From Data to Model Programing: Injecting Structured Priors for Knowledge Extraction

Xiang Ren

Department of Computer Science, USC

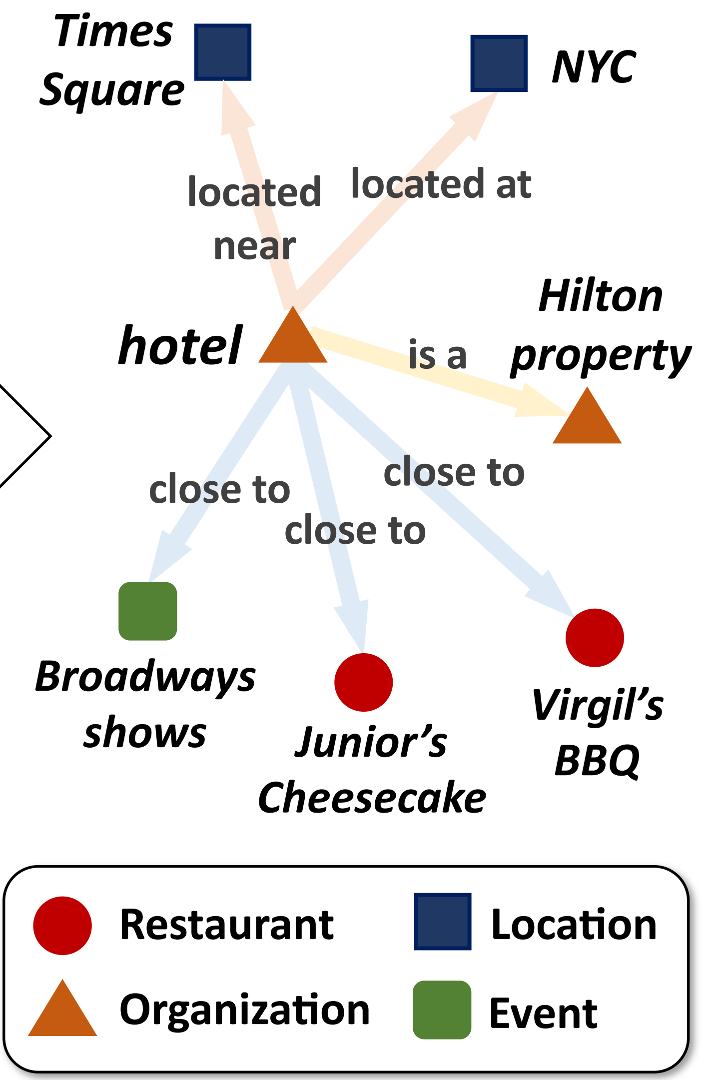
USC Information Science Institute

USC Machine Learning Center

Machine Reading: From Text to Knowledge Structures

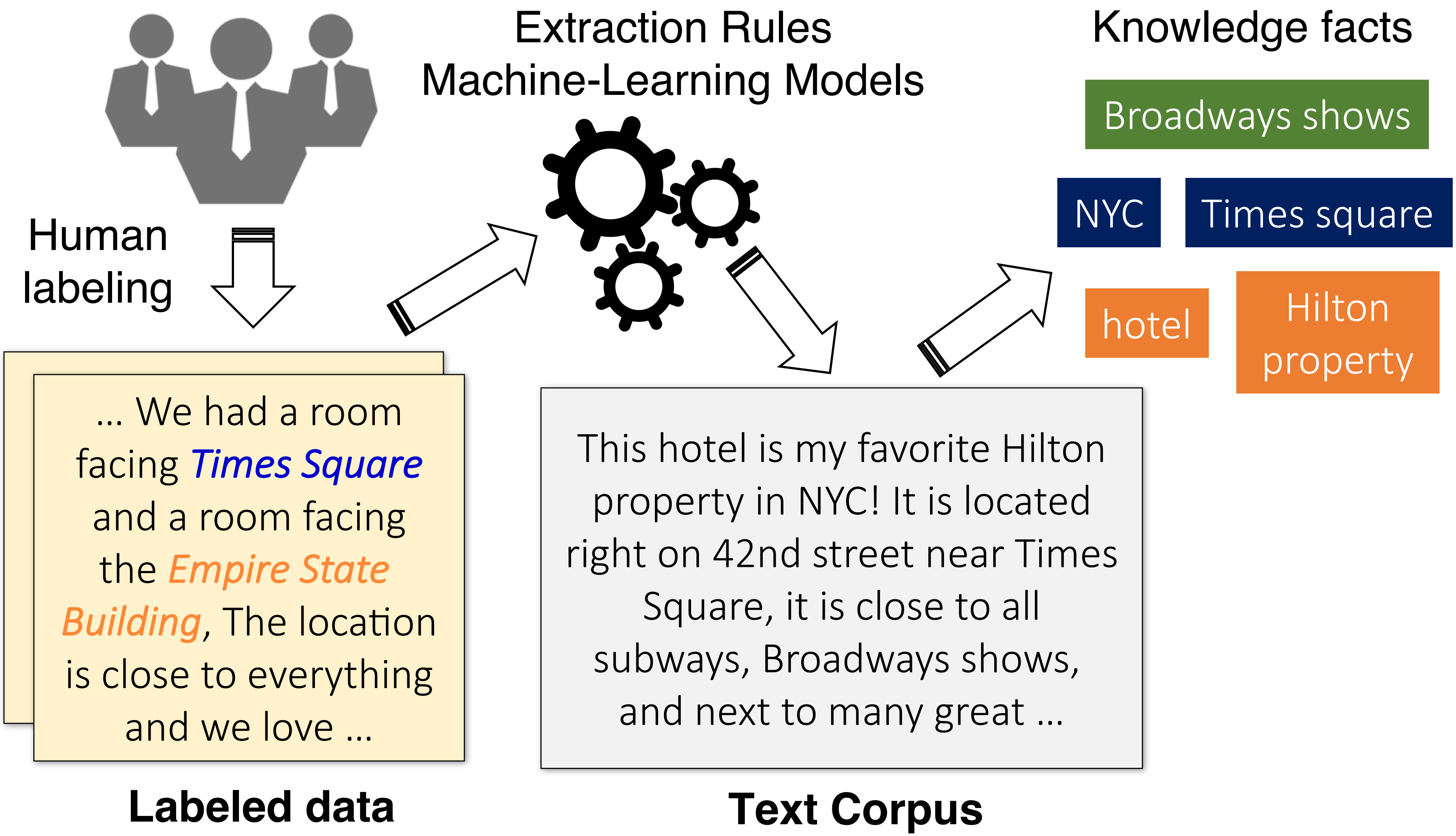
This **hotel** is my favorite **Hilton property** in **NYC**! It is located right on 42nd street near **Times Square**, it is close to all subways, **Broadways shows**, and next to great restaurants like **Junior's Cheesecake**, **Virgil's BBQ** and many others.

