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***?

AI Needs to Understand Relations of Concepts

QA and Semantic Search

(?car, ***produced by***, Mazda)
(?car, ***won***, 24 Hours of Le Mans)

AI Needs to Understand Relations of Concepts

E-commerce

- Relations of Products and Users

Comp. Bio./Med.

- Interactions of (bio)molecules
- Relations of diseases and drugs

IE Is the Backbone of Any Knowledge-driven AI Systems

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/;[6] Hawaiian: [hono ˈlulu]) 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.

Structural Extracts

Knowledge Graphs

Freebase, yago select knowledge, NELL, WIKIDATA, ConceptNet An open, multilingual knowledge graph, DBpedia

Bio-med Ontologies /Data Banks

MeSH, SKEMPI, neXtprot, STRING, PubMed, GENEONTOLOGY Unifying Biology, RCSB PDB PROTEIN DATA BANK

Product Graphs

product graph, ebay, Alibaba.com

IE automatically extracts structural knowledge about concepts and relations

Two Types of IE Tasks

Two types of IE tasks

Lexical IE: Identifying **concepts** (entities, events, terms, etc.) and their **types**.

Annotated documents

- E.g, OntoNotes, ACE, etc…

Mention Detection, Typing, Linking

Relational IE: Identifying **relations** (or properties) of concepts

Relation Extraction

Fragility of IE Models

Fragility in Learning

Wrong Args	Authorities said they ordered the detention of Bruno 's wife , [Dayana Rodrigues]$_{tail:per}$, who was found with [Samudio]$_{head:per}$'s baby .	*per:spouse*	109
Relation Def.	[Zhang Yinjun]$_{tail:per}$, spokesperson with one of China 's largest charity organization , the [China Charity Federation]$_{head:org}$	*org:top_mem.*	96
Entity Type	[Christopher Bollyn]$_{head:per}$ is an [independent]$_{tail:religion}$ journalist	*per:religion*	31

Noisy Training Data

IE (structural) annotation is difficult and often noisy

- 5-8% errors in TACRED & CoNLL03
- <70% IAA in HiEve & IC
- etc.

Data distribution in UFET

Ultra Diverse Labels and Low Training Resources

The extracts are often:

- Diverse and unbalanced
+ Expensive and insufficient

Fragility of IE Models

Fragility in Inference

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

How do we ensure the extracts are **globally consistent**?

Visited? or founded?

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

And **faithful**?

Michael Jordan [PERSON] is a professor at Berkeley [ORG]

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

What about **out-of-distribution** Inputs?

Michael Jordan [PER] is an expert in machine learning .

Statistician?

Comp. neuroscientist?

Unknown Extract?

Goal: Robust IE

The goal of developing a robust IE system

Robustness in Learning

- **Noise robustness**: proactively identifying and mitigating training noise
- **Constraint learning**: capturing logical constraints of labels
- **Debiased training**: mitigating feature shortcuts and balancing training signals
- **Indirectly supervised**: should not rely solely on expensive (direct) supervision

Overcome minimal, noisy and biased supervision

Robustness in Inference

- **Selectiveness**: knowing what is extractable, what is not
- **Constrained inference**: ensuring logically consistent extracts
- **Faithfulness**: does not rely on spurious correlation

Self-contained, selective and faithful extraction.

Agenda

1. Noise-robust IE

2. Faithful IE

3. Logically Consistent IE

4. Indirect Supervision (Brief)

Agenda

1. Noise-robust IE

2. Faithful IE

3. Logically Consistent IE

4. Indirect Supervision (Brief)

Noise In Training and Inference

Training

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

On Tuesday, there was a typhoon-strength (***e*$_1$*:storm***) in Japan. One man got (***e*$_2$*:killed***) and thousands of people were left stranded. Police said an 81-year-old man (***e*$_3$*:died***) in central Toyama when the wind blew over a shed, trapping him underneath. Later this afternoon, with the agency warning of possible tornadoes, Japan Airlines (***e*$_4$*:canceled***) 230 domestic flights, (***e*$_5$*:affecting***) 31,600 passengers.

Reading long documents, annotating complex structures

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
- <70% IAA in HiEve, Intelligence Community, etc.

Inference

In real application, IE models see way larger, more diverse and noisy data than in training

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) is an expert in machine learning .

Statistician?

Comp. neuroscientist?

