



USCViterbi

School of Engineering

Robust and Indirectly Supervised Information Extraction

Muhao Chen

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



How do we make IE models *more reliable*?

Information Extraction (IE)



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

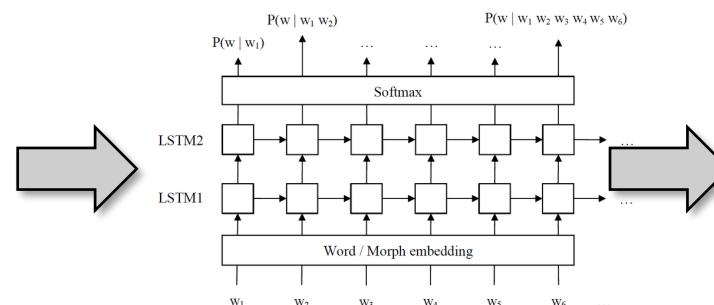
Honolulu

From Wikipedia, the free encyclopedia

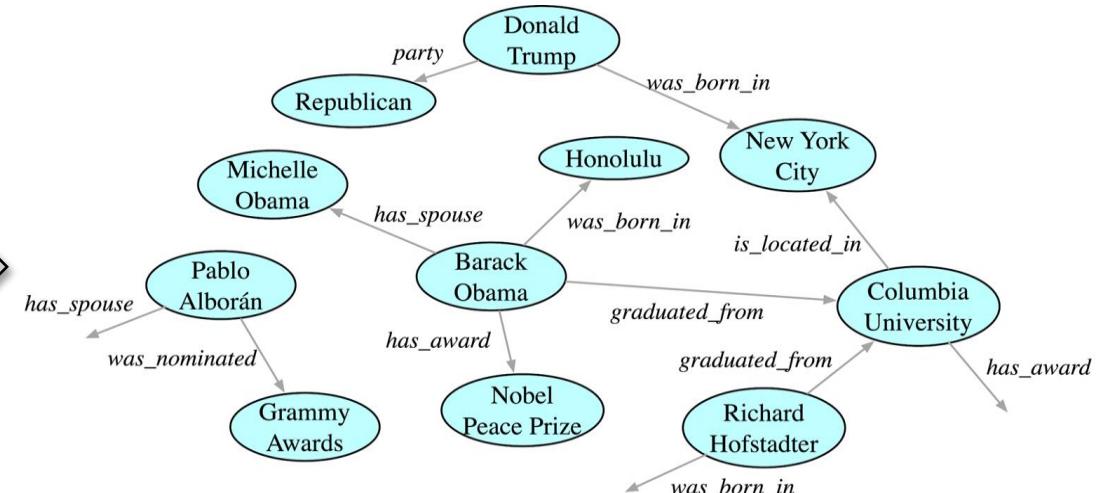
This article is about the largest city and state capital city of Hawaii. Honolulu itself, see [Honolulu County, Hawaii](#). For other uses, see

Honolulu (/ha'no lū'lū/; [7] Hawaiian: [hono'lulu]) is the capital and largest city of the U.S. state of Hawaii, which is located in the Pacific Ocean. It is an unincorporated county seat of the consolidated City and County of Honolulu, situated along the southeast coast of the island of Oahu,[8] and is the westernmost and southernmost major U.S. city. Honolulu is Hawaii's main gateway to the world. It is also a major hub for international business, finance, hospitality, and military defense in both the state and Oceania. The city is characterized by a mix of various Asian, Western, and Pacific cultures, as reflected in its diverse demography, cuisine, and traditions.

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



IE Model/System





IE is Integral to Natural language Understanding

Understanding text depends on the ability to extract information from it

- › Identifying and contextualizing
 - » entities,
 - » quantities (and their scope),
 - » events,
 - » relations, etc.



- › Infer the identities of entities and concepts

- › Facilitate answering questions about the text

In the first quarter, the Vikings scored the first touchdown of the game, a 1-yard pass from Matt Ryan to wide receiver Greg Olsen. ... In the third quarter, the Vikings responded with a 1-yard run by Adrian Peterson. The Bears increased their lead over the Vikings with Jay Cutler's 2-yard TD pass to tight end Desmond Clark. The gap was reduced when Cutler fired a 6-yard TD pass to tight end Visanthe Shiancoe. The Vikings responded with Adrian Peterson's second 1-yard TD run. The Bears responded with Cutler firing a 20-yard TD pass to wide receiver Devin Aromashodu.

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

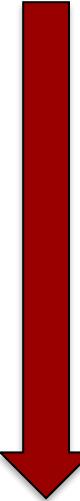
Mayor Rahm Emanuel How much did his challengers raise? toward his bid for a third term – more than five times the total raised by his 10 challengers combined, campaign finance records show.

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

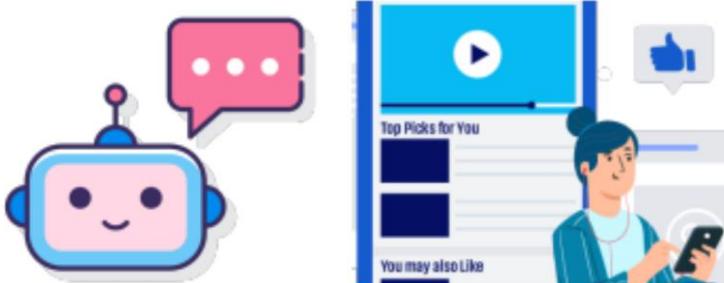
IE Benefits For Content Management (DARPA KMASS Project)



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

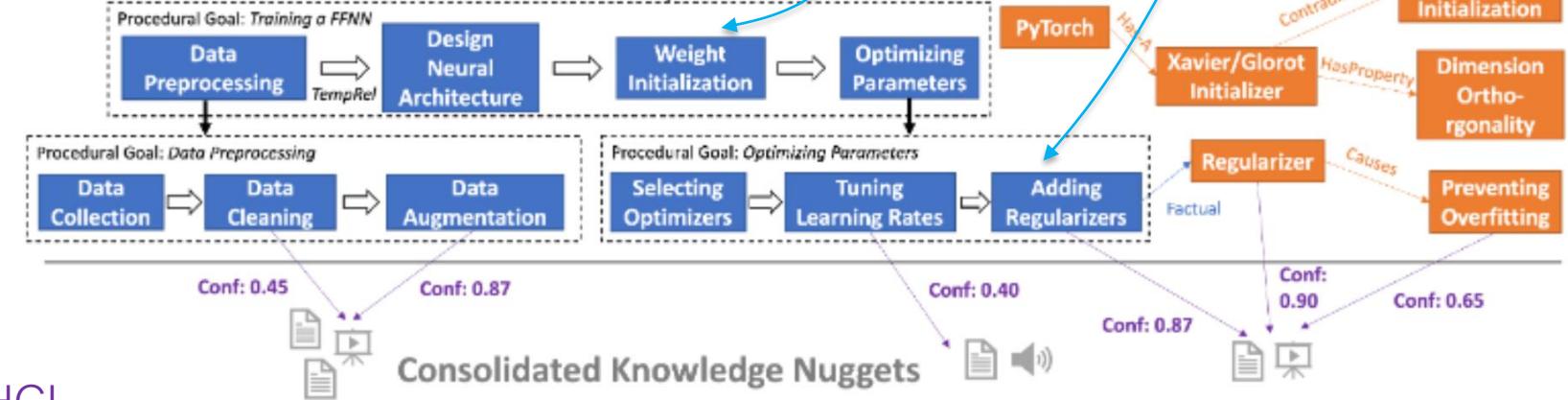
The leftmost layer in this network is called the input layer, and the neurons within the layer are called input neur. The rightmost or output layer contains the output neurons, or, as in this case, a single output neuron. The middle layer is called a hidden layer, since the neurons in this layer are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

Regularization in Deep Learning

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

Procedural Index

Training a Feed-Forward Neural Network

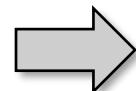




IE Is the Backbone of Nearly All Knowledge-driven Tasks

WIKIPEDIA The New York Times

PubMed bioRxiv



Knowledge Bases

Freebase™

ASER

ConceptNet
An open, multilingual knowledge graph

yAGO
select knowledge

WIKIDATA

DBpedia

Bio/Med Databanks

DRUGBANK

nextprot

MeSH

STRING

GO GENEONTOLOGY
Unifying Biology

PDB
PROTEIN DATA BANK

Knowledge Representation

Common Sense

Narrative Understanding

Commonsense QA
Event Prediction
Intent Prediction

Storytelling
Content Selection
Newsworthiness Detection

bioinformatics
data sequence
protein analysis
genomes simulations
information coding
access codon
rna snp
algorithms
primary motifs
bases
genes
homologous
accession
databases
storage phylogenetics
assembly biological variation
file cash
acids single
proteins
protein
peripheral
information
folds
synteny
WWW
organism
interactions
structure
regions
nlm
pdb
sql
ncbi
genes
homologous
accession
databases
storage phylogenetics
assembly biological variation
file cash
acids single
proteins
protein
peripheral
information
folds
synteny
WWW
organism
interactions
structure

Medical
INFORMATICS

Proteomic Interaction Prediction
Mutation Effect Estimation
Genomic Function Prediction

Diagnostic Prediction
Disease Phenotyping
Drug Repurposing



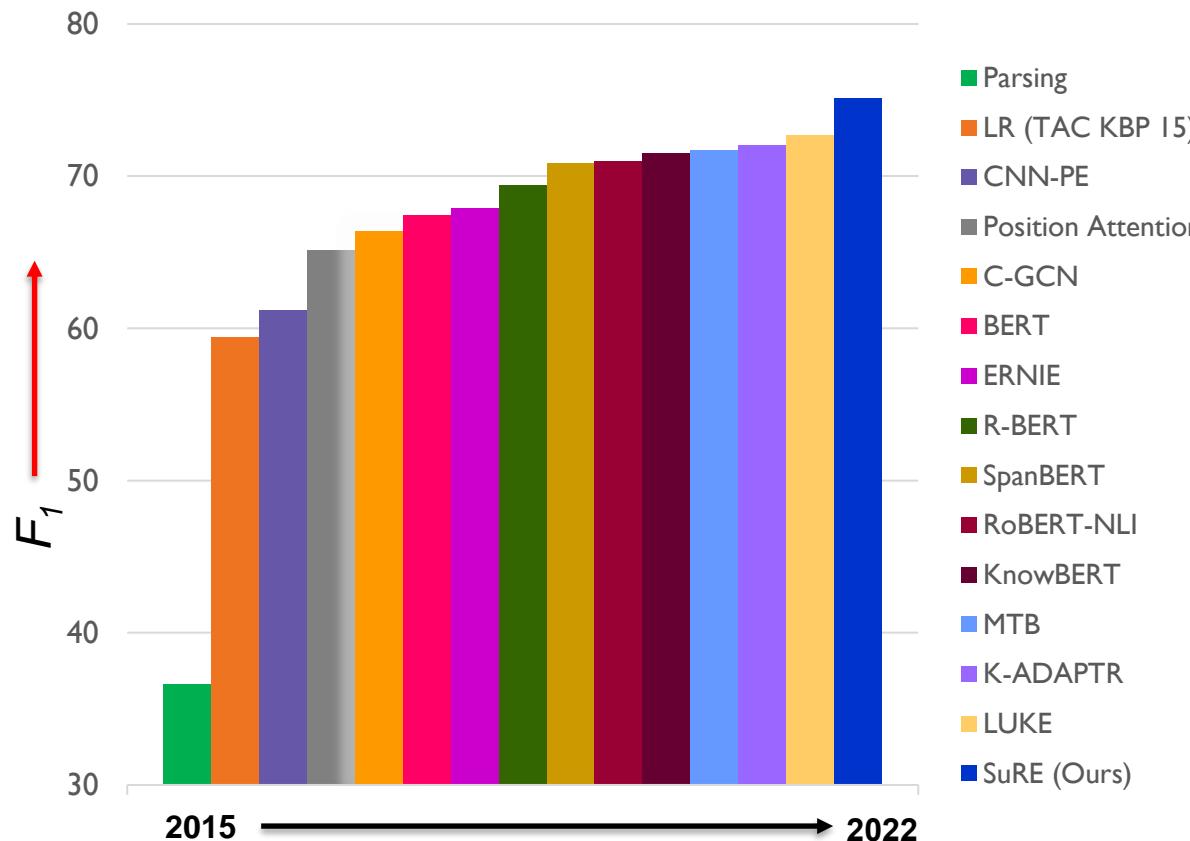
How IE Is Doing Today

On Benchmarks



Rise of the Pre-trained Models

Relation Extraction



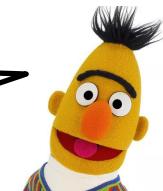
In Reality

Is still brittle, limitedly generalizable, and costly to develop

I stayed in **Treasure Island**
LOC
Type?



Island X



$$\text{pill} + \text{pill} = \text{warning sign}$$

???



I A R P A

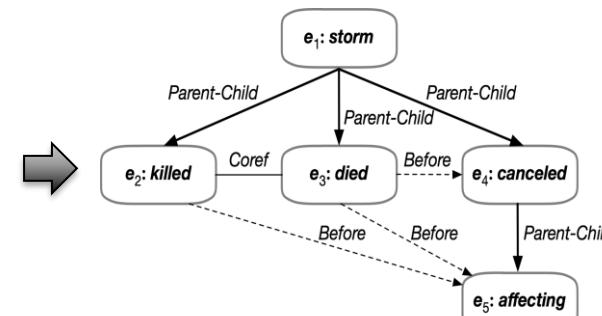
AIDA, KAIROS, BETTER,
KMASS, GAIA: all costing
tens of millions \$.



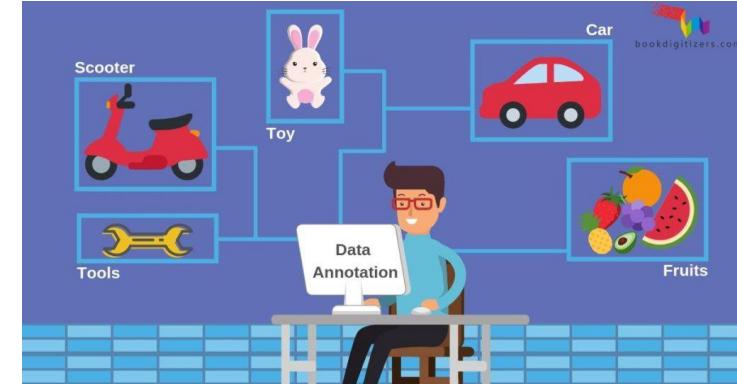
Challenge: Expensive Supervision

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

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



Reading long documents, recognizing complex structures



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

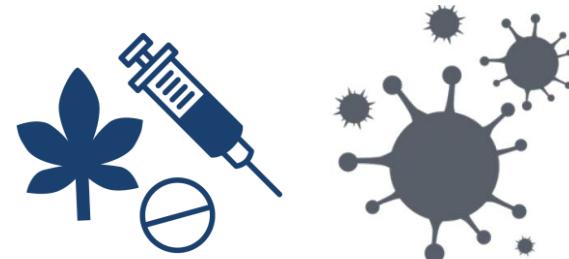
Insufficiency

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

Noise

- **In-correct labels:** e.g. 5-8% errors in TACRED, 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 .

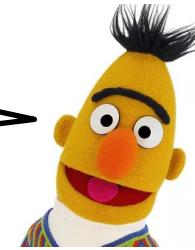
- Prior knowledge
- Biased training

Rel?