-- *TripAdvisor*

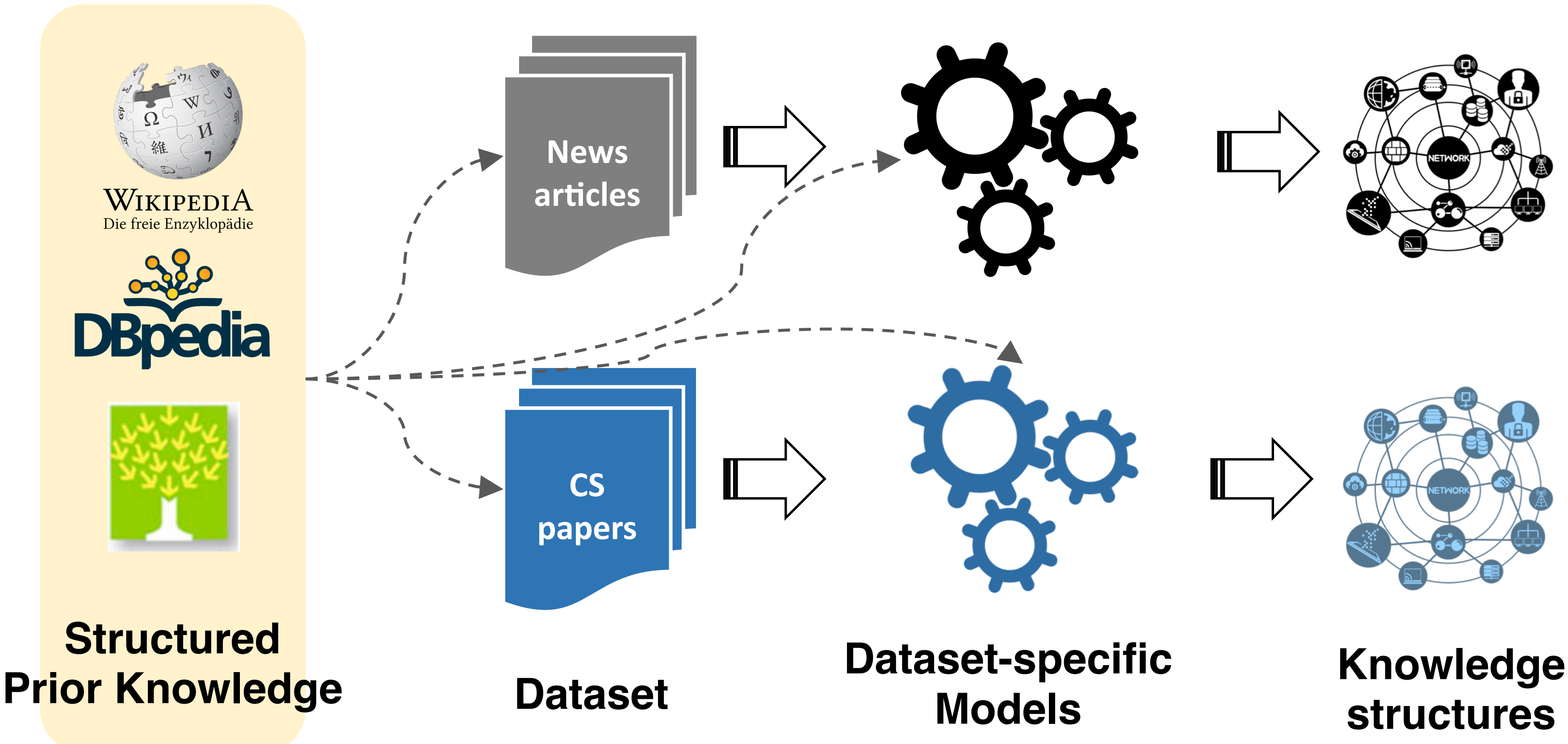


Structured Facts
1. "Typed" entities
2. "Typed" relationships

Prior Art: Machine Reading with *Repeated* Human Annotation Effort



Making Machine Learning ***Cheaper*** *on Knowledge Extraction*



- Enables ***quick*** development of applications over various corpora
- Extracts ***complex*** structures without introducing human errors

