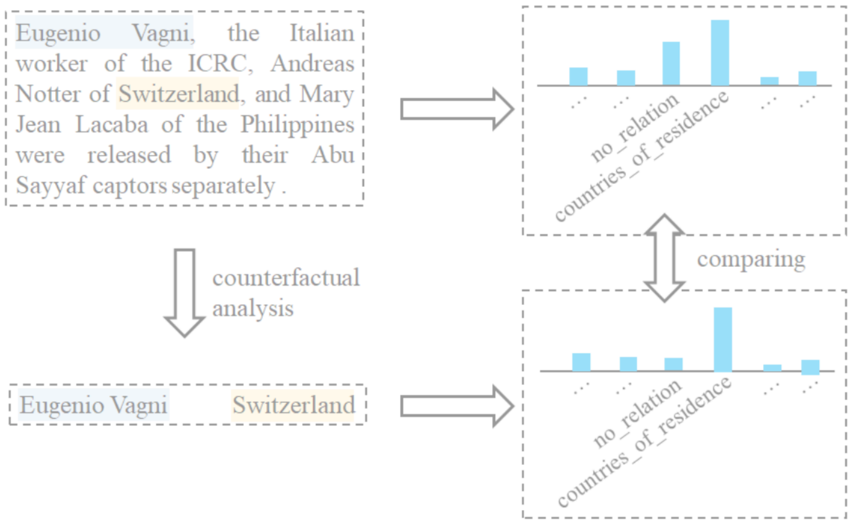
4. Future Directions

Agenda

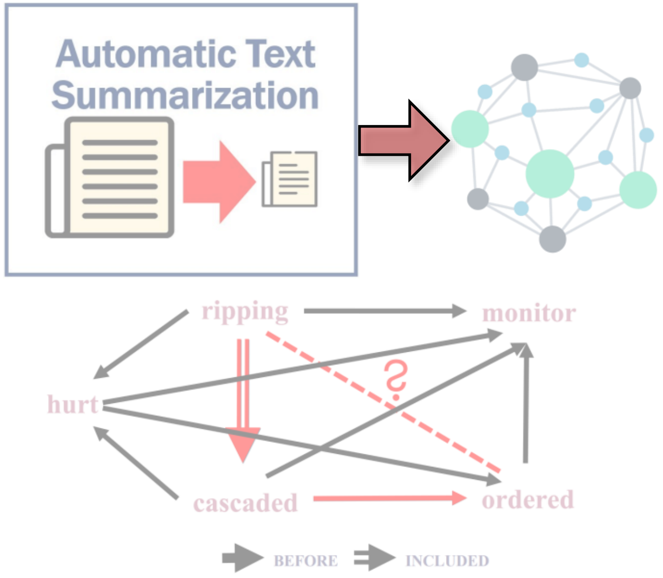
1. Noise-robust IE



2. Faithful IE



3. Indirectly Supervised IE

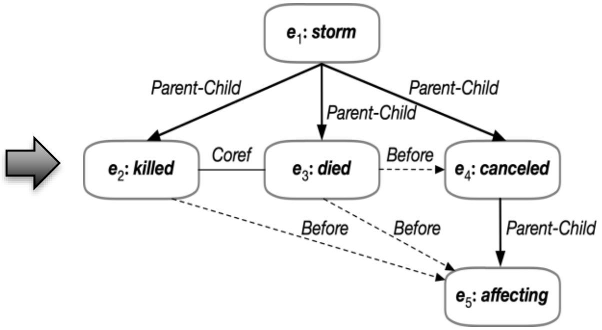


4. Future Directions

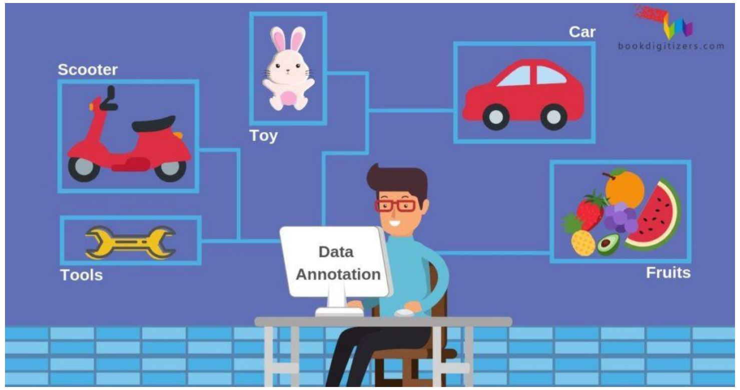
Imperfect Supervision

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

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.



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.

…

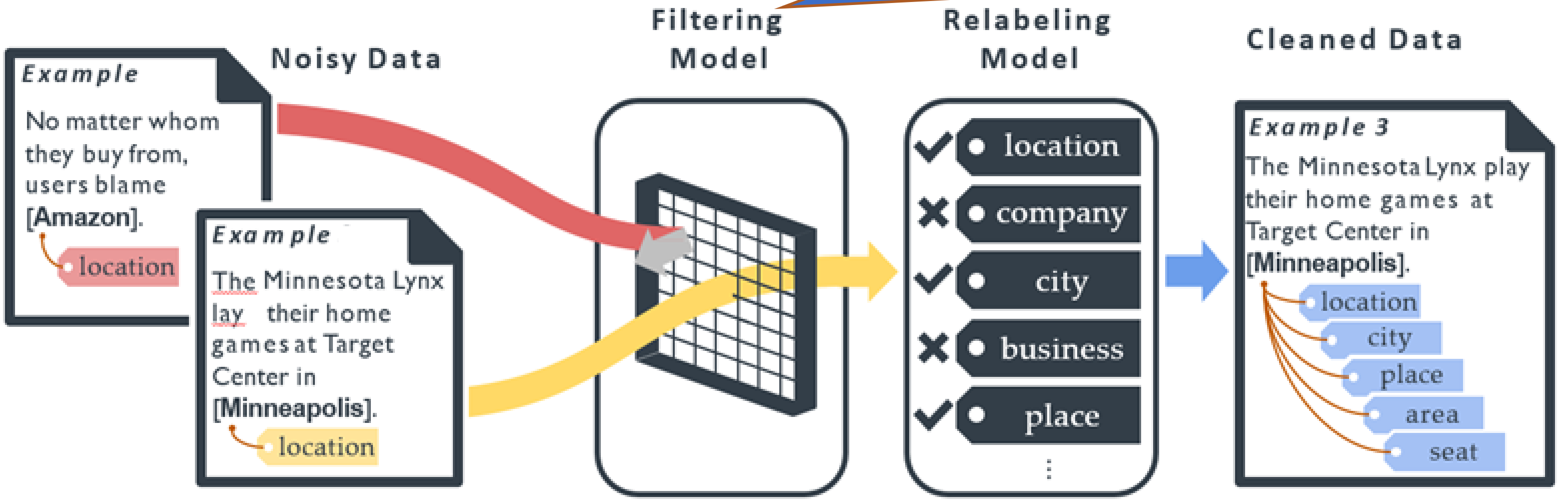
Even moderate training noise leads to significant flaws

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.



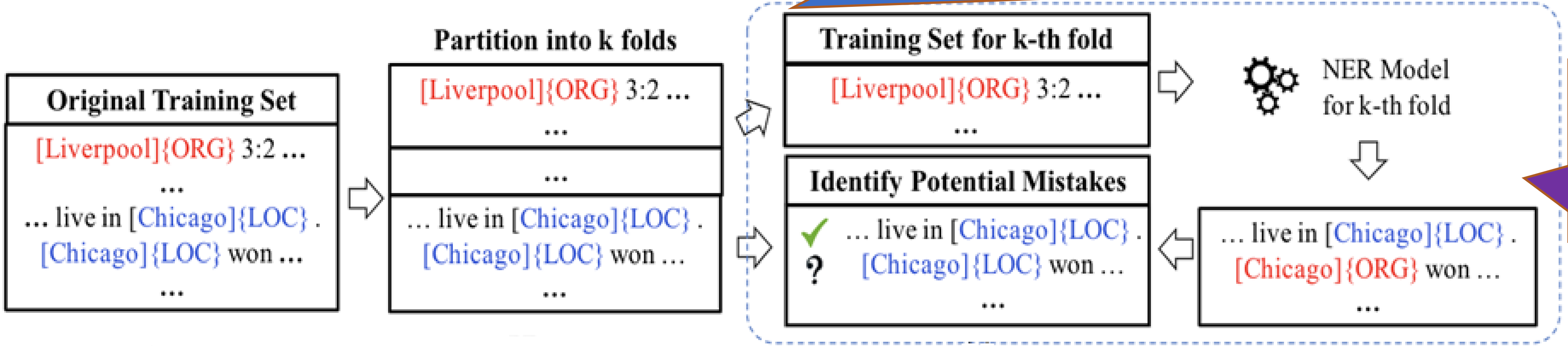
Cost: manually labeling enough clean data is nowhere cheaper.

Onoe and Durrett. *Learning to Denoise Distantly-Labeled Data for Entity Typing*. **NAACL** 2019 (UT Austin)

Swayamdipta+. *Dataset Cartography: Mapping and Diagnosing Datasets with Training Dynamics*. **EMNLP** 2020 (AI2)

Ensemble-based Denoising

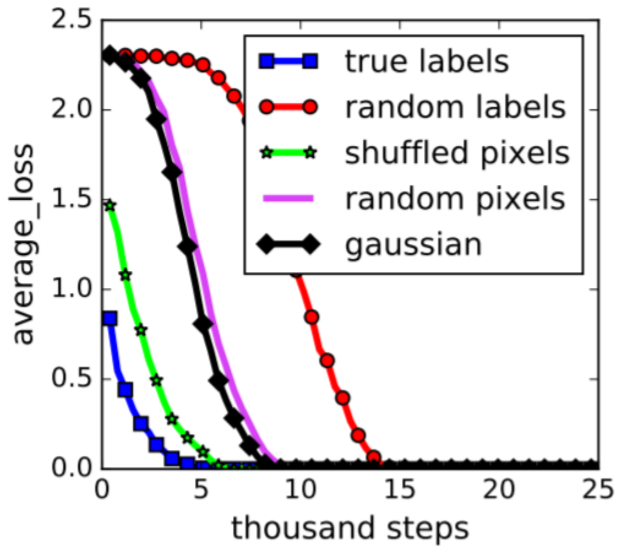
Reweighting k-folds of data based on cross-validation.



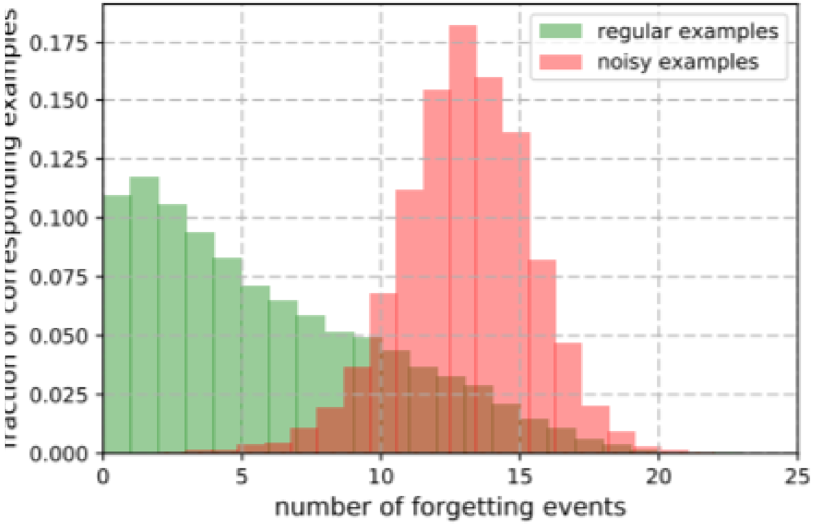
Coarse-grained noise estimation and redundant training and testing.

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

Unsupervised Denoising: Co-regularized Knowledge Distillation



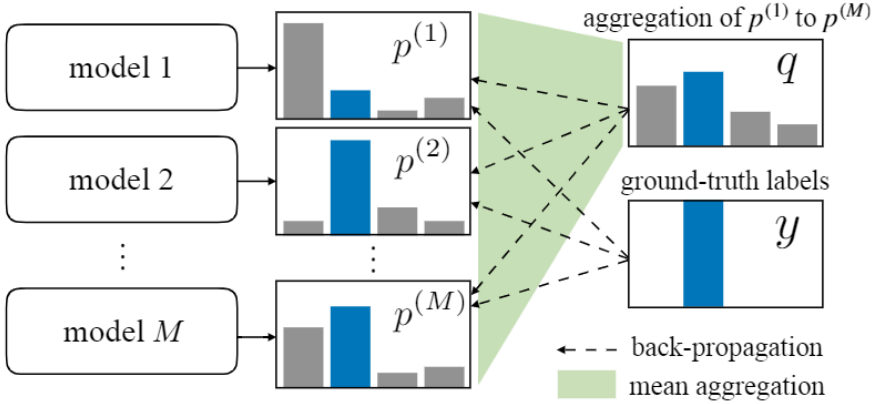
(1)



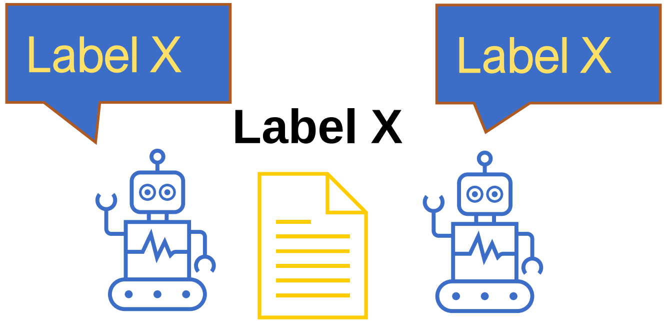
(2)

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

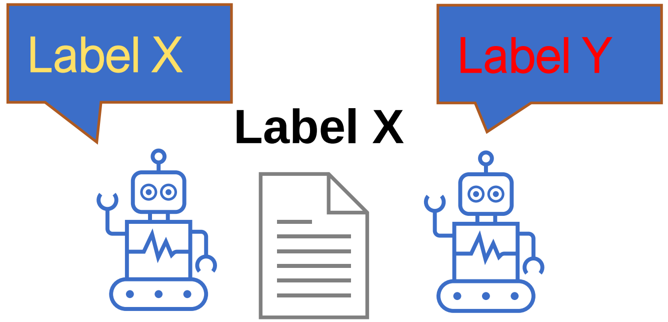
Noisy instances are outliers to the task inductive bias.



