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 Learning and Inference for IE

How do we make IE models ***more reliable***?

Information Extraction (IE)

The process of automatically extracting structural information from unstructured data (e.g. natural language text)

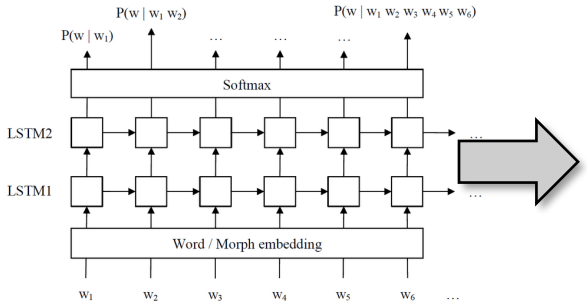
Honolulu

From Wikipedia, the free encyclopedia

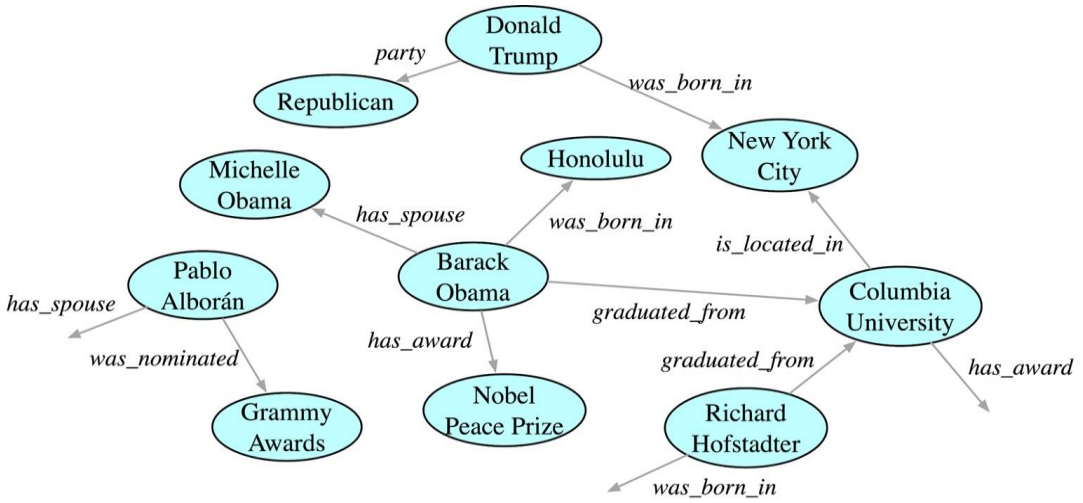
This article is about the largest city and state capital city of Hawai
Honolulu itself, see Honolulu County, Hawaii. For other uses, see

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 O'ahu,[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 Nu'uanu 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.

- Infer the identities of entities and concepts

- Facilitate 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 …ass to tight end Desmond Clark. The gap was reduced when Favre fire… -yard TD pass to tight end Visanthe Shiancoe. The Vikings … wi… …rian Peterson's second 1-yard TD run. The Bears then responded … Cutler firing a 20-yard TD pass to wide receiver Earl Bennett … Bears then won on Jay Cutler's game-winning 39-yard TD pass t… …de receiver Devin Aromashodu.

What is her seizure frequency?

s… 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?

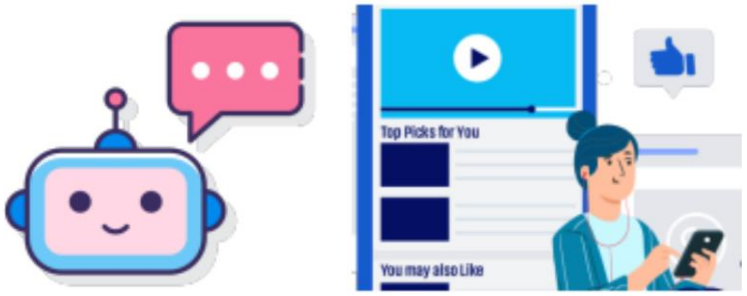
Mayor Rahm Ema… toward his bid for a third term – more than five times the total raised by his 10 challengers combined, campaign finance records show.

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 (DARPA KMASS Project)

Extracting structures about tasks, steps and concepts

A consolidated semantic index for *in-context* content delivery



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.

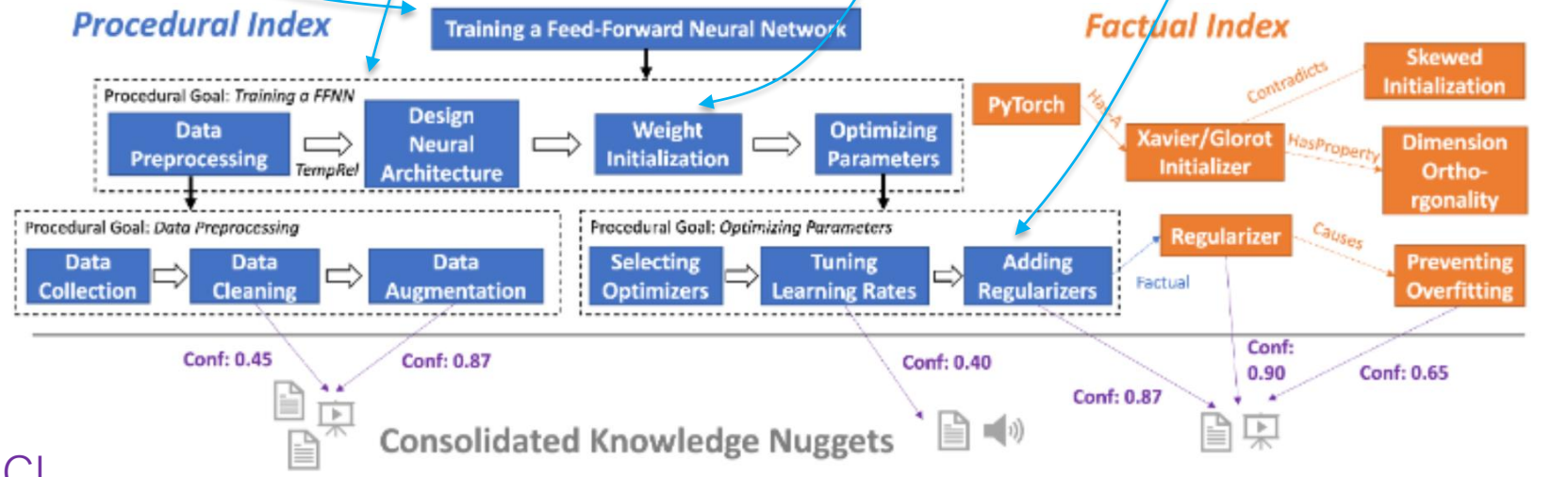
Weight initialization for neural networks

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…

DARPA



IE Is the Backbone of Nearly All Knowledge-driven Tasks

WIKIPEDIA The New York Times

PubMed bioRxiv

Knowledge Bases

Freebase

yago select knowledge

ASER

WIKIDATA

ConceptNet
An open, multilingual knowledge graph

DBpedia

Bio/Med Databanks

DRUGBANK

MeSH

neXtprot

STRING

GENEONTOLOGY Unifying Biology

RCSB PDB PROTEIN DATA BANK

Knowledge Representation

Common Sense

Commonsense QA
Event Prediction
Intent Prediction

Narrative Understanding

Storytelling
Summarization
Newsworthiness Detection

Medical Informatics

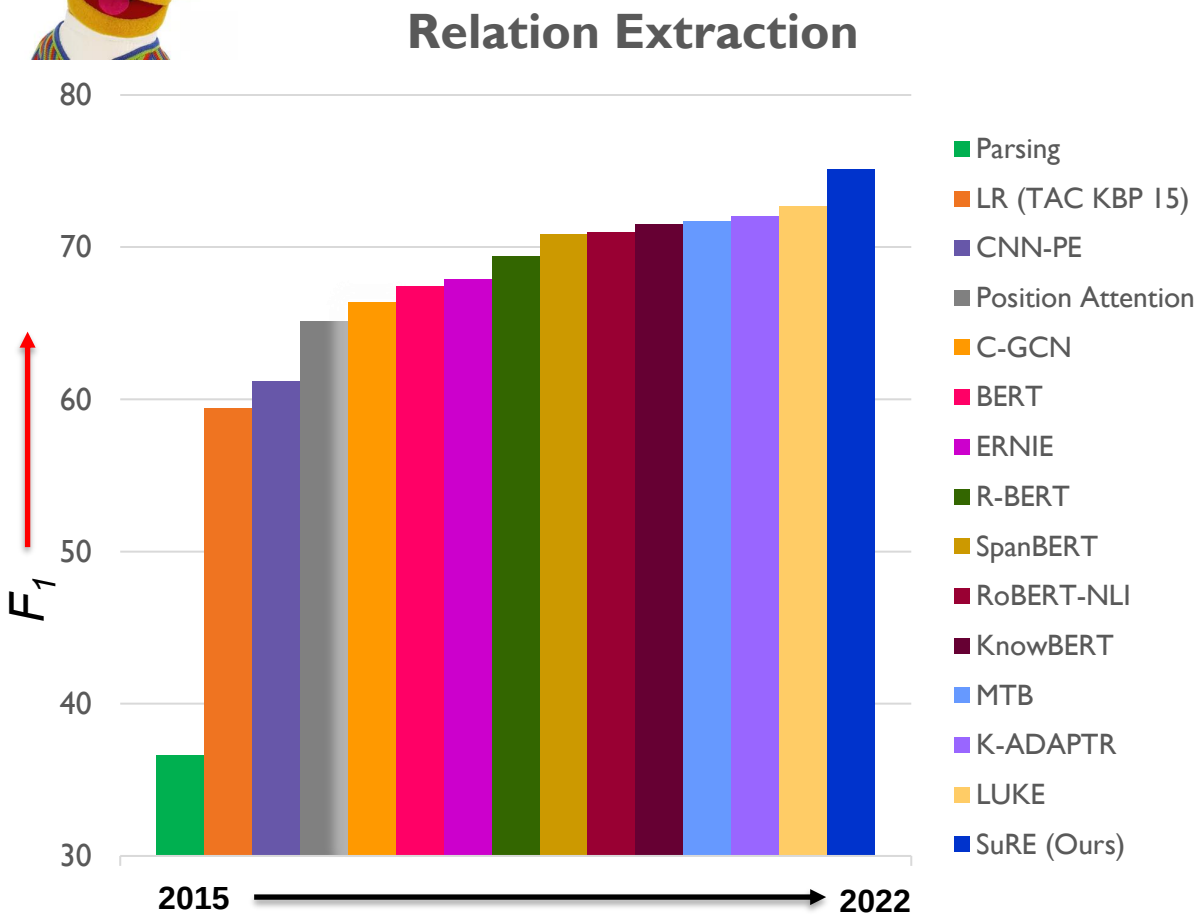
Proteomic Interaction Prediction
Mutation Effect Estimation
Genomic Function Prediction

Diagnostic Prediction
Drug Repurposing
Disease Phenotyping

How IE Is Doing Today

On Benchmarks

Rise of the Pre-trained Models



In Reality

Is still brittle, limitedly generalizable, and costly to develop

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

Type?

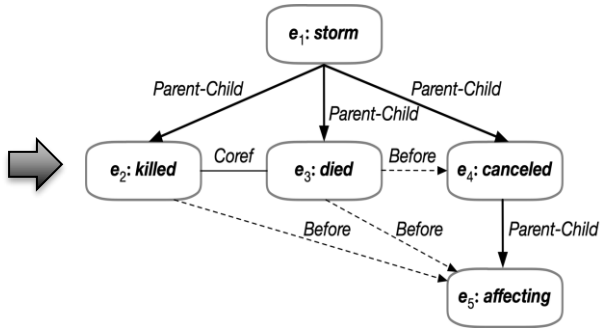
Island X

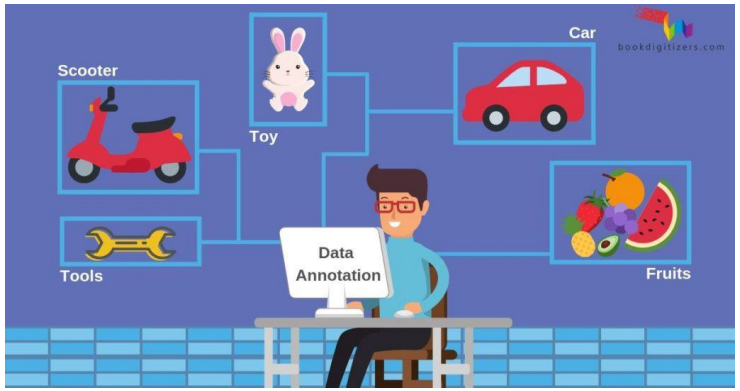
AIDA, KAIROS, BETTER, KMASS, GAIA: all costing tens of millions $.

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, annotating 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. 5-8% errors in TACRED and CoNLL03
- **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 ✗

- Biased training
- Prior knowledge

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

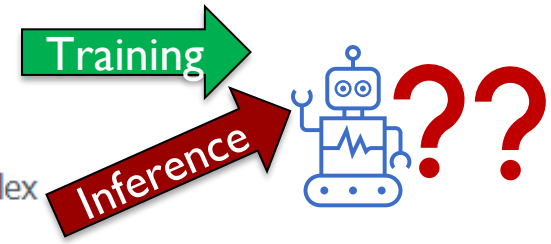
Knowing the Decision Boundary

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



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

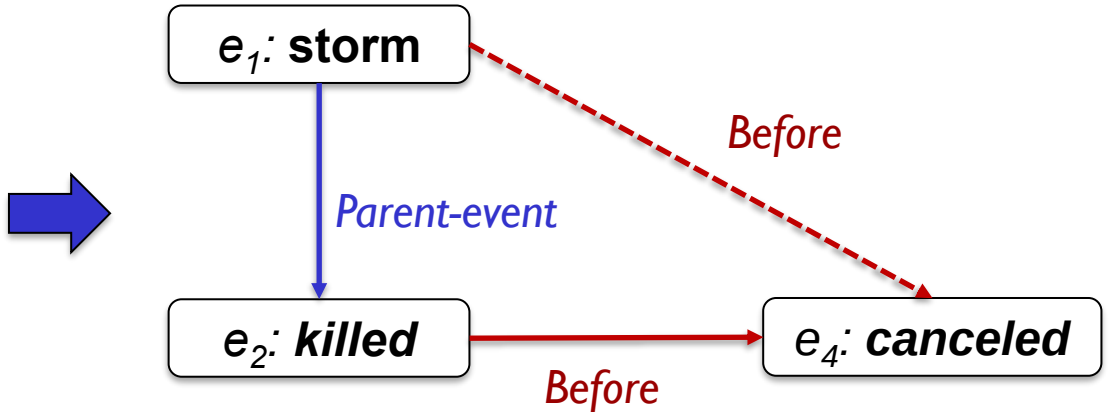
No Rel

Nothing to extract

Challenge: Consistency

Extracts are not standalone.

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-19]

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

Goal: Robust IE

The goal of developing a robust IE system

Robustness in Learning

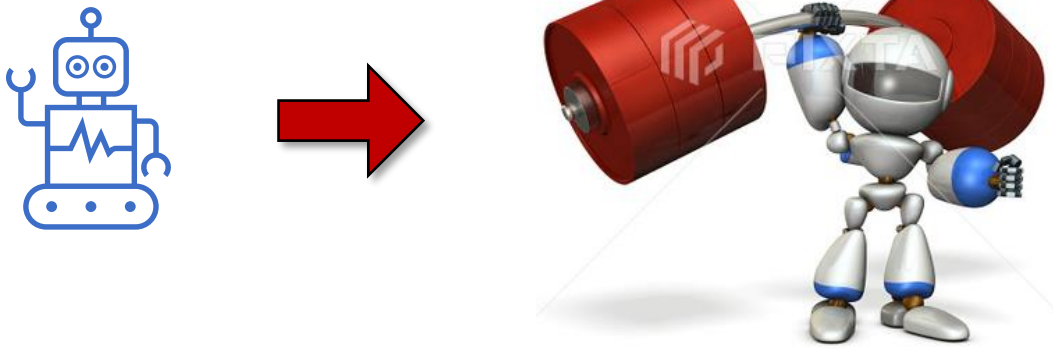
- **Has to be achievable with *minimal and imperfect* supervision**
  - **Dealing with *noise*, *indirect supervision*, *constraints*…**

Start to move away from direct/end-task supervision

Robustness in Inference

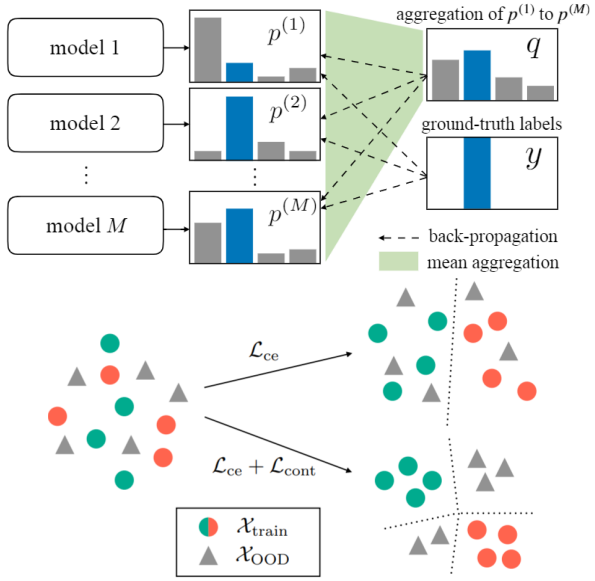
- **Constrained inference**: ensuring logically consistent extraction
- **Faithfulness**: mitigating spurious correlation
- **Selectiveness**: knowing what is extractable, what is not

Self-contained, faithful and selective extraction.