Unknown extraction types

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

No Rel

Nothing to extract

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

③ Relabeling model: repair examples that make through filtering but which still have errors or missing labels (multi-label classification)

④ Cleaned (task) training data

Cost: manually labeling enough clean data can still be expensive.

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

Unsupervised Denoising: Ensemble

① Partition data into k-folds

② Cross-validate the quality of each fold

- 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.

③ Reweight data folds and train the final model

Unsupervised denoising: no longer requires annotated clean data
Cost: needs repeated training and testing of the model for 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

Zhou and Chen. *Learning from Noisy Labels for Entity-Centric Information Extraction*. **EMNLP** 2021

Unsupervised Denoising: Co-regularized Knowledge Distillation

1. Create $M(\geq 2$; 2 is enough) identical 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

Zhou and Chen. *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)

Zhou and Chen. *Learning from Noisy Labels for Entity-Centric Information Extraction*. **EMNLP** 2021

Noise in Inference

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

Open class

- 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

Zhou et al. *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

Agenda

1. Noise-robust IE

2. Faithful IE

3. Logically Consistent IE

4. Indirect Supervision (Brief)

Faithfulness Issues

IE systems may not always **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

Shortcut Prediction: 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 yesterday.

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?

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

Strategy 1: 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

Strategy 2: 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 $Y_{\bar{x}}$

Debiased prediction Y_{final}

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

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

Obtained on dev set

Wang et al. *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.

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

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

Counterfactual Inference

Evaluation on out-of-distribution versions of 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.

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

Faithfulness Issues in Other IE Tasks

Faithfulness in IE is still an underexplored research direction.

Entity Typing and Linking

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

Xu et al. *Does Your Model Classify Entities Reasonably? Diagnosing and Mitigating Spurious Correlations in Entity Typing*. 2022

NER

Original NER Examples

I thank my **Beijing** [GPE] friends and wish everyone a Happy **New Year** [EVENT] .

WIKIDATA

Entity-level Attack

Natural Adversarial Examples (*Entity-only*)

I thank my **Bari** [GPE] friends and wish everyone a Happy **Casimir Pulaski Day** [EVENT] .

Context-level Attack

Natural Adversarial Examples (*Entity* + *Context*)

I ***admire*** my **Bari** [GPE] ***roommates*** and wish everyone a Happy **Casimir Pulaski Day** [EVENT] .

Lin et al. *RockNER: A Simple Method to Create Adversarial Examples for Evaluating the Robustness of Named Entity Recognition Models*. **EMNLP**-21

Agenda

1. Noise-robust IE

2. Faithful IE

3. Logically Consistent IE

4. Indirect Supervision (Brief)

Consistency of IE

How do we ensure the extracts are **globally consistent**?

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.

Take event-event relation extraction as an example

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

Extracts are not independent, but forming a structure with dependencies

- E.g., Temporal relations cannot be a loop
- A main event cannot happen after its subevent

Wang et al. *Joint Constrained Learning for Event-Event Relation Extraction*. **EMNLP 2020**

Logical Constraints Of Relations

Symmetry

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

Conjunction

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

Transitivity

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

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

Why adding logical constraints in learning?

- Learning to provide **globally consistent** predictions
- Providing **indirect supervision** across tasks/learning resources

Li et al. *A Logic-Driven Framework for Consistency of Neural Models.* **EMNLP 2019**
Wang et al. *Joint Constrained Learning for Event-Event Relation Extraction.* **EMNLP 2020**

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	–

Joint Constrained Learning

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

Task loss

Implication and conjunction constraint losses

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

The Joint Constrained Learning Architecture

Constrained learning surpasses SOTA TempRel extraction on MATRES [Ning+, ACL-18] by relatively 3.27% in F_1.

Model	P	R	F_1
CogCompTime (Ning et al., 2018c)	0.616	0.725	0.666
Perceptron (Ning et al., 2018b)	0.660	0.723	0.690
BiLSTM+MAP (Han et al., 2019b)	-	-	0.755
LSTM+CSE+ILP (Ning et al., 2019)	0.713	0.821	0.763
Joint Constrained Learning (ours)	**0.734**	**0.850**	**0.788**

On HiEve [Glavaš+, LREC-14] for subevent extraction, it relatively surpasses previous methods by at least 3.12% in F_1.