Co-regularization Framework



Agree: Clean ✓

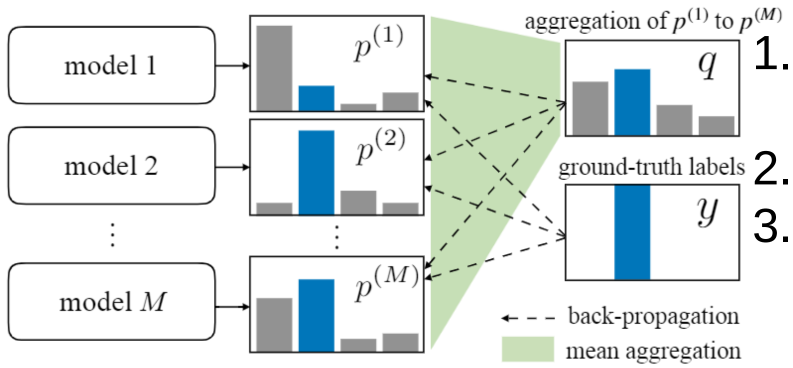


Disagree: Noise ✗

Mutual agreement by models indicates clean/noisy labels

Z&**C**. *Learning from Noisy Labels for Entity-Centric Information Extraction*. **EMNLP** 2021

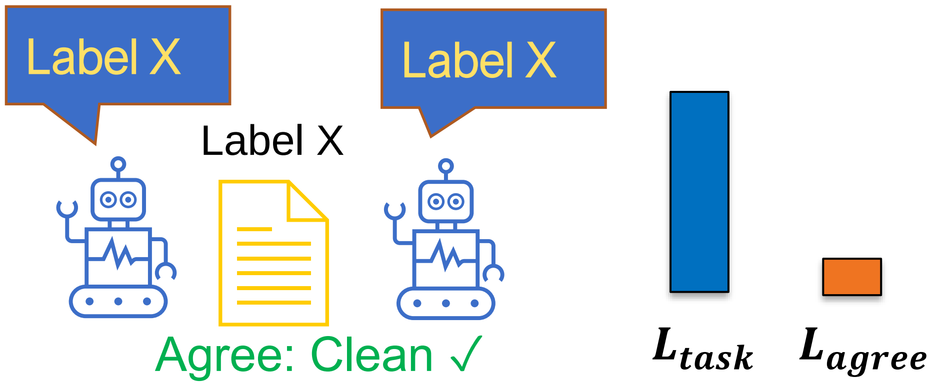
Unsupervised Denoising: Co-regularized Knowledge Distillation

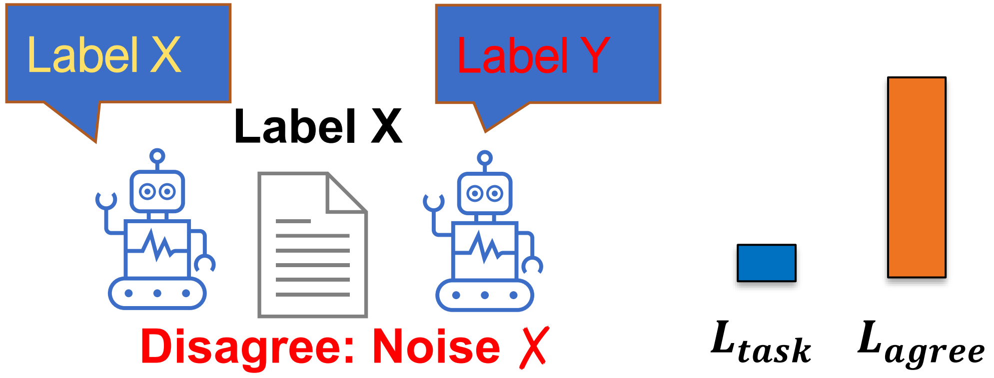


1. Create $M (\geq 2$; 2 is enough) identical models with **different initialization** (w/ **warm up**).
2. Train the models with both **the task loss** and an additional **agreement loss**.
3. Return one of the models.

Cross-entropy L_{task}

K-L divergence between model predictions L_{agree}





On clean data

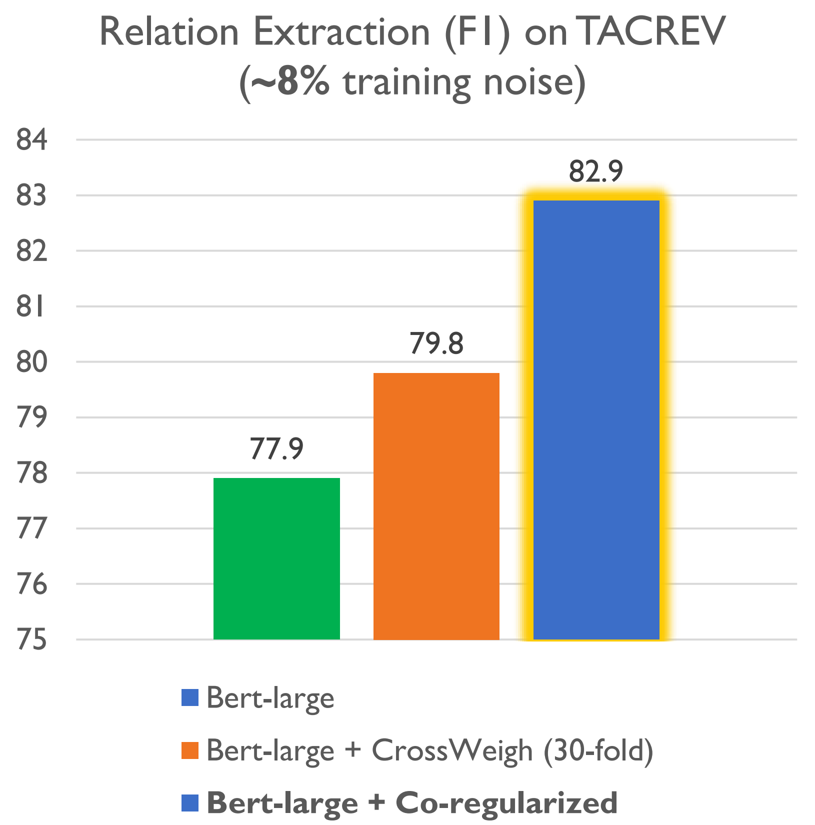
- Lower agreement loss
- Focusing on task optimization

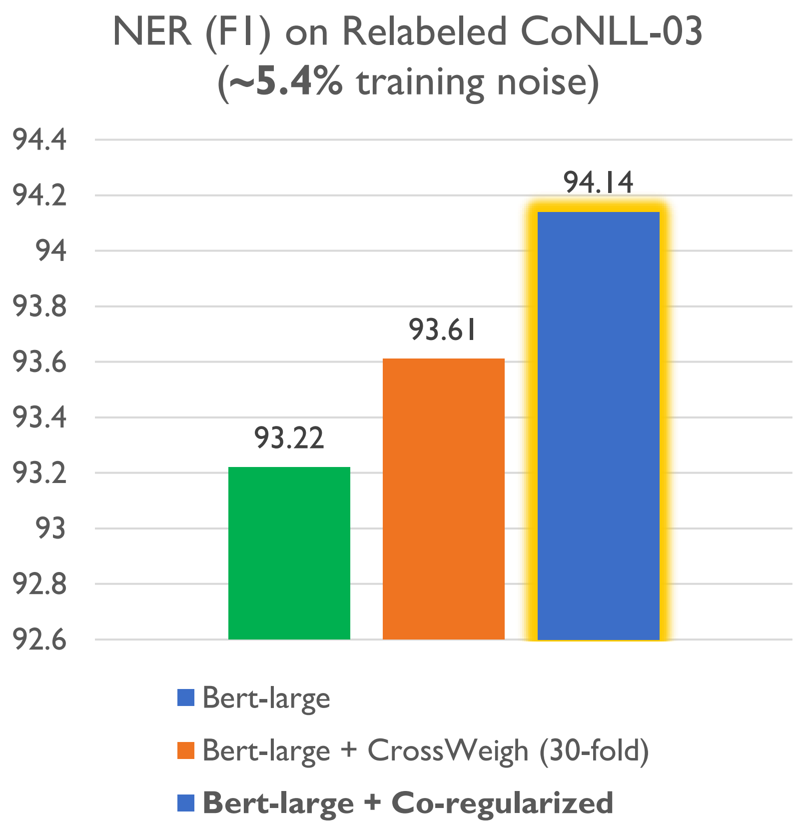
On noisy data

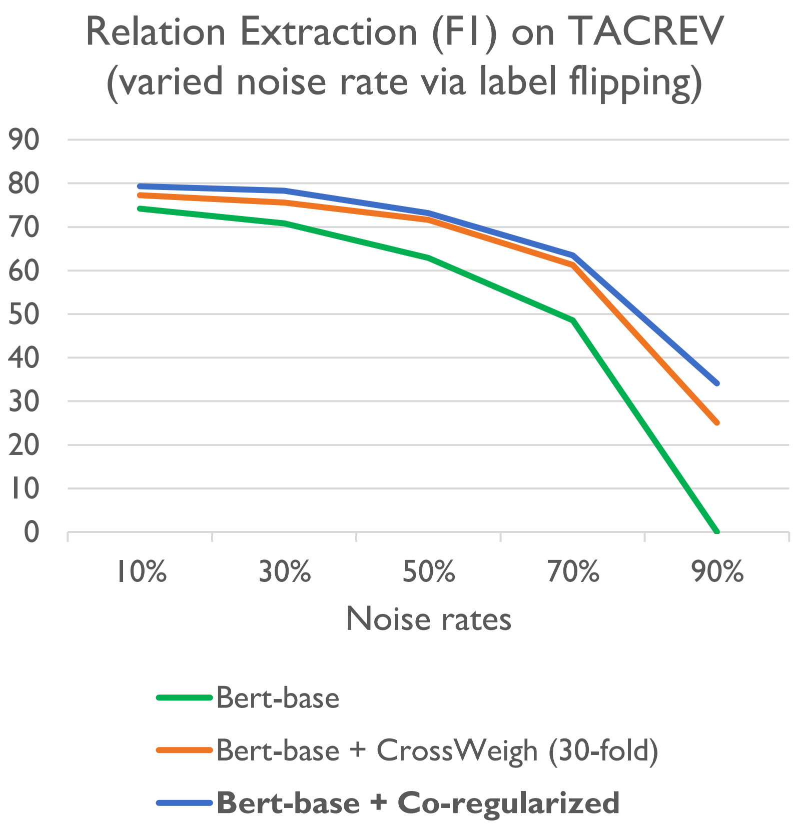
- Higher agreement loss
- Task optimization proactively prevents fitting those data

*Z&**C**. Learning from Noisy Labels for Entity-Centric Information Extraction*. **EMNLP** 2021

Unsupervised Denoising: Co-regularized Knowledge Distillation







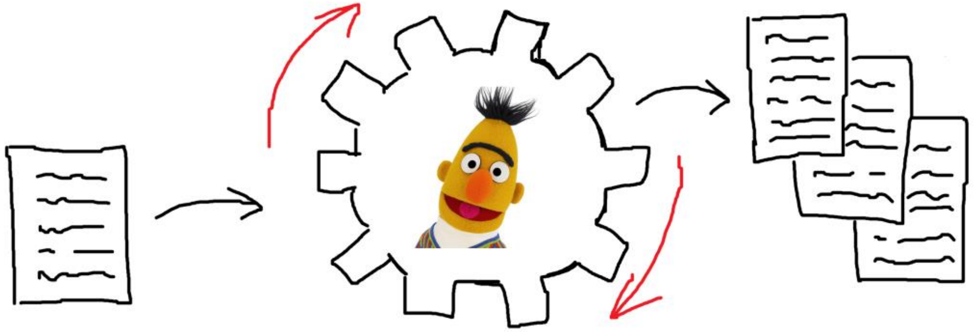
Merits of co-regularized knowledge distillation

- 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)

*Z&**C**. Learning from Noisy Labels for Entity-Centric Information Extraction*. **EMNLP** 2021