Visit ✓

FounderOf X



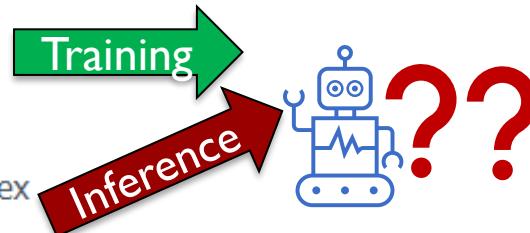
Harmful in many scenarios

- 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

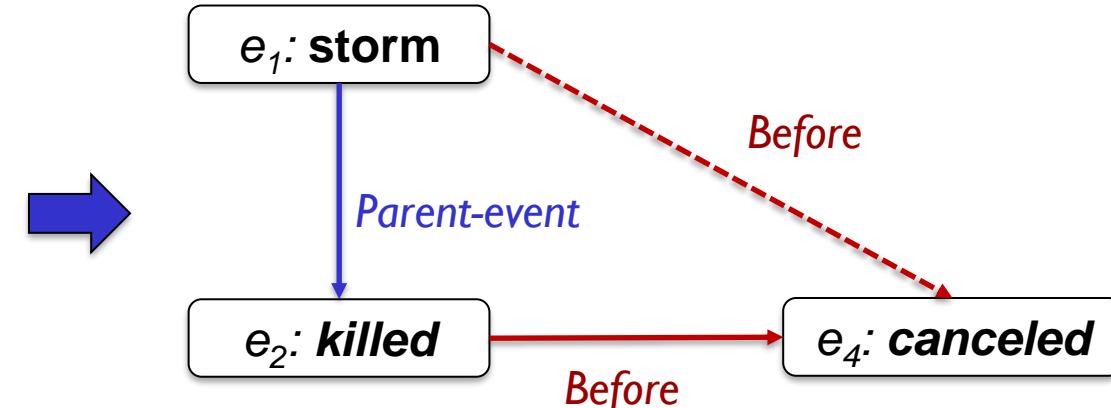


Challenge: Consistency

Extracts are not standalone.

An article about a storm hazard in Japan

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



Extraction Should be *Globally Consistent*

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

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

Implication, Negation ...

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

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



The goal of developing a **robust** IE system

Robustness in Learning

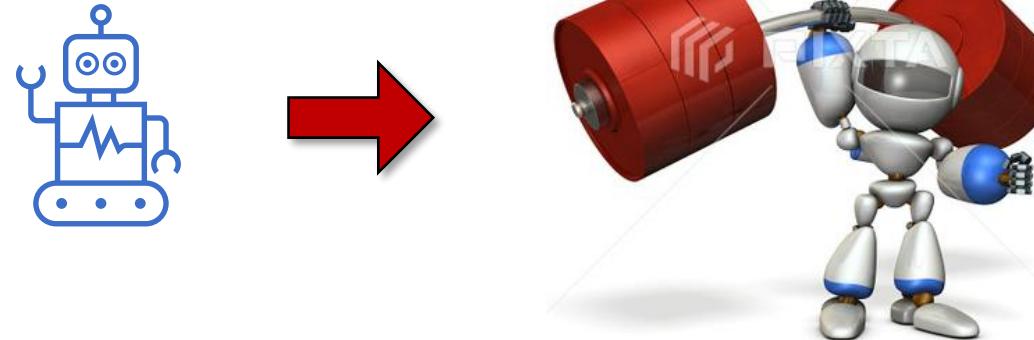
- Has to be achievable with *minimal and imperfect* supervision
 - Dealing with *noise, indirect supervision, constraints...*

Should not solely rely on 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

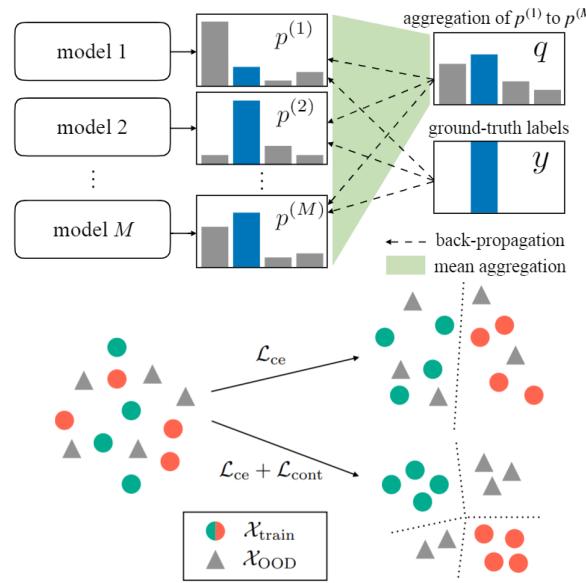
Self-contained, faithful and selective extraction.



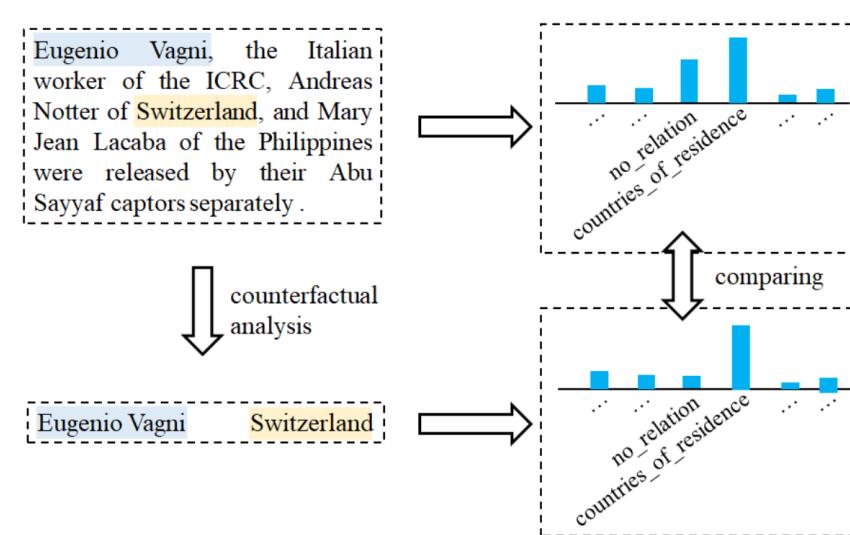
In This Talk



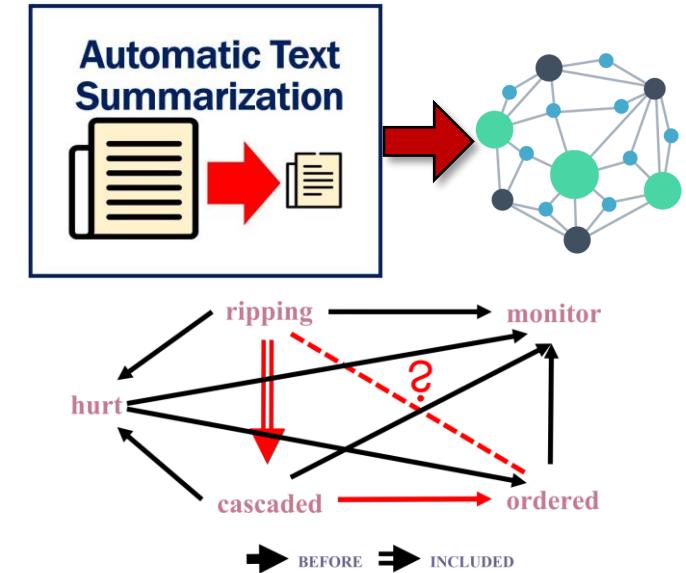
1. Noise-robust IE



2. Faithful IE



3. Indirectly Supervised IE

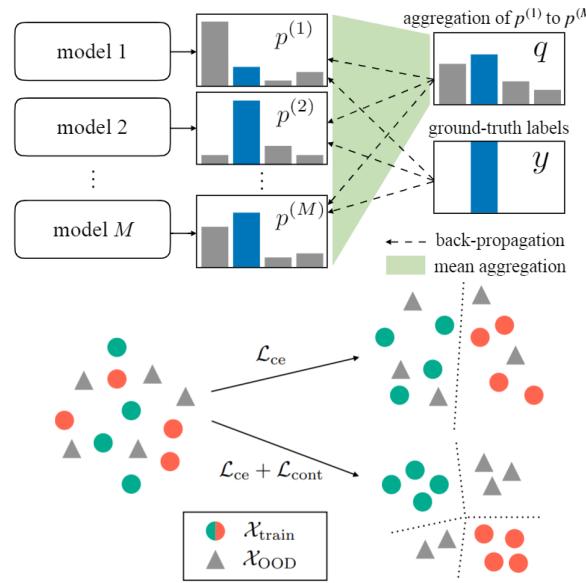


4. Future Directions

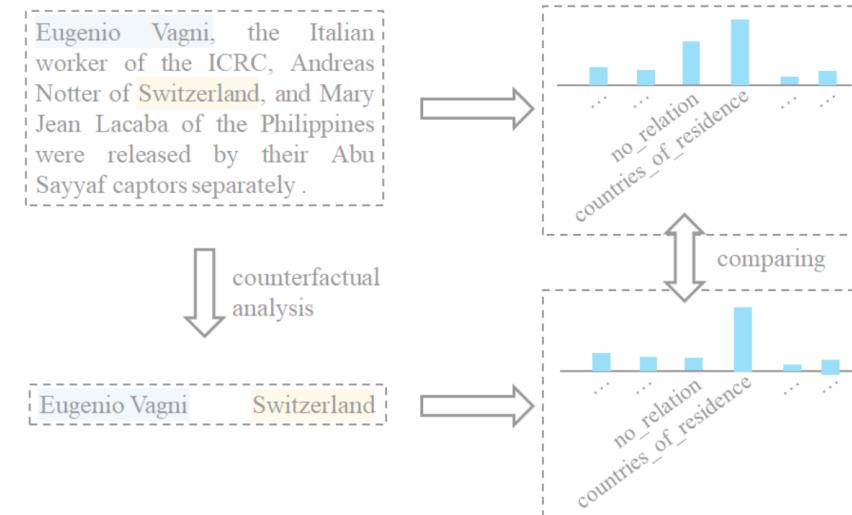




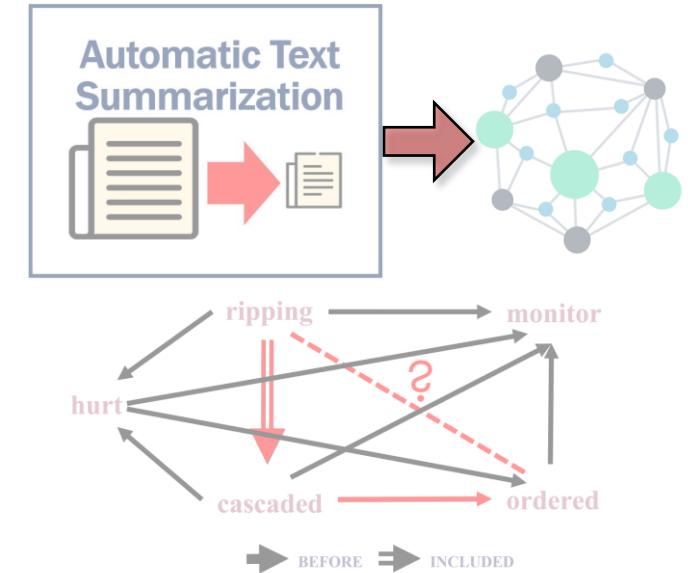
1. Noise-robust IE



2. Faithful IE



3. Indirectly Supervised IE



4. Future Directions



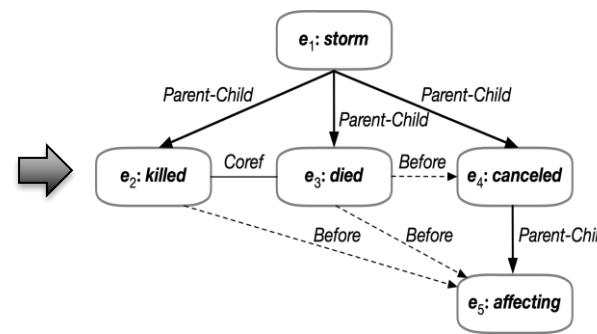


Noise In Training and Inference

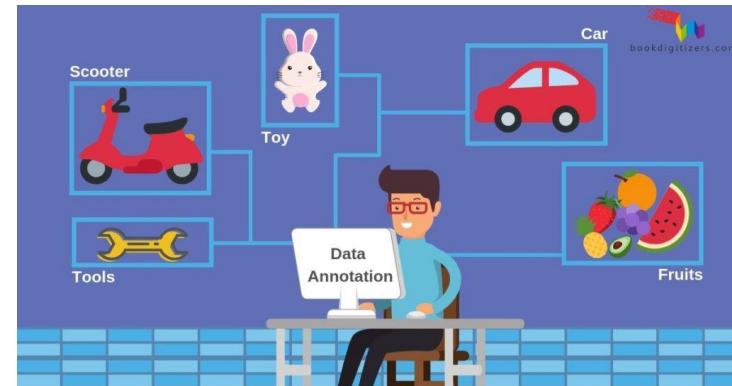
Training

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

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



Reading long documents, annotating complex structures



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

Hence, IE annotations are inevitably noisy. For example:

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

Even moderate noise leads to significant flaws

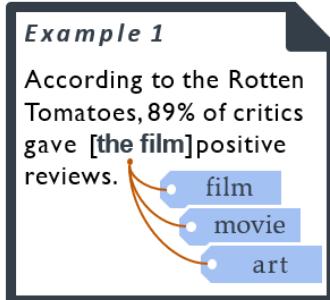
High-performance IE models have to be achievable with *imperfect supervision*



A Glance at Prior Solution: Supervised Denoising

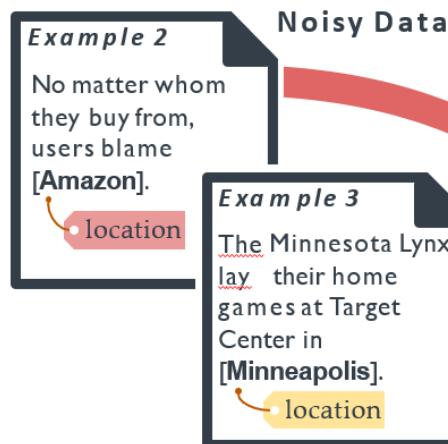
A noise **filtering** or **relabeling** model may be trained, if **clean data** are available.

- ① Labeled clean data and noisy data

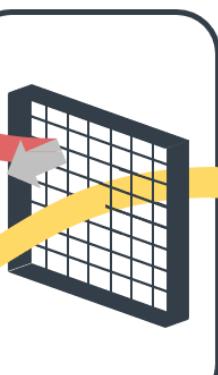


- ② Filtering model (binary classification): decide whether the example should be kept (binary classification)

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



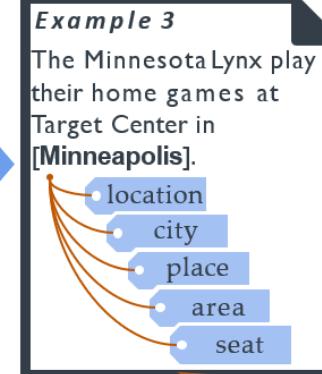
Filtering Model



Relabeling Model



Cleaned Data

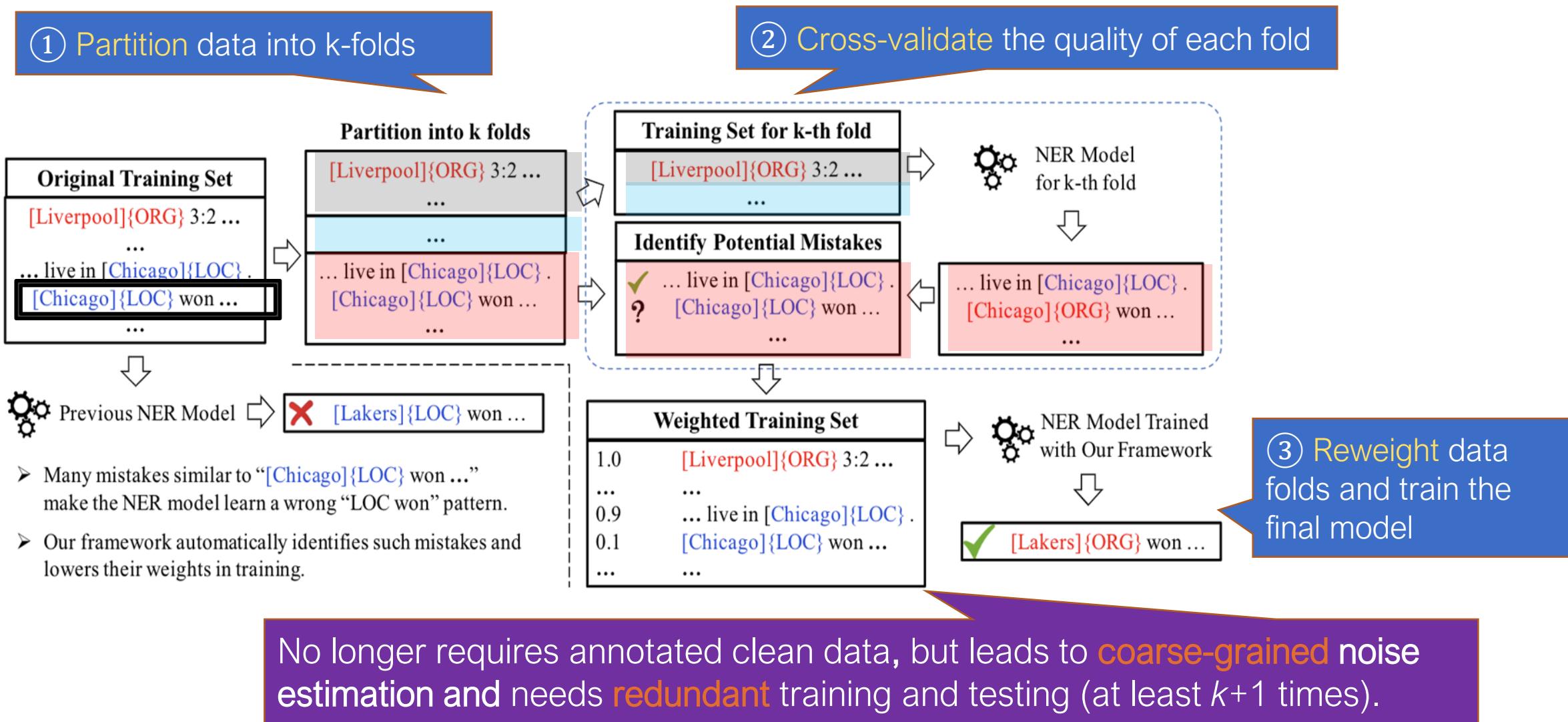


- ④ Cleaned (task) training data

Cost: manually labeling enough clean data is nowhere cheaper.