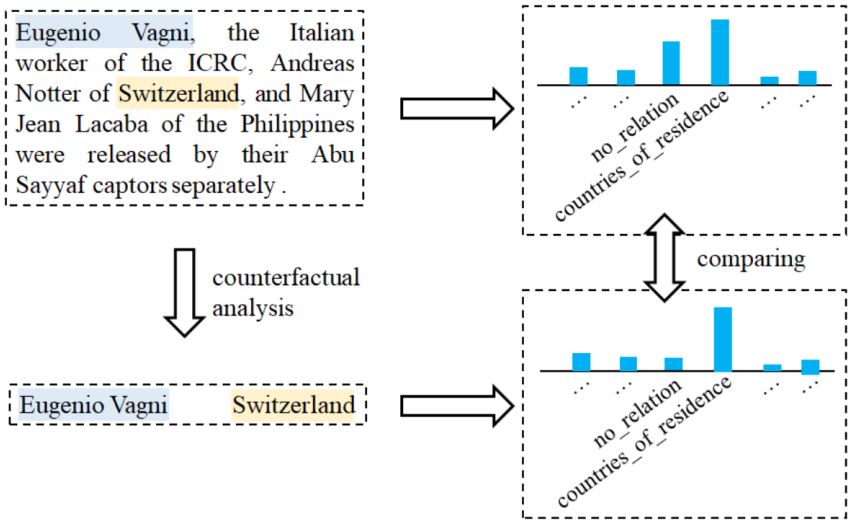


In This Talk

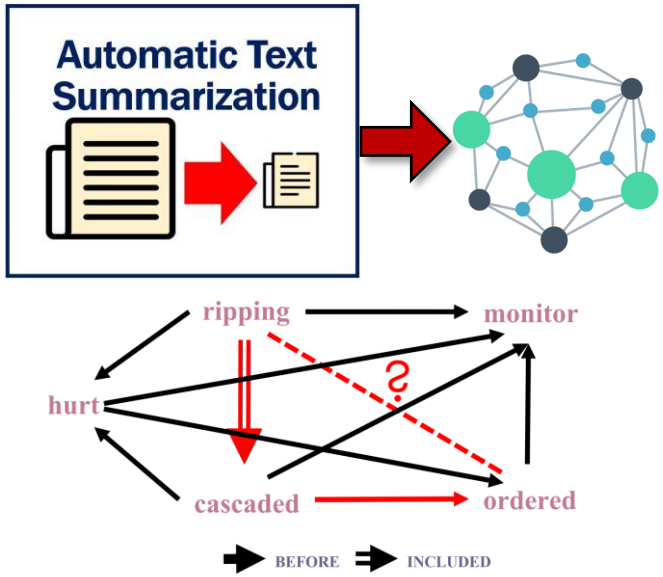
1. Noise-robust IE



2. Faithful IE



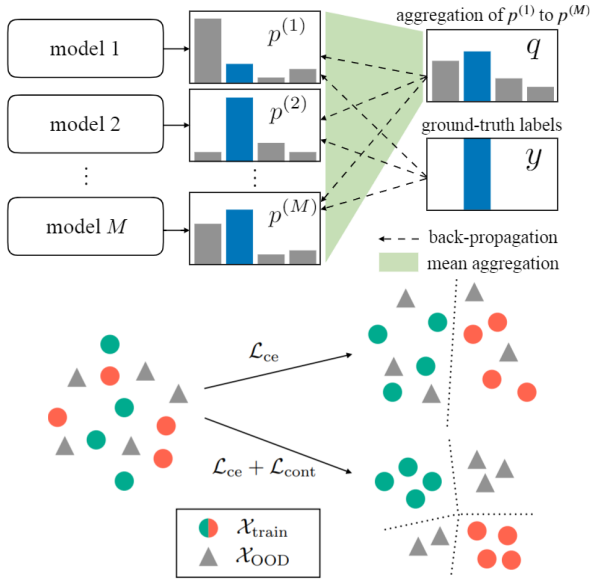
3. Constrained/Indirectly Supervised IE



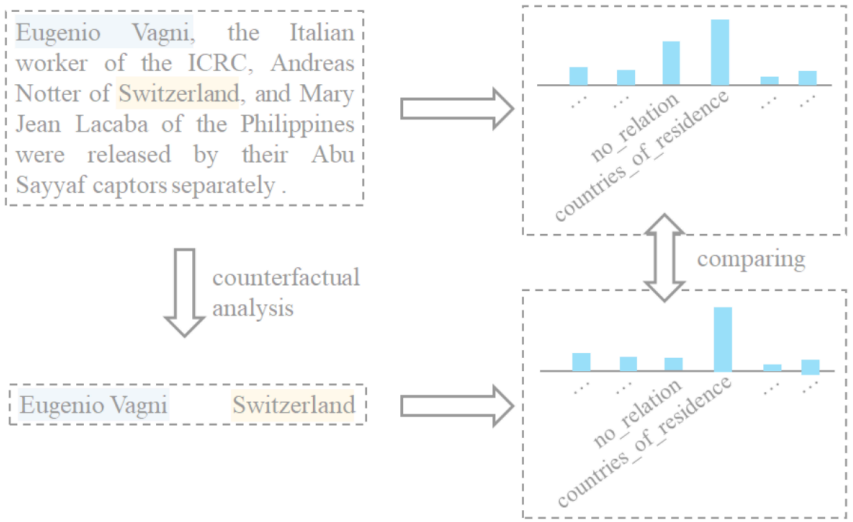
4. Future Directions

In This Talk

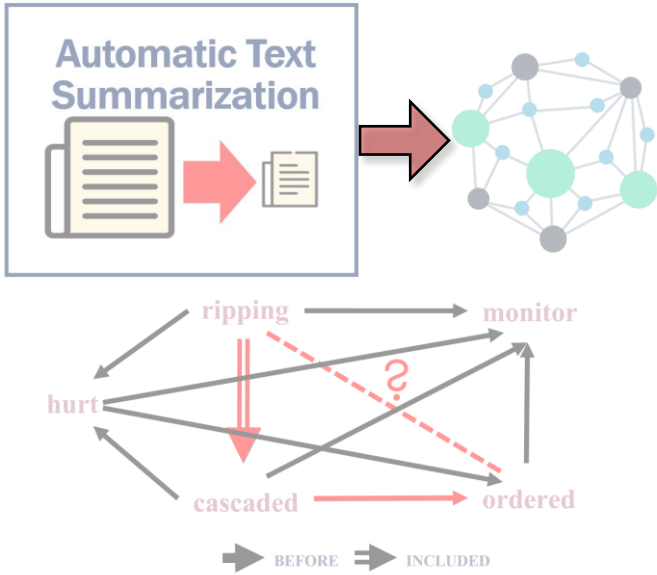
1. Noise-robust IE



2. Faithful IE



3. Constrained/Indirectly Supervised IE

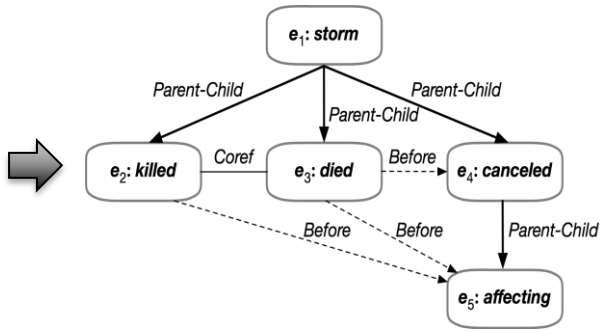


4. Future Directions

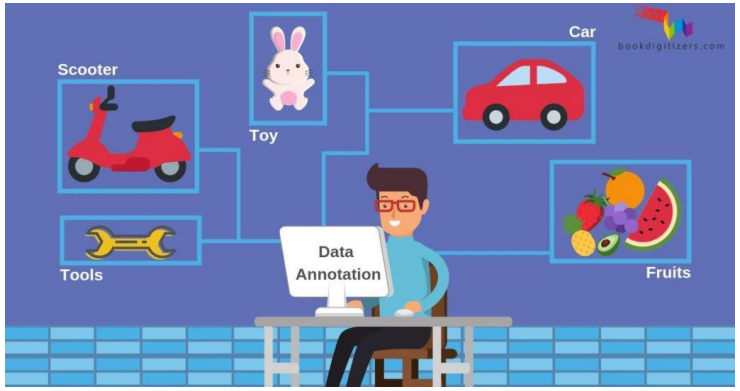
Noise In Training and Inference

Training

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.



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



Reading long documents, annotating complex structures

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

Hence, IE annotations are inevitably noisy. For example:

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

High-performance IE models have to be achievable with *imperfect supervision*

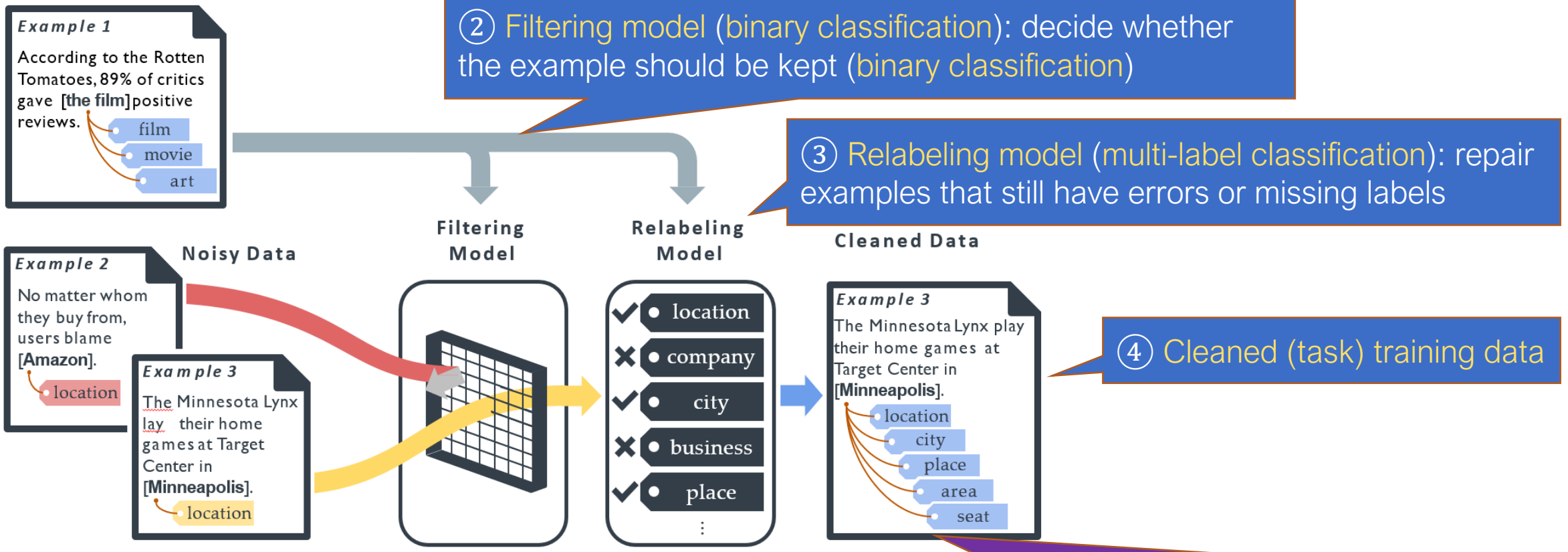
A Glance at Prior Solution: Supervised Denoising

A noise filtering or relabeling model may be trained, if clean data are available.

① Labeled clean data and noisy data

② Filtering model (binary classification): decide whether the example should be kept (binary classification)

③ Relabeling model (multi-label classification): repair examples that still have errors or missing labels

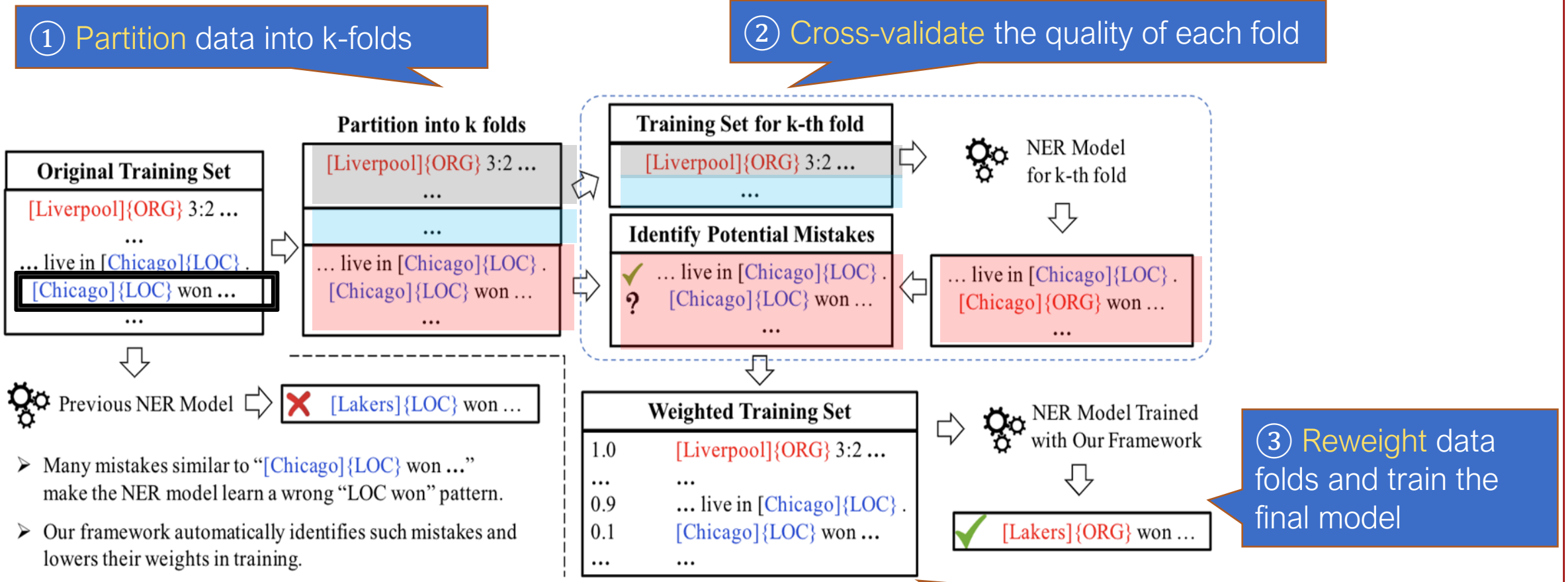


④ Cleaned (task) training data

Cost: manually labeling enough clean data is nowhere cheaper.

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

A Glance at Prior Solution: Unsupervised Denoising with Ensemble

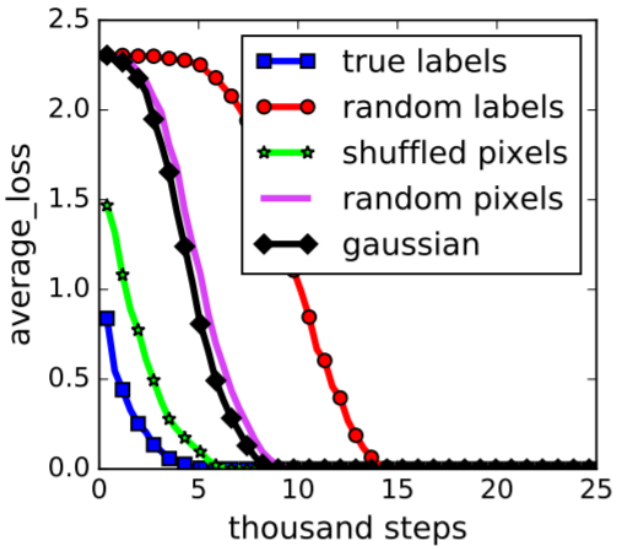


- Many mistakes similar to "[Chicago]{LOC} won …" make the NER model learn a wrong "LOC won" pattern.
- Our framework automatically identifies such mistakes and lowers their weights in training.

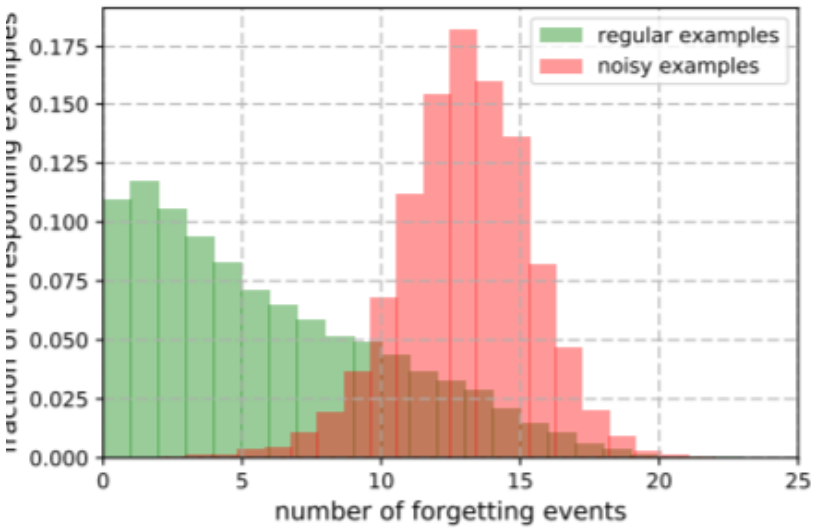
No longer requires annotated clean data, but leads to coarse-grained noise estimation and needs redundant training and testing (at least k+1 times).