Our Continuing Studies

Robust Data Augmentation

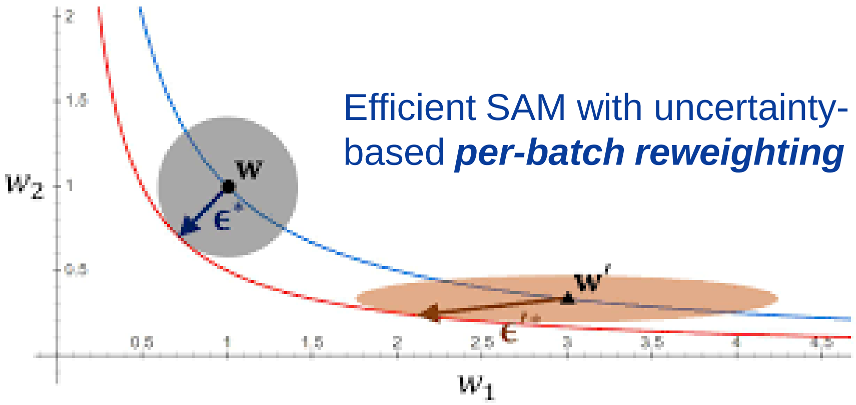


Denoising automatically augmented training data

- 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



Efficient SAM with uncertainty-based ***per-batch reweighting***

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

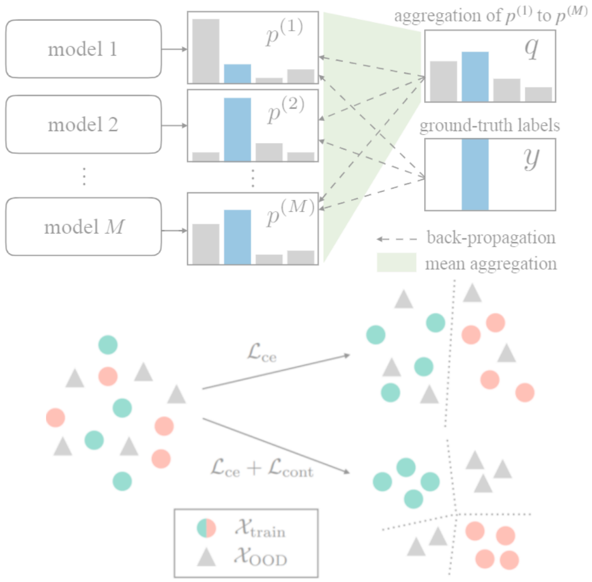
- Improving model robustness by finding flatter minima
- Consistent improvement on IE, textual retrieval, summarization, and NLU tasks

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

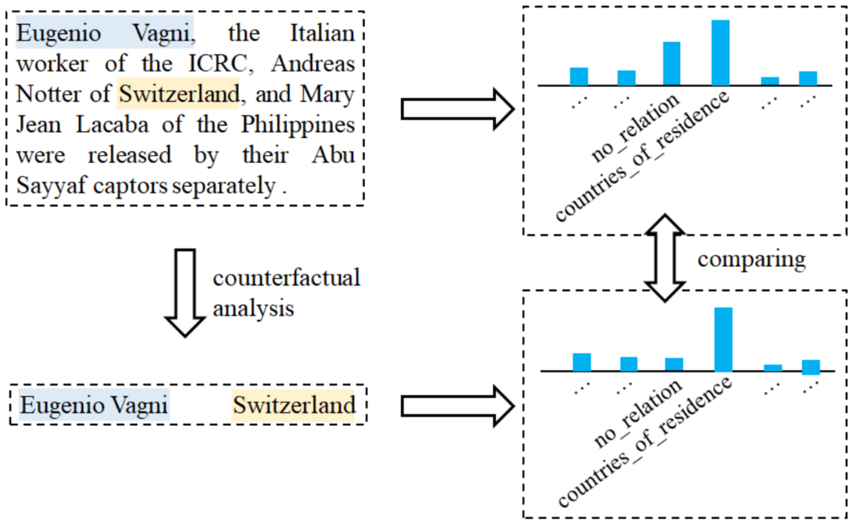
ZLZ***C***. *Sharpness-Aware Minimization with Dynamic Reweighting*. **EMNLP** 2022

Agenda

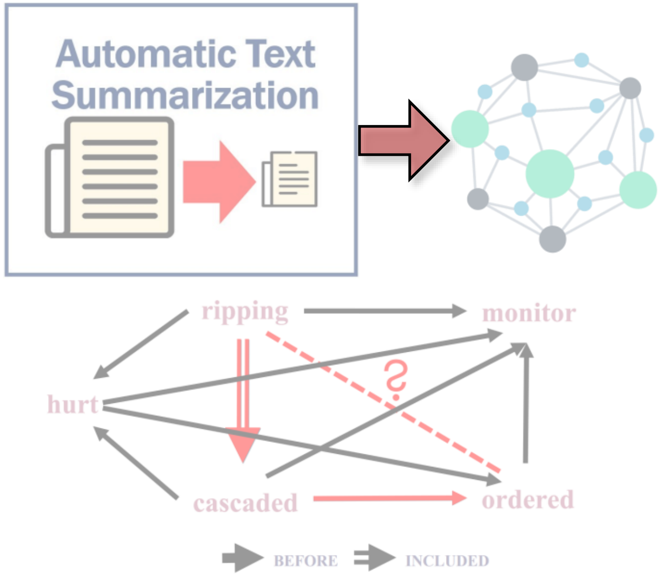
1. Noise-robust IE



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: Bill Gates PER paid a visit to Building 99 MISC of Microsoft ORG yesterday .

Rel?

Visit ✓

FounderOf ✗

According to prior knowledge

Prior knowledge (in PLMs) may lead to unfaithful extraction

Event relation extraction: I went to *event1* see the doctor. However, I got more seriously *event2* sick.

Before? After?

Before ✓

After ✗

Data

According to statistics

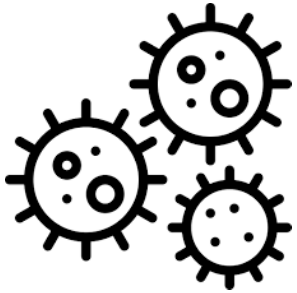
(Statistically) Biased training may lead to unfaithful extraction

Why Faithful IE Is So Important

Drug-drug Interaction

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



Disease-target detection

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**

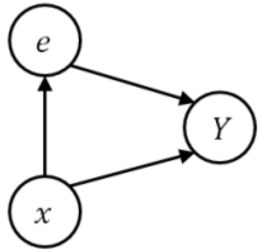
- **Disease phenotype extraction** from medical reports
- **Disaster event extraction** from social media
- **API version compatibility** detection from software documents
- **Travel event extraction** from emails and meeting logs
- …

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*.

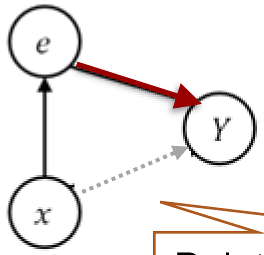


Relations should be inferred based on both mentions and the context

What it may actually do

Bill Gates [illegible] Microsoft [illegible]

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?

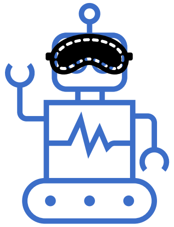
WCZCLLH. *Should We Rely on Entity Mentions for Relation Extraction? Debiasing Relation Extraction with Counterfactual Analysis*. **NAACL** 2022

Debiased Training

Mention masks: mask out entity names with their types

Person paid a visit to Building 99 of Org yesterday.

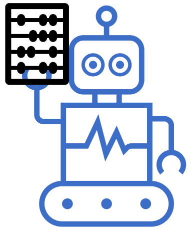
Similarly for *event RE*, we can mask using trigger types and tense



Mask mentions in both training and inference

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

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



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

Upweight hard instances

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

Lin et al. *Focal loss for dense object detection*. **CVPR** 2017 (FAIR)
Liu et al. *Just Train Twice: Improving Group Robustness without Training Group Information*. **ICML** 2021 (Stanford)

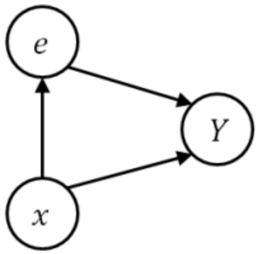
Our Strategy: Counterfactual Inference

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

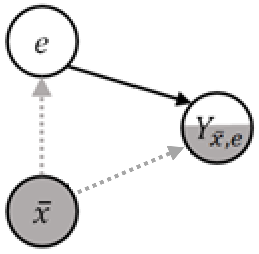
Instance: Bill Gates paid a visit to Building 99 of Microsoft yesterday.

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

Counterfactual instance: Bill Gates Microsoft



Original Prediction



Entity-specific Bias

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

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

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

WCZCLYH. *Should We Rely on Entity Mentions for Relation Extraction? Debiasing Relation Extraction with Counterfactual Analysis*. **NAACL** 2022

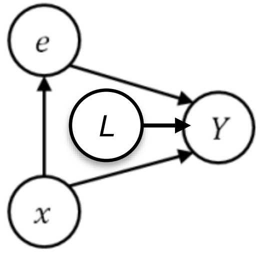
Our Strategy: Counterfactual Inference

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 $f(X) = P(Y|X) = \sum_E P(Y|X, E)P(E|X)$

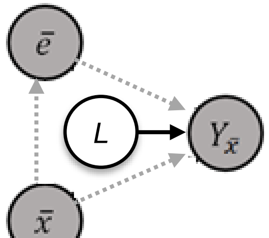
Counterfactual instance: ∅



Original Prediction

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



Label Bias

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

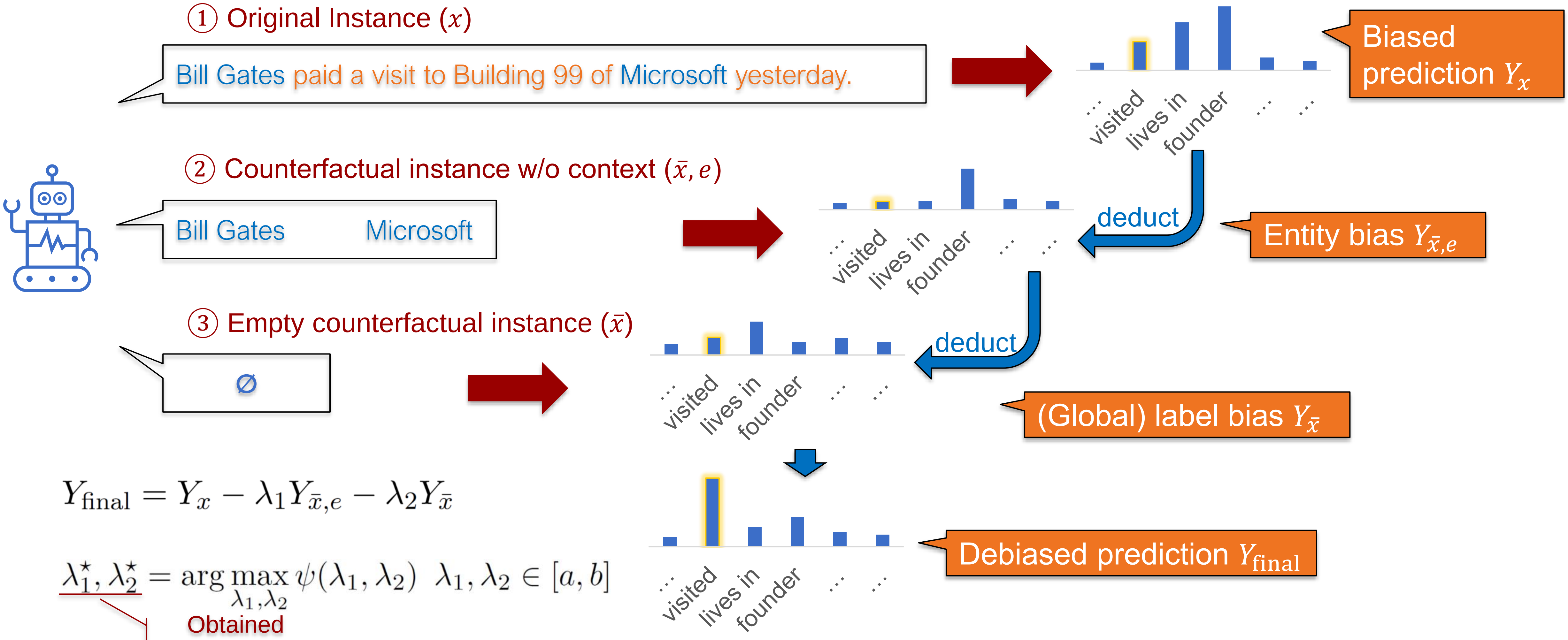
- Also estimates the *Total Effect* [Pearl and Mackenzie, 2018] of *X*