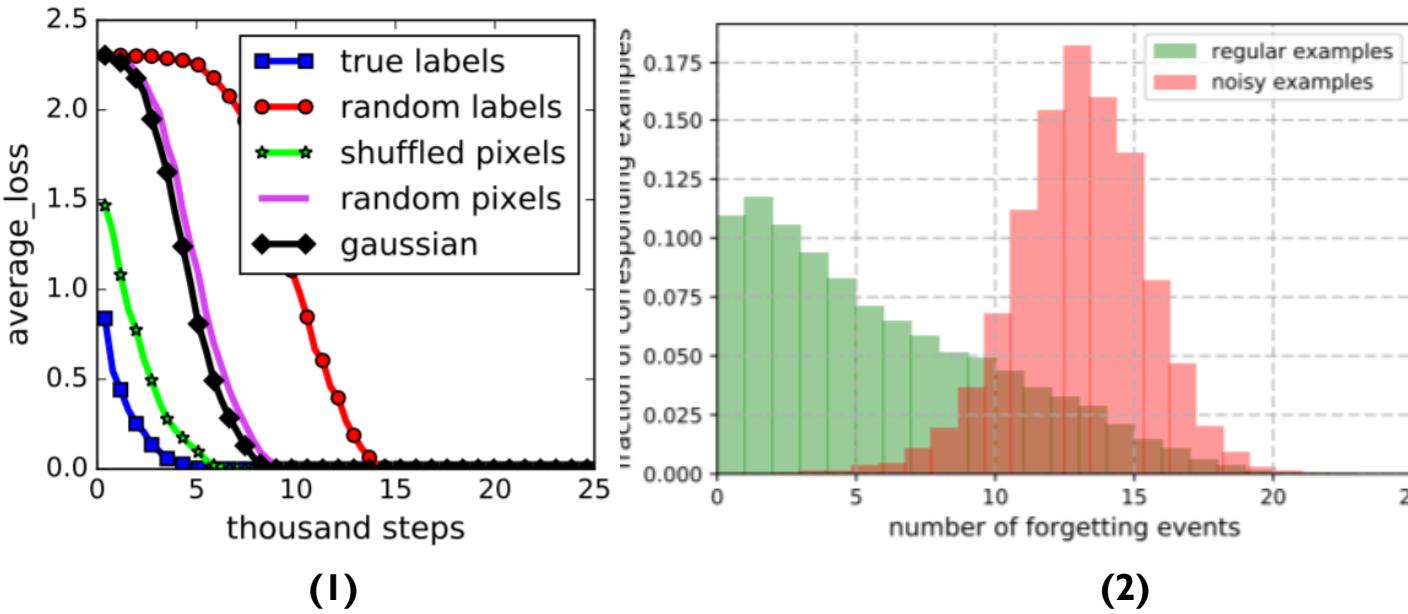


A Glance at Prior Solution: Unsupervised Denoising with Ensemble

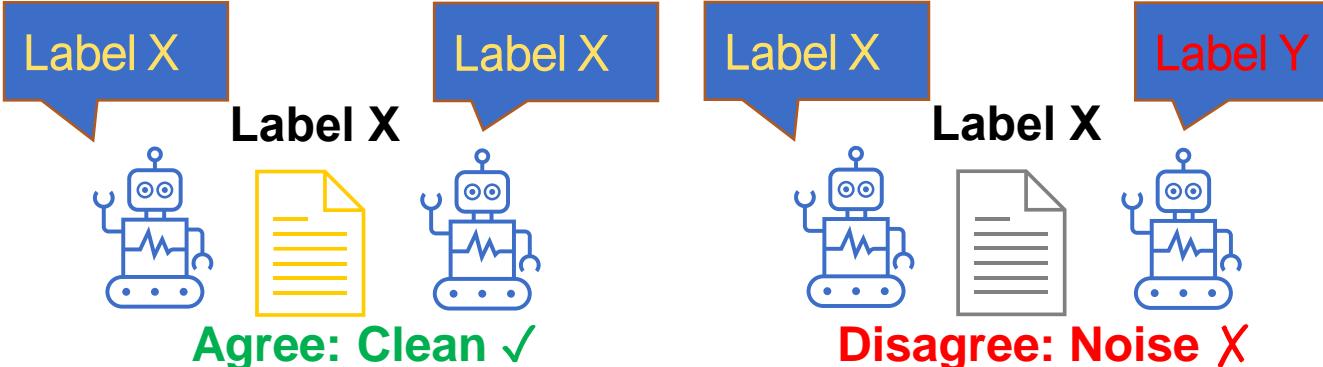




Unsupervised Denoising: Co-regularized Knowledge Distillation



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

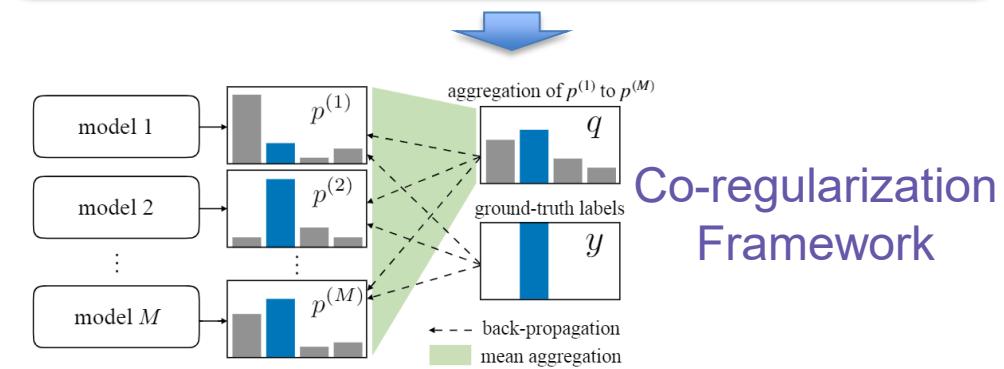


Mutual agreement by models indicates clean/noisy labels

Noisy labels are outliers to the task inductive bias.

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

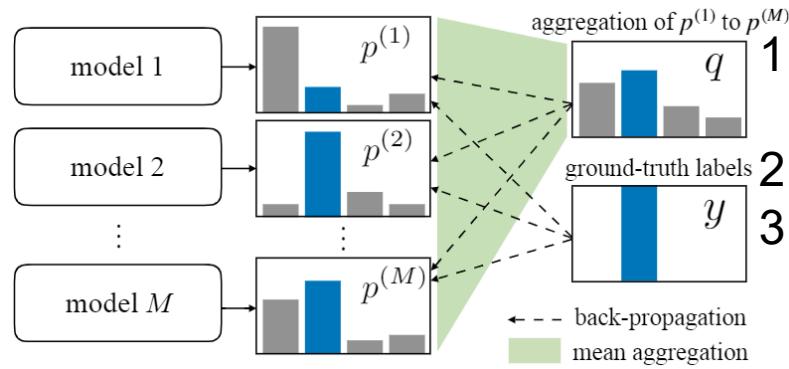
Model prediction is often inconsistent or oscillates on noisy labels in later epochs.



Co-regularization Framework



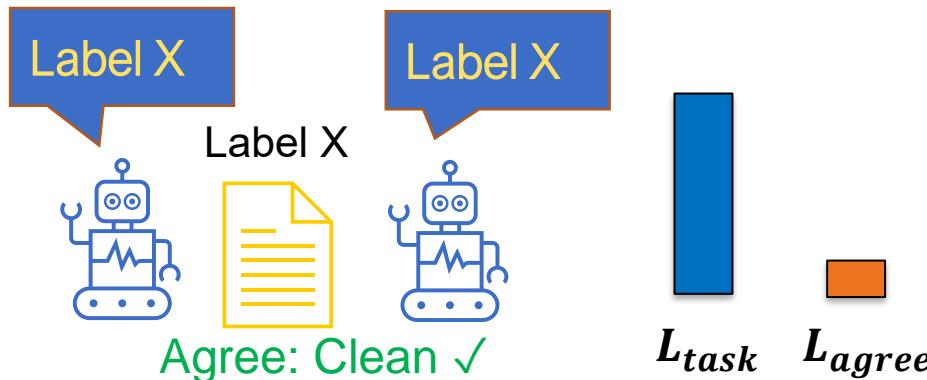
Unsupervised Denoising: Co-regularized Knowledge Distillation



1. Create $M (\geq 2)$; 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.

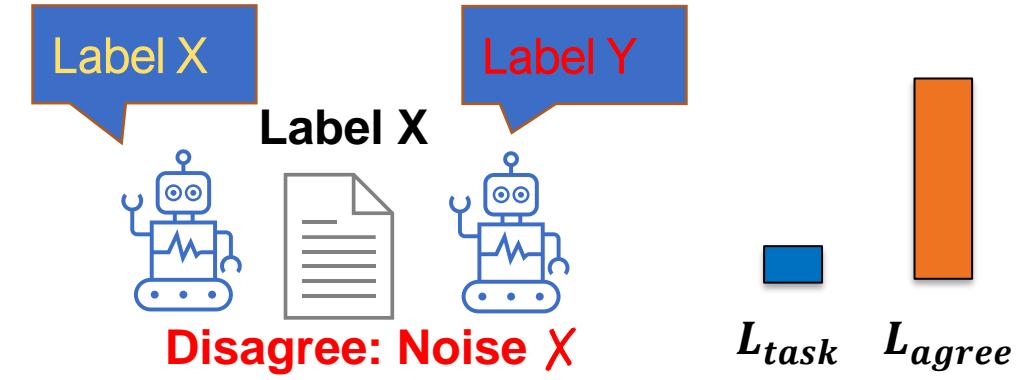
Cross-entropy L_{task}

K-L divergence between model predictions L_{agree}



On clean data

- Lower agreement loss
- Focusing on task optimization



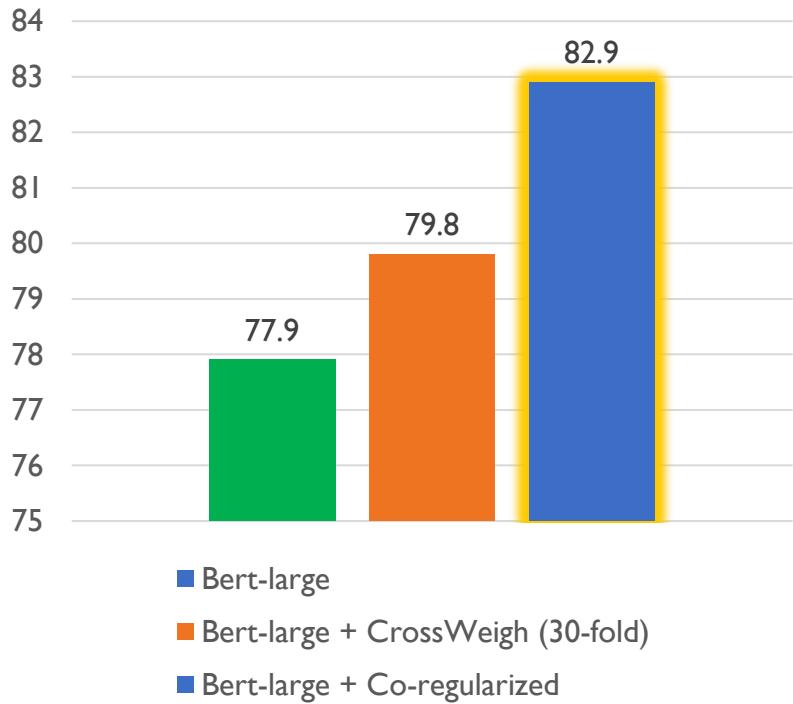
On noisy data

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

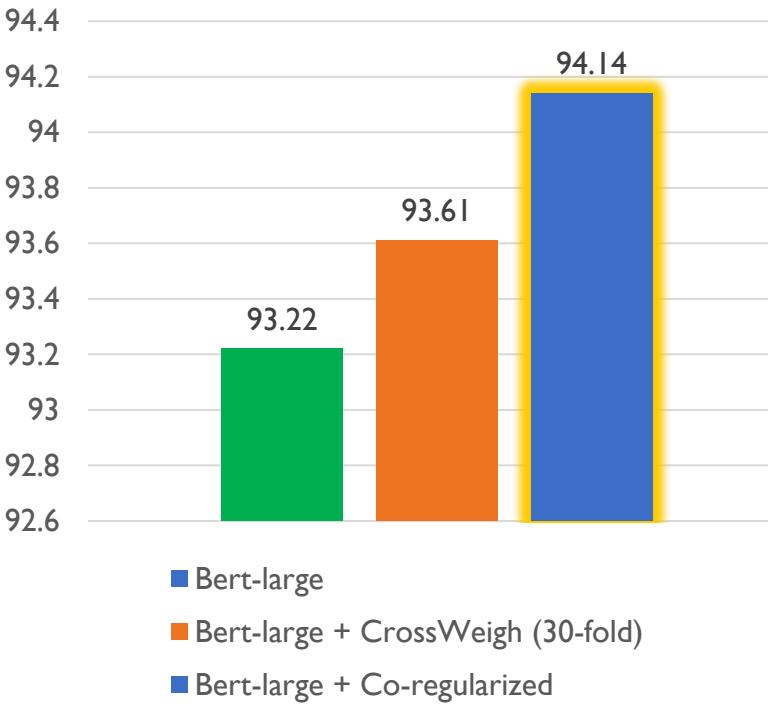
Unsupervised Denoising: Co-regularized Knowledge Distillation



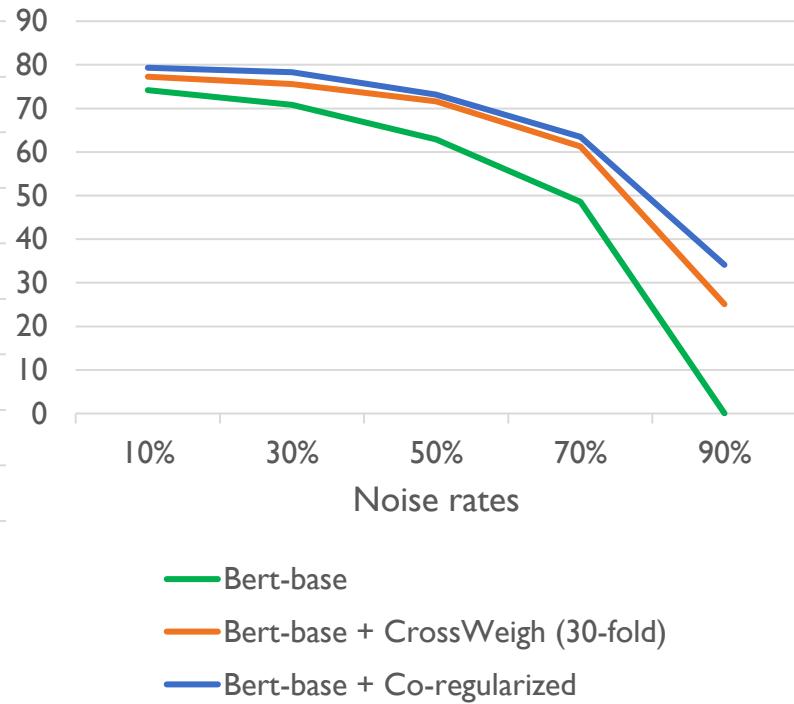
Relation Extraction (F1) on TACREV
(~8% training noise)