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

Unsupervised Denoising: Co-regularized Knowledge Distillation

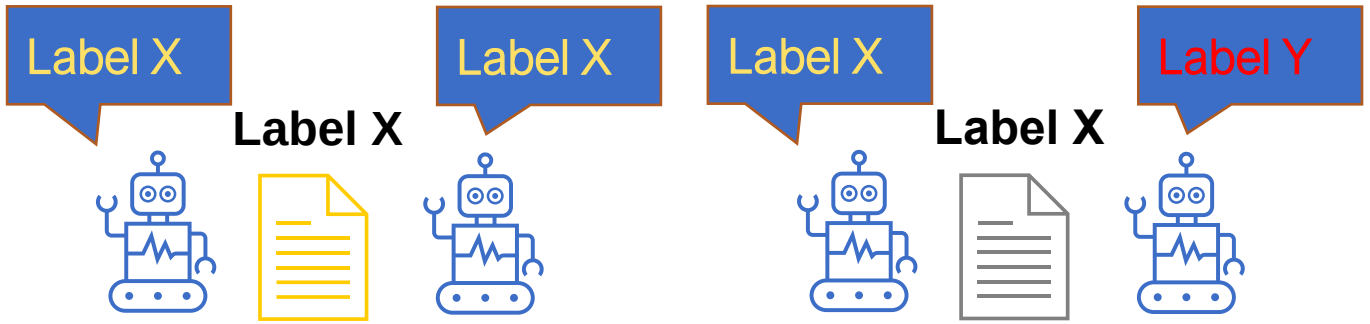


(1)



(2)

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



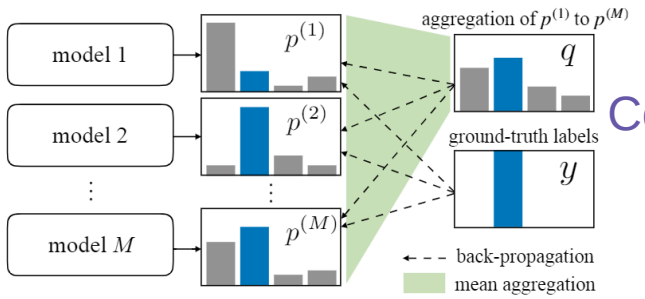
Agree: Clean ✓ **Disagree: Noise ✗**

Mutual agreement by models indicates clean/noisy labels

Noisy labels are outliers to the task inductive bias.

(1) Noisy labels take longer to be learned.
(2) Noisy labels are frequently forgotten.

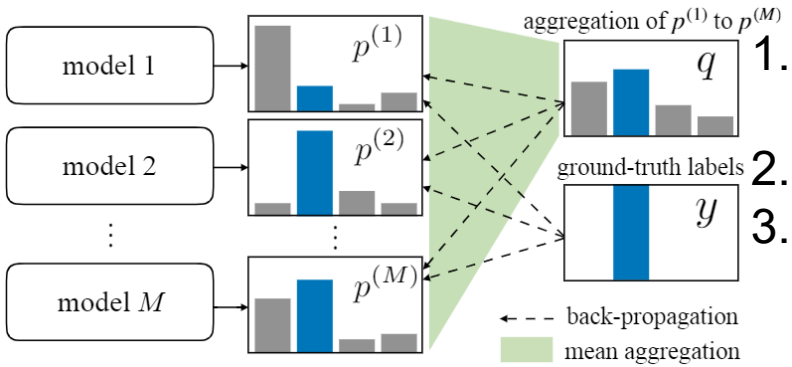
Model prediction is often inconsistent or oscillates on noisy labels in later epochs.



Co-regularization Framework

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

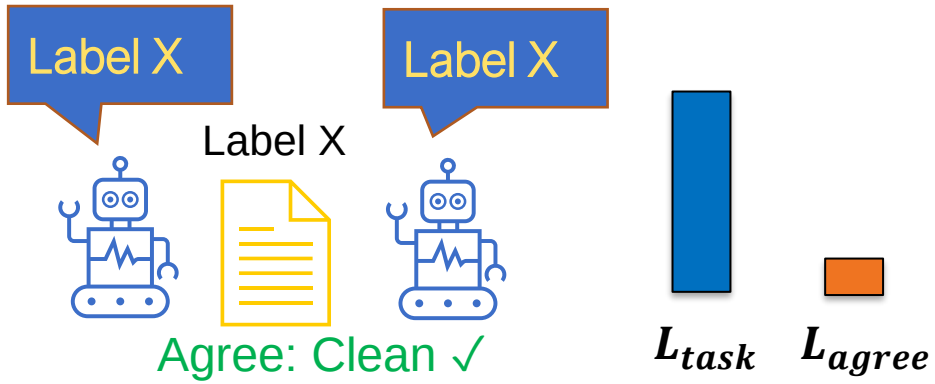
Unsupervised Denoising: Co-regularized Knowledge Distillation



1. Create $\boldsymbol{M}(\geq \mathbf{2}$; 2 is enough) identical neural models with **different initialization**, and **warm up** them using only the **task loss**.
2. Train the models with both **the task loss** and an additional **agreement loss**.
3. Return one of the models.

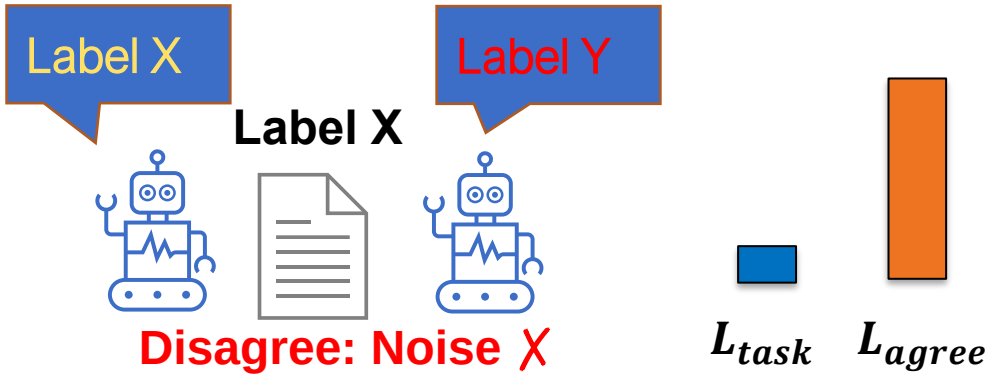
Cross-entropy L_{task}

K-L divergence between model predictions L_{agree}



On clean data

- Lower agreement loss
- Focusing on task optimization

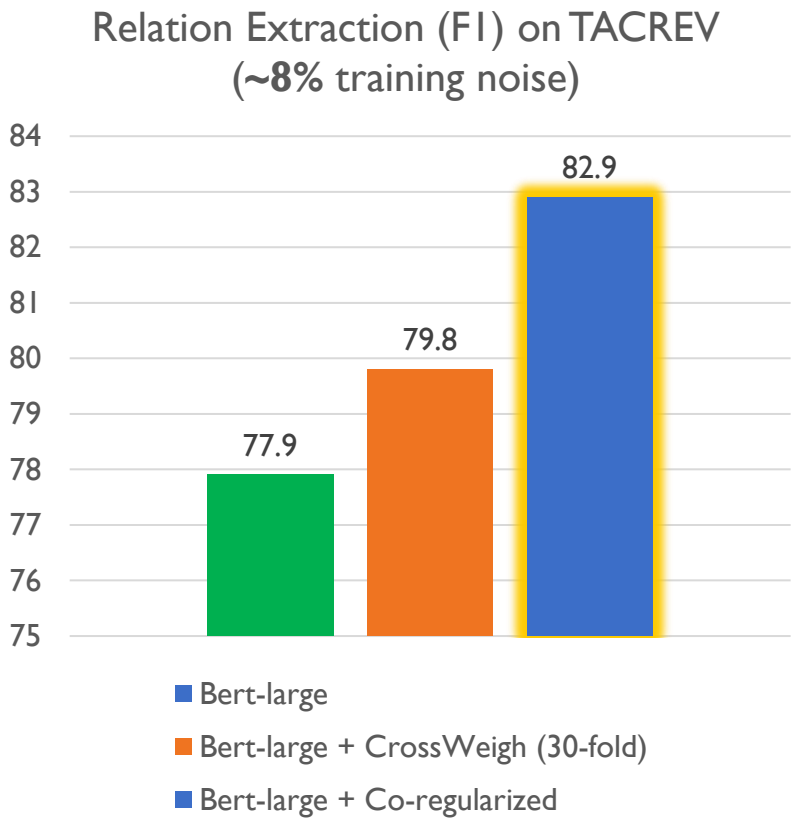


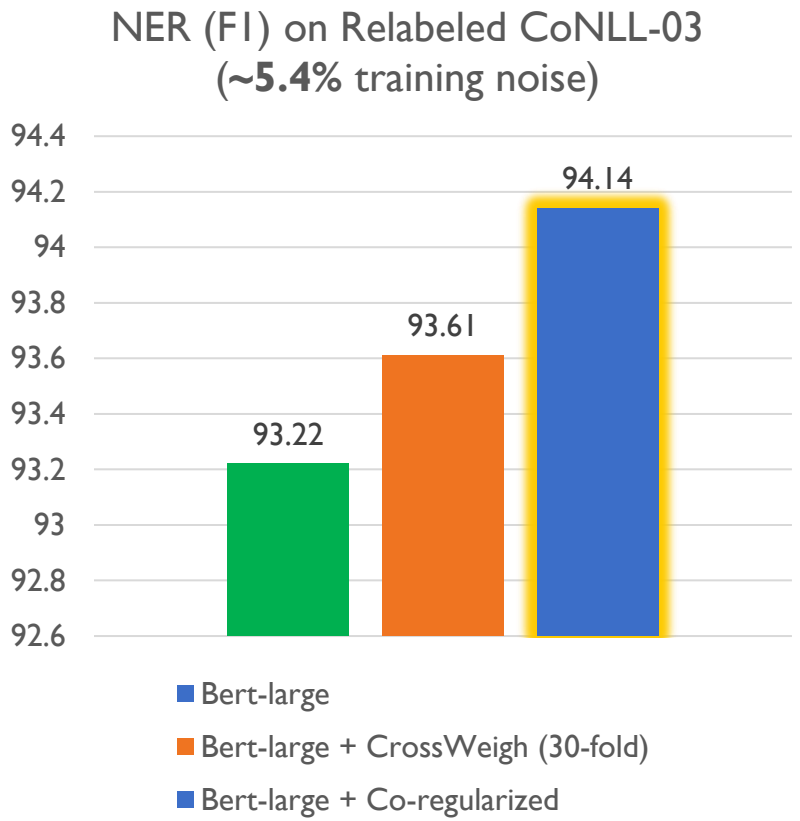
On noisy data

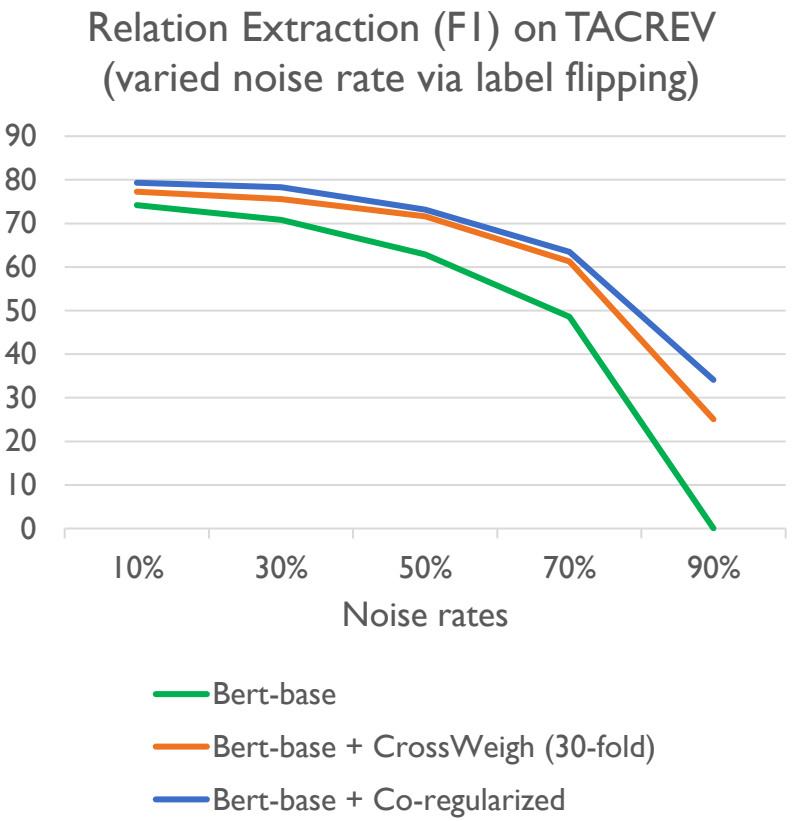
- Higher agreement loss
- Task optimization proactively prevents fitting those data

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

Unsupervised Denoising: Co-regularized Knowledge Distillation







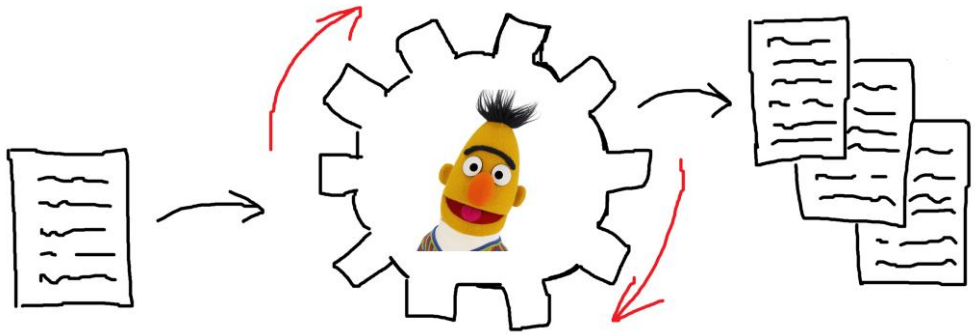
Merits of co-regularized knowledge distillation

- More robust than ensemble (e.g. CrossWeigh), especially when noise rates are higher
- More efficient (only 1-fold of training and no additional inference cost)
- Can be applied to train 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

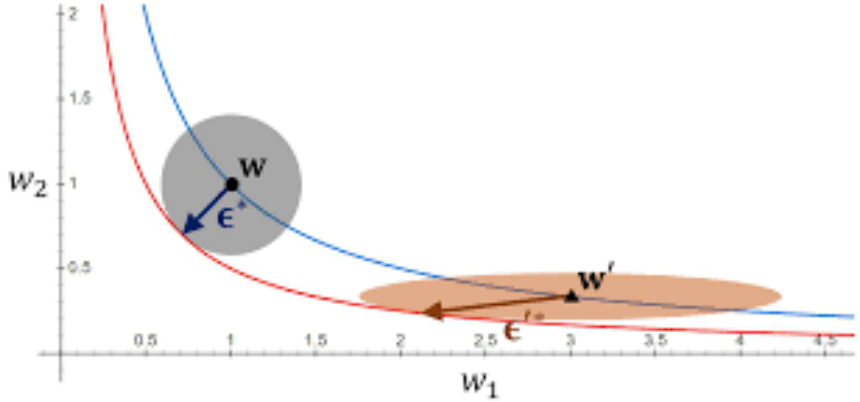


Denoising automatically augmented training data

- Self-regularization with random dropout
- Agreement test between original data and augmentation

5.6% improvement on low-resource text classification (1% TREC + EDA).