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

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

Our Strategy: Counterfactual Inference

Deducting the distilled biases from the original prediction



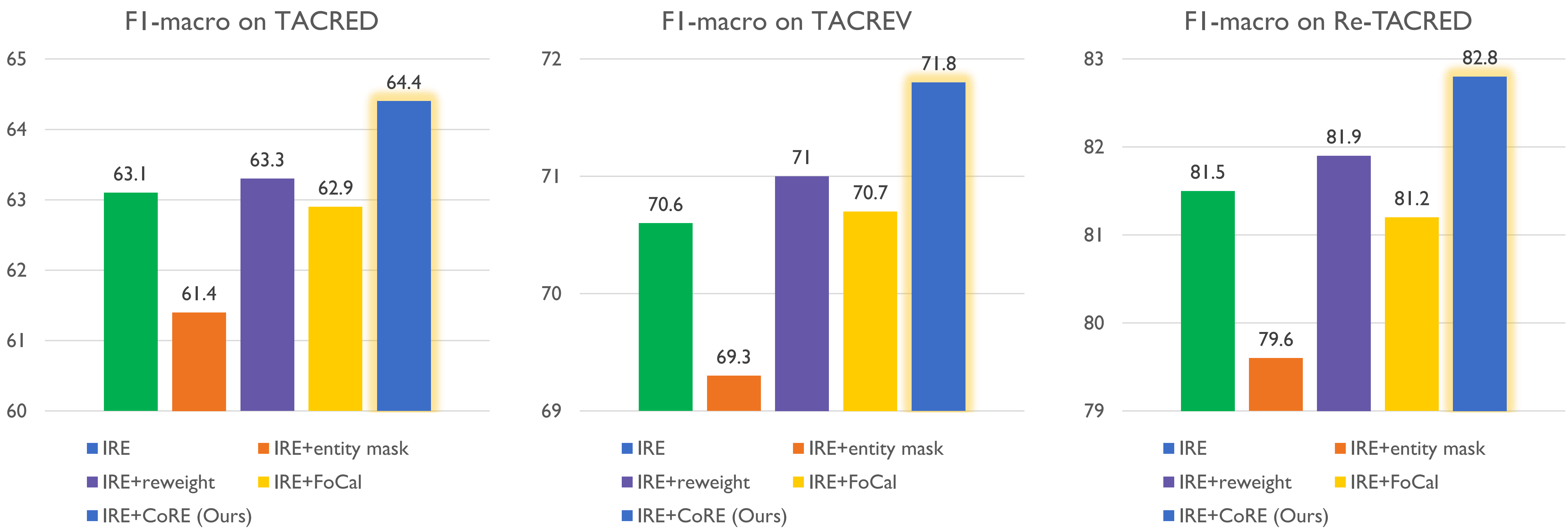
$$Y_{\text{final}} = Y_x - \lambda_1 Y_{\bar{x},e} - \lambda_2 Y_{\bar{x}}$$

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

Obtained on dev set

WCZCLYH. *Should We Rely on Entity Mentions for Relation Extraction? Debiasing Relation Extraction with Counterfactual Analysis*. **NAACL** 2022

Counterfactual Inference



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

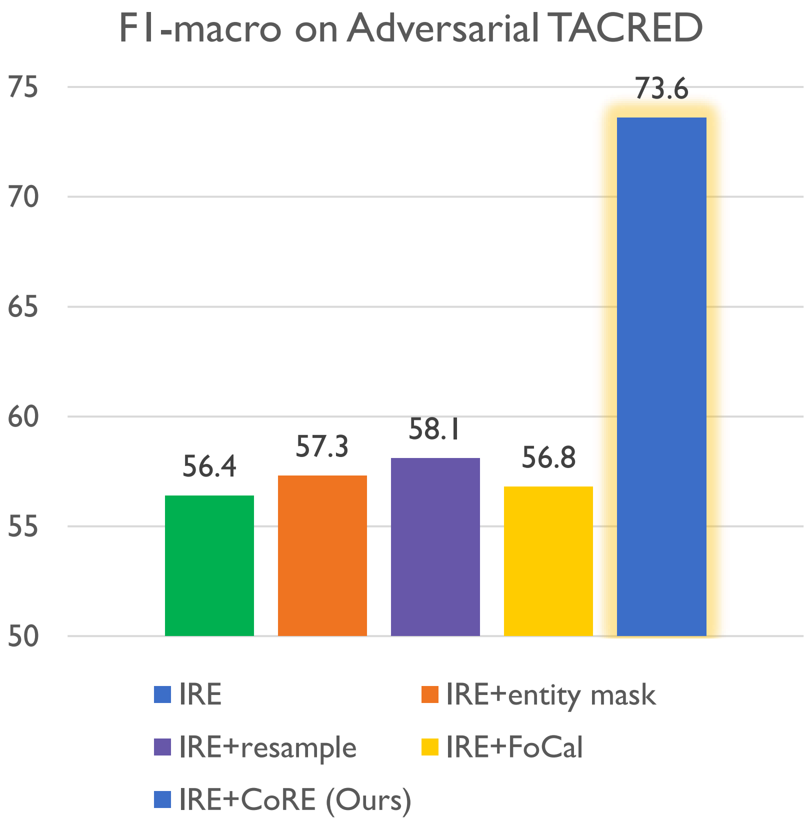
*$\text{IRE}_{\text{RoBERTa}}$ is one of the best-performing sentence-level RE models (Z&**C** AACL 2022). Results are also available for LUKE.

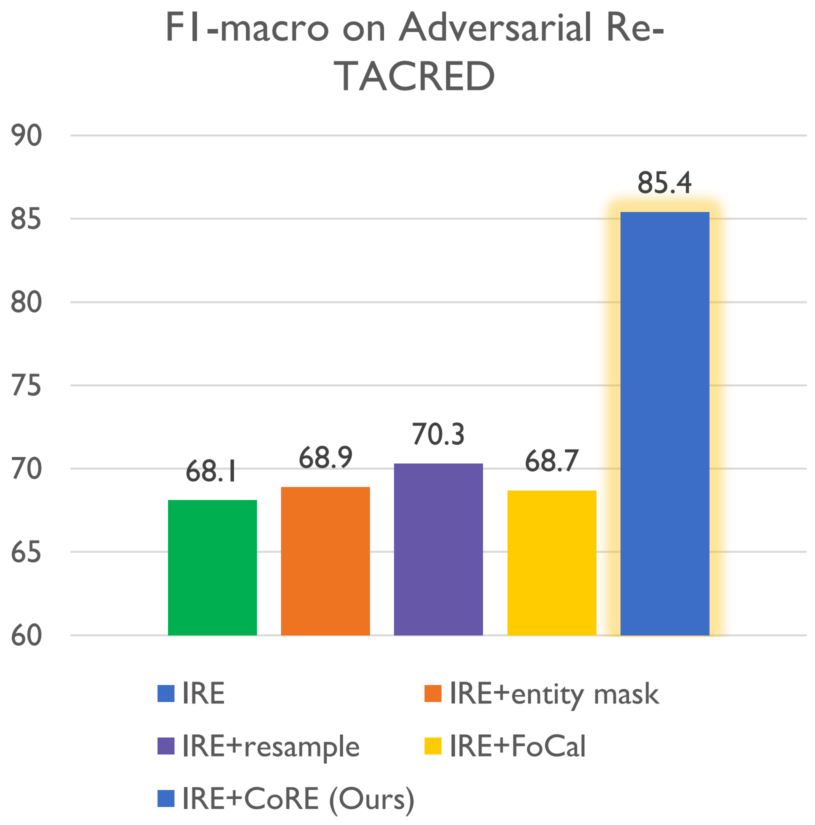
W**C**ZCLYH. *Should We Rely on Entity Mentions for Relation Extraction? Debiasing Relation Extraction with Counterfactual Analysis*. **NAACL** 2022

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.

WCZCLYH. *Should We Rely on Entity Mentions for Relation Extraction? Debiasing Relation Extraction with Counterfactual Analysis*. **NAACL** 2022

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

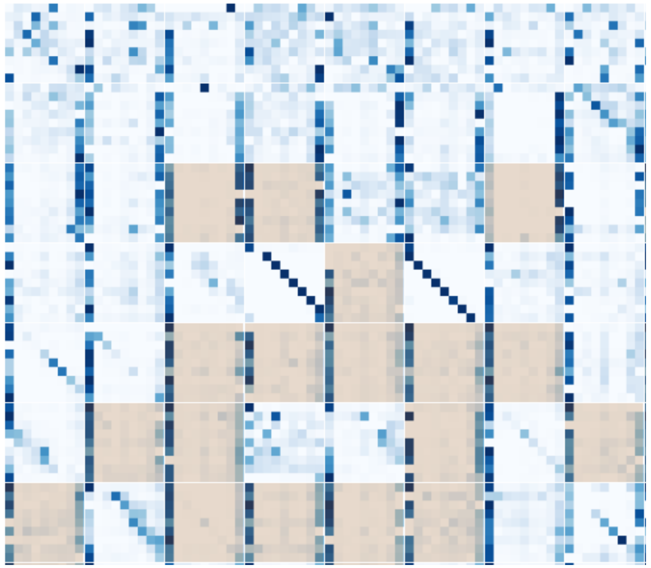
Input: Last week I stayed in ***Treasure Island*** for two nights when visiting Las Vegas.
Gold labels: hotel, resort, location, place
Pred labels: island, land, location, place

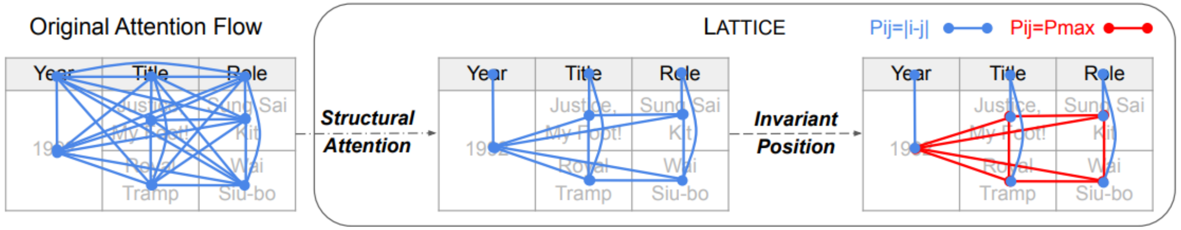
Dependency bias

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



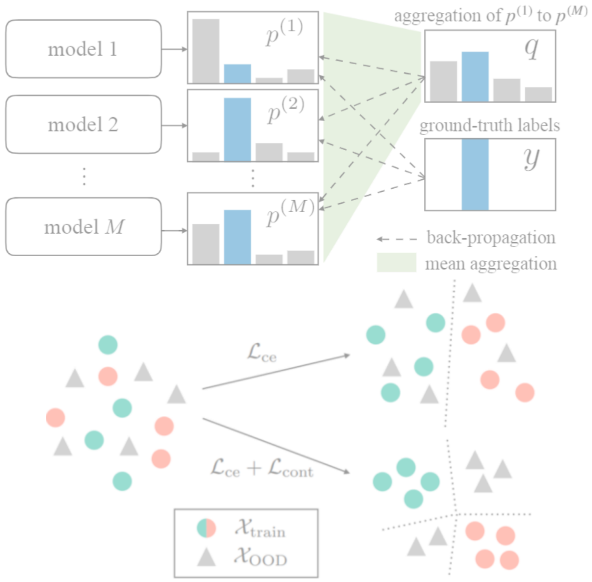


- Attention smoothing, perturbation, constrained PoE
- Feature-equivariance learning

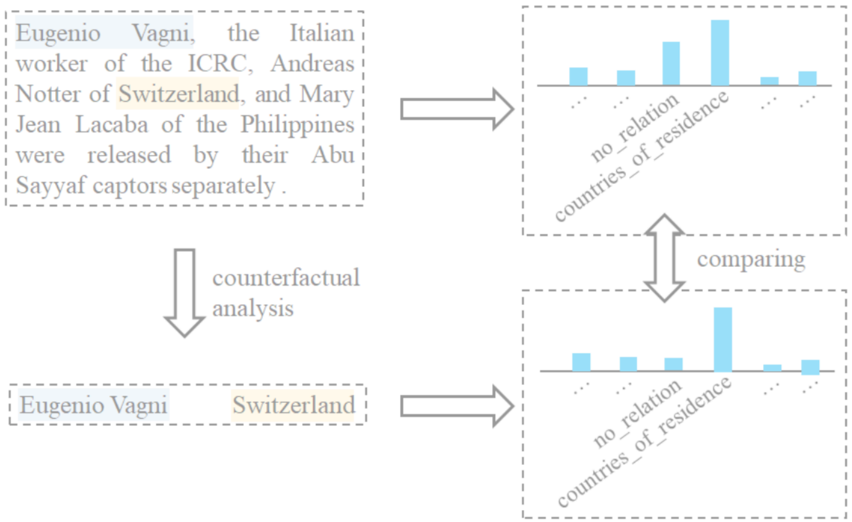
XWLD**C**. *Does Your Model Classify Entities Reasonably? Diagnosing and Mitigating Spurious Correlations in Entity Typing*. **EMNLP** 2022
WXS**C**. *Robust (Controlled) Table-to-Text Generation with Structure-Aware Equivariance Learning*. **NAACL** 2022