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



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

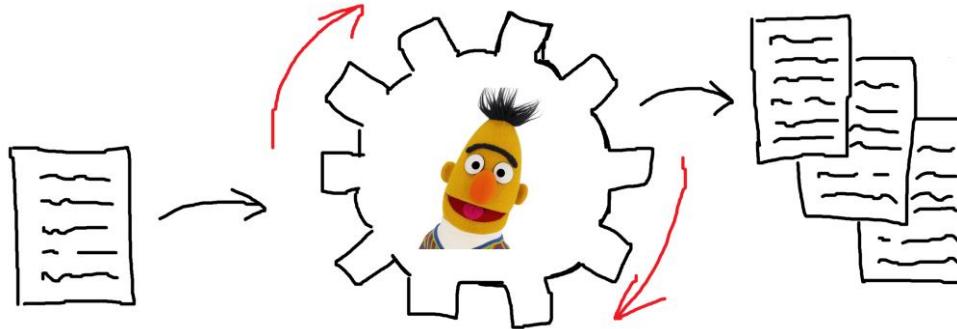


Merits of co-regularized knowledge distillation

- More robust than ensemble (e.g. CrossWeigh), especially when noise rates are higher
- More efficient (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)



Robust Data Augmentation

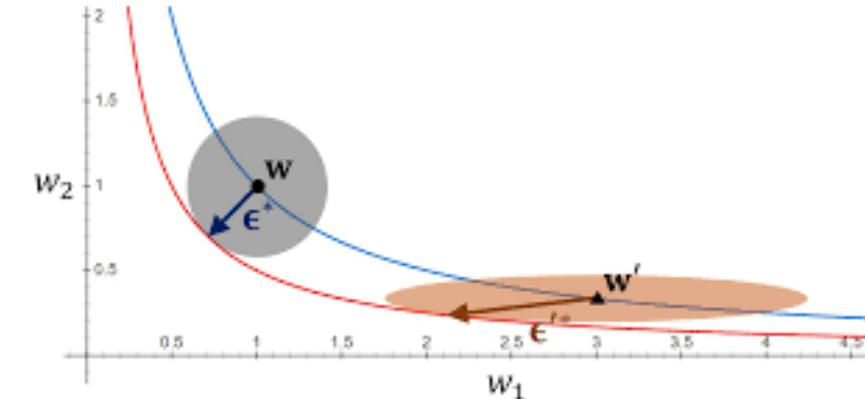


Denoising automatically augmented training data

- Long-term uncertainty measure based on moving gradient
- Agreement test between original data and augmentation

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

Perturbation Robustness



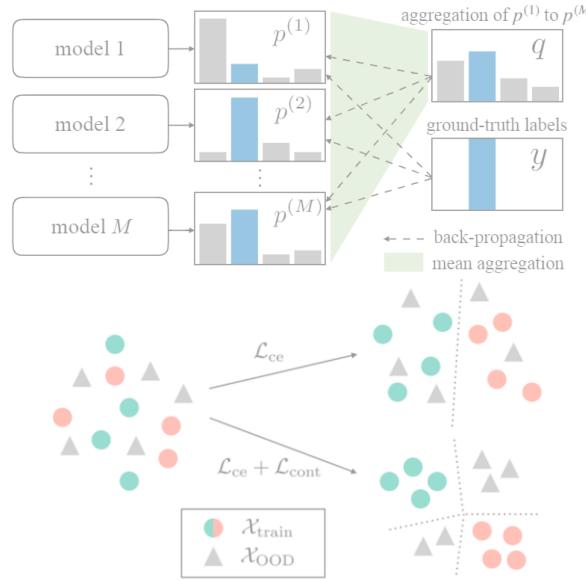
δ -SAM: efficient adversarial parameter perturbation

- 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

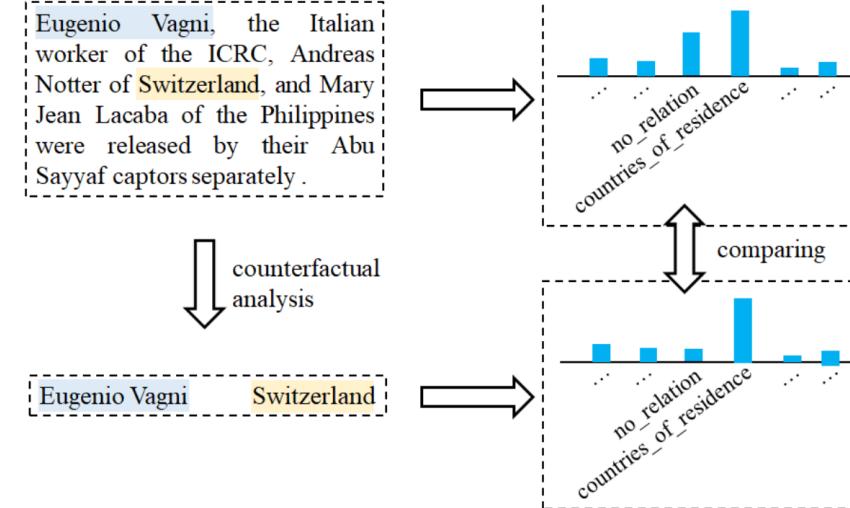
In This Talk



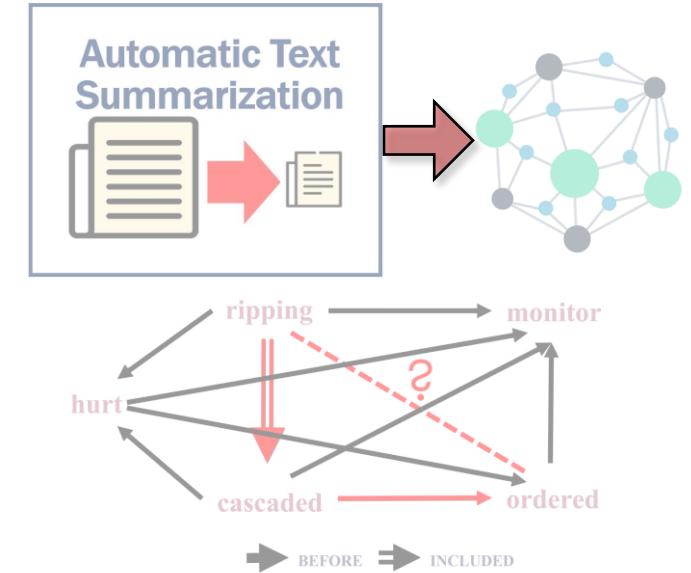
1. Noise-robust IE



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:



Event relation extraction:



Prior knowledge (in PLMs) can lead to biased extraction

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

Before? After?

```
graph LR; S1[see the doctor] --> S2[sick.]
```

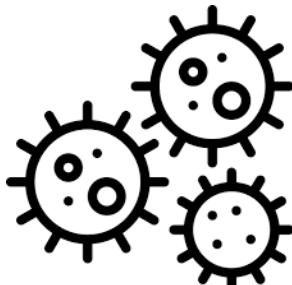


(Statistically) Biased training can lead to biased extraction

Why Faithful IE Is So Important



Drug-drug Interaction



Disease-target detection



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

Interaction; type: mechanism

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

Interaction; type: binding

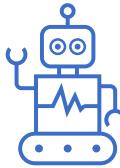
More risky tasks where we couldn't afford any guesses from unfaithful IE

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



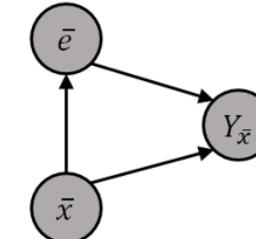
What is Causing the Problem: Take Relation Extraction as An Example

What we hope the IE model to do



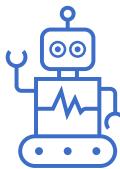
Bill Gates paid a visit to Building 99 of Microsoft yesterday.

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



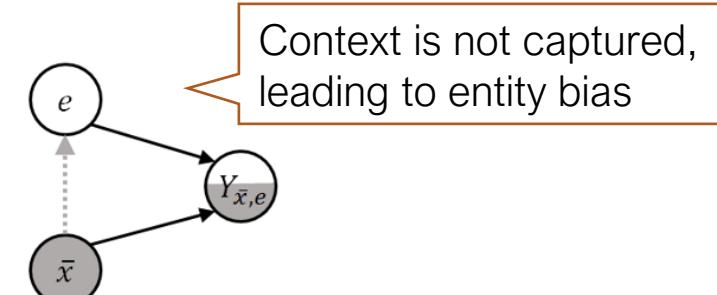
Relations should be inferred based on both mentions and the context

What it may actually do



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

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



Overly relying on entity mentions lead to a shortcut for RE

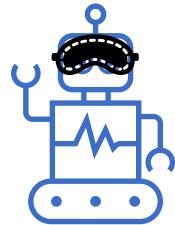
How do we mitigate this **spurious correlation?**



Mention masks: mask out entity names with their types

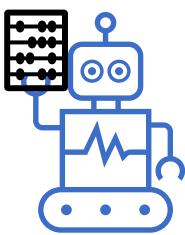
Person paid a visit to Building 99 of Org yesterday.

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



Mask mentions in both training and inference

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



Reweighting instances: FoCal loss, resampling, two-stage optimization, etc.

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

Upweight hard instances

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

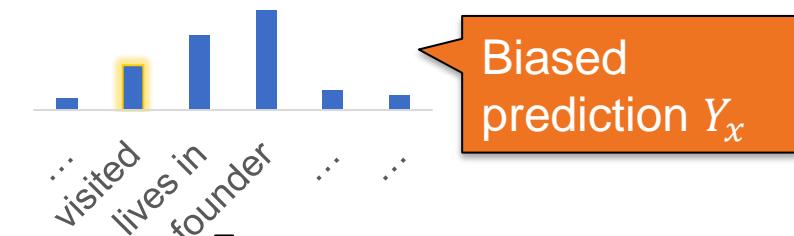


Our Strategy: Counterfactual Inference

Measure the biases using counterfactual instances, then deduct the biases

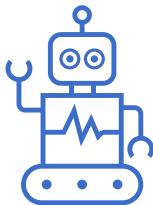
① Original Instance (x)

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



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

Bill Gates Microsoft



③ Empty counterfactual instance (\bar{x})

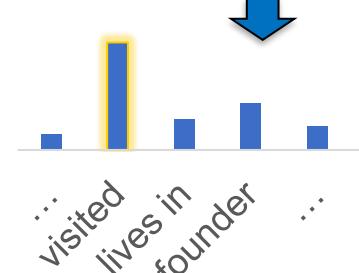
\emptyset



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

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

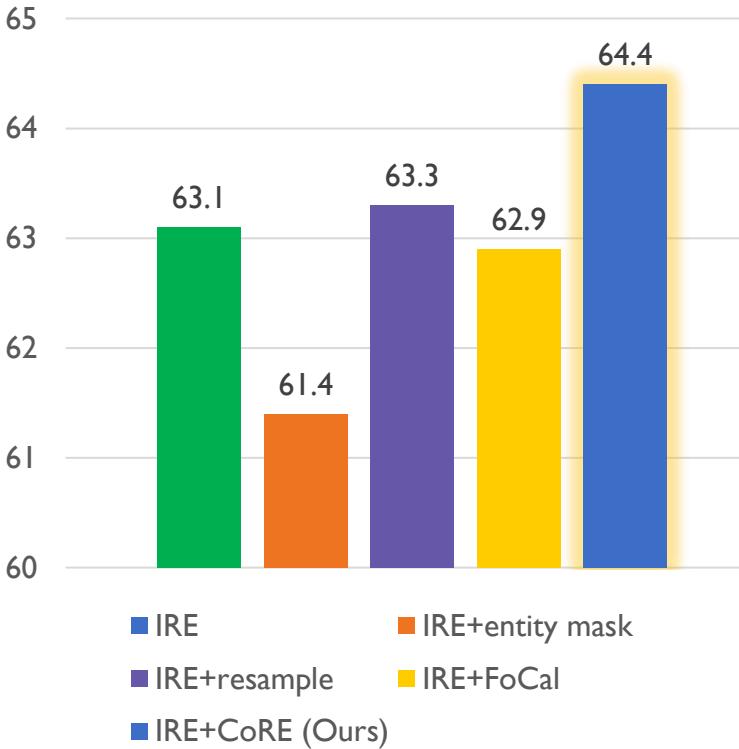
Obtained
on dev set



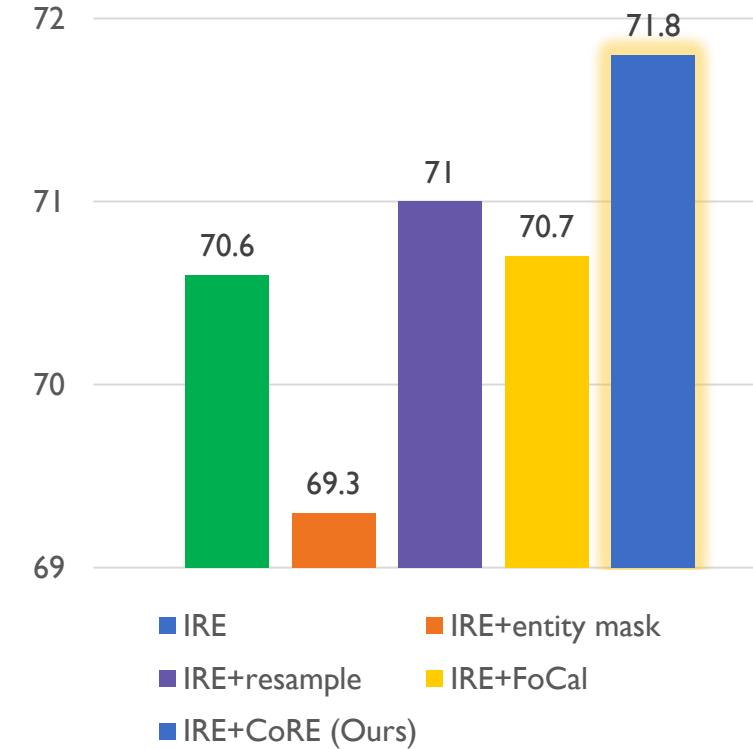
Counterfactual Inference



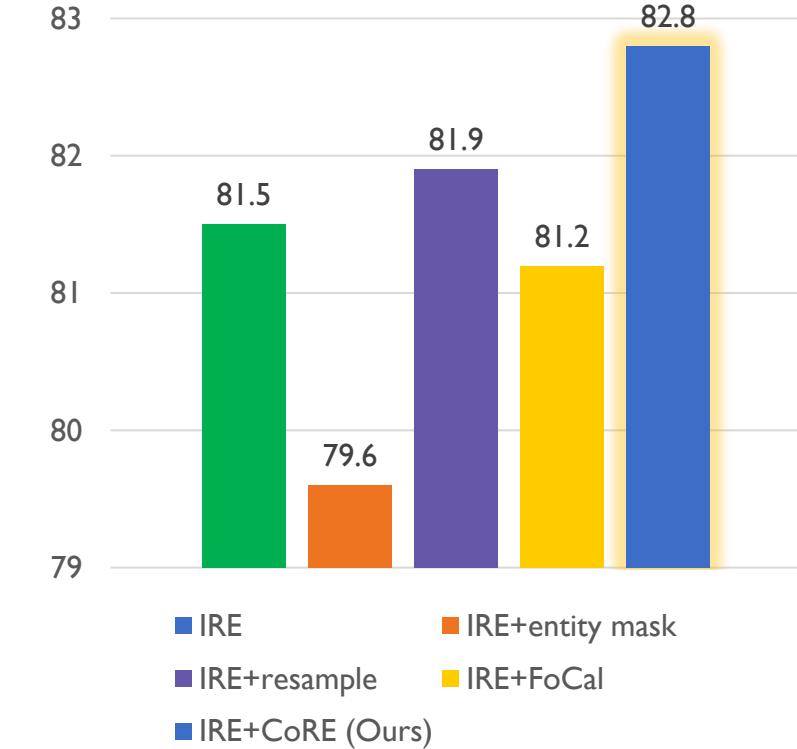
Fl-macro on TACRED



Fl-macro on TACREV



Fl-macro on Re-TACRED



Counterfactual inference leading to more precise and fairer relation extraction.