Perturbation Robustness



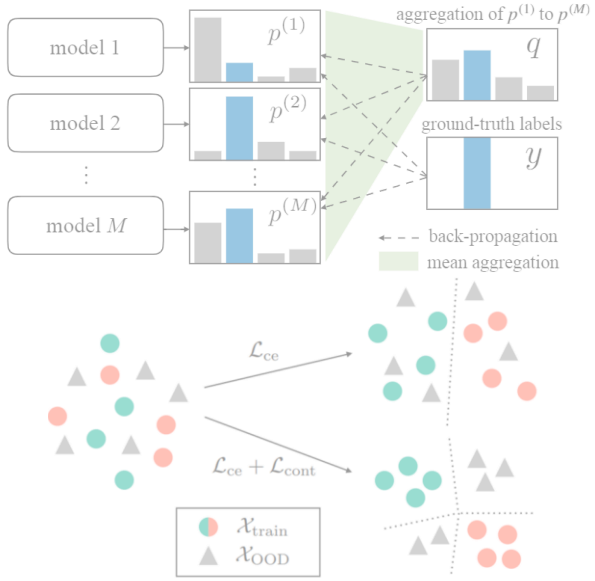
δ-SAM: efficient adversarial parameter perturbation

- Fast and accurate approximation of the computationally prohibitive *per-instance* sharpness-aware minimization (with *per-batch reweighting*)
- Significant improvement of model generalization on textual retrieval, summarization, and NLU tasks

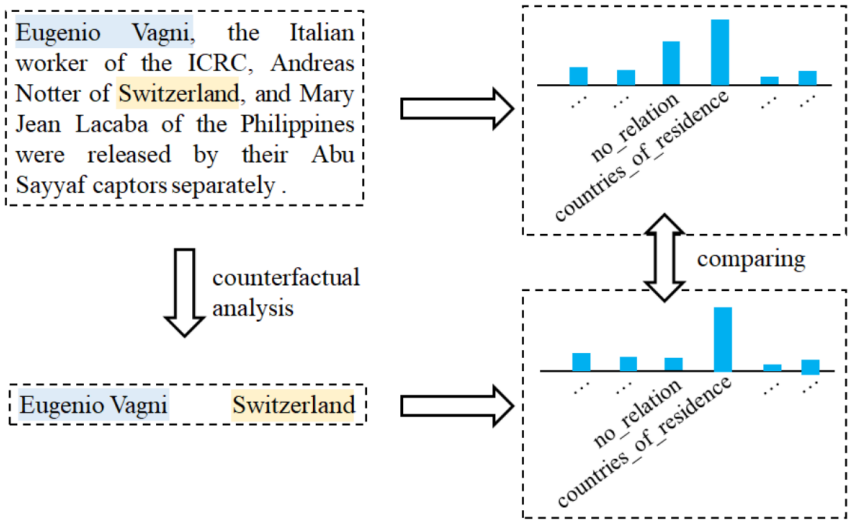
ZLZ***C***. *Sharpness-Aware Minimization with Dynamic Reweighting*. **EMNLP** 2022

In This Talk

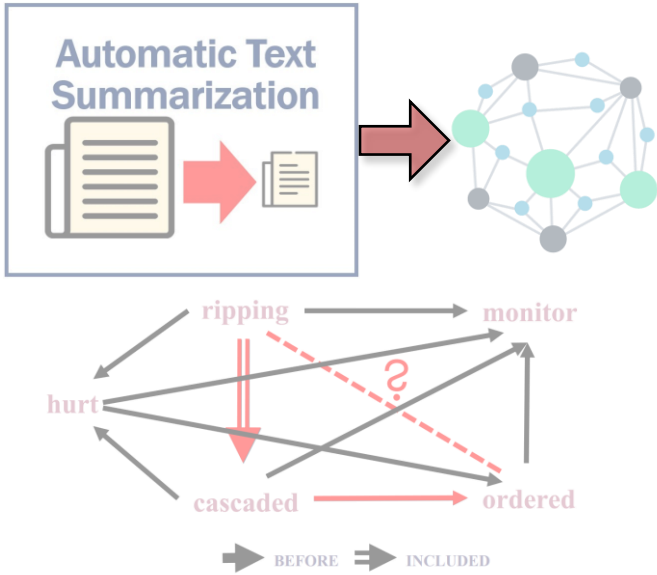
1. Noise-robust IE



2. Faithful IE



3. Constrained/Indirectly Supervised IE



4. Future Directions

Faithfulness Issues

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

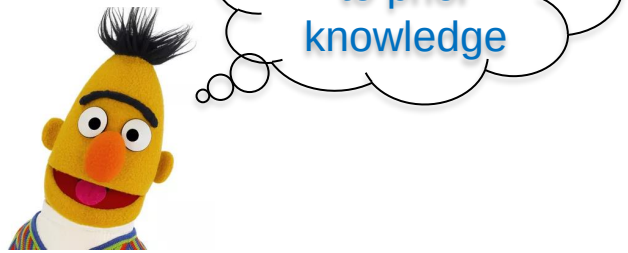
Entity relation extraction:

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

Rel?

Visit ✓

FounderOf ✗



Prior knowledge (in PLMs) can lead to biased extraction

Event relation extraction:

event1 *event2*

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

Before? After?

Before ✓

After ✗

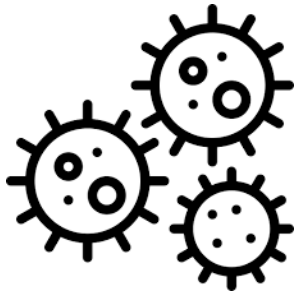
(Statistically) Biased training can lead to biased extraction

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

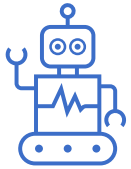
More risky tasks where we couldn't afford any ***guesses*** from **unfaithful IE**

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

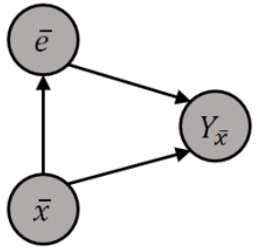
What is Causing the Problem: Take Relation Extraction as An Example

What we hope the IE model to do

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



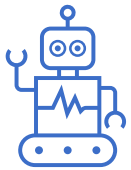
Comprehend the *context*, and induce the mentioned *relation* of *entities*.



Relations should be inferred based on both mentions and the context

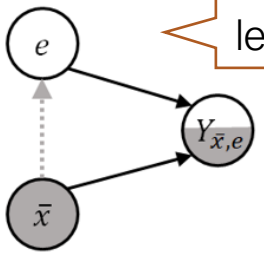
What it may actually do

Bill Gates [illegible] Microsoft [illegible].



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

Context is not captured, leading to entity bias



Overly relying on entity mentions lead to a shortcut for RE

How do we mitigate this spurious correlation?

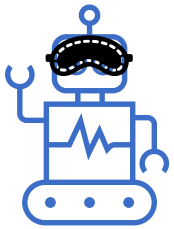
W**C**ZCLLH. *Should We Rely on Entity Mentions for Relation Extraction? Debiasing Relation Extraction with Counterfactual Analysis*. **NAACL** 2022

Debiased Training

Mention masks: mask out entity names with their types

Person paid a visit to Building 99 of Org yesterday.

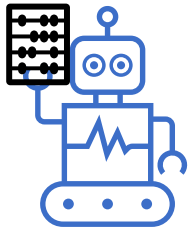
Similarly for *event RE*, we can mask using trigger types and tense



Mask mentions in both training and inference

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

Reweighting instances: FoCal loss, resampling, 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
Liu et al. *Just Train Twice: Improving Group Robustness without Training Group Information*. **ICML** 2021

Our Strategy: Counterfactual Inference

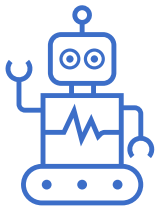
Measure the biases using counterfactual instances, then deduct the biases

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 ($\bar{x}$)

∅

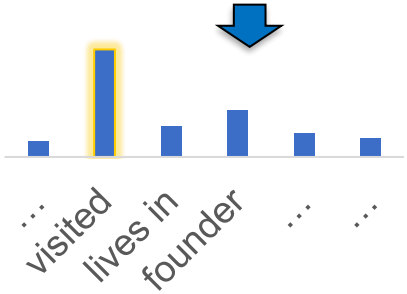
deduct

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

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

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

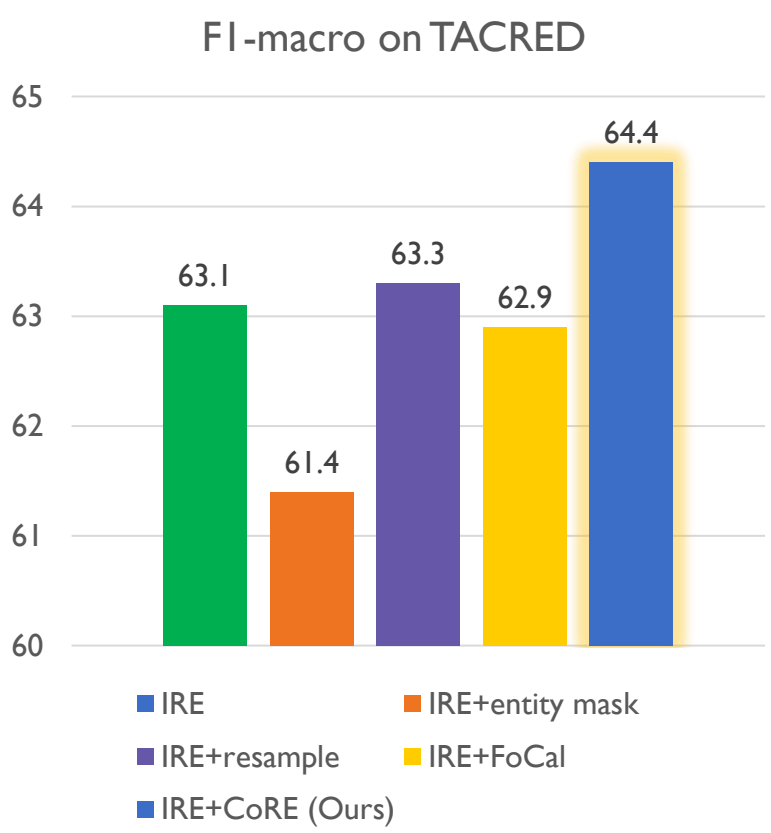
Obtained on dev set

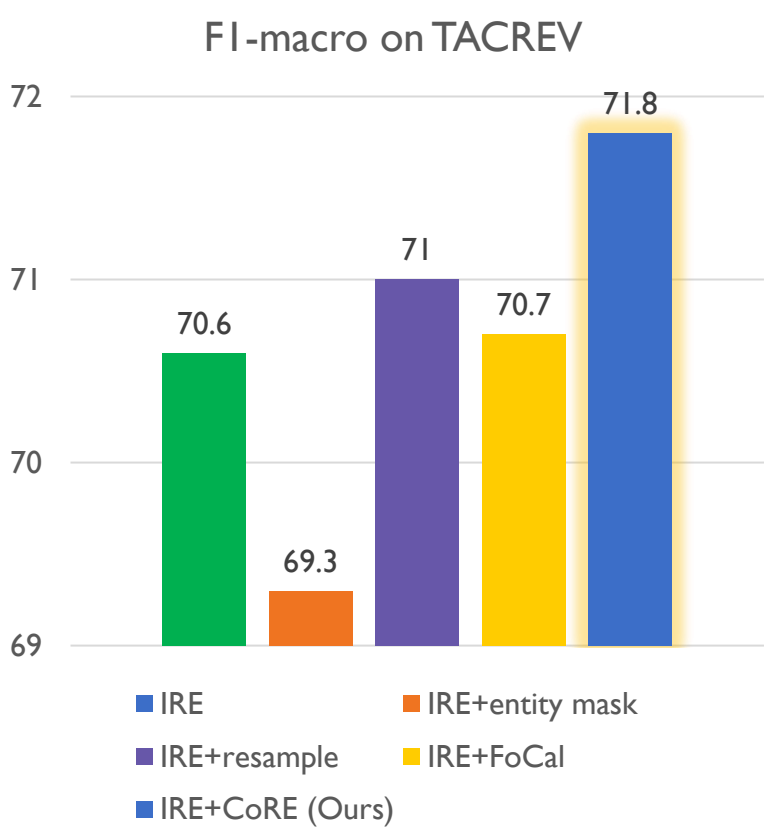


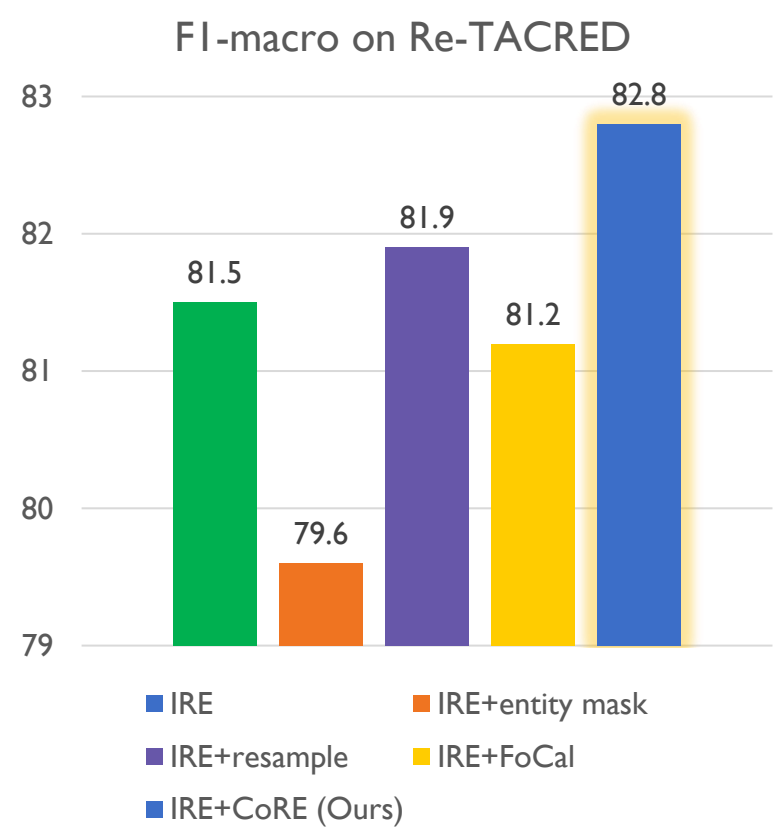
Debiased prediction Y_{final}

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

Counterfactual Inference







Counterfactual inference leading to more precise and fairer relation extraction.

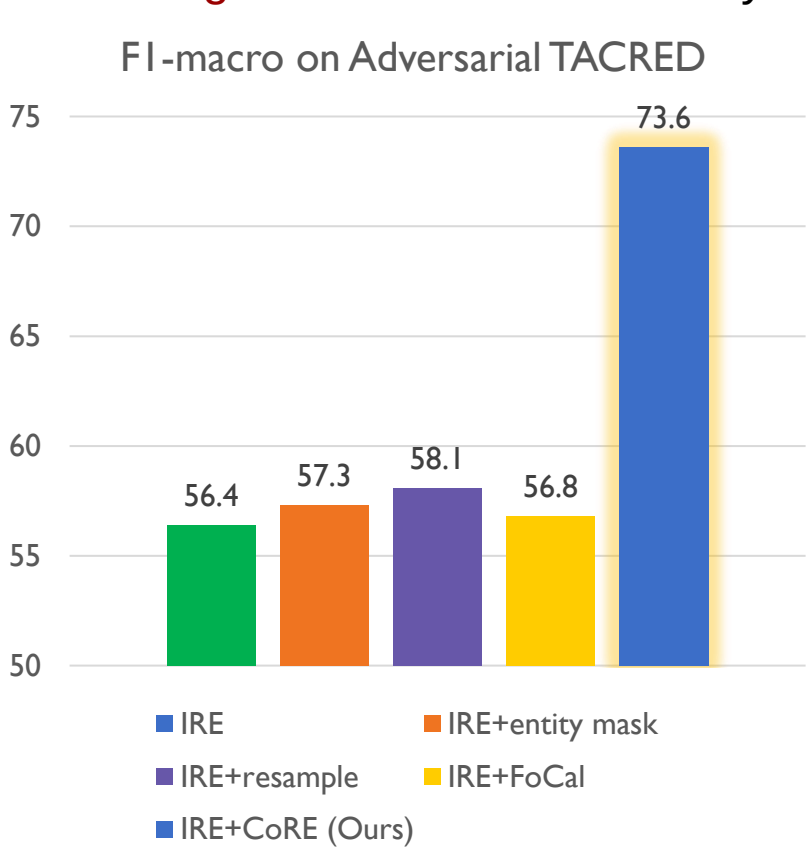
*$IRE_{RoBERTa}$ is one of the best-performing sentence-level RE models (Zhou and Chen 2021). Results are also available for LUKE.

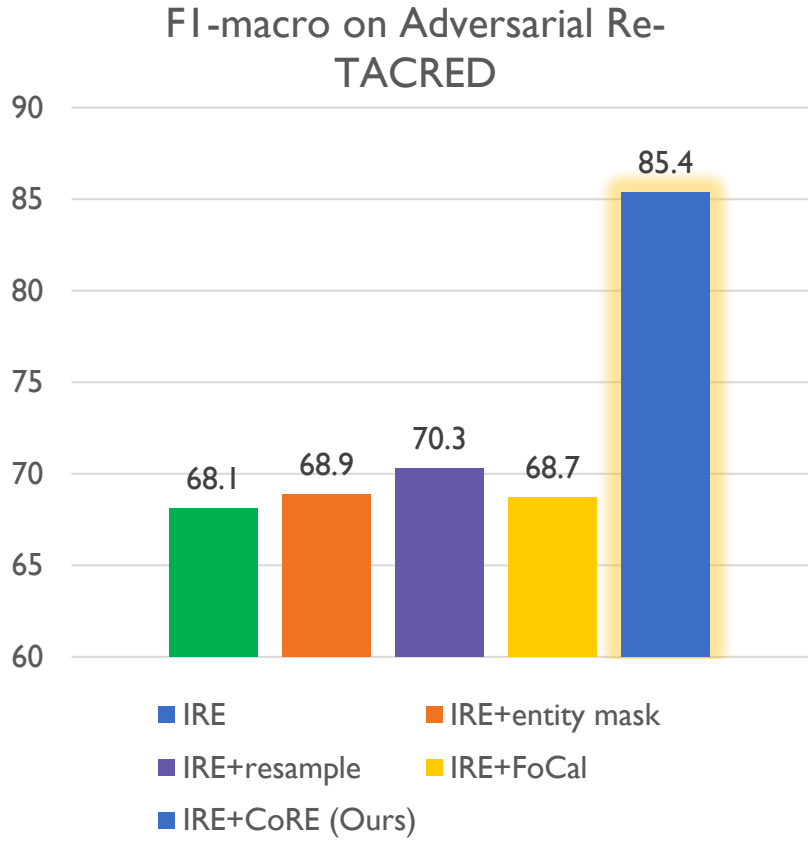
W**C**ZCLLH. *Should We Rely on Entity Mentions for Relation Extraction? Debiasing Relation Extraction with Counterfactual Analysis*. **NAACL** 2022

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.