Agenda

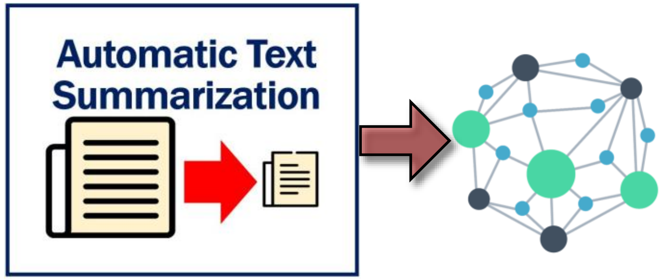
1. Noise-robust IE

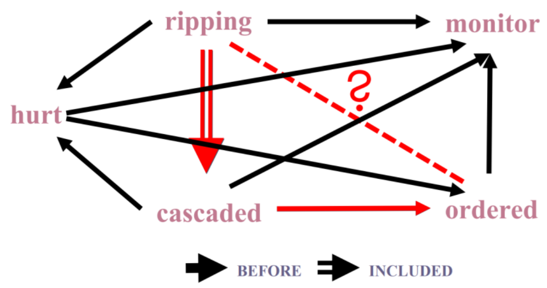


2. Faithful IE



3. Indirectly Supervised IE





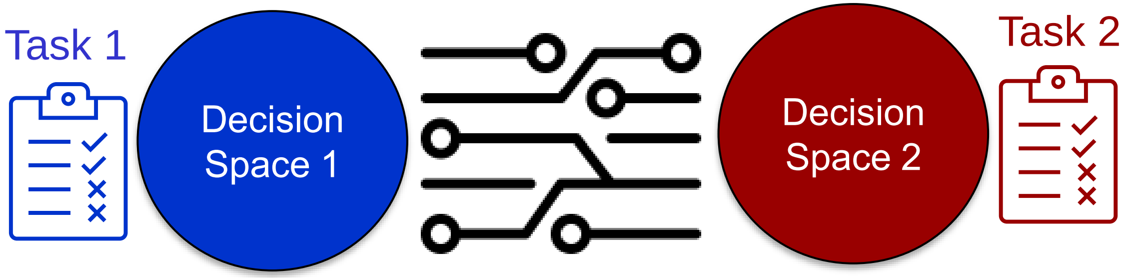
4. Future Directions

Two Forms of Indirect Supervision

IE suffers from insufficient supervision

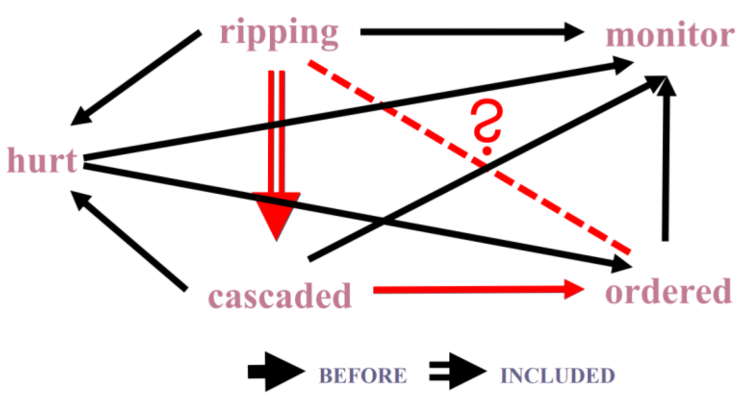
At least one has insufficient supervision

Direct annotation is difficult and expensive

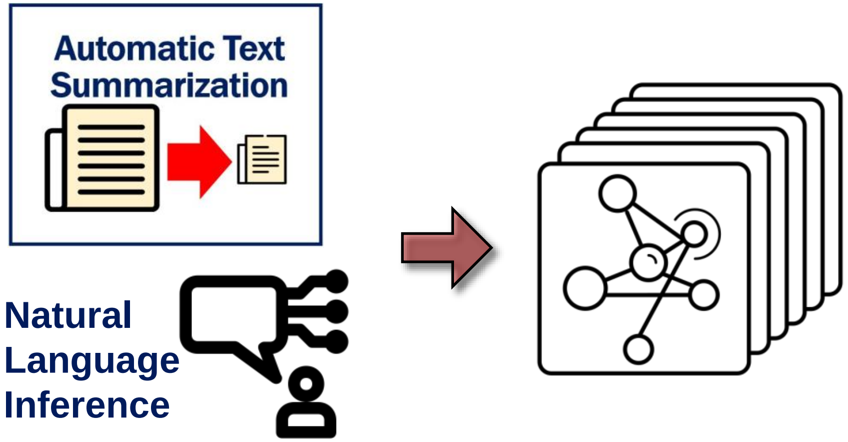


Can we bridge the supervision of different tasks?

Two Forms of Indirect supervision



(Logically) Constrained Learning



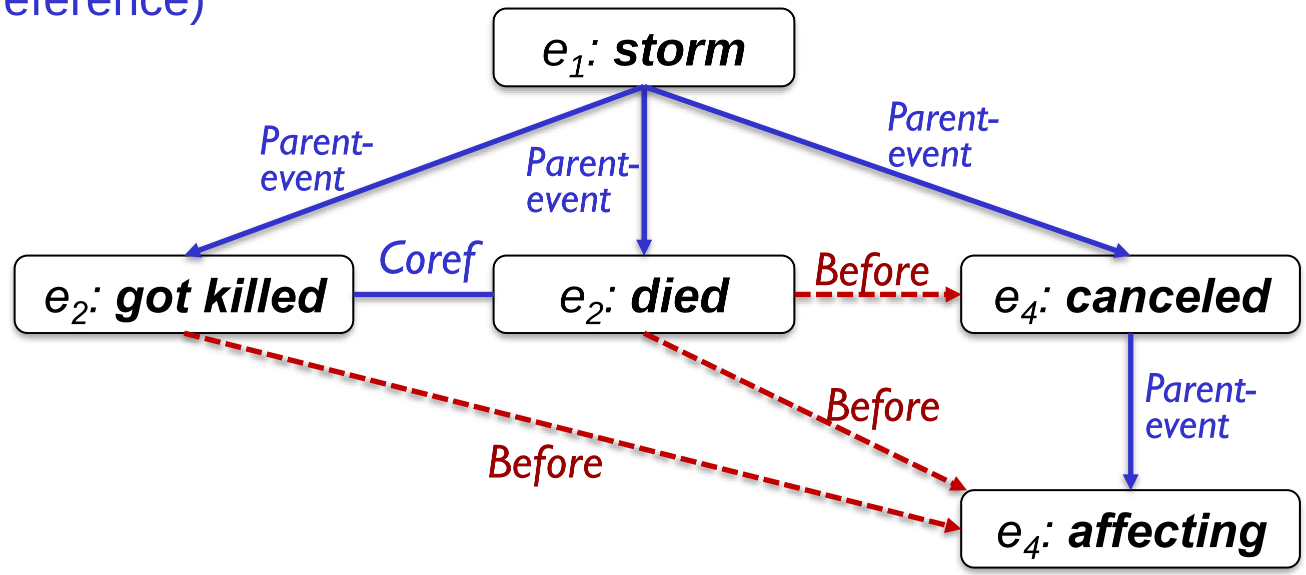
Cross-task Transfer

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

e1:storm is PARENT of *e4:canceled* => *e1:storm* is BEFORE *e4:canceled*

Conjunction

e3:died is BEFORE *e4:canceled* ∧*e4:canceled* is a PARENT of *e5:affecting* =>
e3:died is BEFORE *e5:affecting*

Implication, *Negation*, …

Use logical constraints!

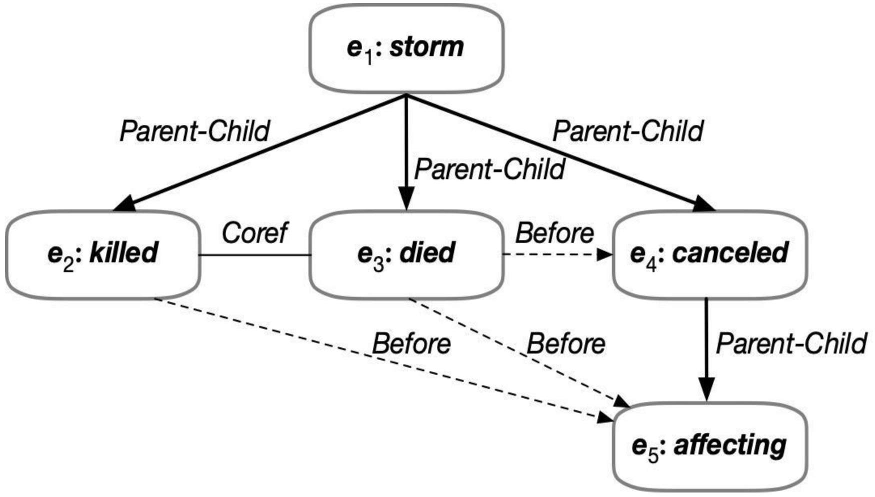
Logical Constraints Of Relations

Dependency of Decisions

Symmetry
e3:died is BEFORE *e4:canceled*
=> *e4:canceled* is AFTER *e3:died*

Conjunction
e3:died is BEFORE *e4:canceled*
∧*e4:canceled* is a PARENT of *e5:affecting*
=> *e3:died* BEFORE *e5:affecting*

Transitivity
e1:storm is PARENT of *e4:canceled*
∧*e4:canceled* is a PARENT of *e5:affecting*
=> *e1:storm* is a PARENT of *e5:affecting*



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

Goal: incorporating logical constraints into neural model training.

- Learning to provide **globally consistent** predictions
- Providing **indirect supervision** across tasks/decision spaces

W**C**ZR. *Joint Constrained Learning for Event-Event Relation Extraction*. **EMNLP 2020**

Incorporating Logical Constraints in A Neural Architecture

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

Using **product *t*-norm** model constraints as differentiable functions

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

Constraints become entropy regularizers

α \ β	PC	CP	CR	NR	BF	AF	EQ	VG
PC	PC, ¬AF	–	PC, ¬AF	¬CP, ¬CR	BF, ¬CP, ¬CR	–	BF, ¬CP, ¬CR	–
CP	–	CP, ¬BF	CP, ¬BF	¬PC, ¬CR	–	AF, ¬PC, ¬CR	AF, ¬PC, ¬CR	–
CR	PC, ¬AF	CP, ¬BF	CR, EQ	NR	BF, ¬CP, ¬CR	AF, ¬PC, ¬CR	EQ	VG
NR	¬CP, ¬CR	¬PC, ¬CR	NR	–	–	–	–	–
BF	BF, ¬CP, ¬CR	–	BF, ¬CP, ¬CR	–	BF, ¬CP, ¬CR	–	BF, ¬CP, ¬CR	¬AF, ¬EQ
AF	–	AF, ¬PC, ¬CR	AF, ¬PC, ¬CR	–	–	AF, ¬PC, ¬CR	AF, ¬PC, ¬CR	¬BF, ¬EQ
EQ	¬AF	¬BF	EQ	–	BF, ¬CP, ¬CR	AF, ¬PC, ¬CR	EQ	VG, ¬CR
VG	–	–	VG, ¬CR	–	¬AF, ¬EQ	¬BF, ¬EQ	VG	–

Capturing 80 constraints in total

W**C**ZR. *Joint Constrained Learning for Event-Event Relation Extraction*. **EMNLP 2020**