Structured Prior Knowledge

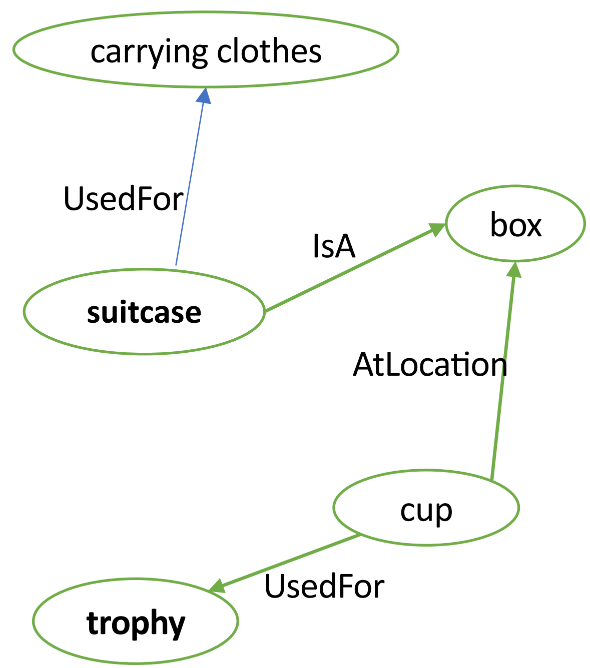
Domain Dictionaries

Entity Type	Canonical Name	Synonyms
Person	Donald Trump	Trump, President Trump, …
…	…	…

Labeling Rules

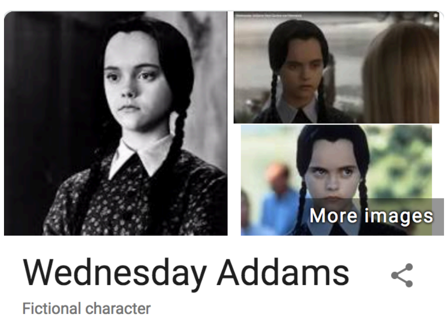
$P1$	(**SUBJ-PER**, *'s children*, **OBJ-PER**)	$\rightarrow$	**PER:CHILDREN**
$P2$	(**SUBJ-PER**, *is known as*, **OBJ-PER**)	$\rightarrow$	**PER:ALTERNATIVE_NAMES**
$P3$	(**SUBJ-ORG**, *was founded by*, **OBJ-PER**)	$\rightarrow$	**ORG:FOUNDED_BY**

Ontologies/Knowledge Graphs



Challenges of Leveraging Structured Knowledge

- *Noise* in the grounding process



Entity Type	Canonical Name	Synonyms
Person	Wednesday Addams	Wednesday, …
…	…	…

~~Person~~

Today is **Wednesday**.

Challenges of Leveraging Structured Knowledge

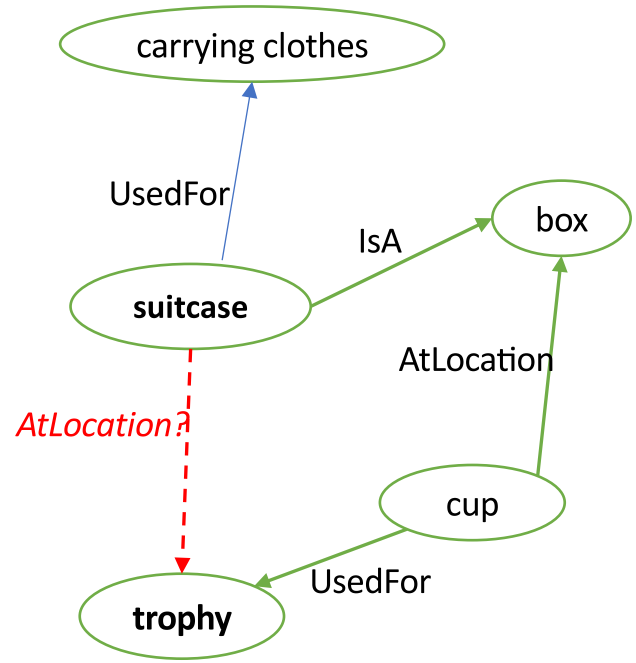
- *Noise* in the grounding process
- *Incompleteness* of the knowledge sources

~~Person~~ ~~Non-Entity~~

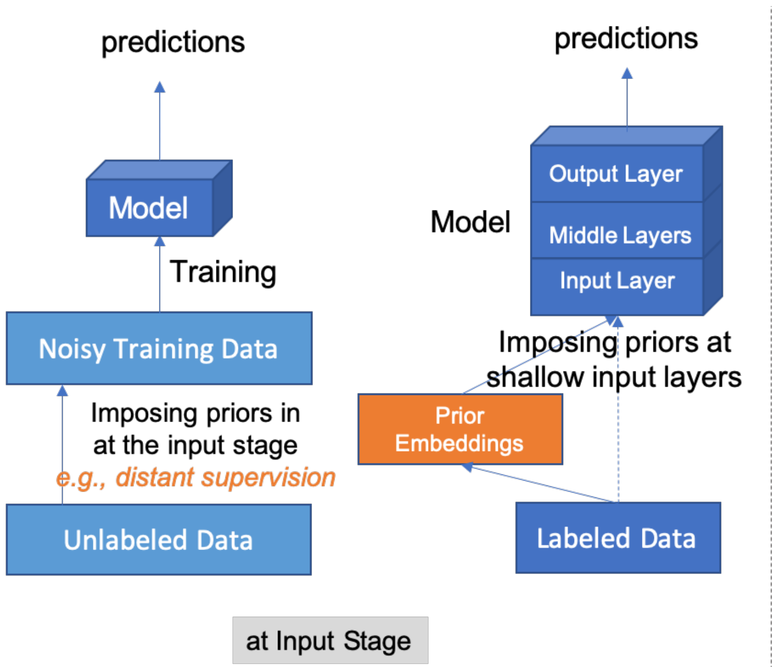
… Obama Administration Office … → … **Obama** Administration Office …