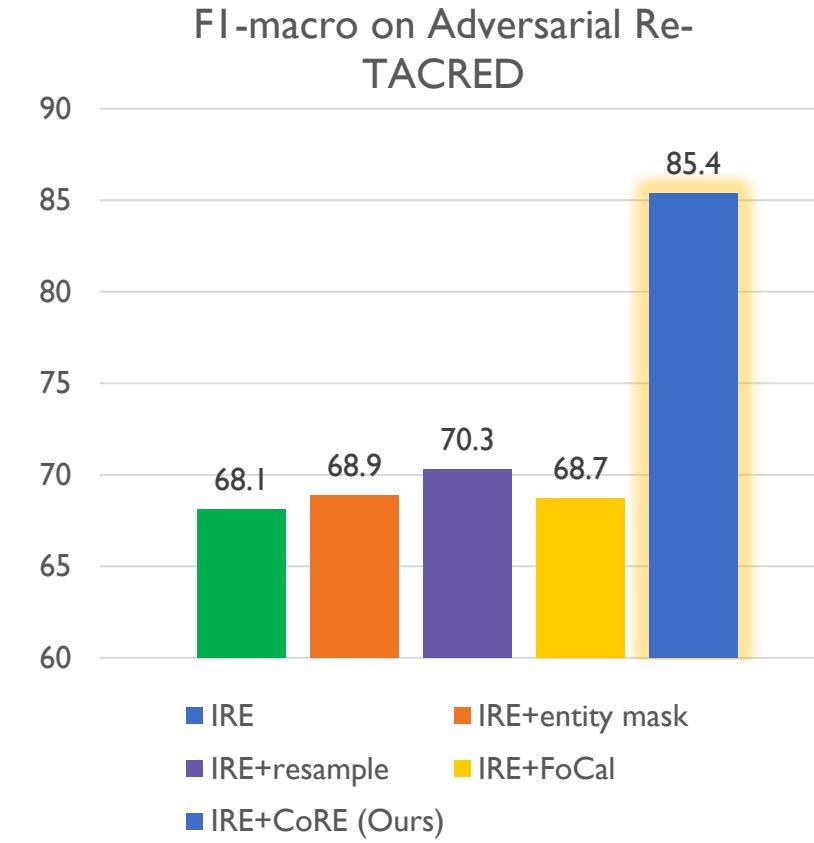
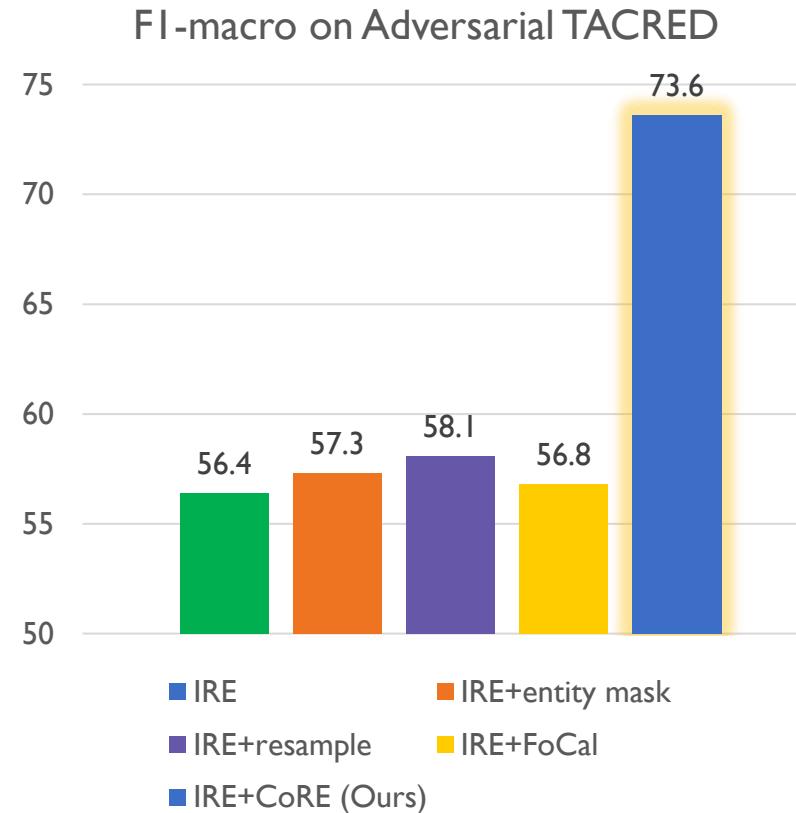
*IRE_{RoBERTa} is one of the best-performing sentence-level RE models (Zhou and Chen 2021). Results are also available for LUKE.

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.



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

XWLDC. 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

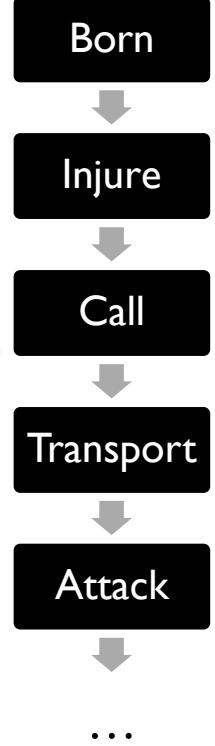
Police also want to speak to Jackson's doctor who witnessed his collapse

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** .

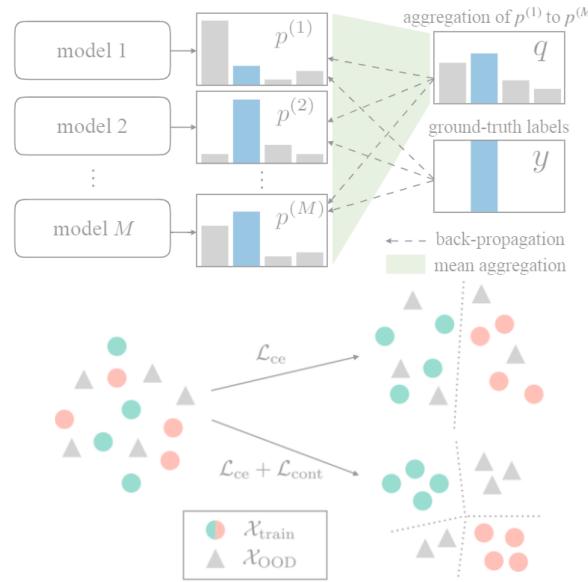


WZDGR C. Extracting or Guessing? Improving Faithfulness in Event Temporal Relation Extraction. 2022

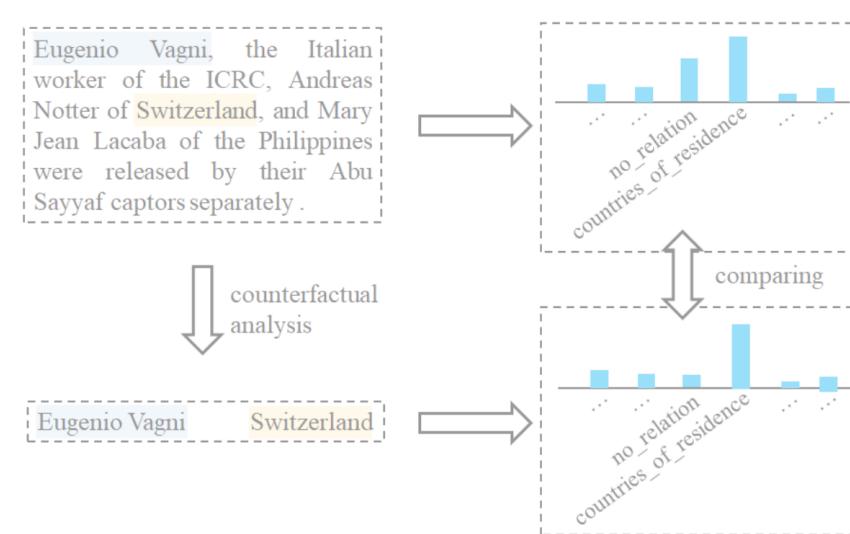
In This Talk



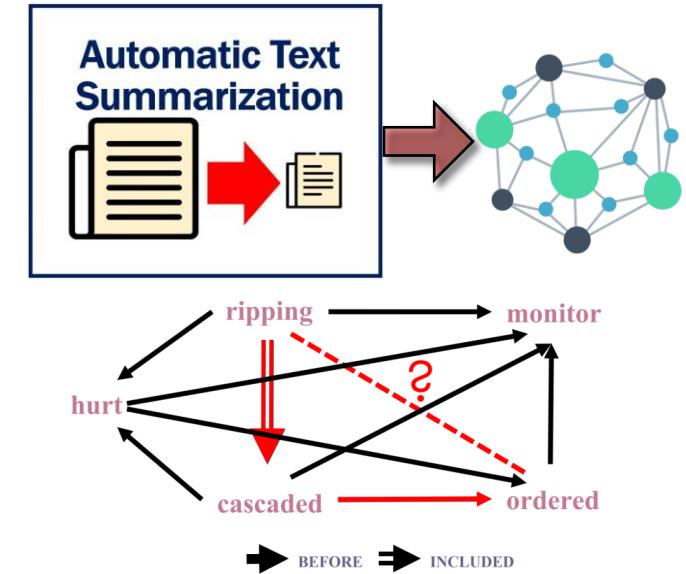
1. Noise-robust IE



2. Faithful IE



3. Indirectly Supervised IE



4. Future Directions



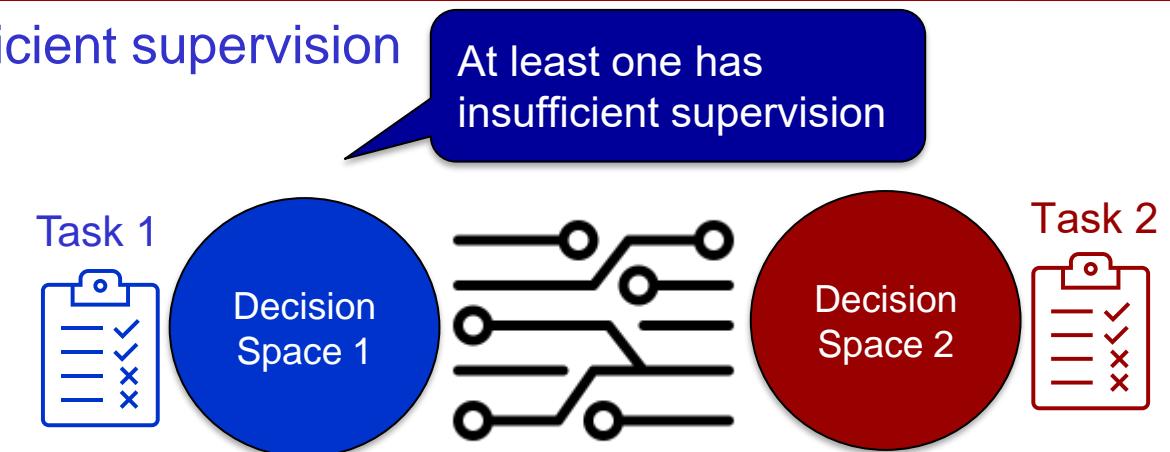


Two Forms of Indirect Supervision



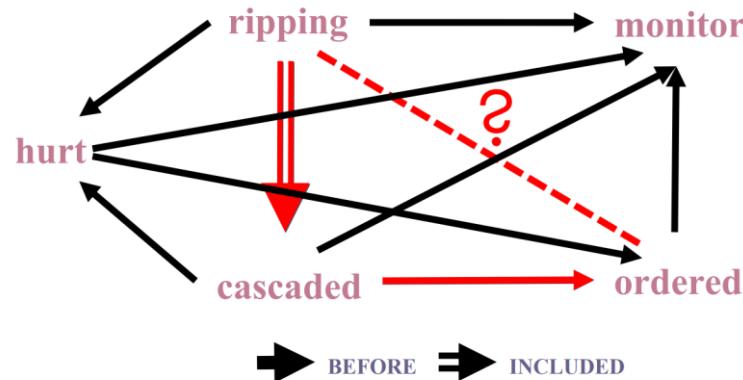
Direct annotation is difficult and expensive

IE suffers from insufficient supervision

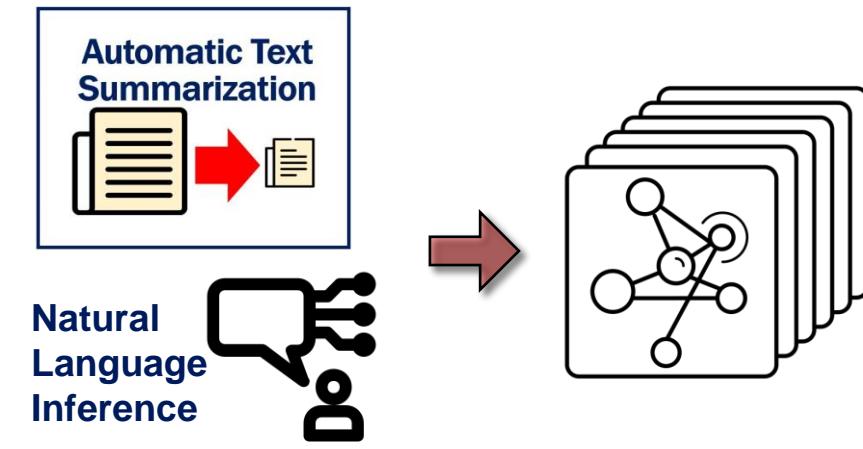


Can we bridge the supervision of different tasks?

Two Forms of Indirect supervision



(Logically) Constrained Learning



Cross-task Transfer

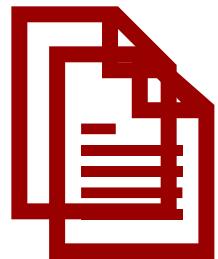
Constrained Learning: Bridging Learning Resources with Logical Constraints



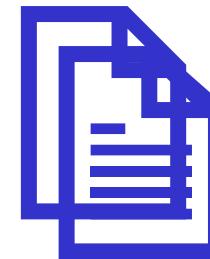
Take Event Relation Extraction as an Example

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

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

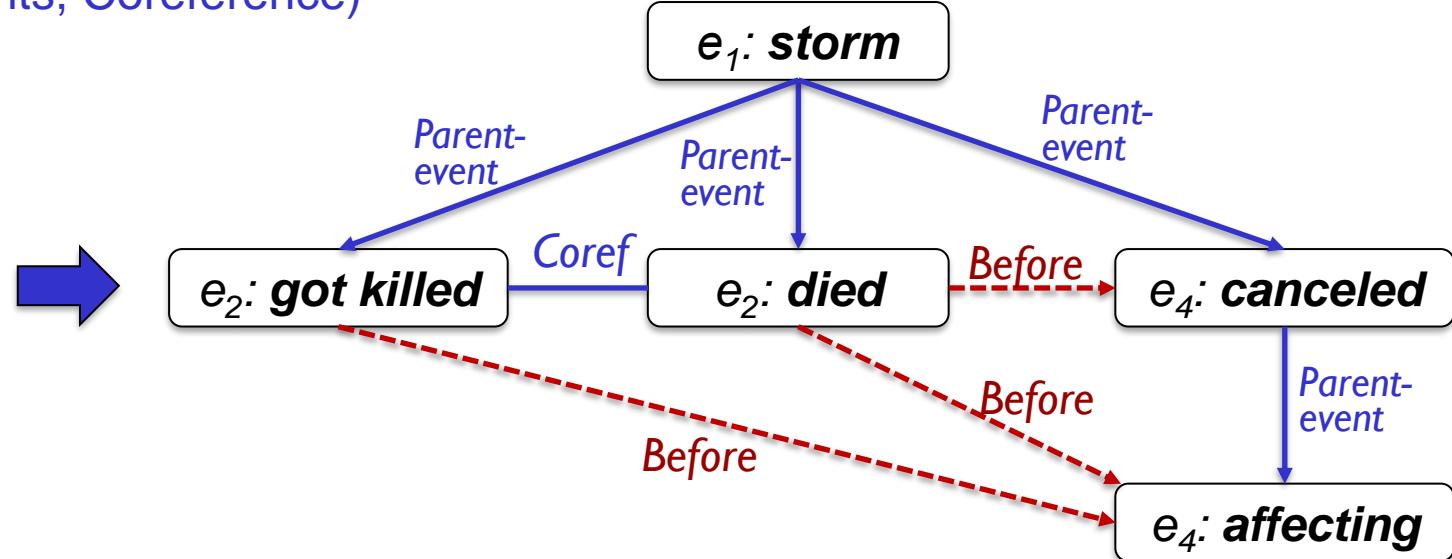


TempRel Corpora
(MATRES, TB-Dense, etc.)