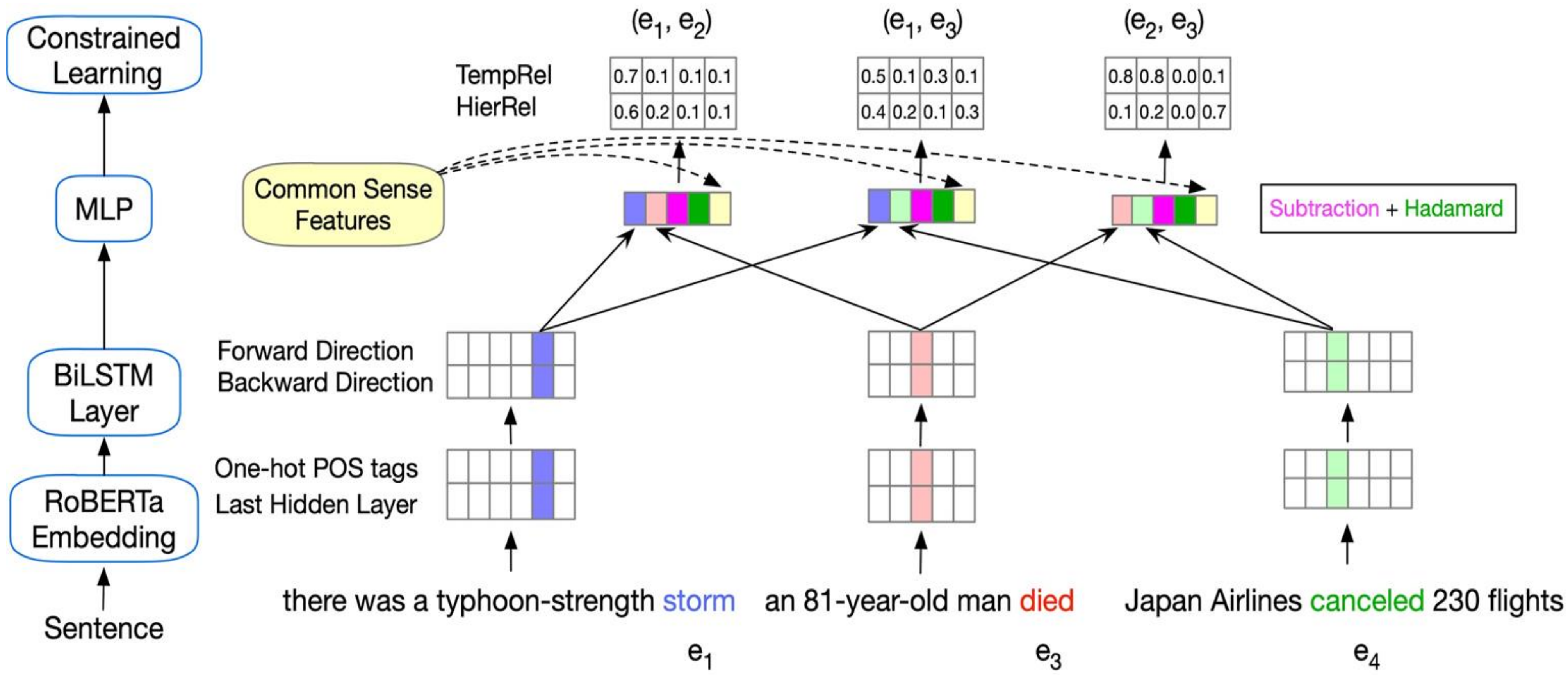
Joint Constrained Learning

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

Task loss

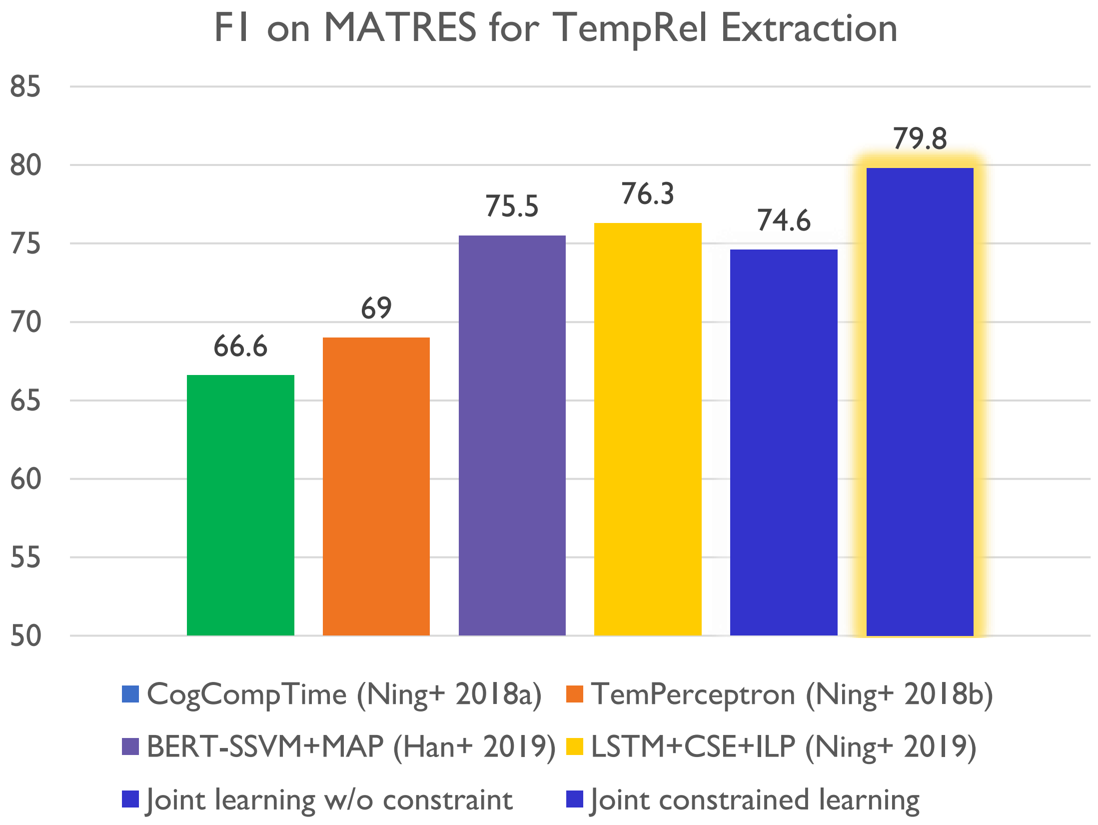
Implication and conjunction constraint losses

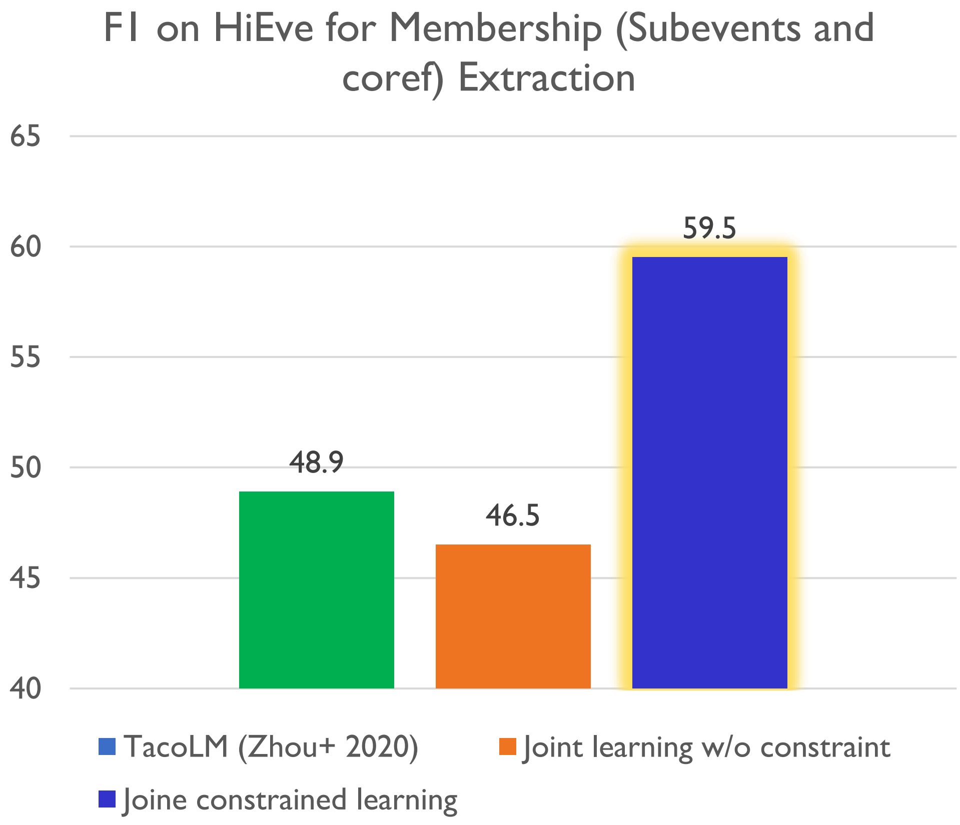
Loss Function: $L = L_A + \lambda_S L_S + \lambda_C L_C$



WCZR. *Joint Constrained Learning for Event-Event Relation Extraction*. **EMNLP 2020**

The Joint Constrained Learning Architecture





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

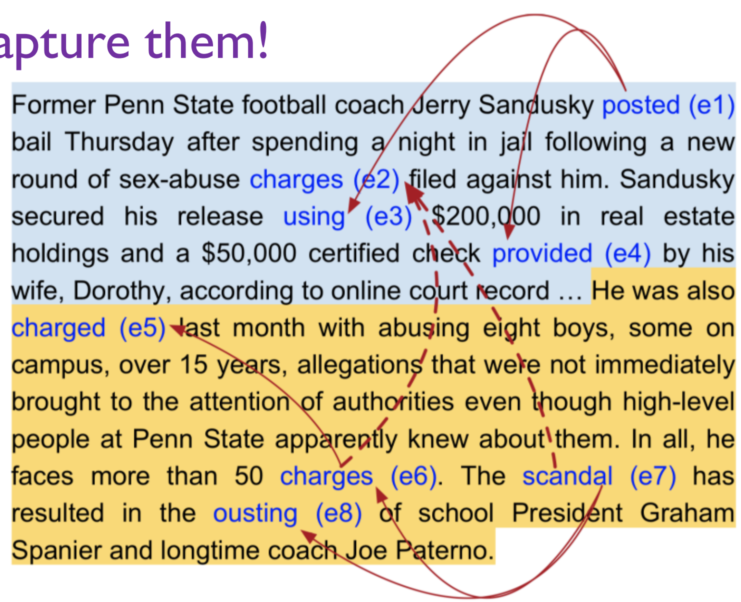
W**C**ZR. *Joint Constrained Learning for Event-Event Relation Extraction*. **EMNLP 2020**

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]



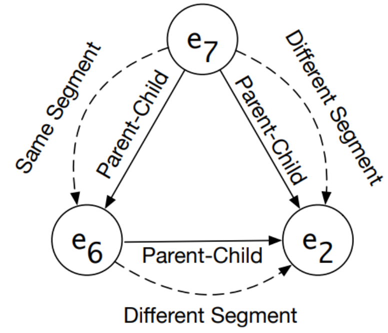
An (implicit) soft logic constraint. How do we capture it?

Constraint Learning

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

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

Estimates probability of a legitimate triangle

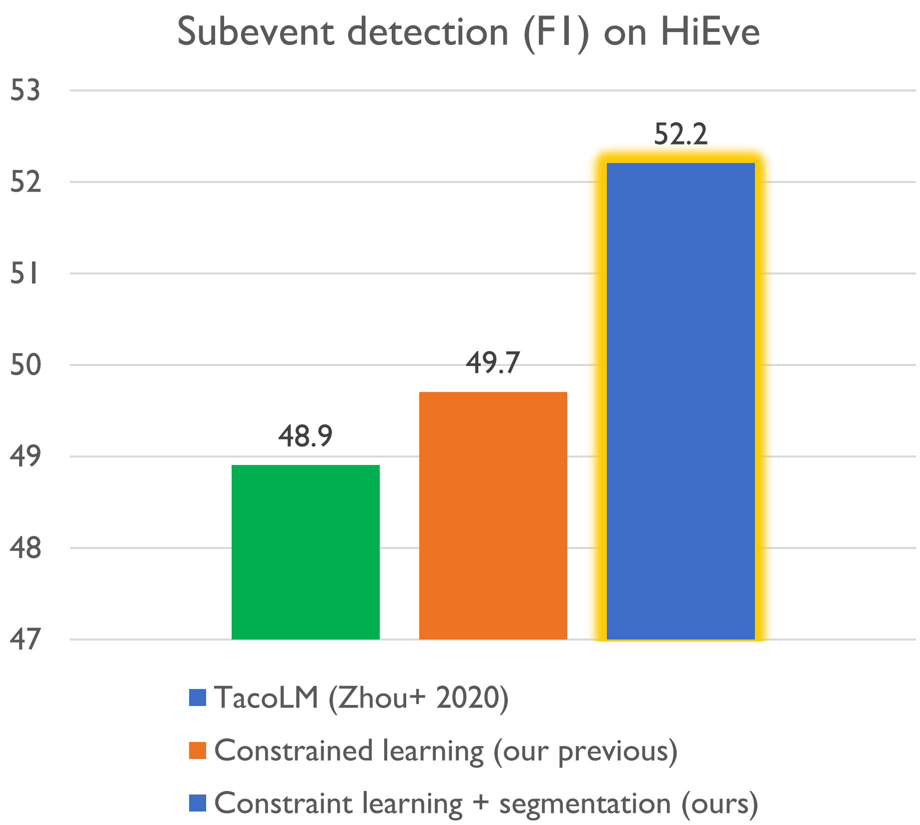


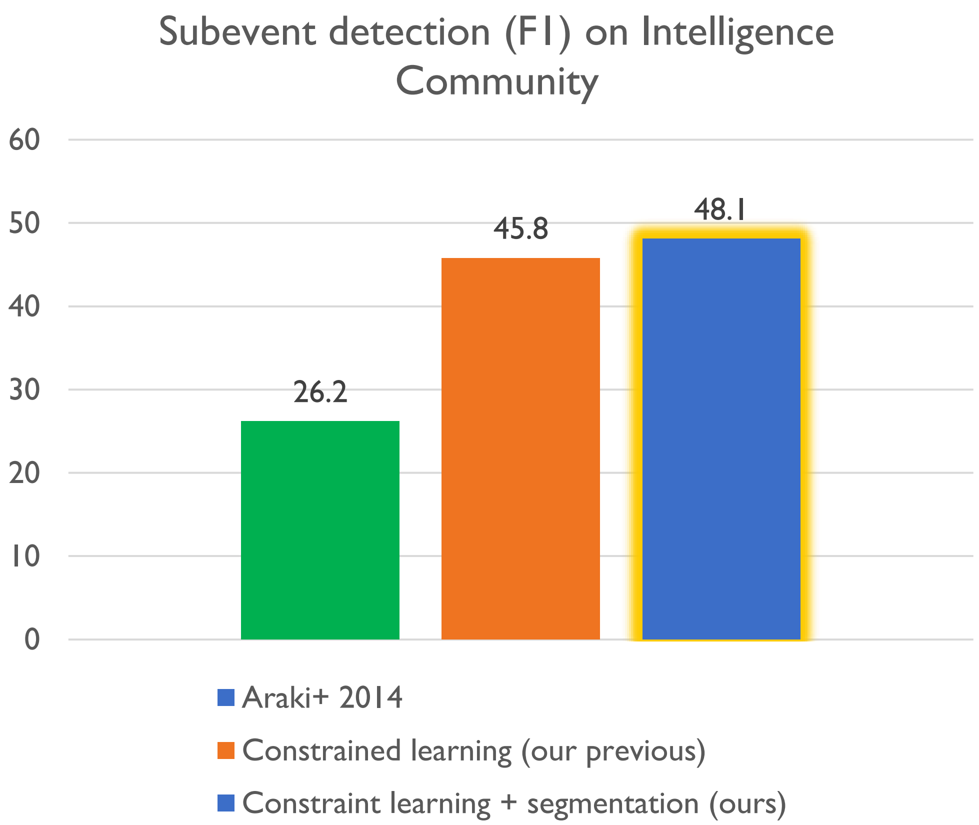
Adding constraint losses according to the the rectifier estimated "truth values" of constraints