Challenges of Leveraging Structured Knowledge

- *Noise* in the grounding process
- *Incompleteness* of the knowledge sources
- *Complex & scalable* reasoning



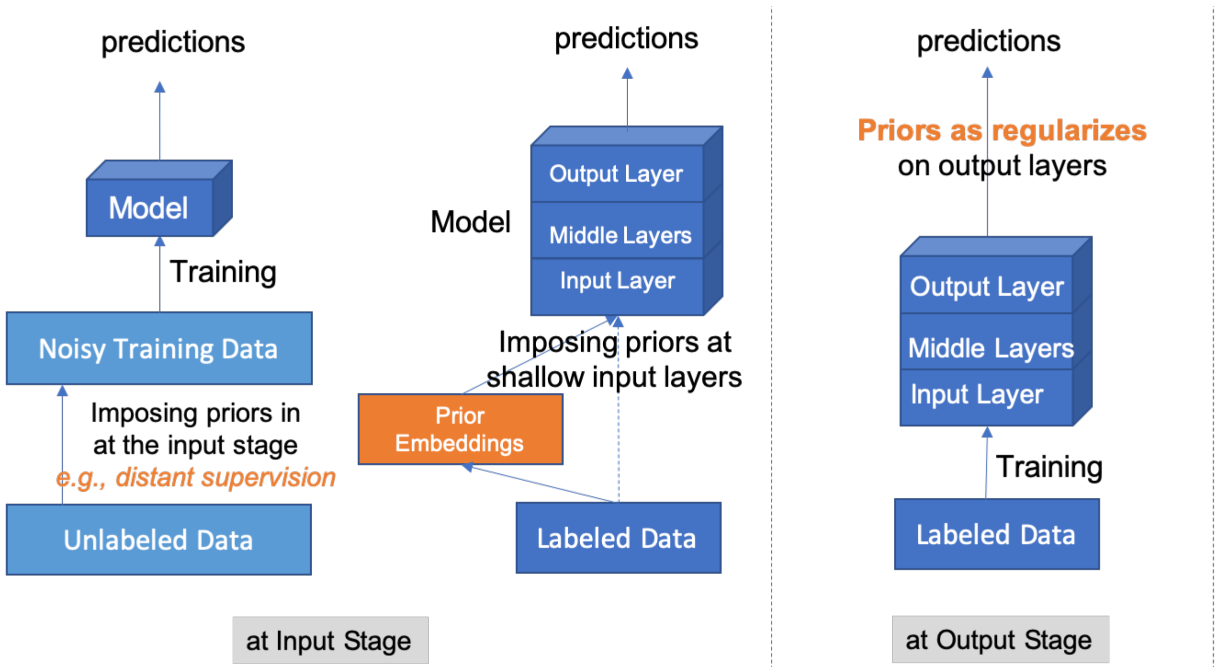
Previous Work & This Talk



Learning named entity tagger from domain dictionary (Shang et al., EMNLP 2018)

Neural rule grounding (Zhou et al., 2019)