Counterfactual inference leads to significantly more faithful relation extraction.

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

Our Continuing Studies

Faithfulness in IE is still an underexplored research direction.

Six Types of Artifacts in Entity Identification

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, overgeneralization
- **Counterfactual data augmentation** to address them all

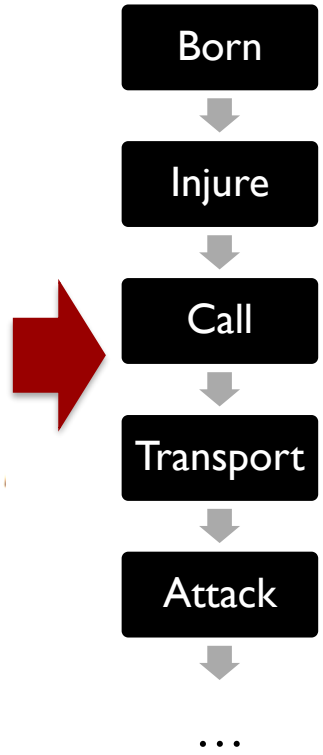
XWLD***C***. *Does Your Model Classify Entities Reasonably? Diagnosing and Mitigating Spurious Correlations in Entity Typing*. **EMNLP** 2022

Faithful Event Timeline Extraction

2009-06-27
There was no sign of foul play in the death of Michael Jackson.....A recording of the telephone call made to emergency services has been released , in which the caller said Jackson was unconscious and had stopped breathing.
Born : August 29, 1958, Gary, Indiana, US.
Also known as : The King of Pop .....,

2009-06-27
Police also want to speak to Jackson's doctor who witnessed his collapse ....
On Friday night thousands of fans came together in London's Trafalgar Square to light candles and sing some of his biggest hits ....

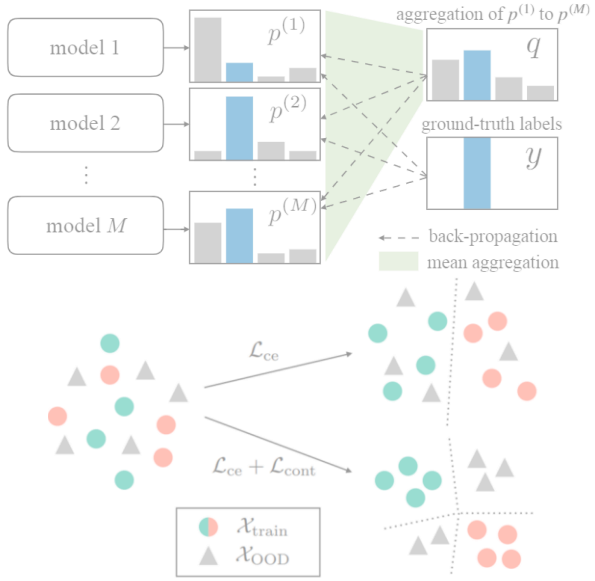
2009-06-28
She said Dr Murray had traveled in ambulance with Jackson after he collapsed last Thursday ...
Dr Murray had been hired by Jackson in May to accompany him as he prepared to embark on a gruelling series of 50 concerts in London in July .



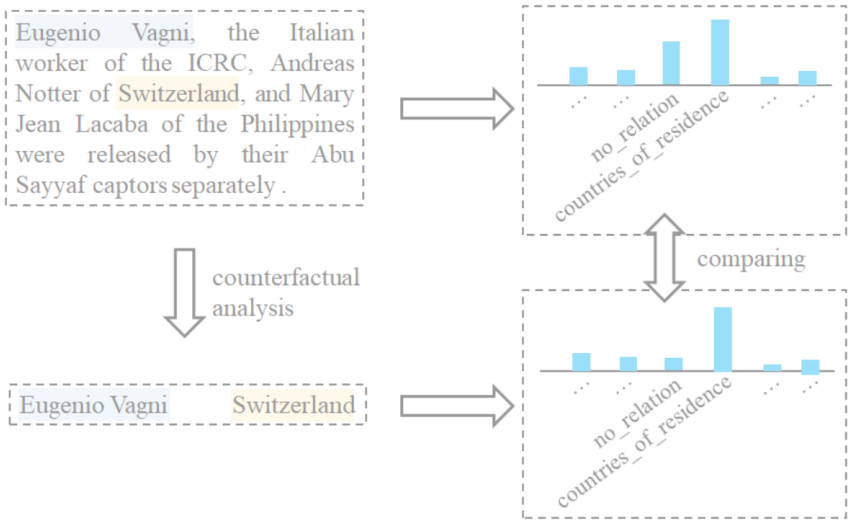
WZDGR***C***. *Extracting or Guessing? Improving Faithfulness in Event Temporal Relation Extraction.* 2022

In This Talk

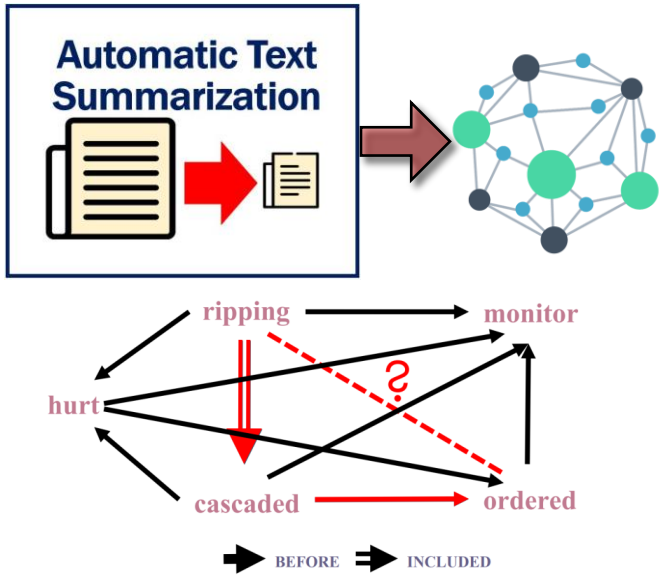
1. Noise-robust IE



2. Faithful IE



3. Constrained/Indirectly Supervised IE

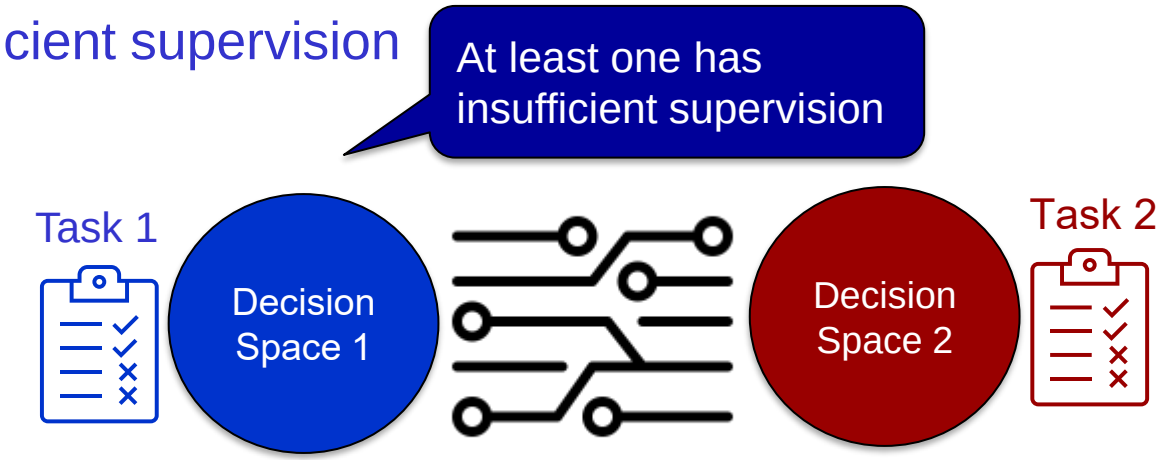


4. Future Directions

Two Forms of Indirect Supervision

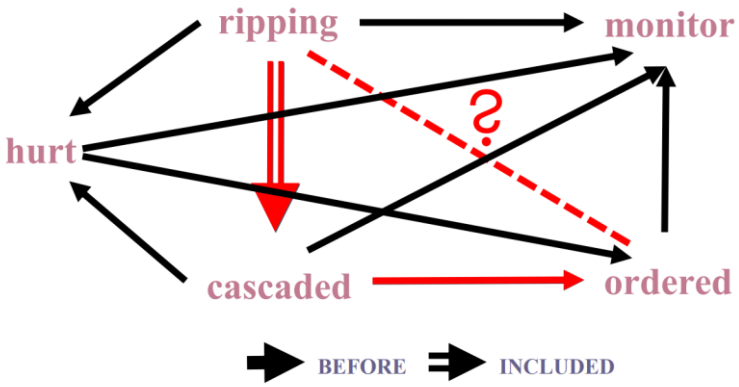
IE suffers from insufficient supervision

Direct annotation is difficult and expensive

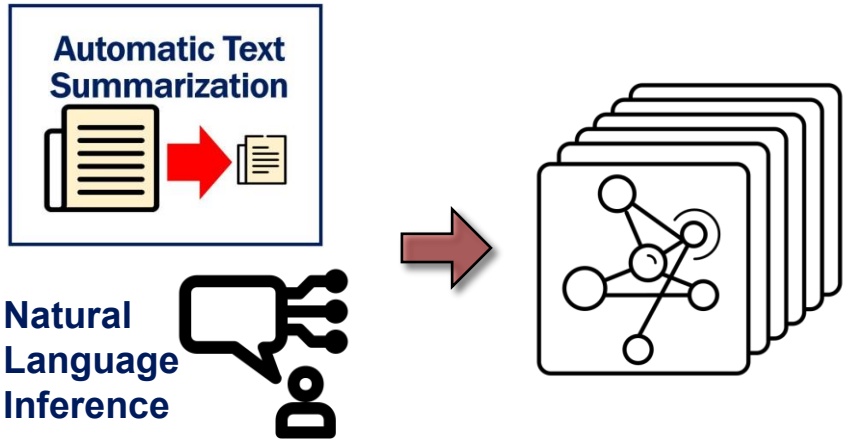


Could we bridge their decision spaces and supervision signals?

Two Forms of Indirect supervision



(Logically) Constrained Learning



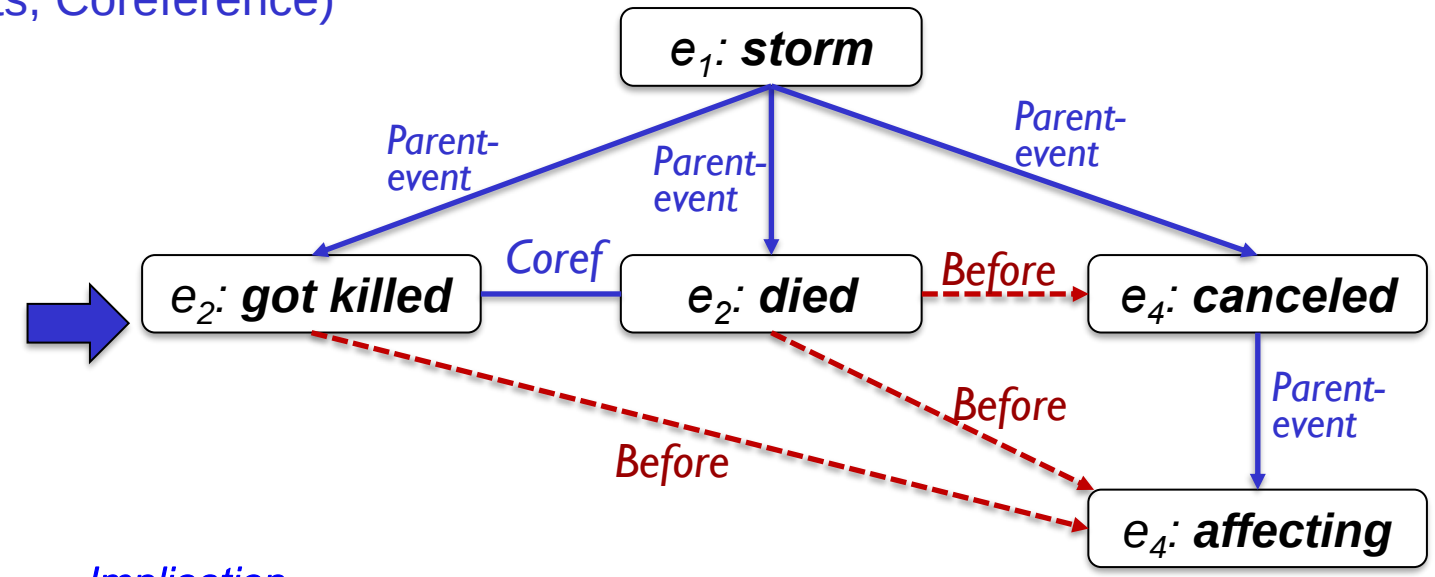
Cross-task Transfer

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*

Use logical constraints!

Logical Constraints Of Relations

Symmetry

e3:died is BEFORE *e4:canceled*

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

Transitivity

e1:storm is PARENT of *e4:canceled*

∧*e4:canceled* is a PARENT of *e5:affecting*

=> *e1:storm* is a PARENT of *e5:affecting*

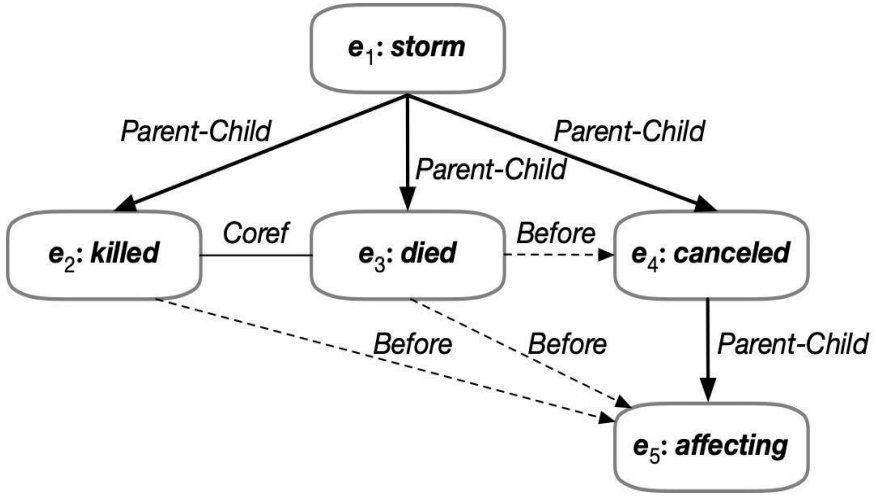
Conjunction

e3:died is BEFORE *e4:canceled*

∧*e4:canceled* is a PARENT of *e5:affecting*

=> *e3:died* BEFORE *e5:affecting*

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



Goal: letting the neural model capture the logical constraints.

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

Incorporating Logical Constraints in A Neural Architecture

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

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

- L_A Task Loss: $\top \rightarrow r(e_1, e_2) \quad \boxed{\rightarrow} \quad -w_r \log r_{(e_1,e_2)}$
- L_S Implication Loss: $\alpha(e_1, e_2) \leftrightarrow \bar{\alpha}(e_2, e_1) \quad \boxed{\rightarrow} \quad |\log \alpha_{(e_1,e_2)} - \log \bar{\alpha}_{(e_2,e_1)}|$
- L_C Conjunction Loss: $\alpha(e_1, e_2) \wedge \beta(e_2, e_3) \rightarrow \gamma(e_1, e_3) \quad \boxed{\rightarrow} \quad \log \alpha_{(e_1,e_2)} + \log \beta_{(e_2,e_3)} - \log \gamma_{(e_1,e_3)}$

  $\alpha(e_1, e_2) \wedge \beta(e_2, e_3) \rightarrow \neg\delta(e_1, e_3) \quad \boxed{\rightarrow} \quad \log \alpha_{(e_1,e_2)} + \log \beta_{(e_2,e_3)} - \log(1 - \delta_{(e_1,e_3)})$
- 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	–