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

WCR. *Learning Constraints and Descriptive Segmentation for Subevent Detection.* **EMNLP 2021**

Automatically Learning Constraints





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

WCR. *Learning Constraints and Descriptive Segmentation for Subevent Detection*. **EMNLP 2021**

Indirect supervision via Cross-task Transfer

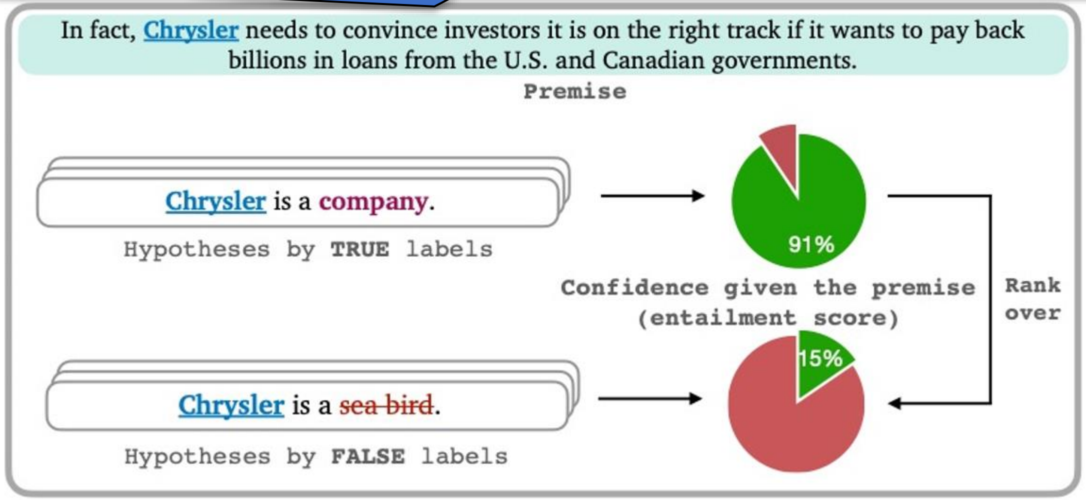
Ultra-fine Entity Typing

- **>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

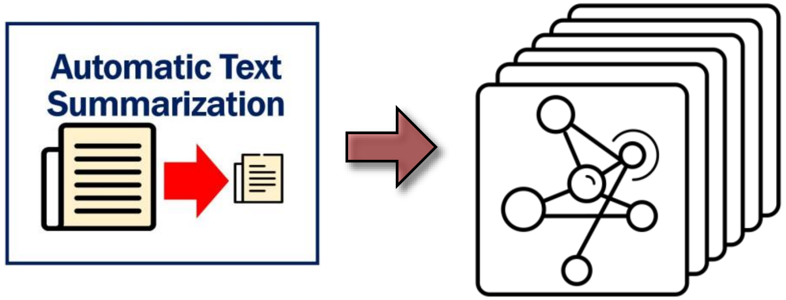
Indirect Supervision from Natural language Inference



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

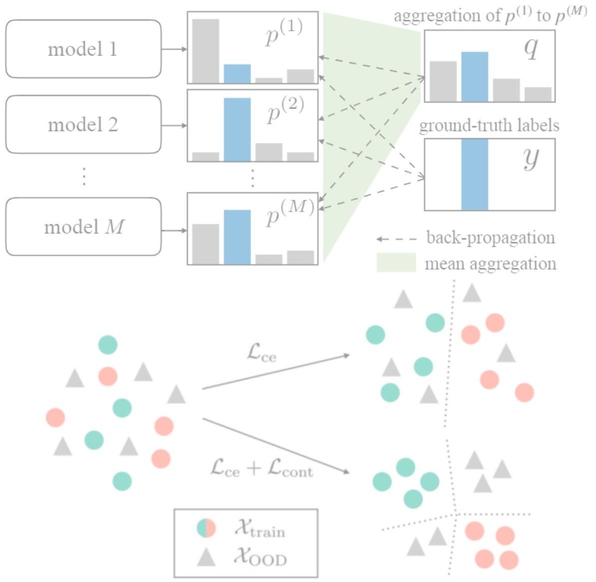
Close to SOTA performance using only 5% of training data on TACRED

LY***C***. Ultra-fine Entity Typing with Indirect Supervision from Natural Language Inference. **TACL** 2022

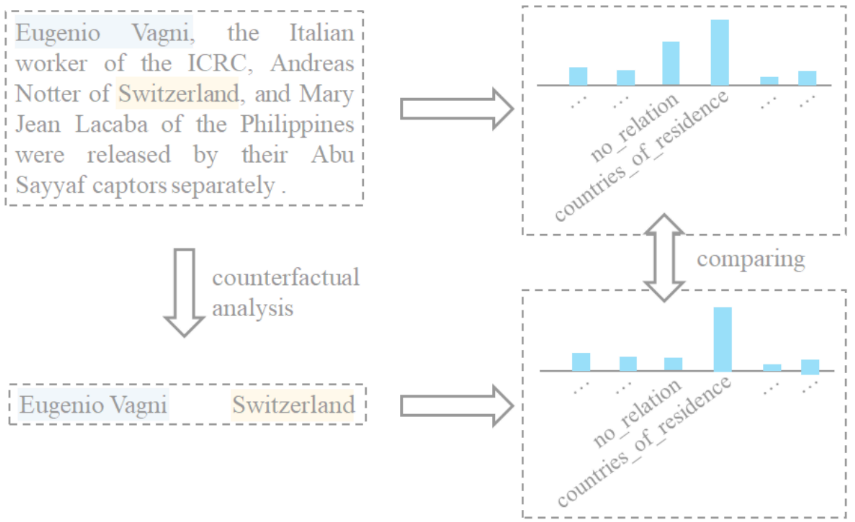
LHZM***C***. Summarization as Indirect Supervision for Relation Extraction. **Findings of EMNLP** 2022

Agenda

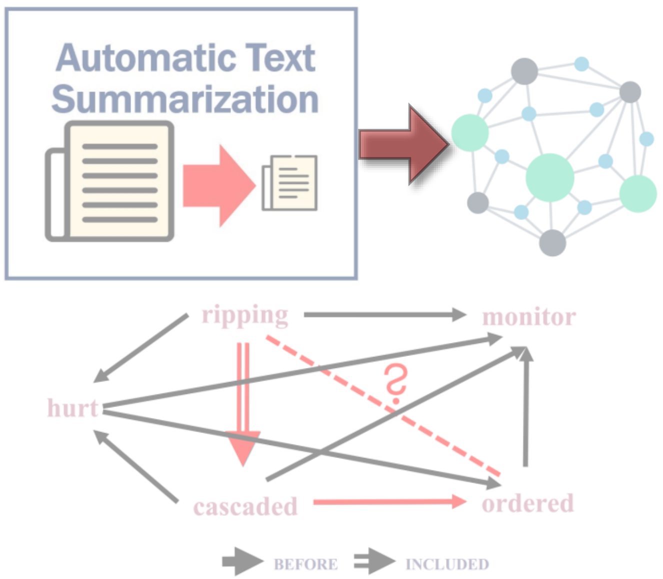
1. Noise-robust IE



2. Faithful IE



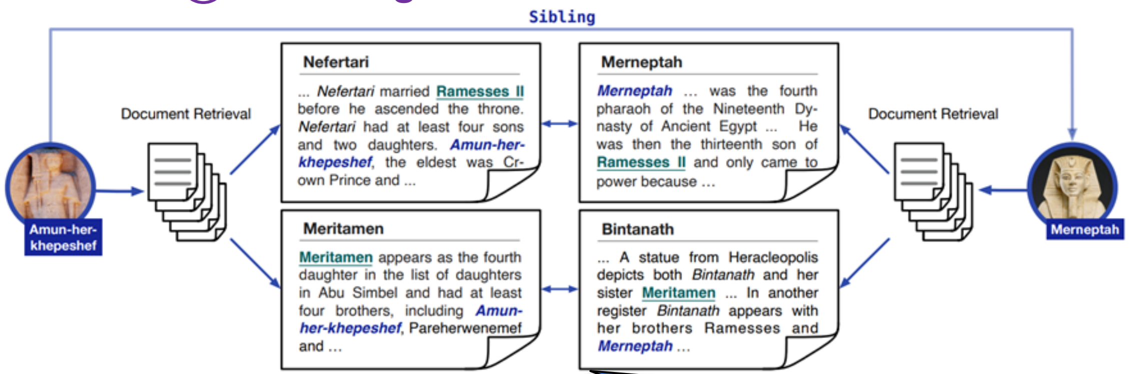
3. Indirectly Supervised IE



4. Future Directions

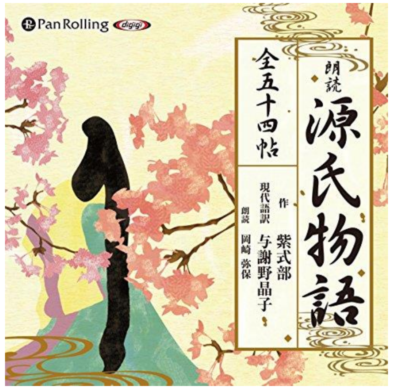
Thinking Across Documents

① Inducing relations across documents



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

② Consolidating unevenly distributed knowledge



DBpedia → Novel

チビペディア → Monogatari (story), Love story, Royal family story, Realistic novel, Ancient literature

③ From understanding "what the text says" to "what is happening"

Many more challenges to IE

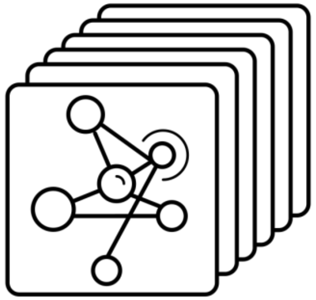
- Multi-hop reasoning
- Consolidation
- Tracking information pollution
- Long-form document modeling
- Mitigating frequency biases
- …

A long way to go

Accountable Indirect Supervision

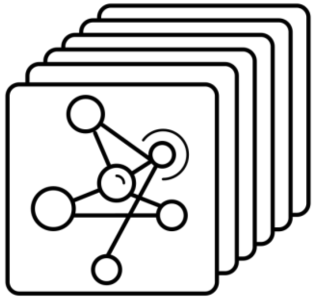
Decision choices for indirectly supervised learning needs to be accountable

Measuring the Affinity of Indirect Supervision Signals



How informative is the source-task data?

What is an efficient and precise measurement?

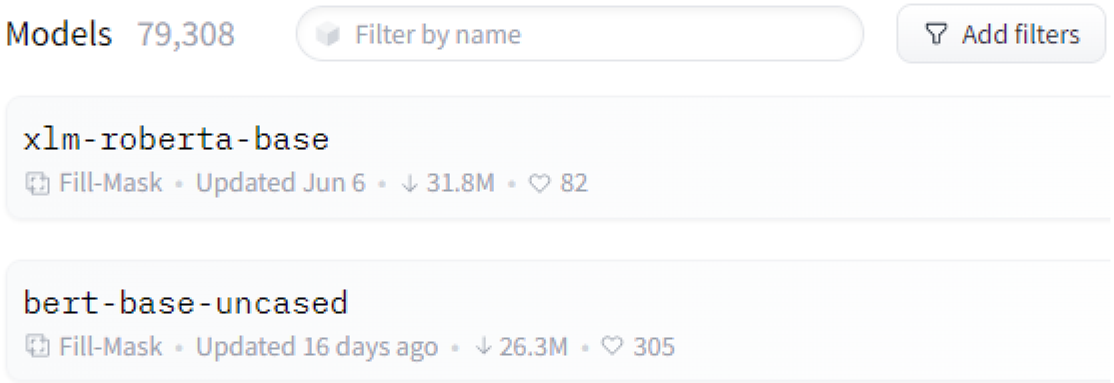


How compatible is the pretrained model?

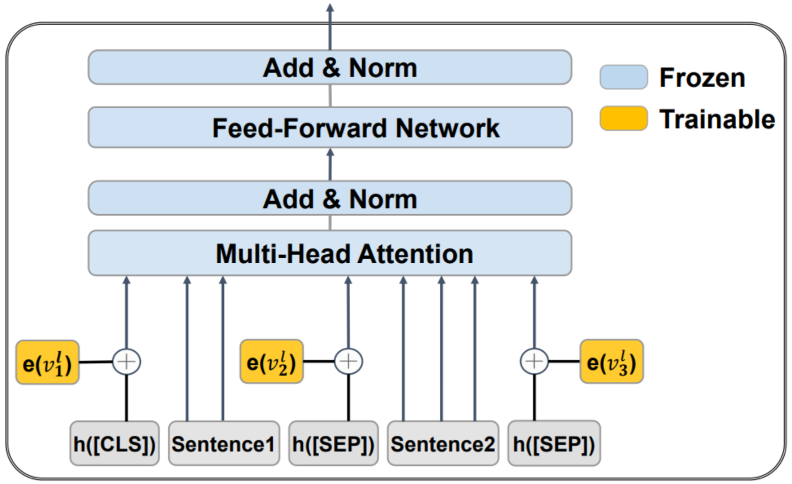
Gradient-based instance attribution?