Membership Corpora
(HiEve, ECB+, etc.)

Could we connect these supervision data?



Implication

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

Conjunction

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

Use logical constraints!



Logical Constraints Of Relations

Symmetry

$e3:died$ is BEFORE $e4:canceled$
 $\Rightarrow e4:canceled$ is AFTER $e3:died$

Conjunction

$e3:died$ is BEFORE $e4:canceled$
 $\wedge e4:canceled$ is a PARENT of $e5:affected$
 $\Rightarrow e3:died$ BEFORE $e5:affected$

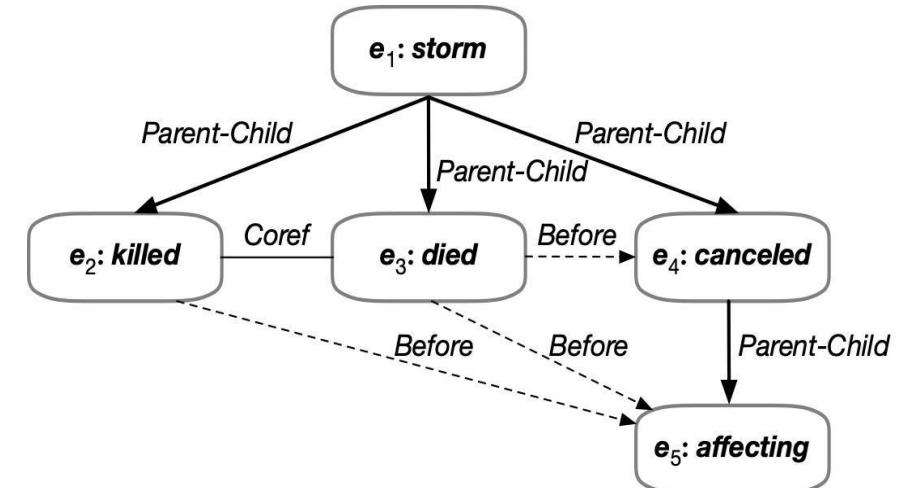
(we also consider **Implication** and **Negation**)

Goal: letting the neural model capture the logical constraints.

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

Transitivity

$e1:storm$ is PARENT of $e4:canceled$
 $\wedge e4:canceled$ is a PARENT of $e5:affected$
 $\Rightarrow e1:storm$ is a PARENT of $e5:affected$





Incorporating Logical Constraints in A Neural Architecture

Using product t -norm model constraints as differentiable functions

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

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

Constraints become entropy regularizers

$\alpha \backslash \beta$	PC	CP	CR	NR	BF	AF	EQ	VG
α	PC, \neg AF	—	PC, \neg AF	\neg CP, \neg CR	BF, \neg CP, \neg CR	—	BF, \neg CP, \neg CR	—
PC	—	CP, \neg BF	CP, \neg BF	\neg PC, \neg CR	—	AF, \neg PC, \neg CR	AF, \neg PC, \neg CR	—
CP	PC, \neg AF	CP, \neg BF	CR, EQ	NR	BF, \neg CP, \neg CR	AF, \neg PC, \neg CR	EQ	VG
CR	\neg CP, \neg CR	\neg PC, \neg CR	NR	—	—	—	—	—
NR	BF, \neg CP, \neg CR	—	BF, \neg CP, \neg CR	—	BF, \neg CP, \neg CR	—	BF, \neg CP, \neg CR	\neg AF, \neg EQ
BF	—	AF, \neg PC, \neg CR	AF, \neg PC, \neg CR	—	—	AF, \neg PC, \neg CR	AF, \neg PC, \neg CR	\neg BF, \neg EQ
AF	—	—	—	—	—	AF, \neg PC, \neg CR	AF, \neg PC, \neg CR	\neg BF, \neg EQ
EQ	\neg AF	\neg BF	EQ	—	BF, \neg CP, \neg CR	AF, \neg PC, \neg CR	EQ	VG, \neg CR
VG	—	—	VG, \neg CR	—	\neg AF, \neg EQ	\neg BF, \neg EQ	VG	—



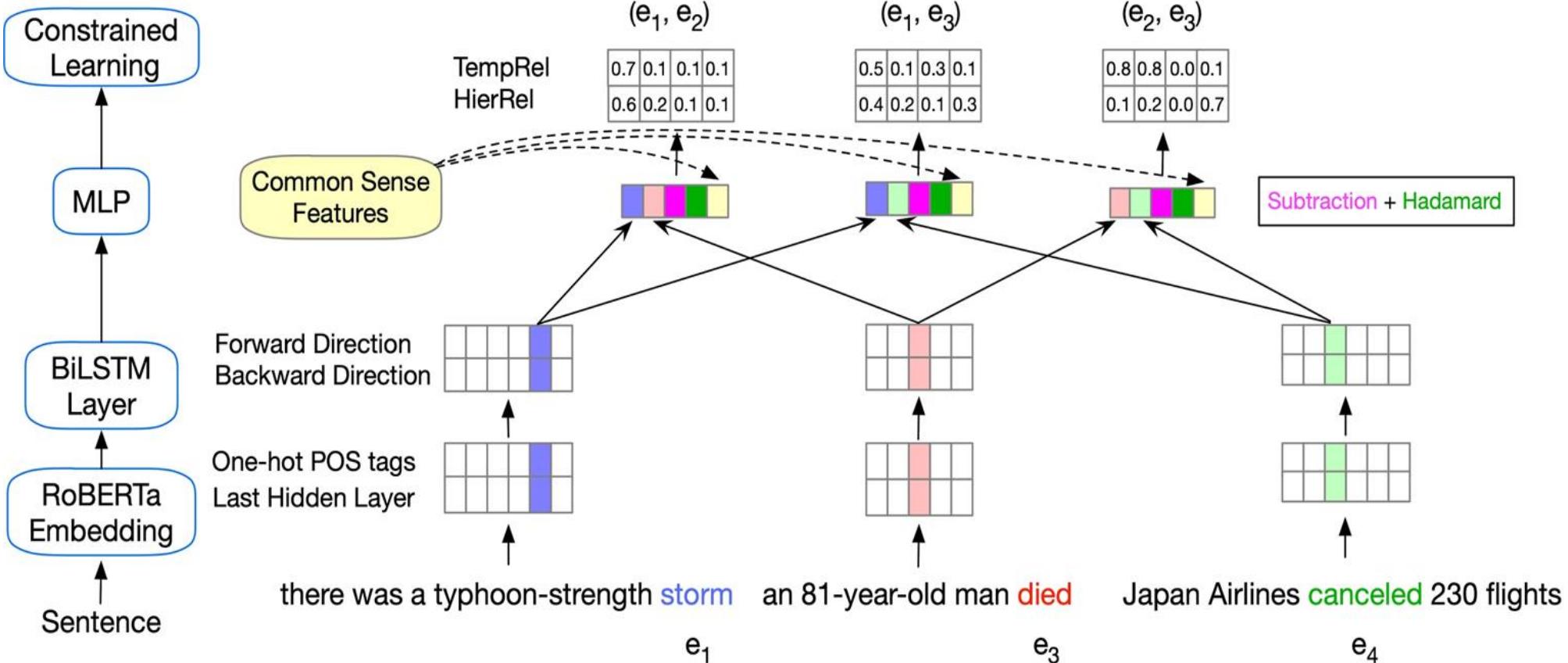
Joint Constrained Learning

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

Task loss

Implication and conjunction constraint losses

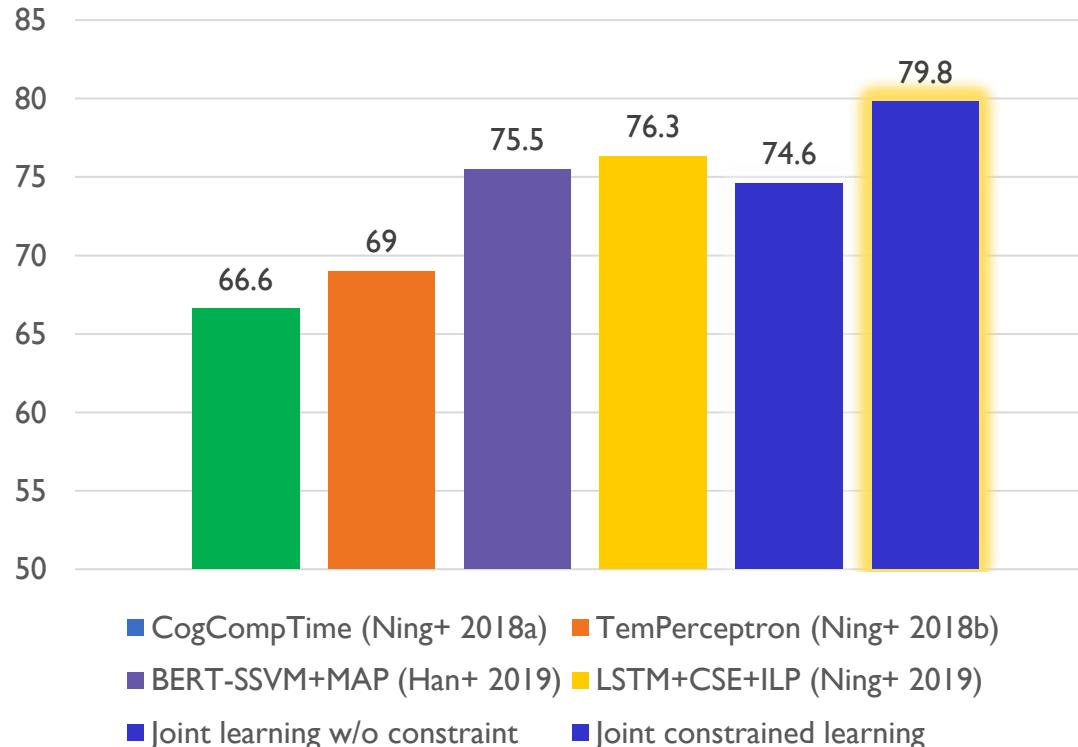
$$\text{Loss Function: } L = L_A + \lambda_S L_S + \lambda_C L_C$$



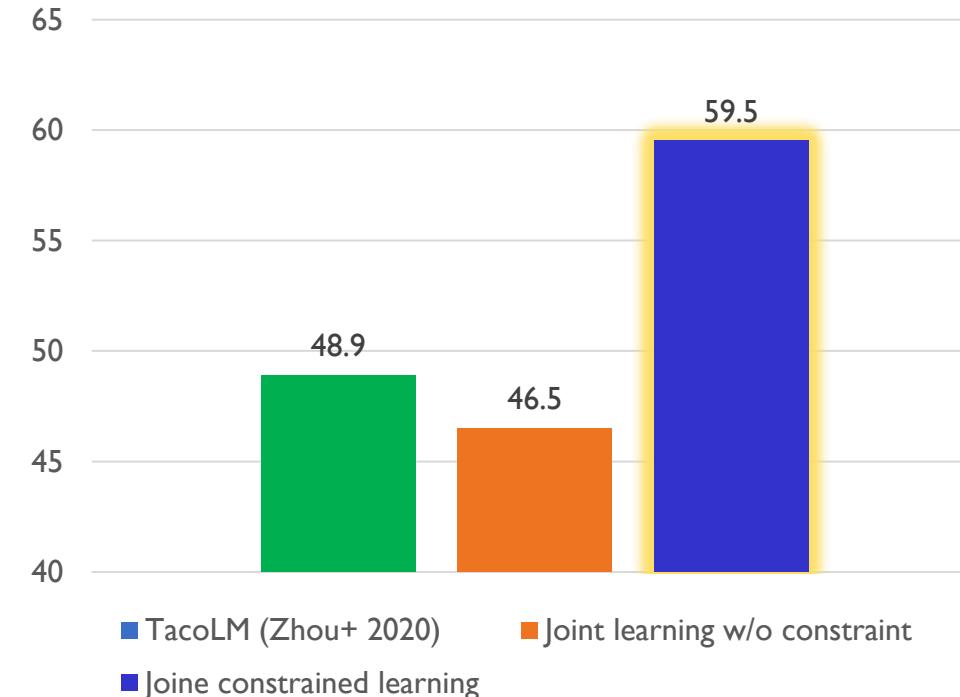
The Joint Constrained Learning Architecture



F1 on MATRES for TempRel Extraction



F1 on HiEve for Membership (Subevents and coref) Extraction



Key Observations

- Constraints are a natural bridge for learning resources with different sets of relations
- Adding constraints in learning is sufficient to enforce logical consistency of outputs, surpassing ILP in inference (w/ constrained learning) by 2.6-12.3% in ACC

Automatically Learning Constraints



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

Event-event relations are related to narrative 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?