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

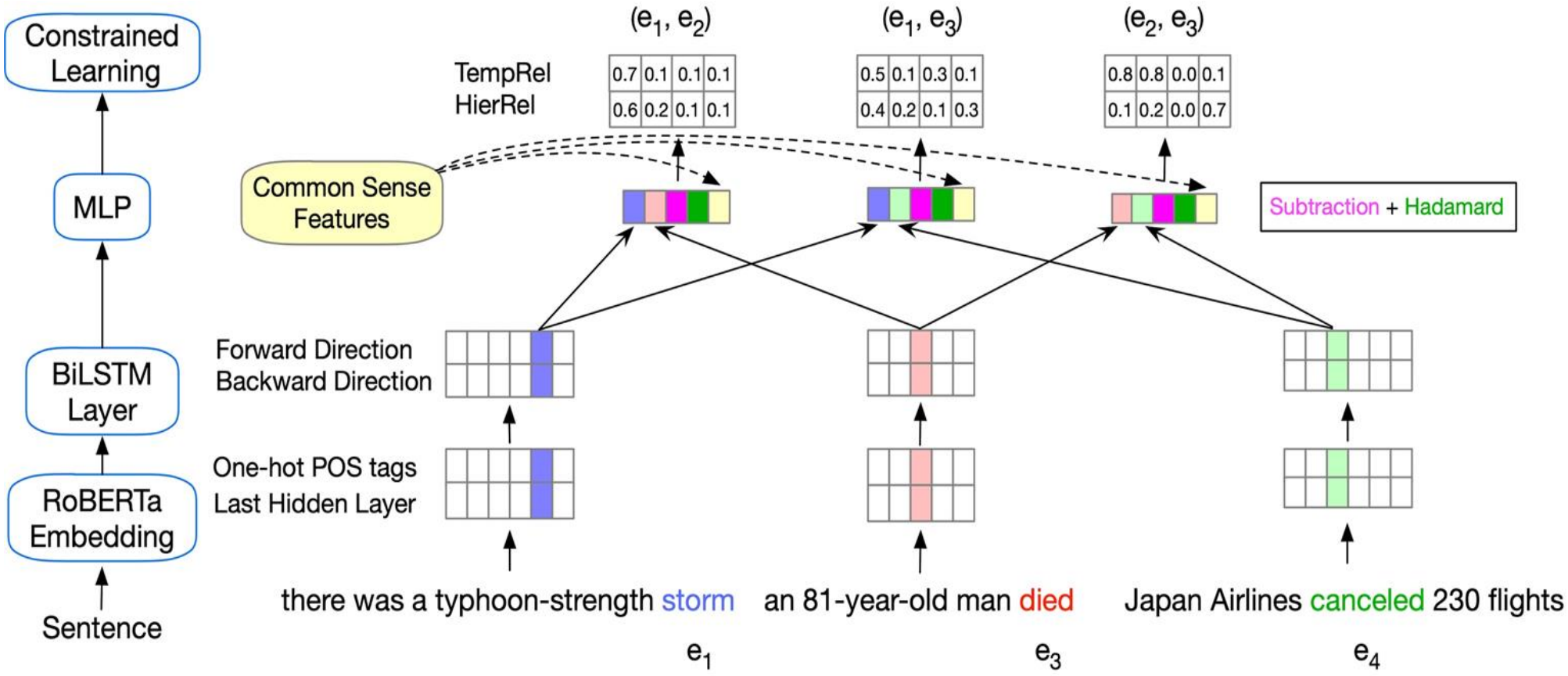
Joint Constrained Learning

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

Task loss

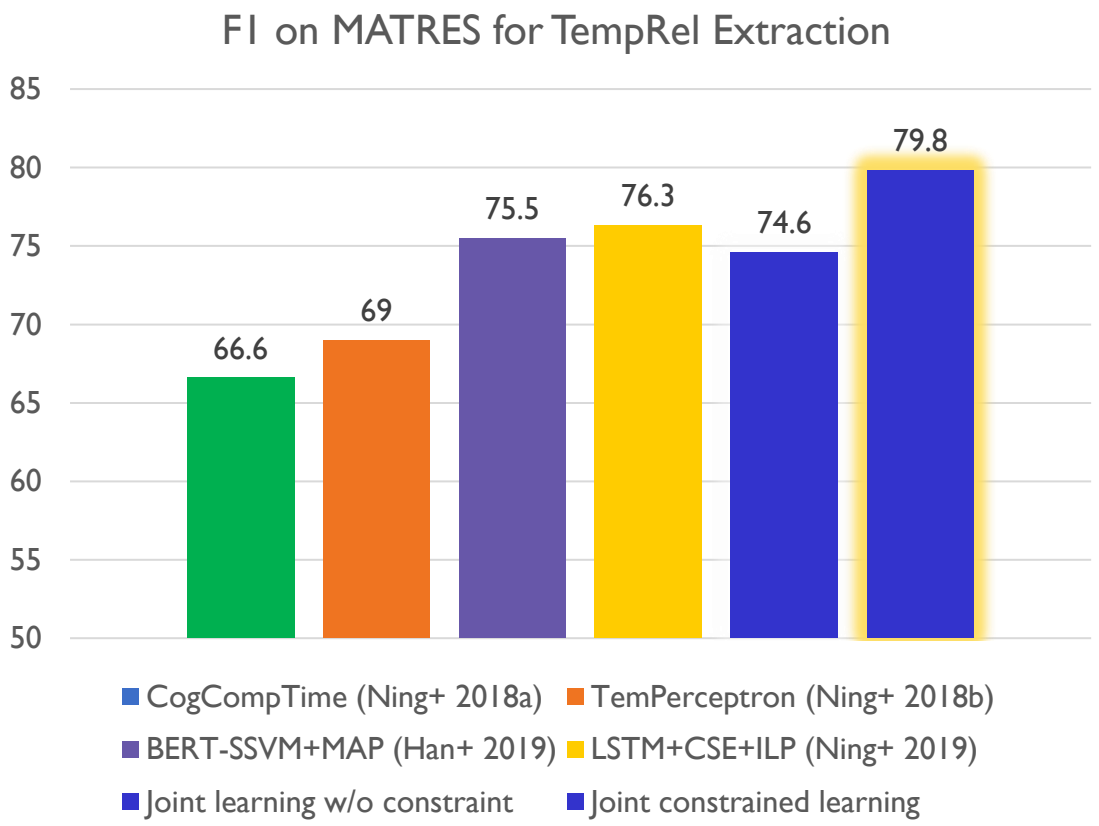
Implication and conjunction constraint losses

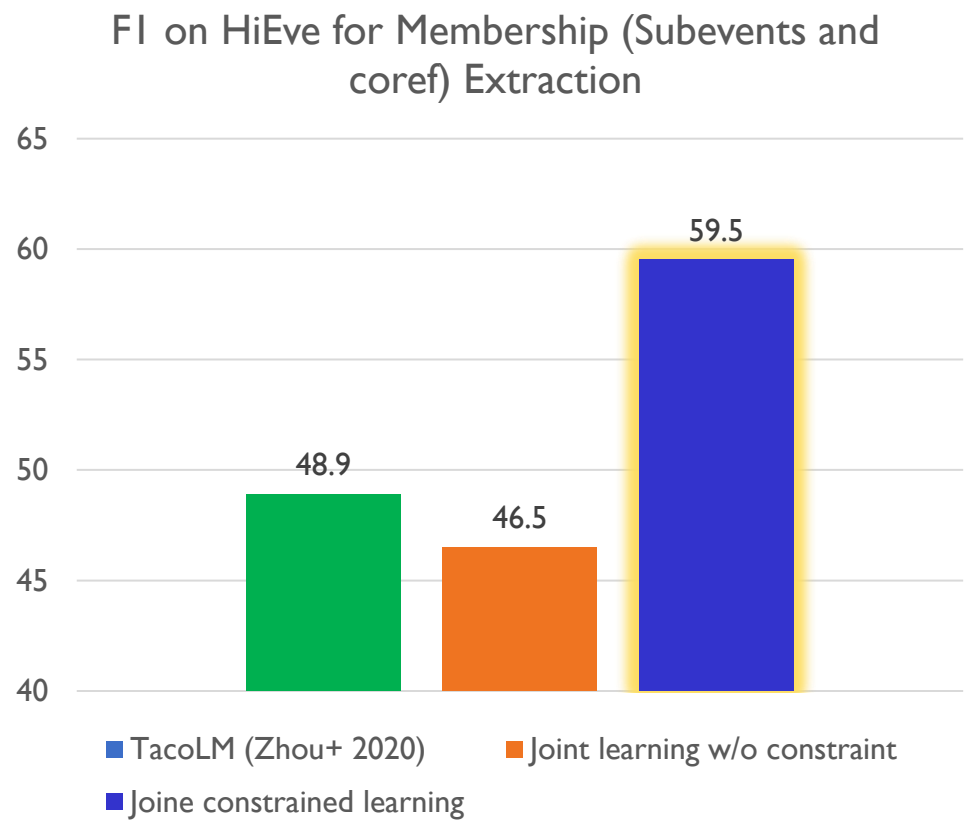
Loss Function: $L = L_A + \lambda_S L_S + \lambda_C L_C$



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

The Joint Constrained Learning Architecture





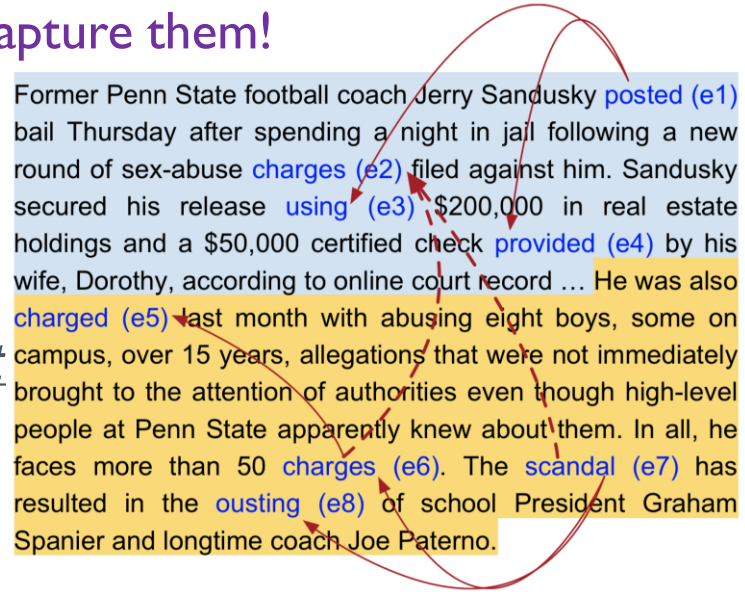
Key Observations

- Constraints are a natural bridge for learning resources with different sets of relations
- Adding constraints in learning is sufficient to enforce logical consistency of outputs, surpassing ILP in inference (w/ 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 segments

- **Text segmentation** [Lukasik+ EMNLP-20]: identifying standalone subdocument pieces
- *Subevent relations happen much more often* ***within the same narrative segment***

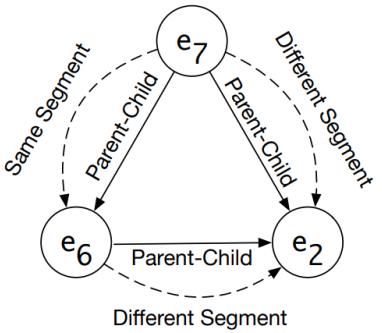
A hard-to-articulate soft probabilistic constraint. How do we capture it?

Constraint Learning

Training a single-layer rectifier network on all ``triangles'' of the training data

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

Estimates probabilities of conjunctive constraints

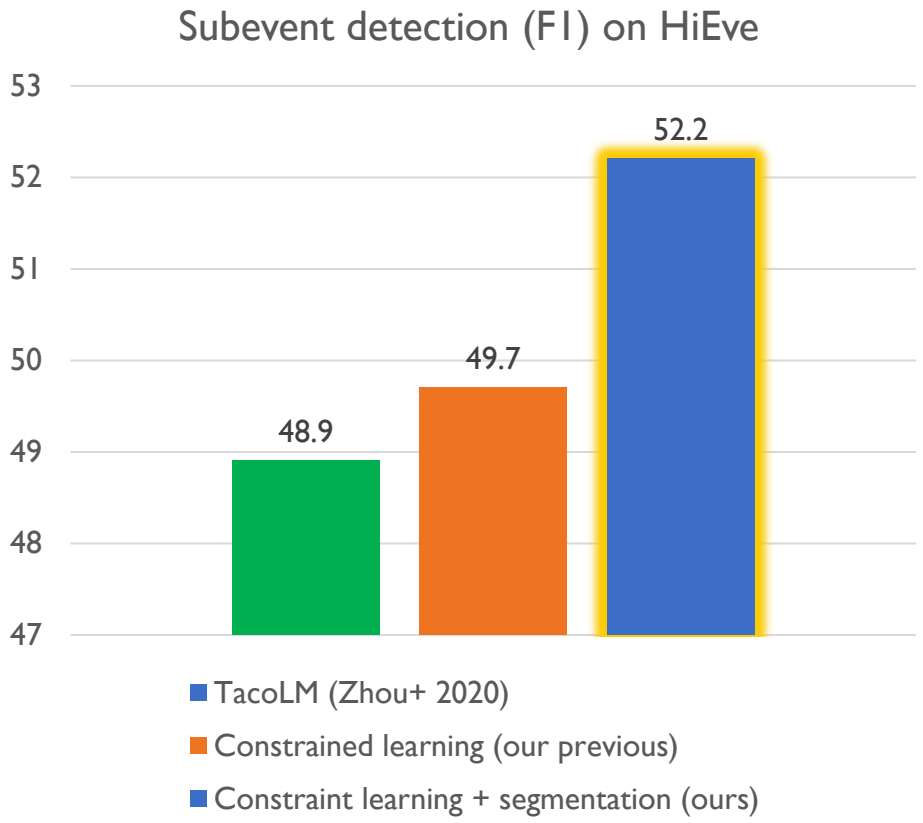


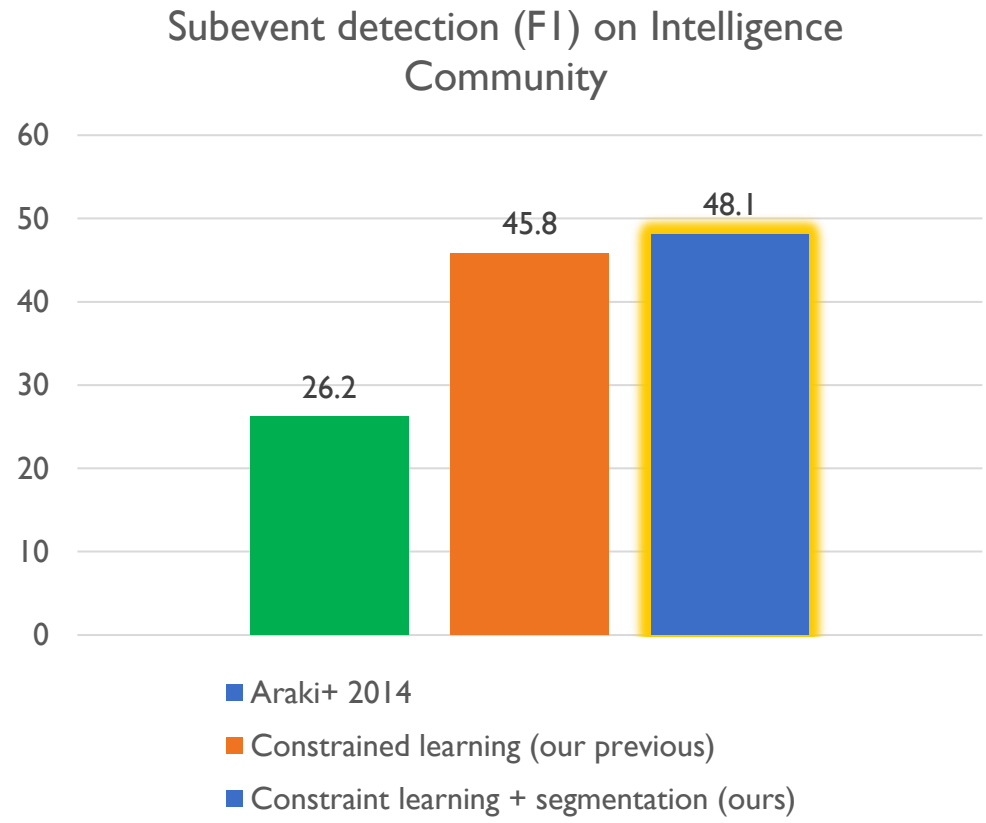
Adding the rectifier estimated constraint probability as a regularization loss in task training

$$L_{cons} = -log\left(Sigmoid\left(1 - \sum_{k=1}^{N} ReLU(\mathbf{w}_k \cdot \boldsymbol{\psi} + b_k)\right)\right)$$

W**C**R. *Learning Constraints and Descriptive Segmentation for Subevent Detection*. **EMNLP 2021**

Automatically Learning Constraints





Constraint learning automatically captures soft constraints, and allow narrative segmentation to be introduced as a form of indirect supervision.