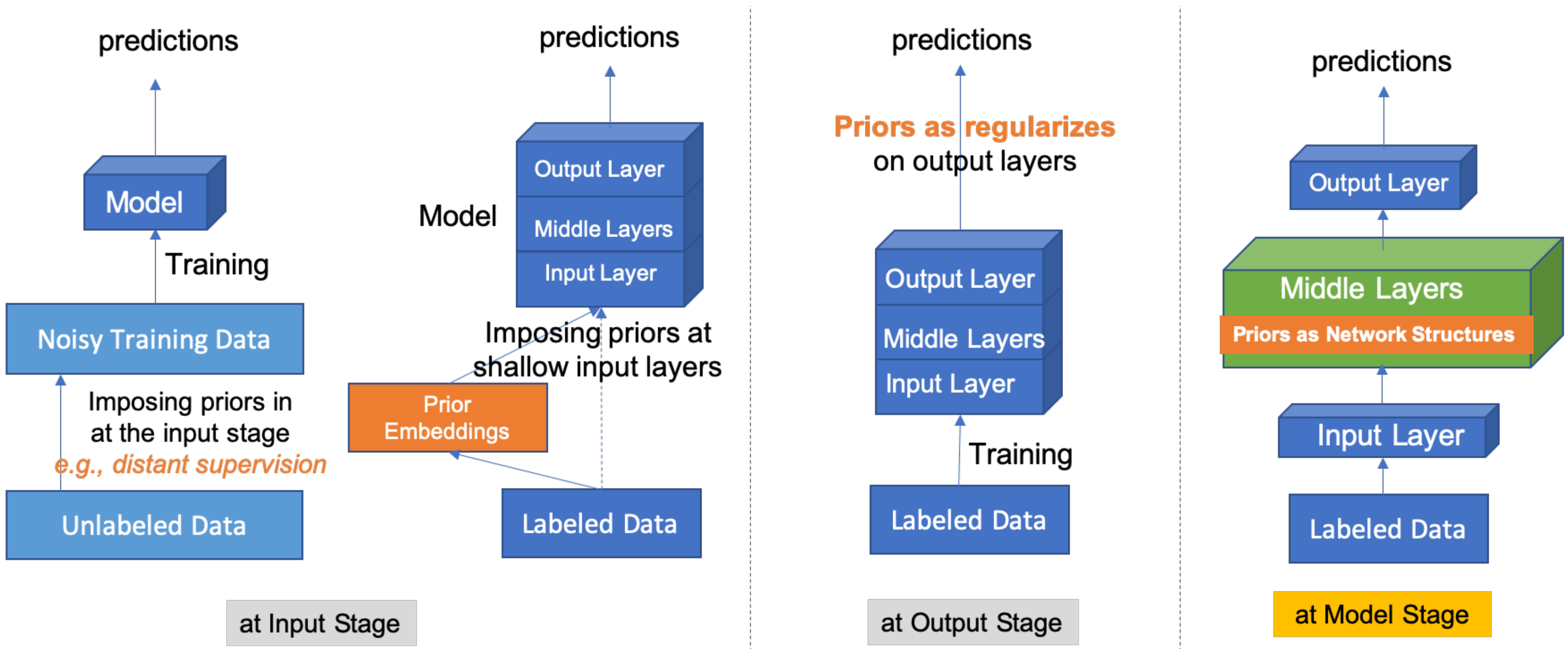
Previous Work & This Talk



Learning named entity tagger from domain dictionary (Shang et al., EMNLP 2018)

Neural rule grounding (Zhou et al., 2019)

Previous Work & This Talk



Learning named entity tagger from domain dictionary (Shang et al., EMNLP 2018)

Neural rule grounding (Zhou et al., 2019)

KagNet: Learning to Answer Commonsense Questions with Knowledge-aware Graph Networks (Lin et al., 2019)

Learning Named Entity Tagger using *Domain-Specific Dictionary*

EMNLP 2018

Joint work with Jingbo Shang, Lucas Liu, Xiaotao Gu

Sequence Tagging: Problem

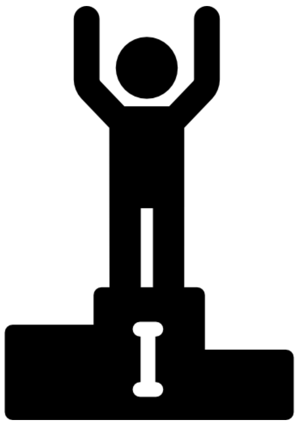
Every sentence needs to be annotated ***token by token***.

```
INPUT:  Jim       bought 300 shares of  Acme  Corp.     in  2006
LABEL: [Jim]:PER bought 300 shares of [Acme Corp.]:ORG in [2006]:Time
```

Token-level labels by human annotator

```
BIO:    B-PER     O      O   O      O   B-ORG I-ORG     O   B-Time
```

Challenge: Expensive & Slow on Creating Token-level Training Data



Achieved new SoTA on multiple sequence tagging benchmarks with LM-LSTM-CRF architecture (Liu et al., 2018)

Expensive to adapt to specific domains (e.g., biomedical, business, finance).

Can we generate ***high-precision, high-recall*** annotations ***automatically*** from domain dictionaries?

(Liu et al., AAAI 2018)

Can We Train Effective Sequence Tagger with Distant Supervision?

~~INPUT: Jim bo~~
LABEL: [Jim]PER bo
BIO: B-PER O
~~BIOES: S-PER O~~

No line-by-line annotations,
Learn named entity tagger
with *distant supervision*.

~~. in 2006 .~~
]ORG in [2006]Time .
G O B-Time O
~~G O S-Time O~~

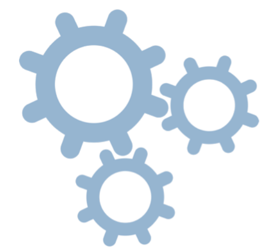
Unlabeled corpus

+

Entity Type	Canonical Name	Synonyms
Person	Donald Trump	Trump, President Trump, …
…	…	…

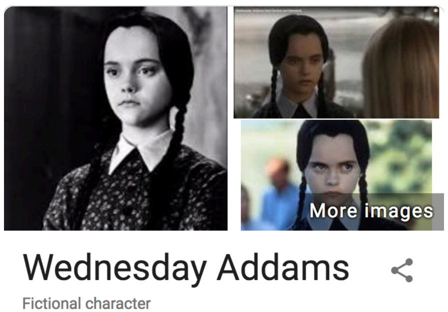
Entity Dictionary

"prior knowledge at the input level"



Seq tagging model

Distant Supervision: Issues with Simple Dictionary Matching



Entity Type	Canonical Name	Synonyms
Person	Wednesday Addams	Wednesday, …
…	…	…

~~Person~~

Today is **Wednesday**.

Name ambiguity & context-agnostic matching → ***false positive***

… Obama Administration Office …

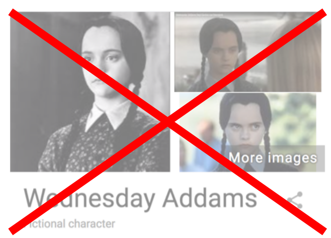
~~Person~~ Non-~~Entity~~

… **Obama** Administration Office …

Incomplete dictionary → ***false positive & false negative***

AutoNER: Label Filtering & Augmentation

- *Removes "irrelevant" entities (and their synonyms)* whose canonical names never show up in the corpus



Today is Wednesday.