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

Constraint Learning

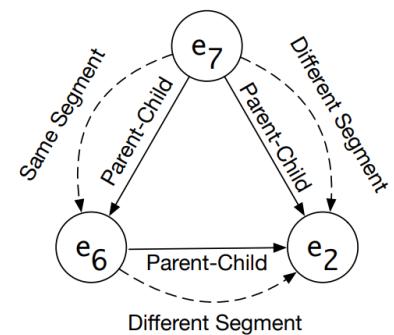
Training a single-layer rectifier network on all ``triangles'' of the training data

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

Estimates probabilities of conjunctive constraints

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

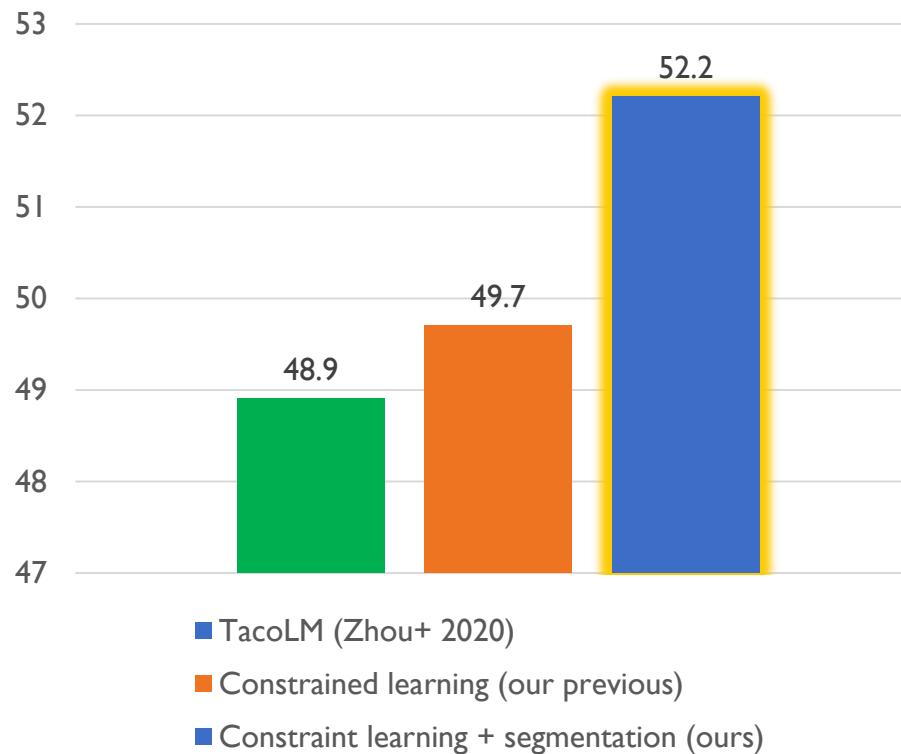
$$L_{cons} = -\log \left(\text{Sigmoid} \left(1 - \sum_{k=1}^N \text{ReLU}(\mathbf{w}_k \cdot \psi + b_k) \right) \right)$$



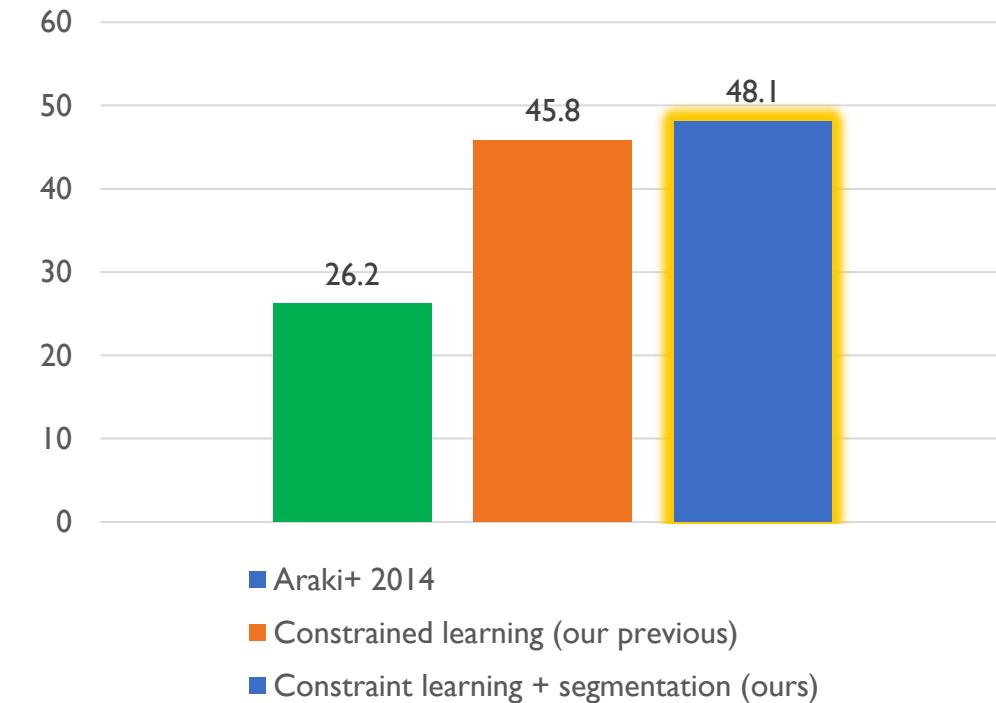
Automatically Learning Constraints



Subevent detection (F1) on HiEve



Subevent detection (F1) on Intelligence Community



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



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.

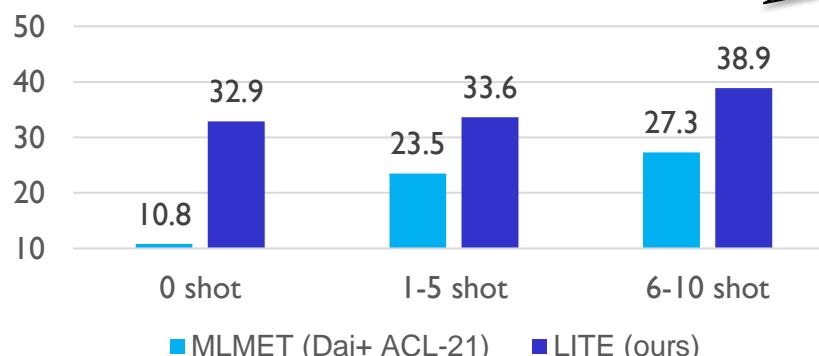


film, art, movie, show, entertainment, creation

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.



Reformulating as NLI + learning to rank.

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

Premise

Chrysler is a **company**.

Hypotheses by **TRUE** labels

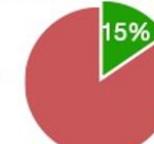


91%

Confidence given the premise (entailment score)

Chrysler is a **sea-bird**.

Hypotheses by **FALSE** labels

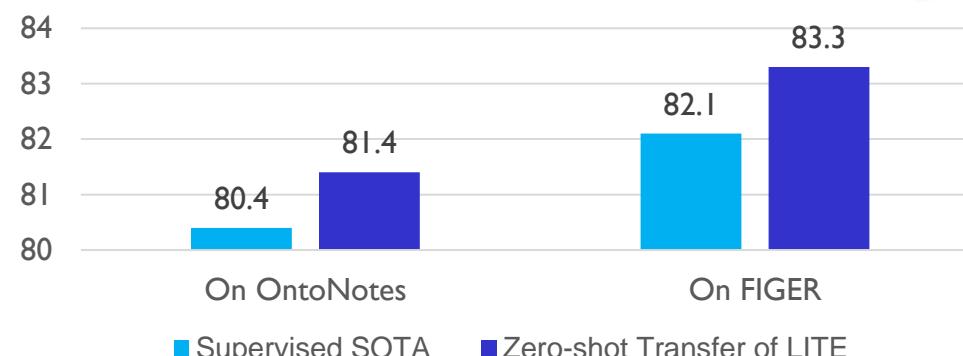


15%

Rank over

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

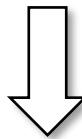
Outperforming full-shot SOTA with 0-shot transfer.



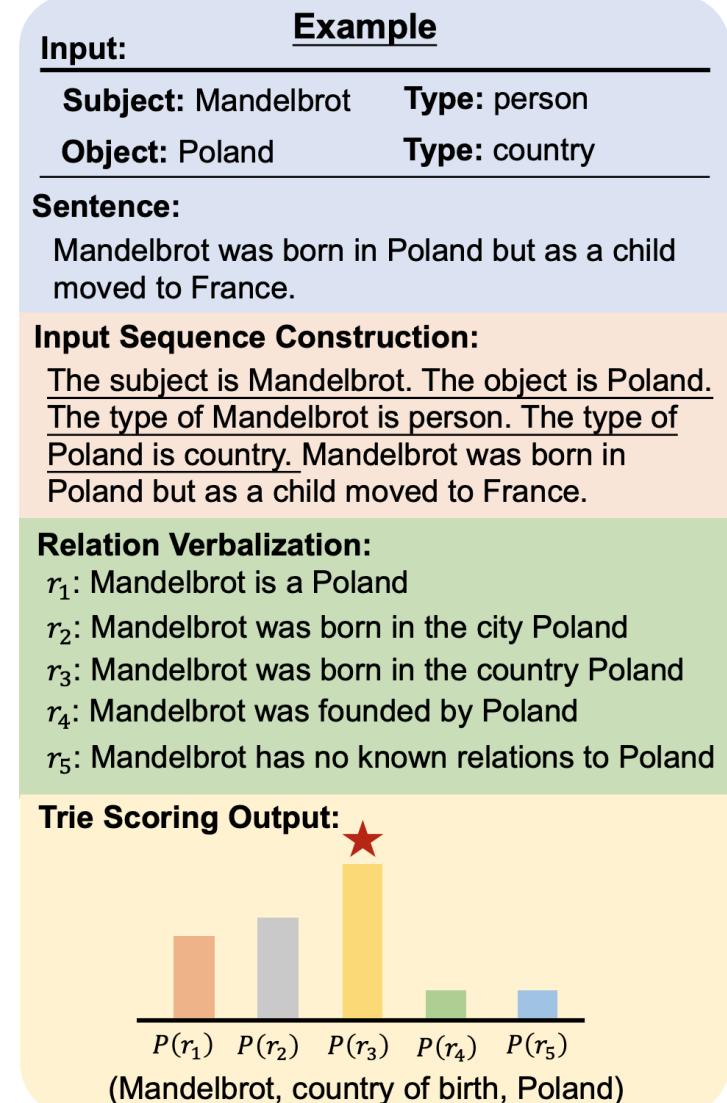
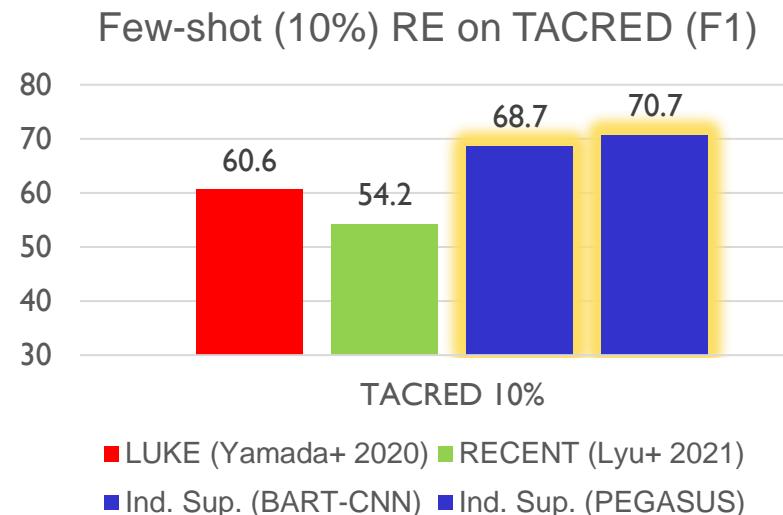
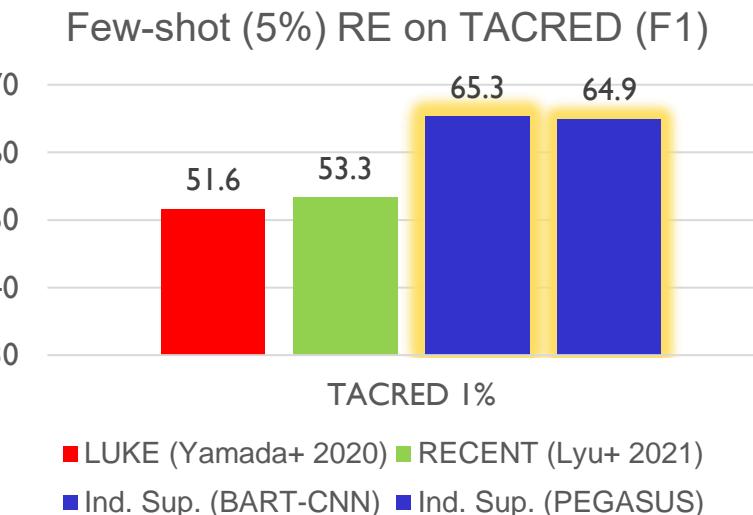
Indirect Supervision with Summarization



- 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

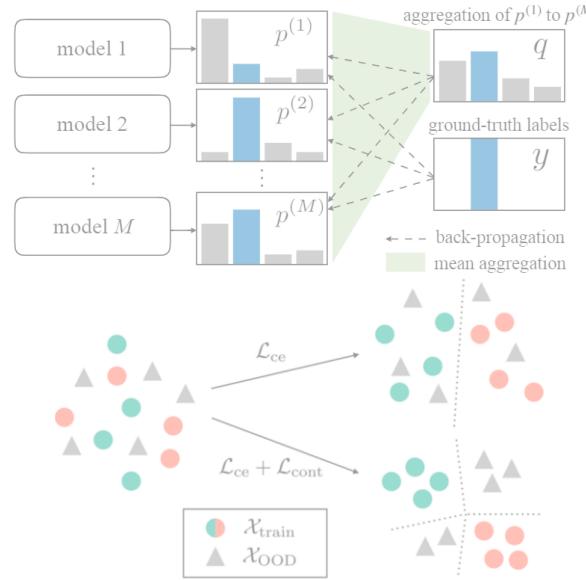


Indirect supervision from abstractive summarization corpora (CNN/Dailymail, Xsum) leads to generalizable RE

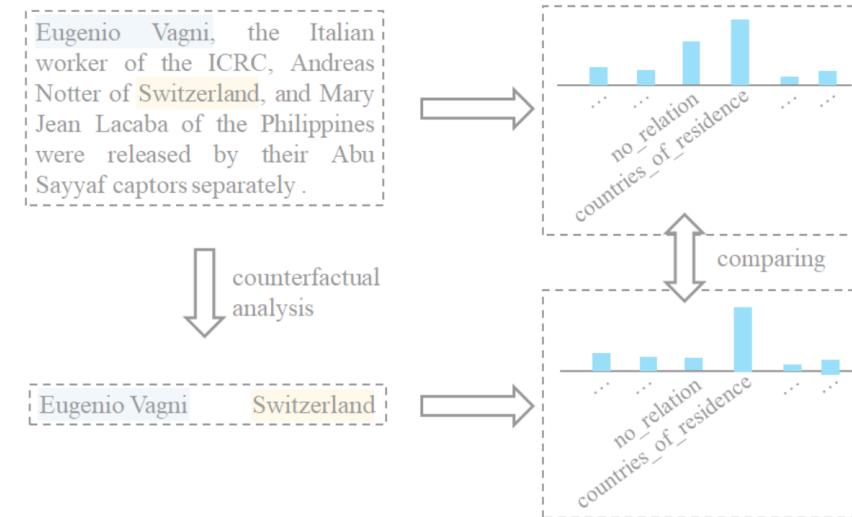




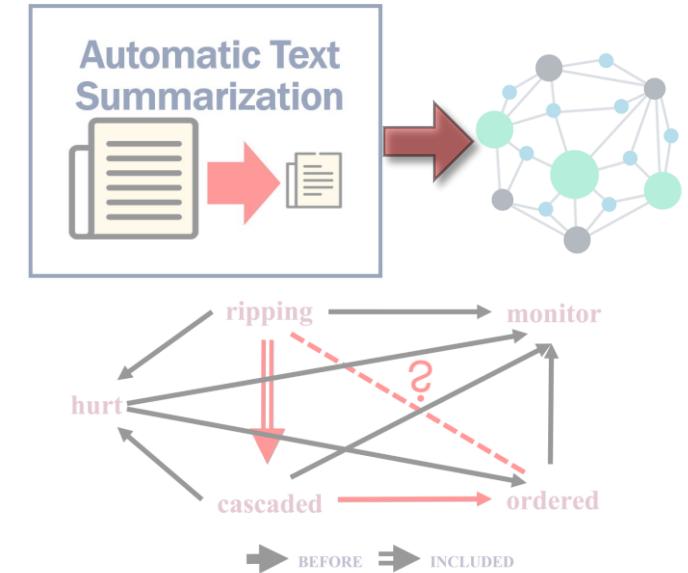
1. Noise-robust IE



2. Faithful IE



3. Indirectly Supervised IE



4. Future Directions

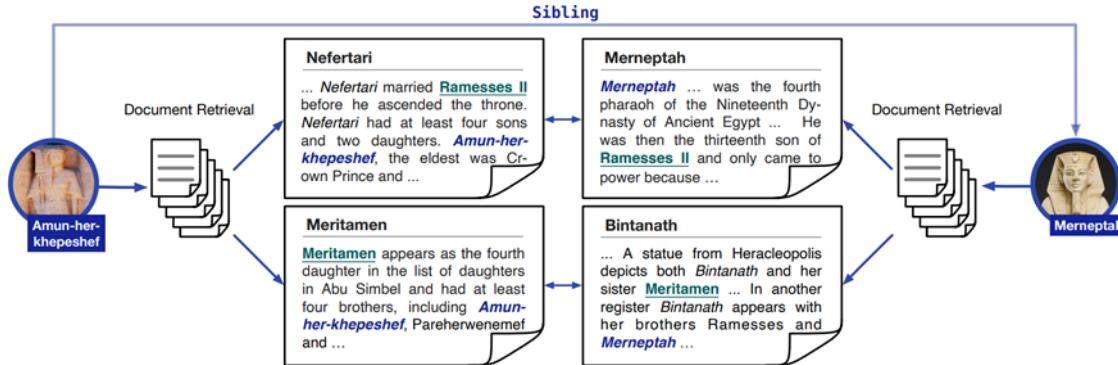




Thinking Across Documents

~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



Novel

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

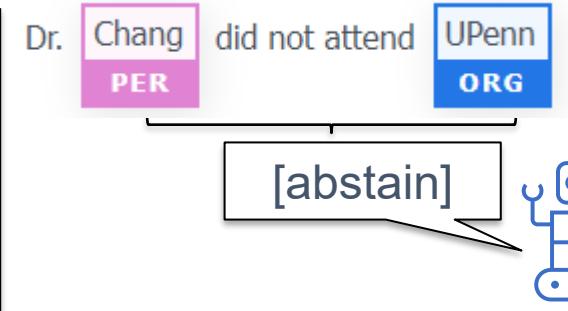
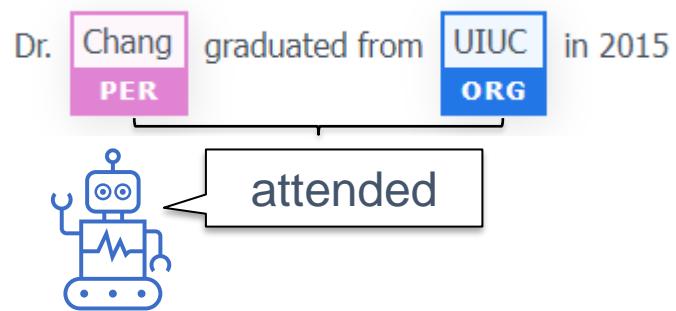
- Multi-hop reasoning
- Consolidation
- Tracking information pollution
- Long-form document modeling
- Mitigating frequency biases
- ...