Model	F_1 score		
	PC	CP	Avg.
StructLR (Glavaš et al., 2014)	0.522	**0.634**	0.577
TACOLM (Zhou et al., 2020a)	0.485	0.494	0.489
Joint Constrained Learning (ours)	**0.625**	0.564	**0.595**

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

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

Estimates probabilities of conjunctive constraints

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

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

Wang et al. *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.

Agenda

1. Noise-robust IE

2. Logically Consistent IE

3. Faithful IE

4. Indirect Supervision (Brief)

Direct Supervision Is Expensive

Direct supervision can be overly **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]

Can we leverage **indirect** supervision from other **resource-rich** tasks?

Natural Language Inference

NLI premise (i.e., input text)	NLI hypothesis
Donald Trump served as the 45th president of the United States from 2017 to 2021.	(✓) **Donald Trump** is a politician (✗) **Donald Trump** is a journalist
Donald Trump served as the 45th president of the **United States** from 2017 to 2021.	(✓) **Donald Trump** is citizen of **United States** (✗) **Donald Trump** is the CEO of **United States**
In [1997]$_{time}$, the [company]$_{sub}$ [hired]$_{trigger}$ [John D. Idol]$_{obj}$ to take over Bill Thang as the new chief executive.	(✓) **John D. Idol** was hired. (✓) **John D. Idol** was hired in 1997. (✗) **Bill Thang** was hired.

Abstractive Summarization

Both tasks have rich resources (in million scales), and refines information from long text

Indirect Supervision from NLI

Ultra-fine Entity Typing (UFET; Choi et al. EMNLP-18)

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

↓

film, art, movie, show, entertainment, creation

- **>10K** free-form types
- Only **2,000** clean training cases

Prior methods as classifiers

- At most mid-40 in F1 [Dai et al. ACL-21]
- Not adaptable to unseen types

Our precise and generalizable solution: **NLI**

- **Entity-mentioning sentences as premises**
- **Ranking type descriptions as hypotheses**

Li et al. Ultra-fine Entity Typing with Indirect Supervision from Natural Language Inference. TACL 2022

Indirect Supervision from NLI

① Type verbalization

- **Human-designed templates (three kinds)**
- **T5 generation: SENT <X> ENT <Y> TYPE <Z>**

② NLI + learning to rank (with entailment scores)

- **Pre-training on MNLI**
- **Fine-tuning as learning-to-rank**
- **Easily adaptable to an open type space**

Our method (LITE) offers the current SOTA UFET performance (>50% F1 for >10K types).

Li et al. Ultra-fine Entity Typing with Indirect Supervision from Natural Language Inference. TACL 2022

Indirect Supervision from NLI

And it is very generalizable

Much better zero- and few-shot performance than previous SOTA (data augmentation-based)

Model		OntoNotes		FIGER	
		macro-F1	micro-F1	macro-F1	micro-F1
Hierarchy-Typing (Chen et al., 2020b)		73.0	68.1	83.0	79.8
Box4Types (Onoe and Durrett, 2020)		77.3	70.9	79.4	75.0
DSAM (Hu et al., 2020)		83.1	78.2	83.3	81.5
SEPREM (Xu et al., 2021)		–	–	86.1	82.1
MLMET (Dai et al., 2021)		85.4	80.4	–	–
LITE	pre-trained on NLI+UFET	**86.6**	**81.4**	80.1	74.7
	NLI+task-specific training	86.4	80.9	**86.7**	**83.3**

Even outperforming supervised SOTA with zero-shot transfer to OntoNotes and FIGER

- **(12.5%-16.8% disjoint labels)**

Li et al. Ultra-fine Entity Typing with Indirect Supervision from Natural Language Inference. TACL 2022

Indirect Supervision with Summarization

- Summarization tasks align with the goal of relation extraction (RE)
  - » If we view extracts as one kind of **salient information**
- Lu et al. (2022): summarization as indirect supervision for RE
  - » Format input subject/object types as natural language
  - » Ask a summarization model for (constrained) generation of verbalized relations

Pre-trained on CNN/Dailymail (Hermann et al., 2015) and XSum (Narayan et al., 2021)

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)

Lu et al. Summarization as Indirect Supervision for Relation Extraction. EMNLP 2022 In Submission (ARR reviewed)

Open Research Direction: Consolidating Extracts

Extracts are local (differ in contexts), but knowledge is global (unique and consistent)

Several relevant tasks on text

- Fact verification
- Answer consolidation

Q: Is coffee good for your health?

Same answers

Coffee can make you slim down.