- Introduces *out-of-dictionary high-quality phrases** as entities of "unknown" type

… Obama Administration Office …

… **Obama Administration Office** …

(Shang et al., EMNLP 2018) *(Shang et al., SIGMOD 2015)

AutoNER: “Tie-or-Break” Schema

- **Label the relationship of two consecutive tokens:**
  - **Tie**, when the two tokens are matched to the same entity
  - **Unknown**, if at least one of the tokens belongs to an *out-of-dictionary phrase*
  - **Break**, otherwise.

	*Today is **Wednesday***	*Today is Wednesday.*
BIOES	O O S-PER	O O O
“Tie-or-Break”	Break Break	Break Break

(Shang et al., EMNLP 2018)

"Tie-or-Break" Encoding Schema

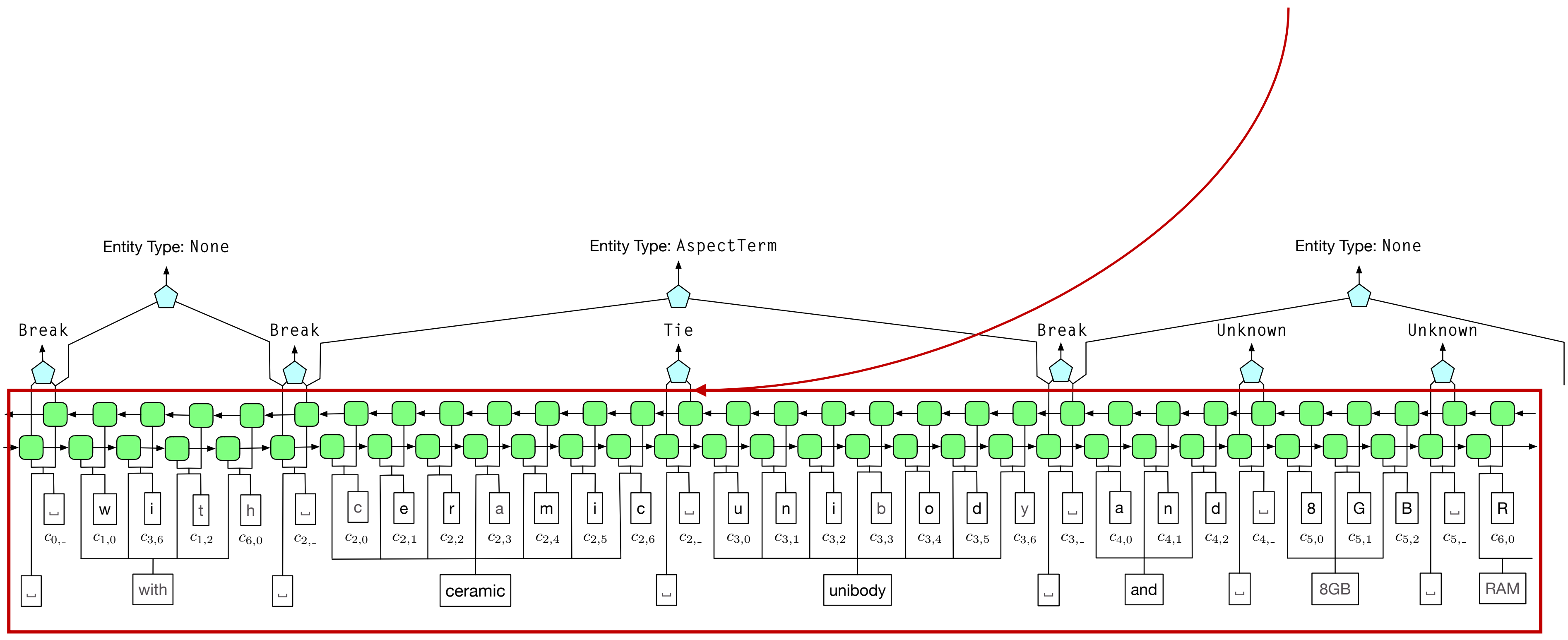
- **Label the relationship of two consecutive tokens:**
  - **Tie**, when the two tokens are matched to the same entity
  - **Unknown**, if at least one of the tokens belongs to an *out-of-dictionary phrase*
  - **Break**, otherwise.

	Ceramic body and 8GB RAM	*Ceramic body and 8GB RAM*
BIOES	B-ASP E-ASP O O O	B-ASP E-ASP O O O
"Tie-or-Break"	Tie Break Break Break	Tie Break Break Unknown

(Shang et al., EMNLP 2018)