W**C**R. *Learning Constraints and Descriptive Segmentation for Subevent Detection*. **EMNLP 2021**

Indirect Supervision from NLI

Ultra-fine Entity Typing

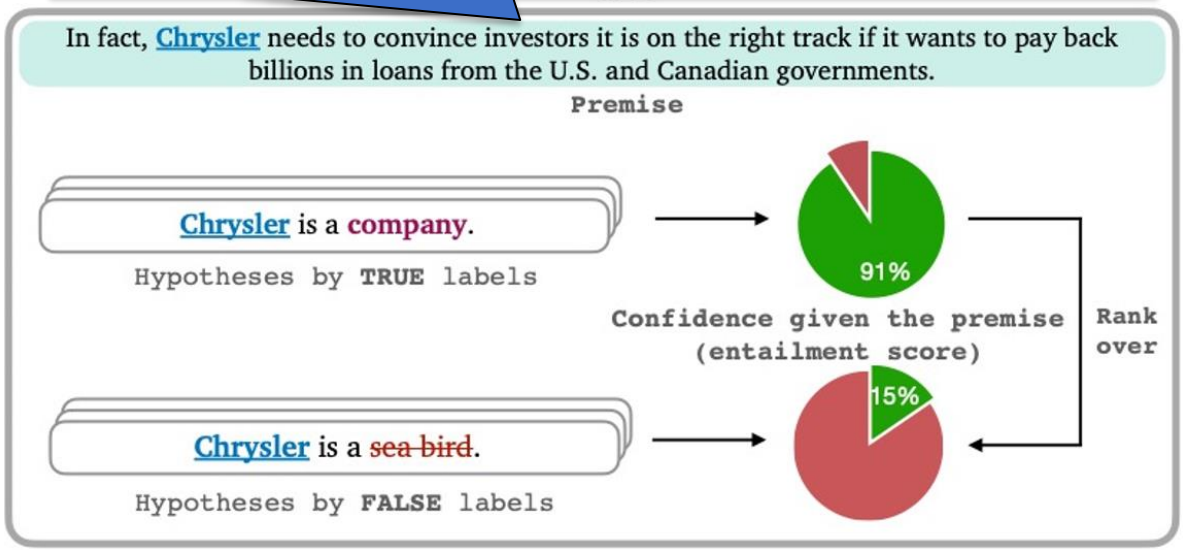
Inferring extremely diverse and fine-grained identities

- **>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

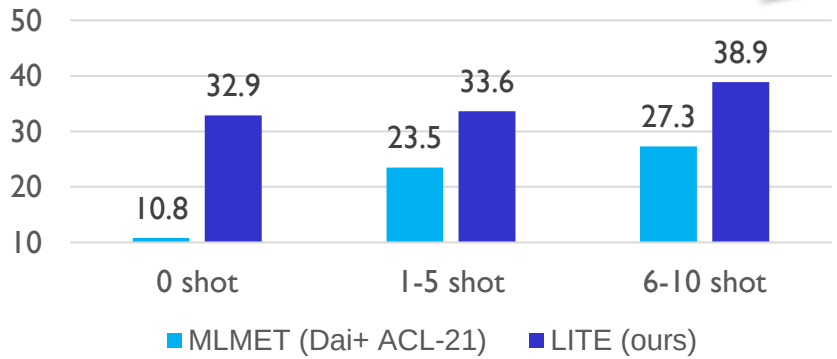
Reformulating as **NLI +learning to rank.**



Our system (LITE) is the current SOTA and the first to reach >50% F1 (for >10k types) on UFET

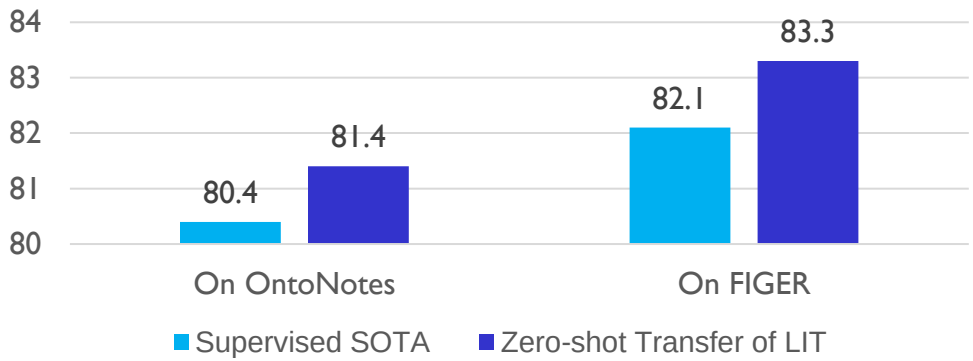
Zero and Few-shot performance vs. previous SOTA

Generalize well on unseen types.



Zero-shot Transfer vs. Previous Supervised SOTA (Trained w/ ~100k labeled instance)

Outperforming full-shot SOTA with 0-shot transfer.

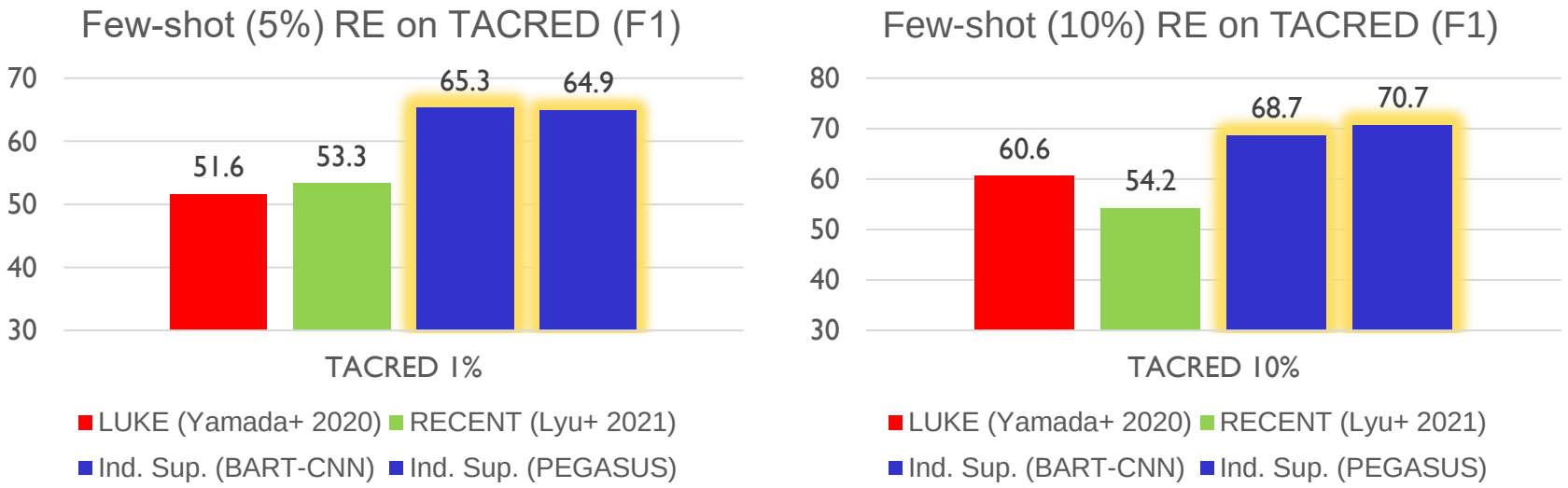


LY**C**. Ultra-fine Entity Typing with Indirect Supervision from Natural Language Inference. **TACL** 2022

Indirect Supervision with Summarization

- Summarization preserves **salient information in longer text**
  - Extracts can be viewed as **one kind of salient information**
  - Ask a summarization model for **constrained generation** of verbalized relations

⇩ Indirect supervision from abstractive summarization corpora (CNN/Dailymail, Xsum) leads to generalizable RE



Example

Input:

Subject: Mandelbrot **Type:** person

Object: Poland **Type:** country

Sentence:

Mandelbrot was born in Poland but as a child moved to France.

Input Sequence Construction:

The subject is Mandelbrot. The object is Poland. The type of Mandelbrot is person. The type of Poland is country. Mandelbrot was born in Poland but as a child moved to France.

Relation Verbalization:

r_1: Mandelbrot is a Poland

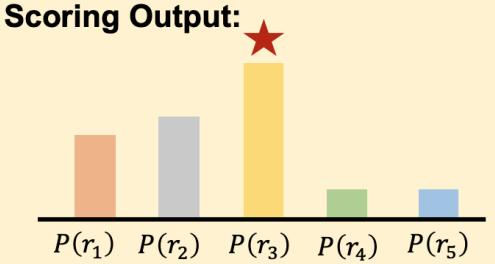
r_2: Mandelbrot was born in the city Poland

r_3: Mandelbrot was born in the country Poland

r_4: Mandelbrot was founded by Poland

r_5: Mandelbrot has no known relations to Poland

Trie Scoring Output:

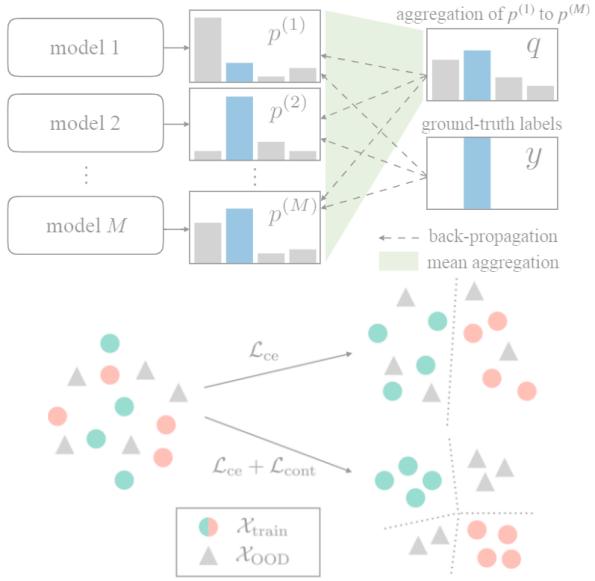


(Mandelbrot, country of birth, Poland)

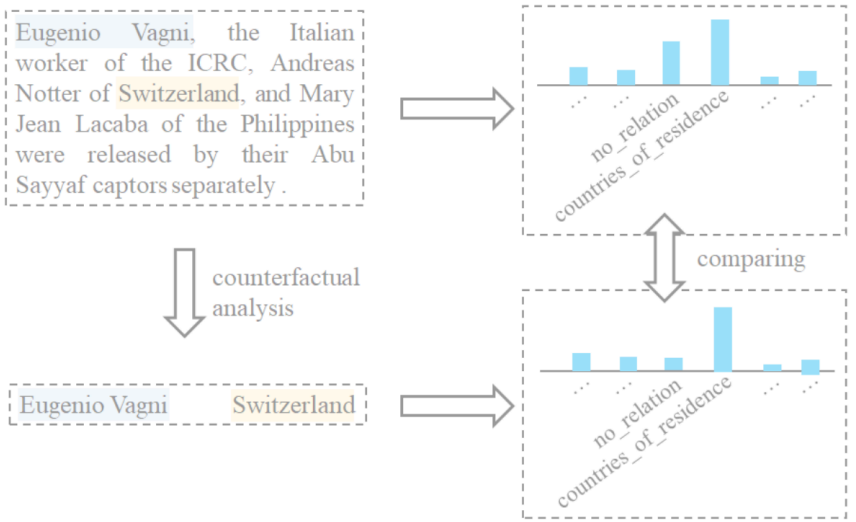
LHZMC. Summarization as Indirect Supervision for Relation Extraction. **EMNLP** 2022

In This Talk

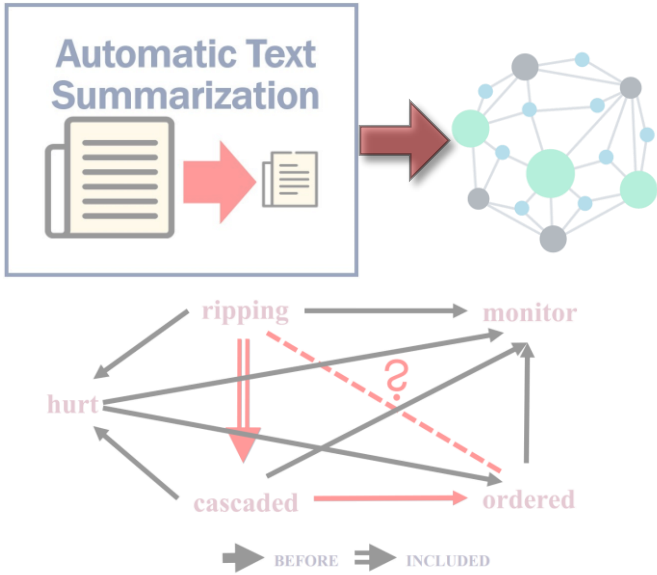
1. Noise-robust IE



2. Faithful IE



3. Constrained/Indirectly Supervised IE

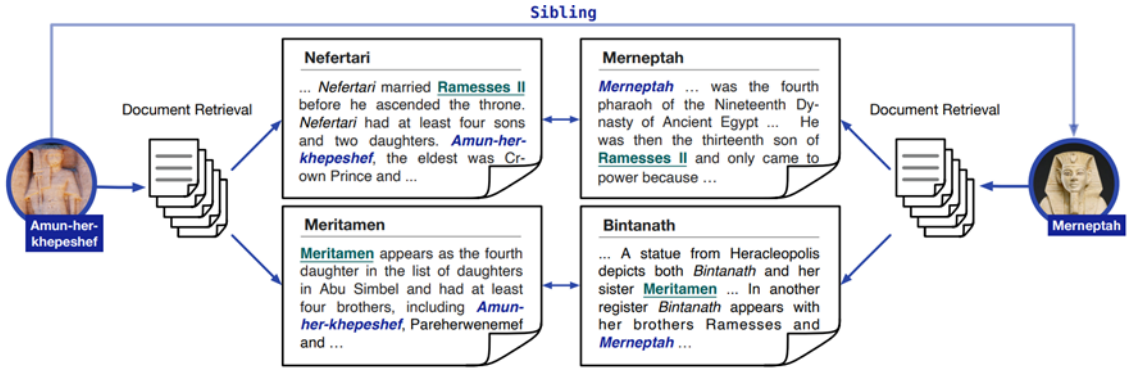


4. Future Directions

Thinking Across Documents

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

① Inducing relations across documents



- Cross-document RE
- Timeline extraction
- Task process induction

② Consolidating unevenly distributed knowledge

DBpedia — Novel

ヂビペディア

- 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
- Resolving redundancy and logical inconsistency
- Tracking information pollution for trustworthiness
- Long-form document modeling
- Mitigating frequency biases
- …

☜ A long way to go

Selective Extraction

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

Dr. Chang (PER) graduated from UIUC (ORG) in 2015 .

attended

Dr. Chang (PER) did not attend UPenn (ORG) .

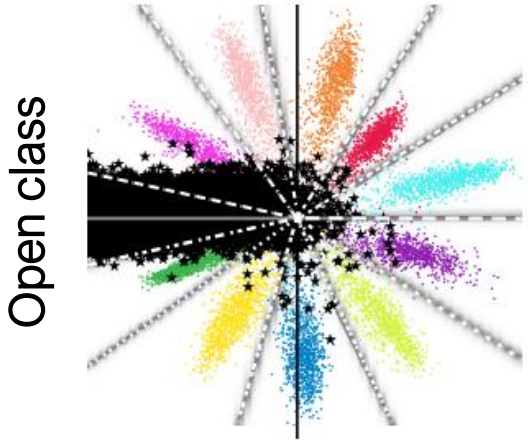
[abstain]

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

How to make inference more selective?

A supervised approach can be a choice

- Classify exceptions into an open class/background set
- However, exceptions can never be close to exhaustive in training data



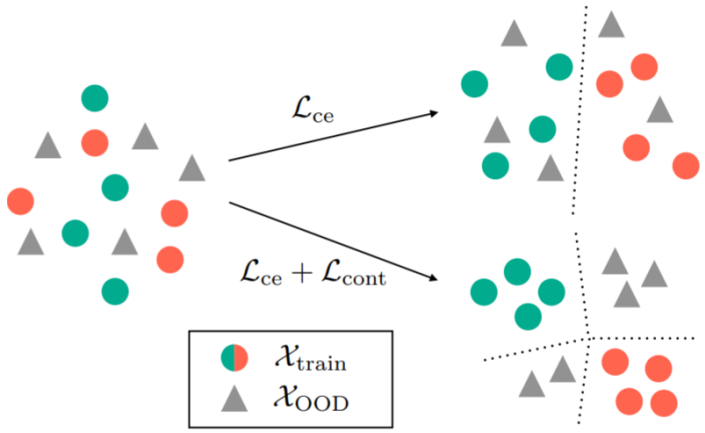
- Task training data
- Annotated exceptions

Dhamija et al. *Reducing network agnostophobia*. **NeurIPS** 2018

Learning to Abstain without Annotated "Abstention"?

This is still an underexplored area, but there are at least two lines of strategies

Unsupervised out-of-distribution (OOD) detection



Increase inter-class discrepancy ⇒ Better OOD detection

Creating compact representations with (margin-based) contrastive learning

- Indirectly making OOD instances as "background" representation

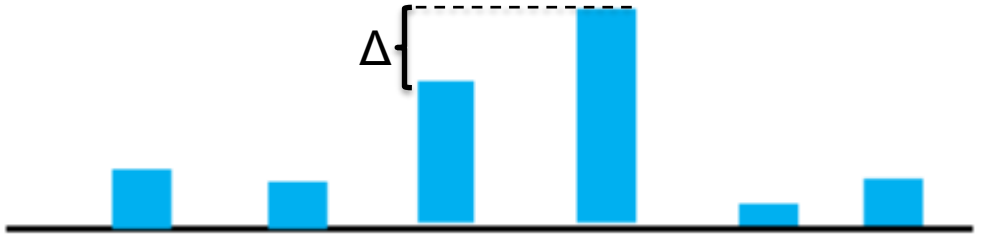
Inference with Mahalanobis distance

- High-order distance measures improve OOD detection

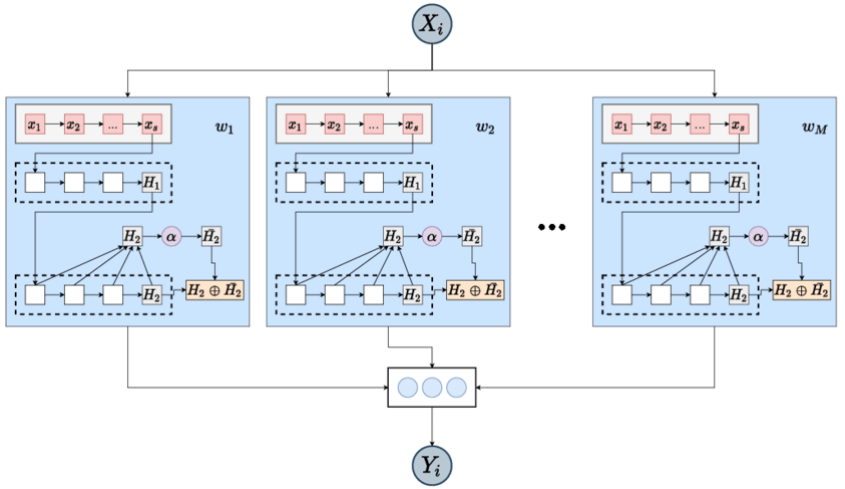
ZL**C**. *Contrastive Out-of-Distribution Detection for Pretrained Transformers*. **EMNLP** 2021

Estimating the uncertainty of prediction

Softmax response: difference between top two class predictions



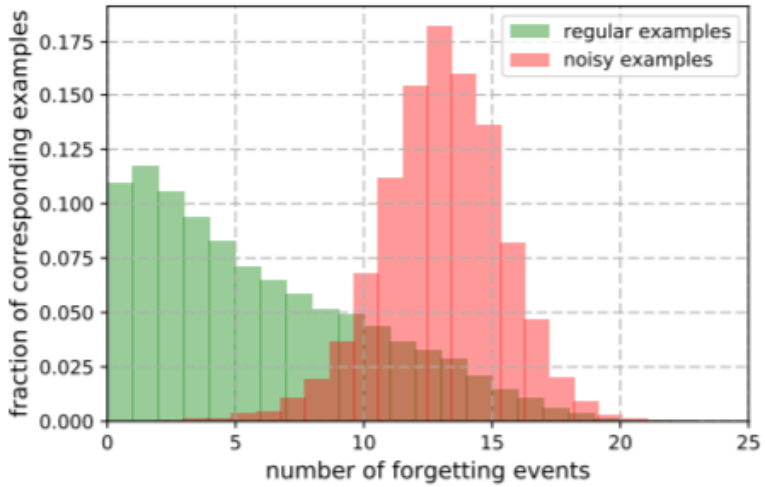
Prediction variance in Monte-Carlo dropout



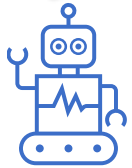
Xin et al. *The Art of Abstention: Selective Prediction and Error Regularization for Natural Language Processing*. **ACL** 2021

Harnessing the Deep IE Models

① Differentiating between noisy and hard instances

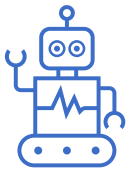


Hard or ***incorrect***?