Coffee can help with weight loss.

Coffee can relieve headache.

How do those technologies consolidate structural extracts?

Knowledge alignment across languages

DBpedia — Novel

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

Zhou et al. *Answer Consolidation: Formulation and Benchmarking*. NAACL 2022
Thorne et al. *FEVER: a large-scale dataset for Fact Extraction and VERification*. NAACL 2018

Chen et al. *Multilingual Knowledge Graph Completion via Ensemble Knowledge Transfer*. EMNLP: Findings 2020
Zhou et al. *Prix-LM: Pretraining for Multilingual Knowledge Base Construction*. ACL 2022

Open Research Direction: Perturbation Robustness

Semantic Perturbation

Last week , Bill Gates (PER) paid a visit to Microsoft Building 99 (MISC) .

??

Last week , Microsoft Building 99 (MISC) had an important visit made by Bill Gates (PER)

??

I heard from Bill Gates (PER) himself that he paid a visit to the Microsoft Building 99 (MISC) last week .

??

Parameter Perturbation

(a) w/o Flooding (b) w/ Flooding

- Qin et al. *Improving Entity and Relation Understanding for Pre-trained Language Models via Contrastive Learning*. ACL 2021
- Huang et al. *Disentangling semantics and syntax in sentence embeddings with pre-trained language models*. NAACL 2021

- Foret et al. *Sharpness-aware minimization for efficiently improving generalization*. ICLR 2020
- Ishida et al. *Do We Need Zero Training Loss After Achieving Zero Training Error?* ICML 2020

Open Research Direction: 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?

Large models still do not support quantitative reasoning well

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

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

References

- Onoe and Durrett. Learning to Denoise Distantly-Labeled Data for Entity Typing. NAACL 2019
- Wang et al. CrossWeigh: Training named entity tagger from imperfect annotations. EMNLP 2019
- Zhou and Chen. Learning from Noisy Labels for Entity-Centric Information Extraction. EMNLP 2021
- Toneva et al. An empirical study of example forgetting during deep neural network learning. ICLR 2019
- Dhamija et al. Reducing network agnostophobia. NeurIPS 2018
- Zhou et al. Contrastive Out-of-Distribution Detection for Pretrained Transformers. EMNLP 2021
- Xin et al. The Art of Abstention: Selective Prediction and Error Regularization for Natural Language Processing. ACL 2021
- Qian et al. Counterfactual Inference for Text Classification Debiasing. ACL 2021
- Wang et al. hould We Rely on Entity Mentions for Relation Extraction? Debiasing Relation Extraction with Counterfactual Analysis. NAACL 2022
- Li et al. A Logic-Driven Framework for Consistency of Neural Models. EMNLP 2019
- Wang et al. Joint Constrained Learning for Event-Event Relation Extraction. EMNLP 2020
- Pan et al. Learning Constraints for Structured Prediction Using Rectifier Networks. ACL 2020
- Wang et al. Learning Constraints and Descriptive Segmentation for Subevent Detection. EMNLP 2021
- Thorne et al. FEVER: a large-scale dataset for Fact Extraction and VERification. NAACL 2018
- Zhou et al. Answer Consolidation: Formulation and Benchmarking. NAACL 2022
- Zhou et al. Prix-LM: Pretraining for Multilingual Knowledge Base Construction. ACL 2022
- Qin et al. Improving Entity and Relation Understanding for Pre-trained Language Models via Contrastive Learning. ACL 2021
- Huang et al. Disentangling semantics and syntax in sentence embeddings with pre-trained language models. NAACL 2021
- Foret et al. Sharpness-aware minimization for efficiently improving generalization. ICLR 2020
- Ishida et al. Do We Need Zero Training Loss After Achieving Zero Training Error? ICML 2020
- Chen et al. Counterfactual Inference for Text Classification Debiasing. ACL 2021
- Wang et al. Should We Rely on Entity Mentions for Relation Extraction? Debiasing Relation Extraction with Counterfactual Analysis. NAACL 2022
- Lin et al. Focal loss for dense object detection. CVPR 2017
- Zhou and Chen. An Improved Baseline for Sentence-level Relation Extraction. 2021
- Lin et al. A Simple Method to Create Adversarial Examples for Evaluating the Robustness of Named Entity Recognition Models. EMNLP 2021
- Zhang et al. Do Language Embeddings Capture Scales? EMNLP: Findings 2020

Thank You