AutoNER: Multi-task Prediction of Entity *Spans* & *Types*

- char-BiLSTM for learning contextualized representation $\mathbf{u}_i$

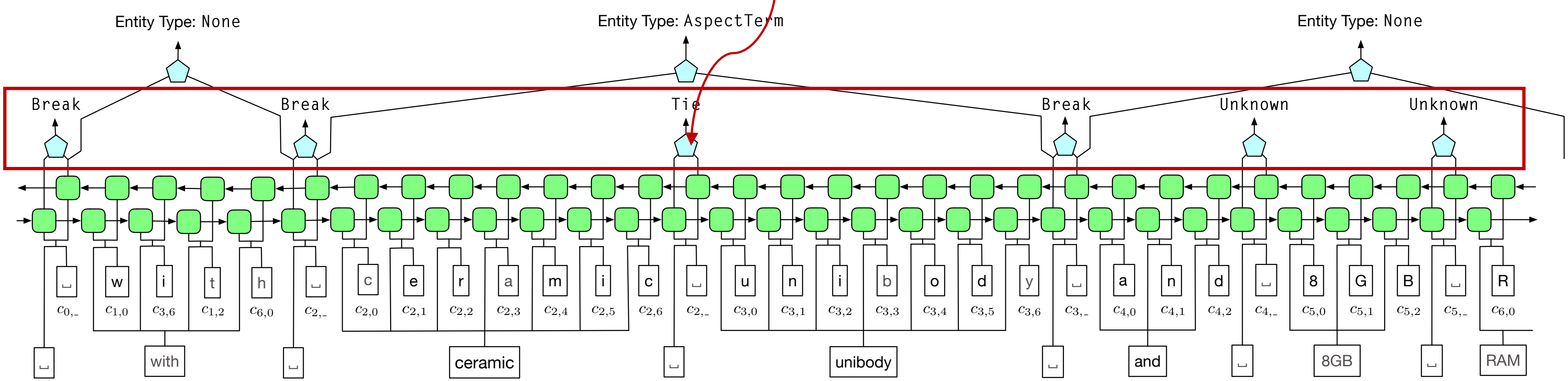


(Shang et al., EMNLP 2018)

AutoNER: Multi-task Prediction of Entity *Spans* & *Types*

- char-BiLSTM for learning contextualized representation $\mathbf{u}_i$
- 1st classification layer – “tie” or “break”

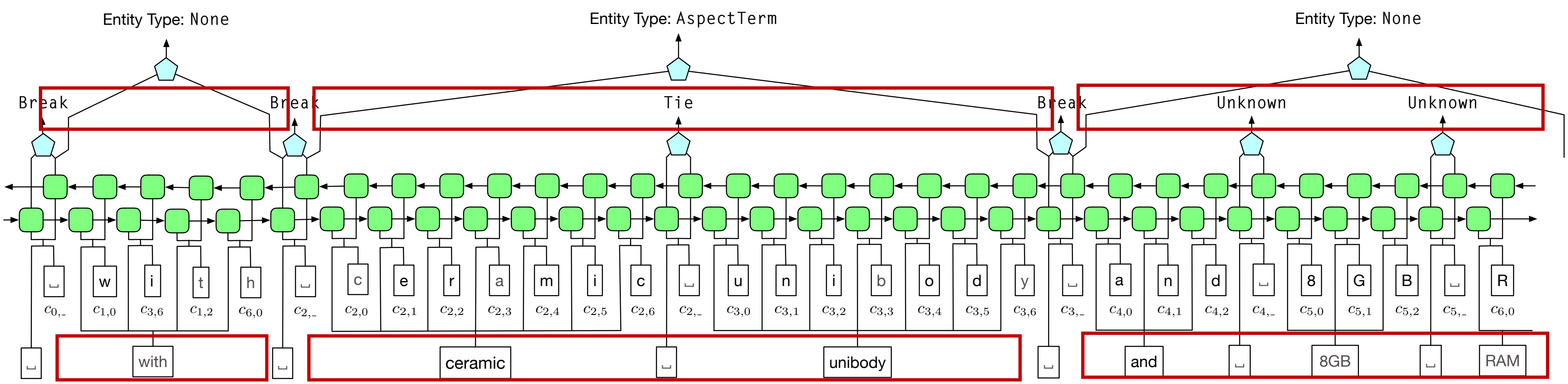
$$\boxed{p(y_i = \texttt{Break} \mid \mathbf{u}_i) = \sigma(\mathbf{w}^T \mathbf{u}_i)} \qquad \mathcal{L}_{\text{span}} = \sum_{i \mid y_i \neq \texttt{Unknown}} l(y_i, p(y_i = \texttt{Break} \mid \mathbf{u}_i))$$



(Shang et al., EMNLP 2018)

AutoNER: Multi-task Prediction of Entity *Spans* & *Types*

- char-BiLSTM for learning contextualized representation
- 1st classification layer – “tie” or “break”
- *candidate entity spans* – merge token(s) between two “break”s