② Learning techniques that improve faithfulness

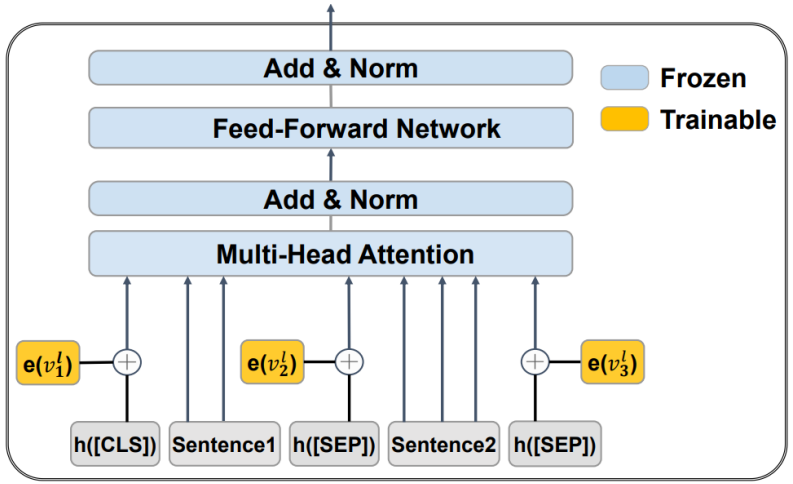
Bill Gates [illegible] Microsoft [illegible].



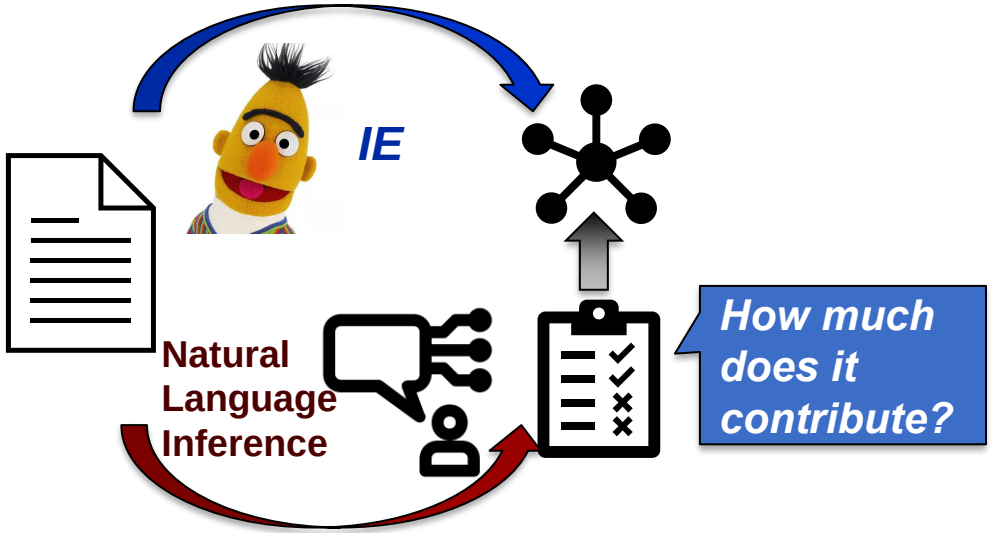
Automatic feature debiasing *w/o prior knowledge.*

Attention smoothing *or* ***perturbation****.*

③ Pruning PLM subnetworks for different IE tasks



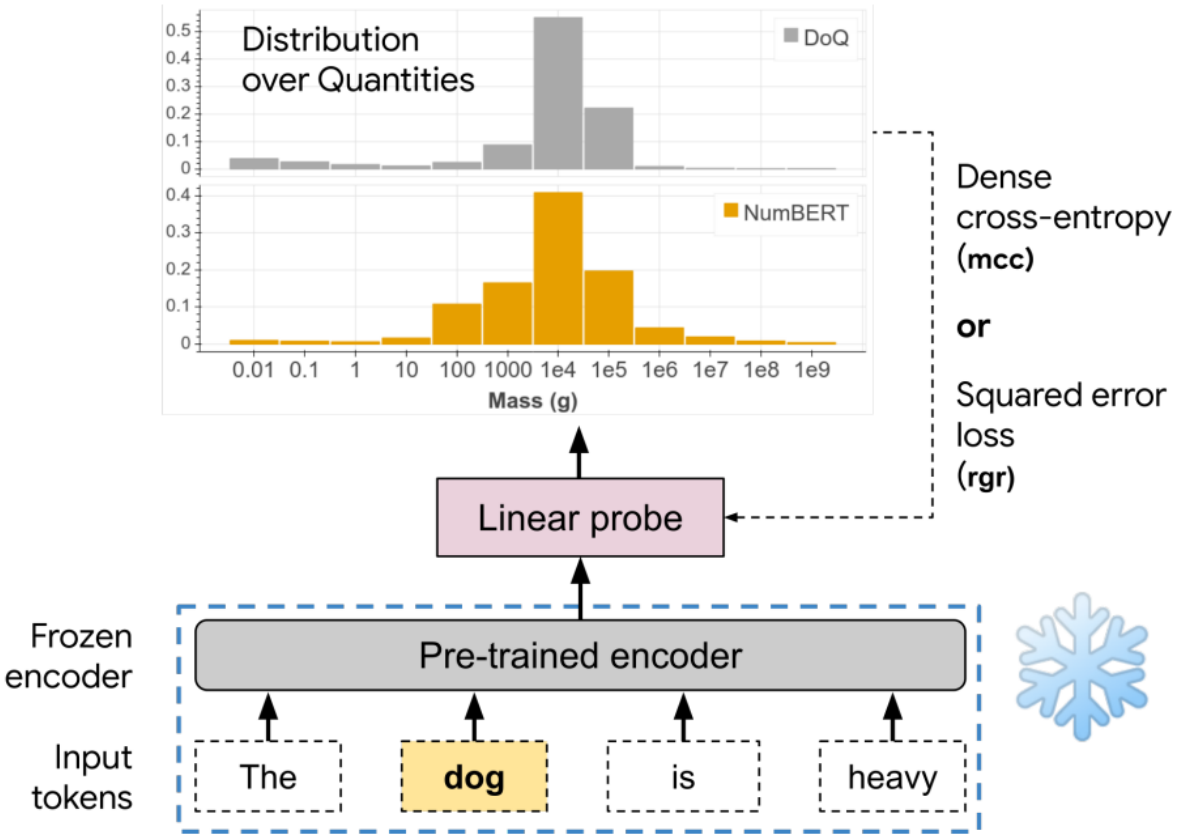
④ Capturing the relations of different IE tasks



YHZ**C**. Parameter-Efficient Tuning with Special Token Adaptation. **EMNLP** 2022

Quantitative Extraction

Extracting quantities



Temporal verification

Medical Reports

… The patient has been constantly smoking in the past year …

Has the patient smoked in the past month?

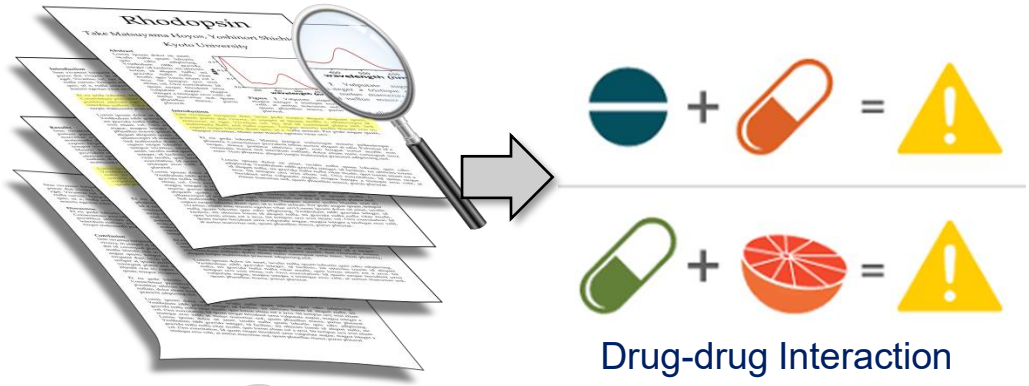
UNIFIED-QA :

Large models still do not support quantitative reasoning well

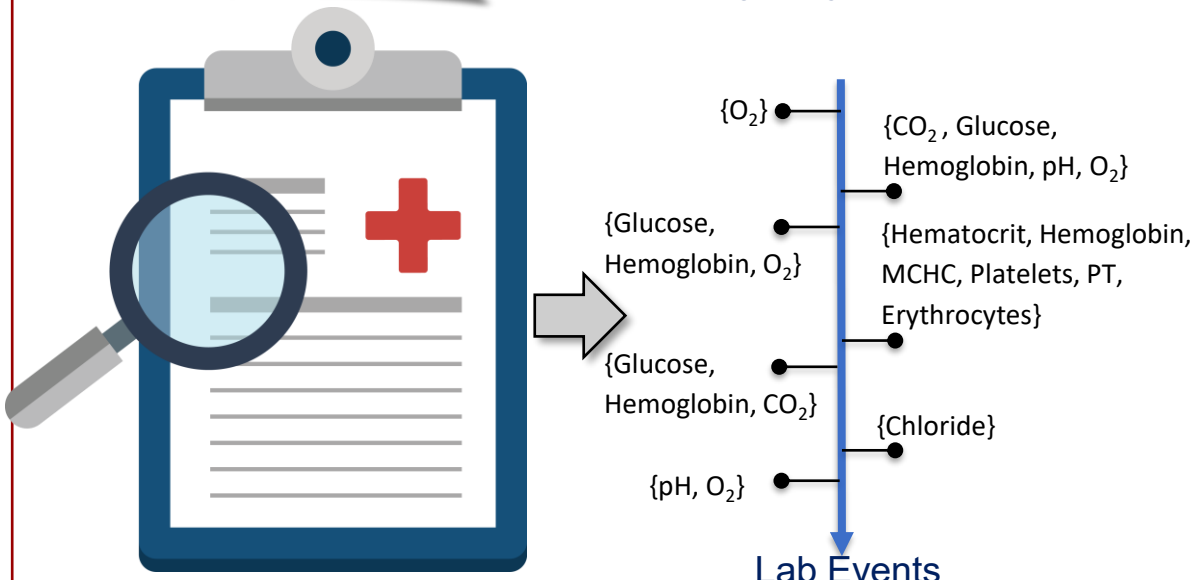
Zhang et al. *Do Language Embeddings Capture Scales?* EMNLP: Findings 2020

IE for the Common Good

Medicine and healthcare

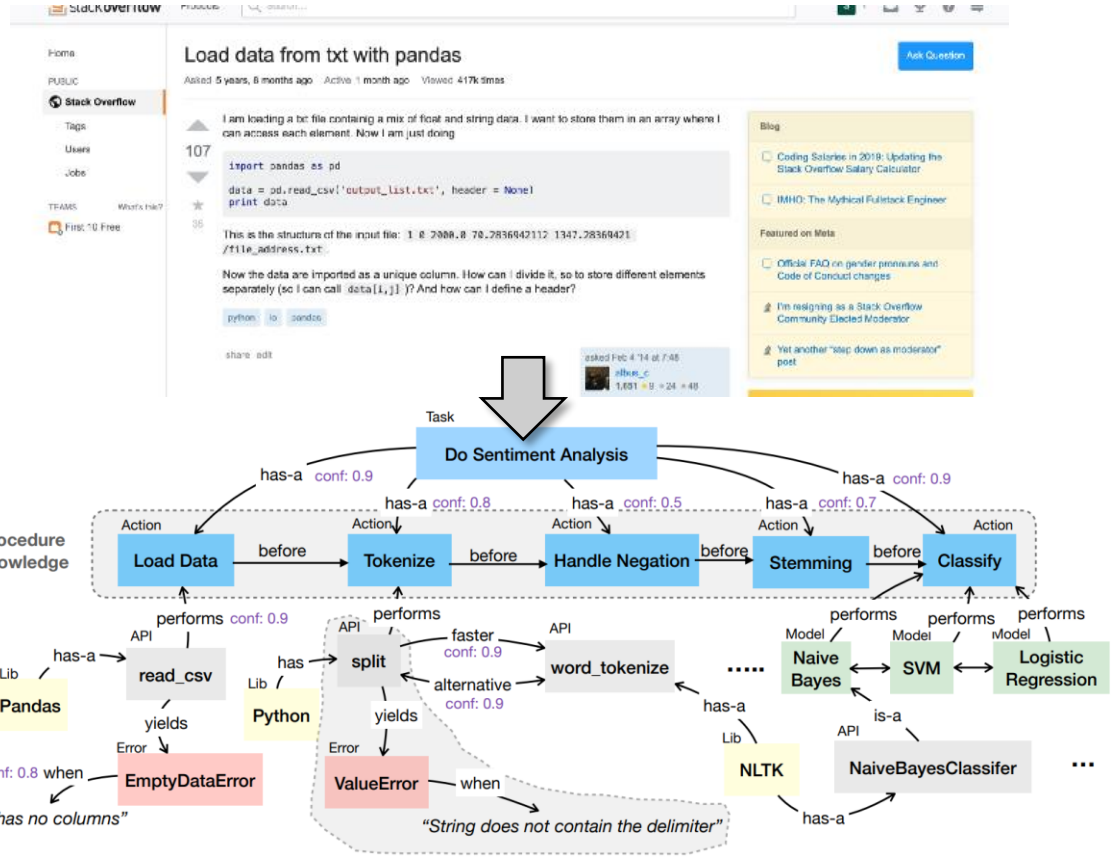


Drug-drug Interaction



Lab Events

Programming education



Low-resource domains that particularly

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

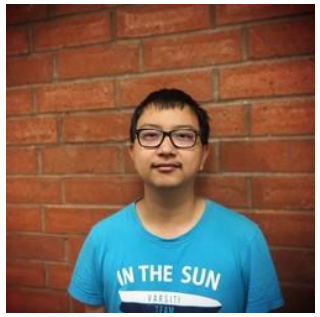
Microsoft Research

David Geffen School of Medicine

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

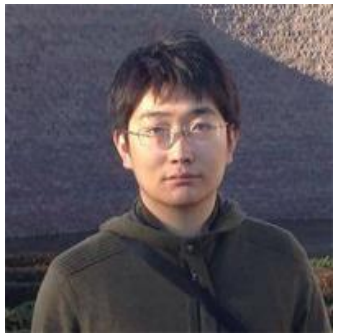
Penn University of Pennsylvania
Carnegie Mellon University
NUS National University of Singapore
University of Cambridge
amazon
Microsoft Research
UCLA David Geffen School of Medicine

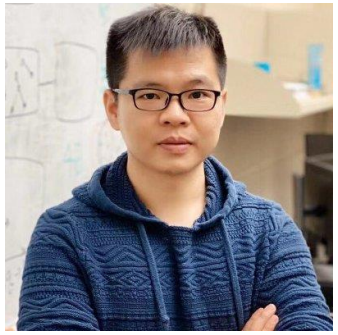
Sponsors

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

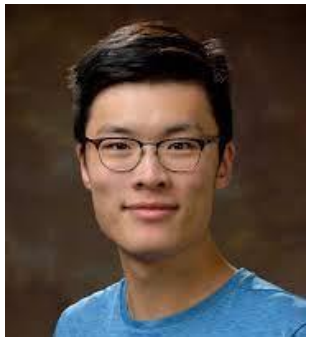
USC Viterbi School of Engineering | VIRGINIA TECH | UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN | Penn | amazon

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

Thank You