Finding Indirect Supervision from the Model Hub

Hugging Face

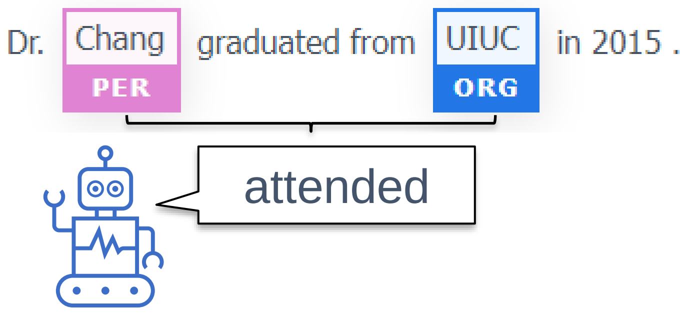


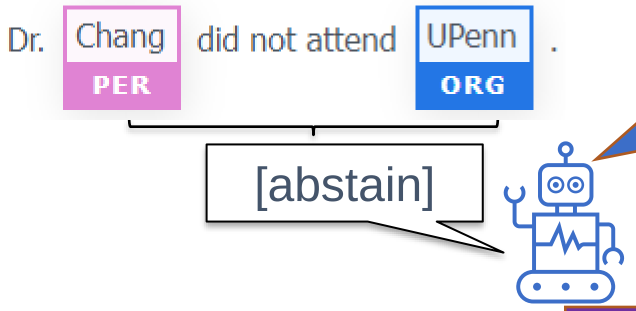
Efficient Adaptation of Large Models to Indirect Supervision



Selective Extraction

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

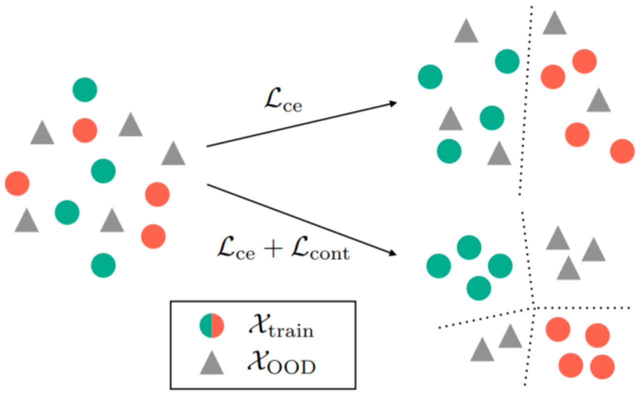




IE models can be exposed to many exception cases in real-world application.

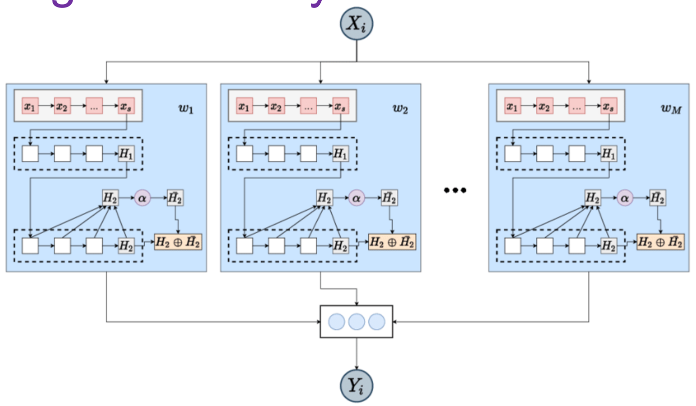
How to know its decision boundary.

Unsupervised out-of-distribution (OOD) detection



Increase inter-class discrepancy ⇒ Better OOD detection

Estimating Uncertainty for Prediciton

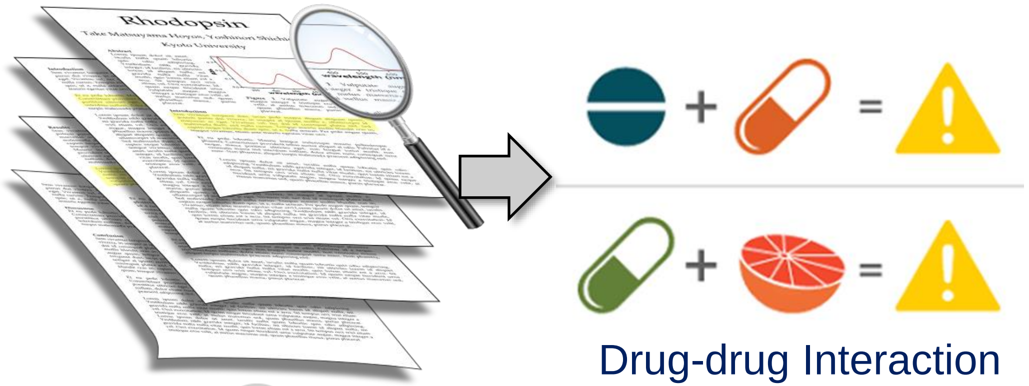


Softmax Response, Monte-Carlo Dropout, etc.

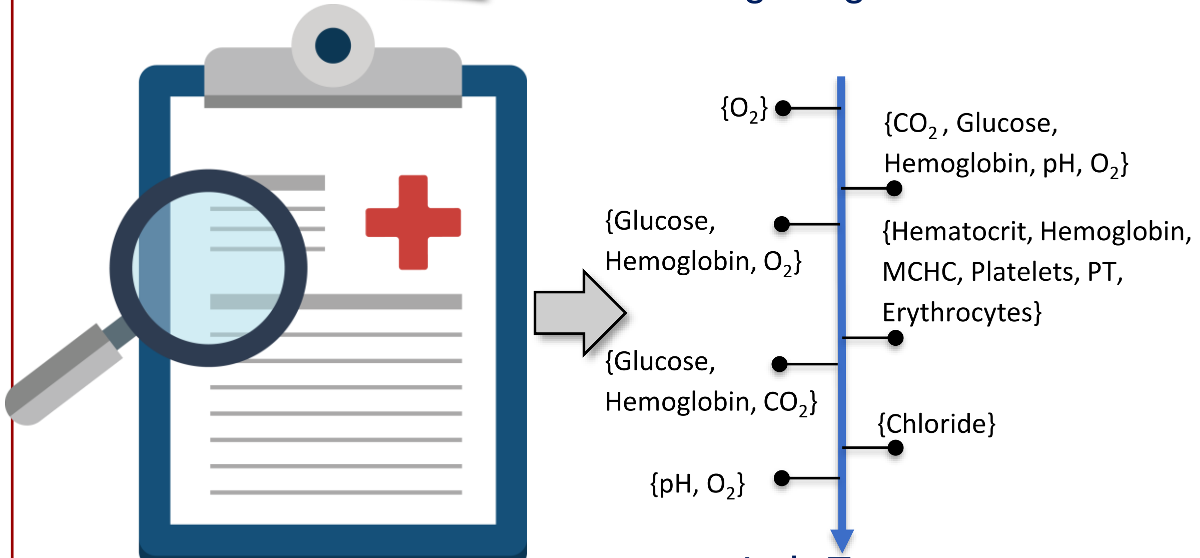
ZL**C**. *Contrastive Out-of-Distribution Detection for Pretrained Transformers*. **EMNLP** 2021

IE for the Common Good

Medicine and Healthcare



Drug-drug Interaction

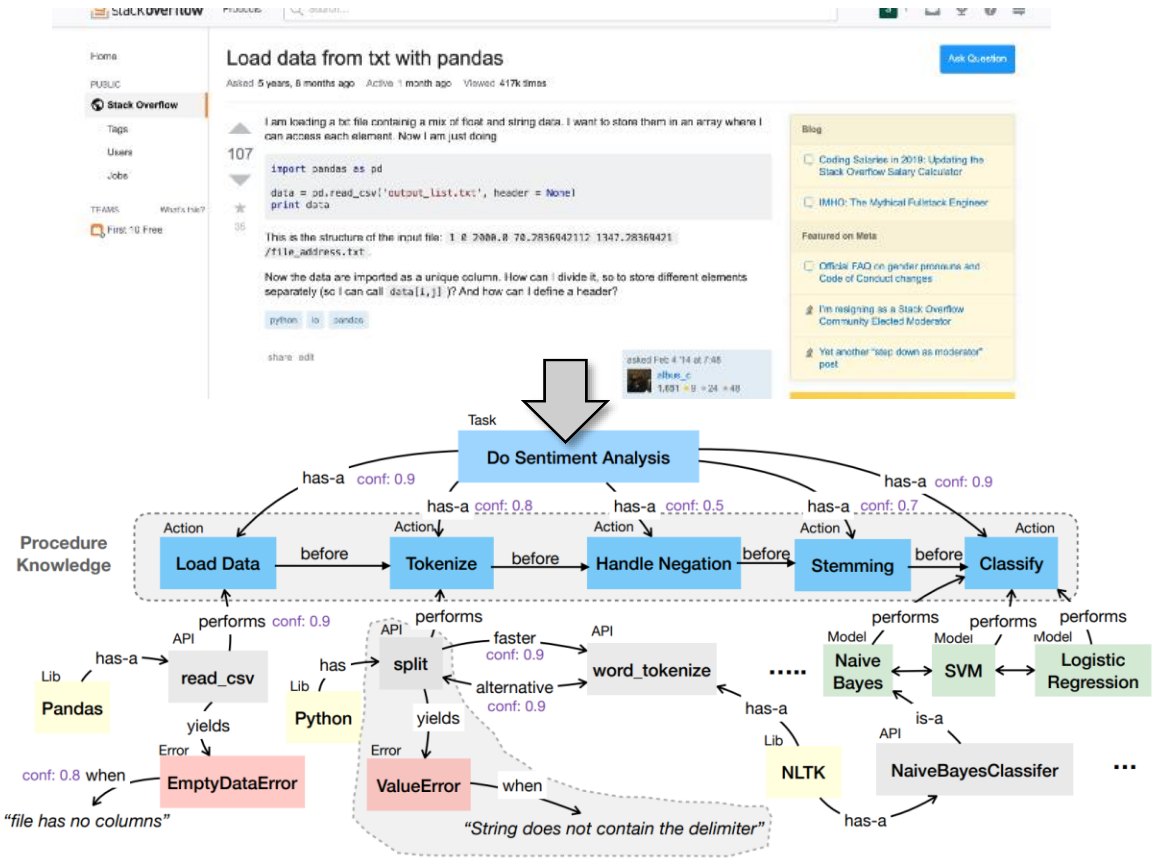


Lab Events

Microsoft Research

David Geffen School of Medicine

Programming Education



Low-resource domains that particularly

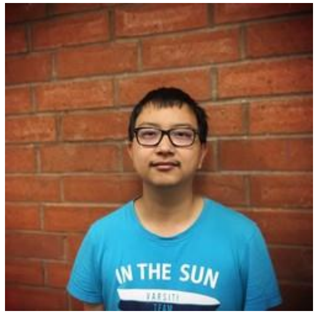
- need **indirect supervision** and **constrained learning**;
- suffer from **noise and faithfulness issues**.

Acknowledgement

Language Understanding and Knowledge Acquisition Lab

LUKA

Student Researchers



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

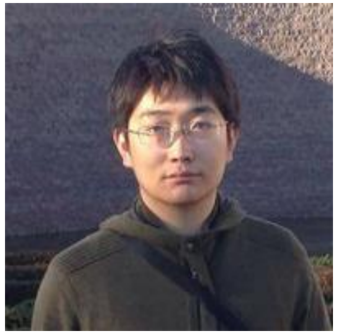
Penn, University of Pennsylvania
Carnegie Mellon University
NUS, National University of Singapore
University of Cambridge
amazon
Microsoft Research
UCLA David Geffen School of Medicine

Sponsors

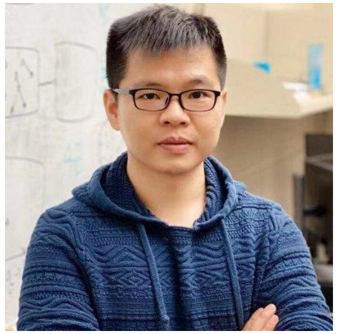
NSF
DARPA
amazon
University of Maryland Applied Research Laboratory for Intelligence and Security
cisco

USC Viterbi School of Engineering | VIRGINIA TECH | UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN | Penn | amazon

New Frontiers of Information Extraction

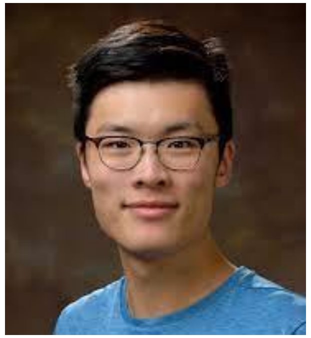


Muhao Chen



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

USC Viterbi
School of Engineering

Thank You