👉 A long way to go



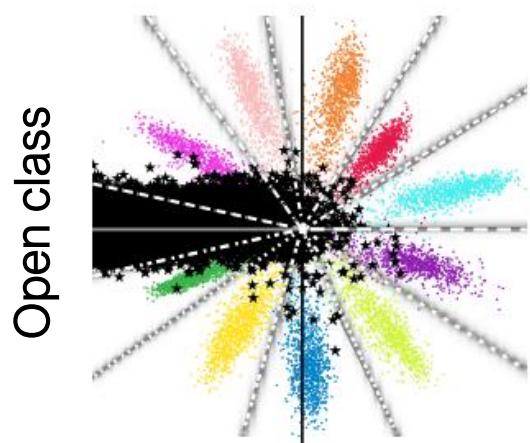
Selective Extraction

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 make inference more **selective**?



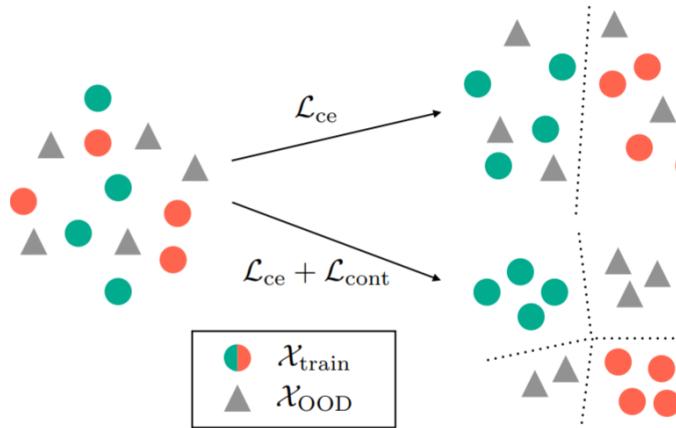
- A supervised approach can be a choice
- Classify exceptions into an open class/background set
 - However, exceptions can never be close to exhaustive in training data
 - Task training data
 - Annotated exceptions



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 \Rightarrow Better OOD detection

Creating compact representations with (margin-based) contrastive learning

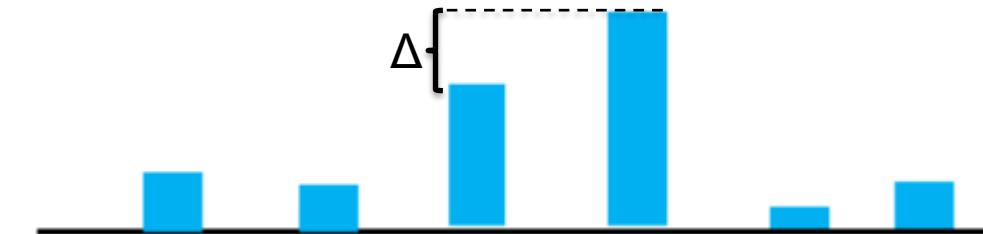
- Indirectly making OOD instances as “background” representation

Inference with Mahalanobis distance

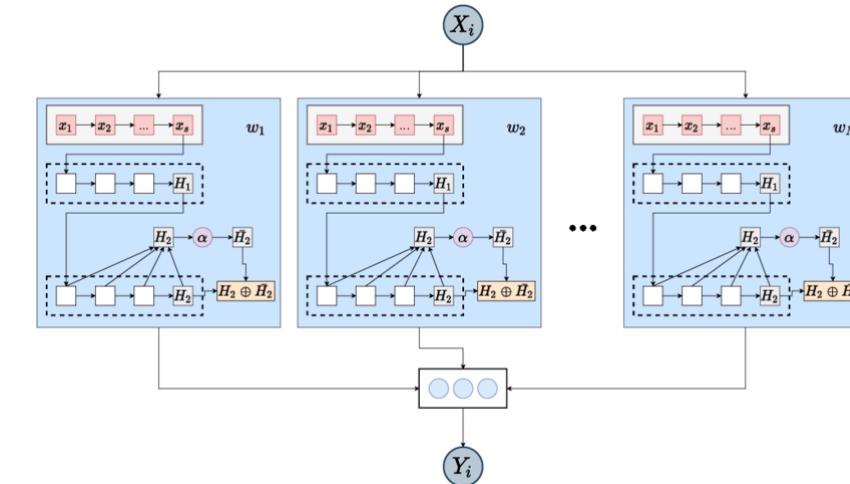
- High-order distance measures improve OOD detection

Estimating the uncertainty of prediction

Softmax response: difference between top two class predictions



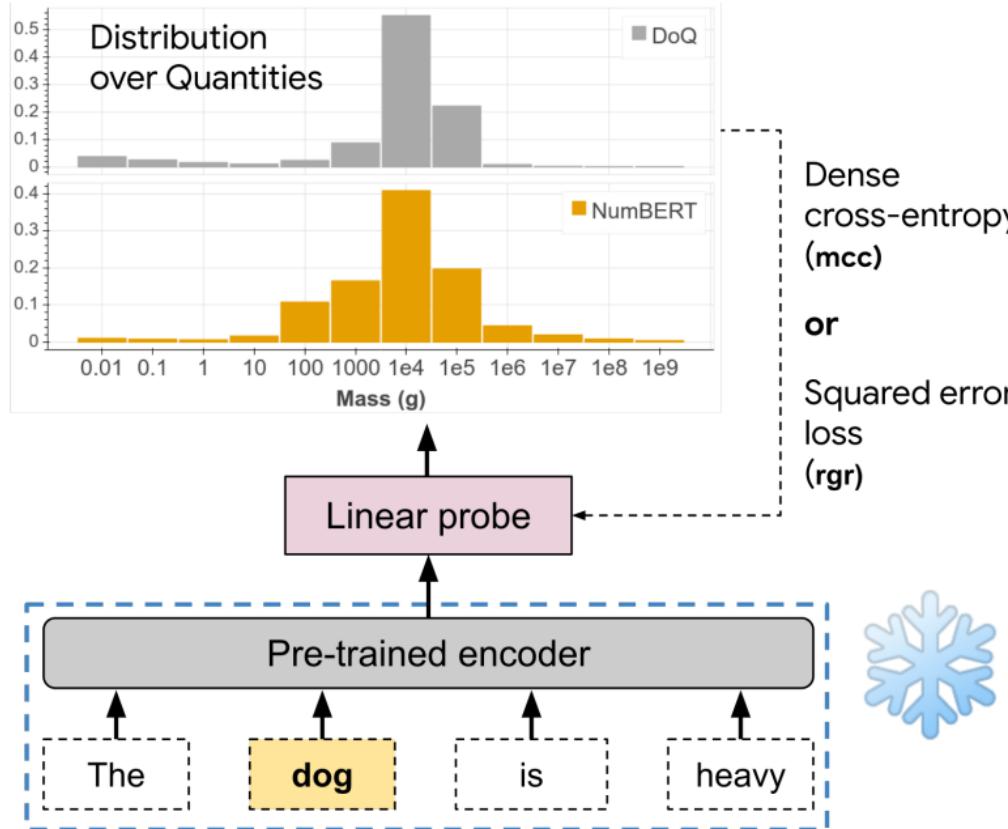
Prediction variance in Monte-Carlo dropout





Quantitative Extraction

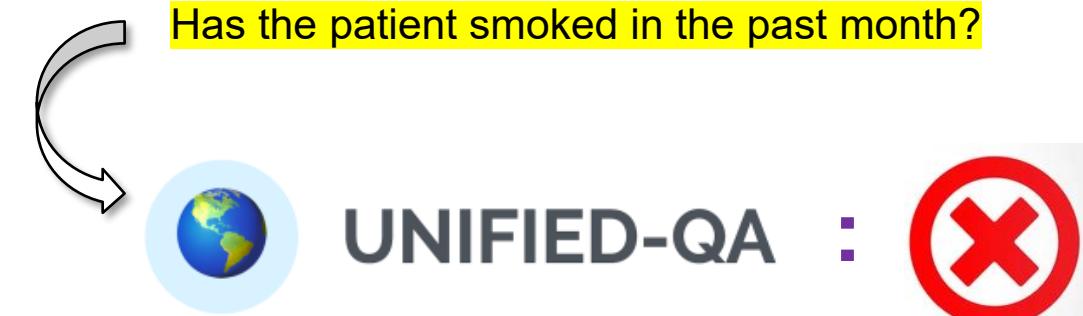
Extracting quantities



Temporal verification

Medical Reports

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

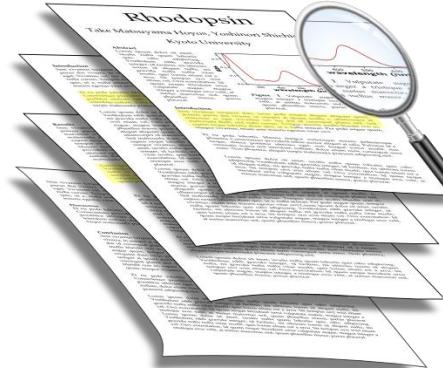


Large models still do not support quantitative reasoning well

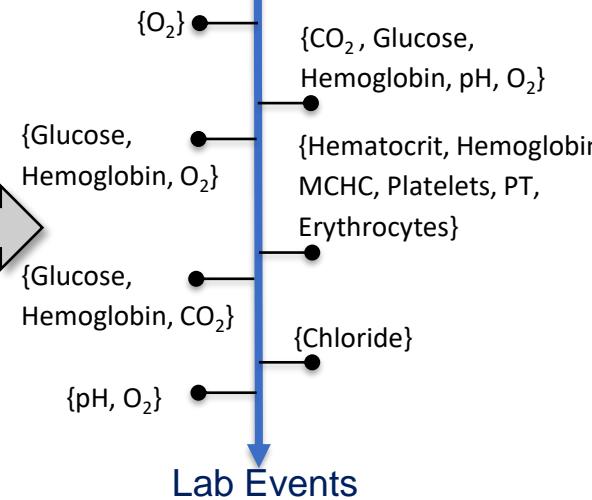
IE for the Common Good



Medicine and healthcare



Drug-drug Interaction



Programming education

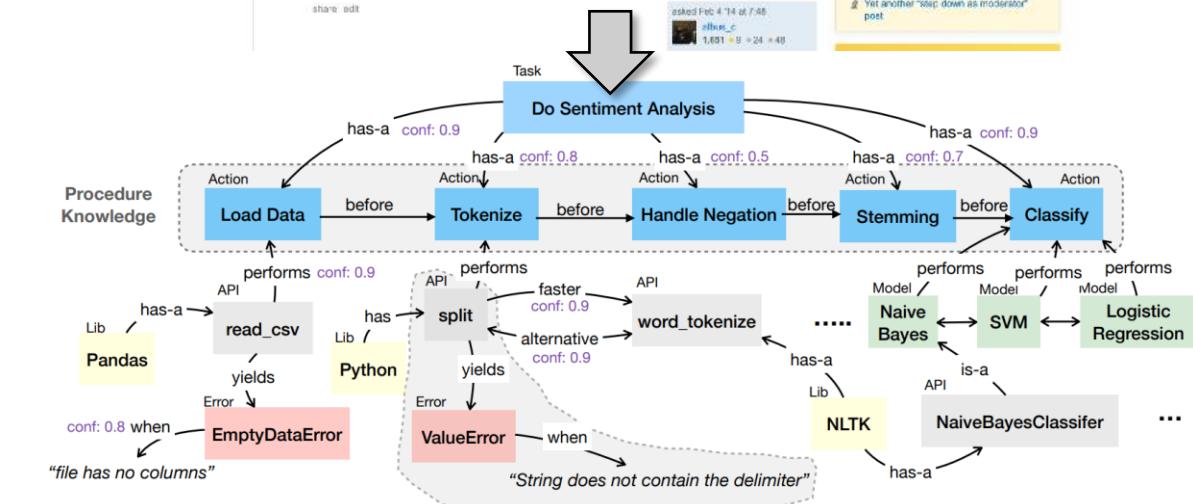
Load data from txt with pandas
Asked 5 years, 6 months ago Active 1 month ago Viewed 417K times

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

This is the structure of the input file: 1 E 2800.8 78.2836942112 1347.2836942112 /file,address,txt
Now the data are imported as a unique column. How can I divide it, so to store different elements separately (so I can call `data[1,i]`)? And how can I define a header?

python io pandas

share edit asked Feb 4 '14 at 7:48 alvaro_c 1/631 9 24 48



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



Acknowledgement

Student Researchers

Language Understanding and Knowledge Acquisition Lab



Wenxuan Zhou
(PhD Student)

Bangzheng Li
(Undergrad →
PhD Student)

Nancy Xu
(PhD Student)

Eric Qasemi
(PhD Student)

Fei Wang
(MS →
PhD Student)

James Y. Huang
(PhD Student)

Keming Lu
(MS Student)

Collaborating Institutes



Carnegie
Mellon
University



NUS

National University
of Singapore



UNIVERSITY OF
CAMBRIDGE



Microsoft
Research



David Geffen
School of Medicine



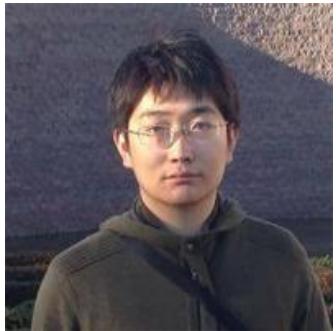
amazon

APPLIED RESEARCH LABORATORY FOR
INTELLIGENCE
AND SECURITY

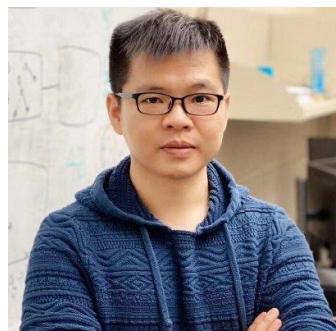
cisco

Sponsors

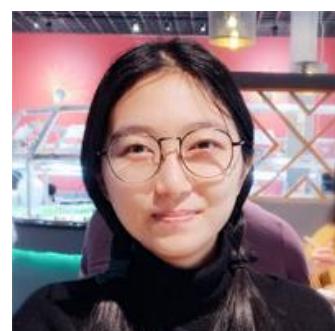
New Frontiers of Information Extraction



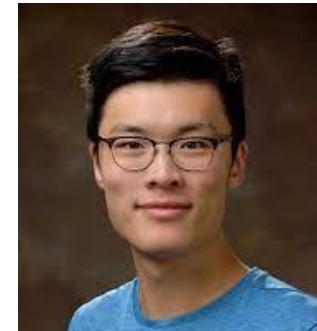
Muhao Chen



Lifu Huang



Manling Li



Ben Zhou



Heng Ji



Dan Roth

Contents

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

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

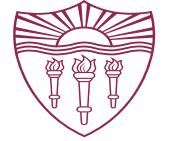


NAACL 2022

July 2022

NAACL Tutorials

New Frontiers of Information Extraction



Thank You