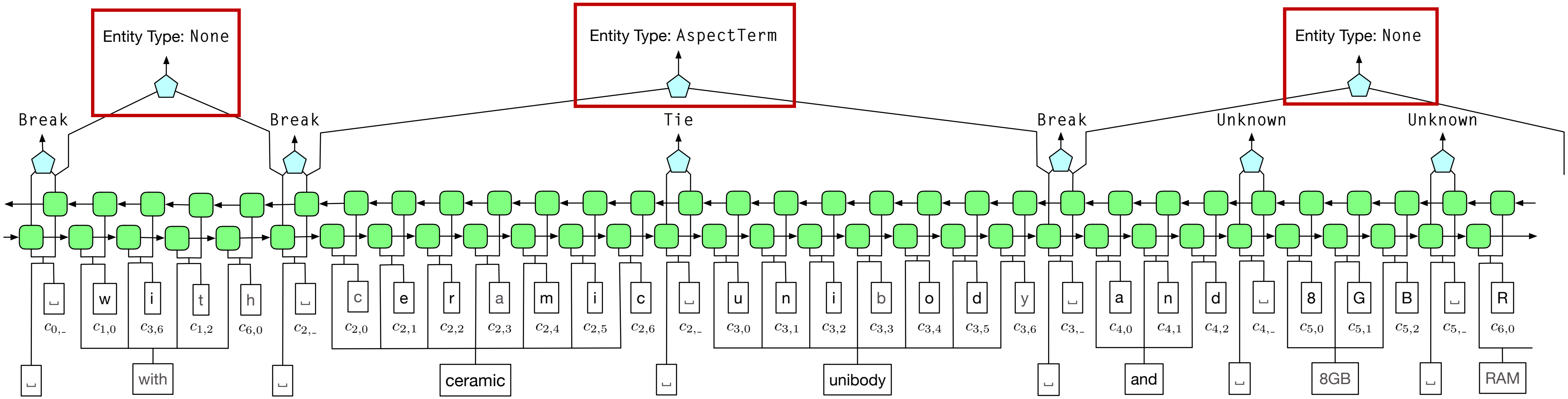


(Shang et al., EMNLP 2018)

AutoNER: Multi-task Prediction of Entity *Spans* & *Types*

- 2nd classification layer – determine entity types

multi-class cross-entropy



(Shang et al., EMNLP 2018)

Results on Biomedical Domain

- BC5CDR NER dataset: **chemical & disease**
- Fuzzy-LSTM-CRF: models tokens with “unknown” label
- AutoNER: *close to model trained on clean labeled data*

Method	Precision	Recall	F1
Dictionary Matching (DM)*	93.93	58.35	71.98
Fuzzy-LSTM-CRF (DM + label cleaning & augmentation)	88.27	76.75	82.11
AutoNER	88.96	81.00	**84.80**
LM-LSTM-CRF on gold-standard	88.84	85.16	<u>86.96</u>

*CTD Chemical and Disease vocabularies: 322,882 Chemical and Disease entity names.

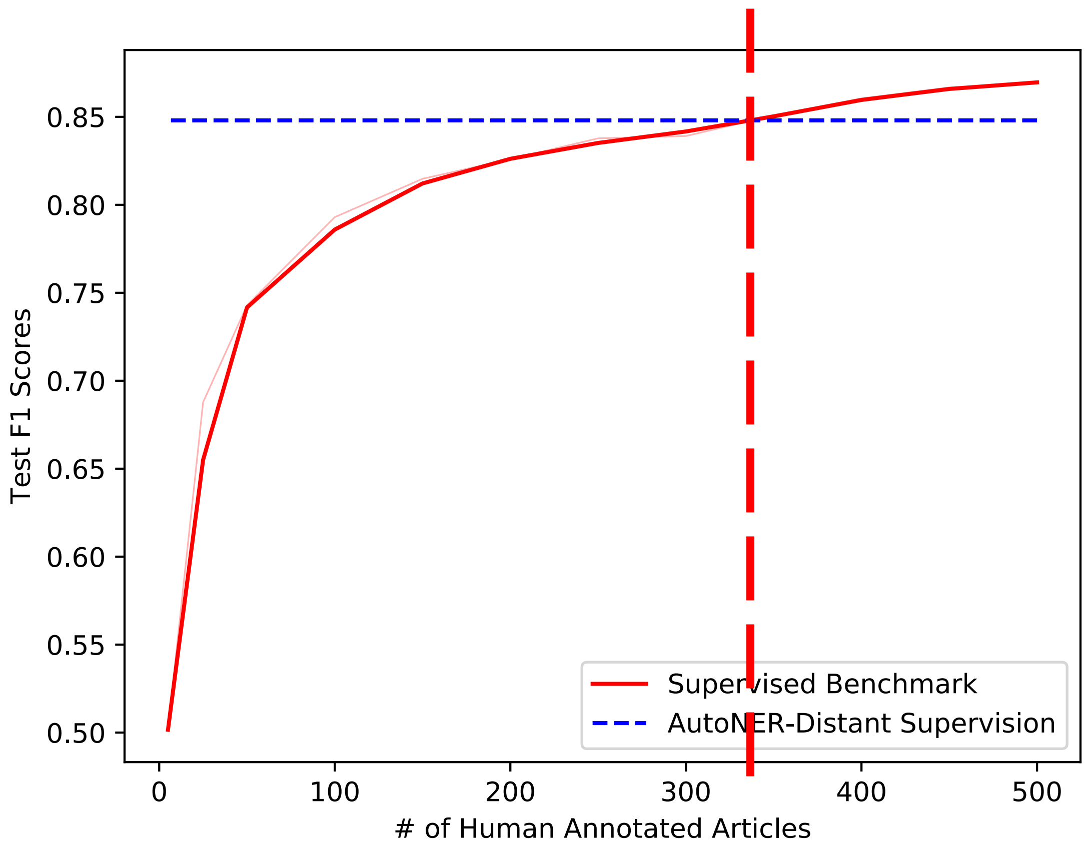
Results on Tech Review Domain

- LaptopReview NER dataset: **aspect terms**
- Models are harder to generalize
- Still a significant gap to *model trained on clean labeled data*

Method	Precision	Recall	F1
Dictionary Matching (DM)*	90.68	44.65	59.84
Fuzzy-LSTM-CRF (DM + label cleaning & augmentation)	85.08	47.09	60.63
AutoNER	72.27	59.79	**65.44**
LM-LSTM-CRF on gold-standard	84.80	66.51	<u>74.55</u>

*13,457 computer terms crawled from a public website.

AutoNER: Effectiveness on Leveraging Domain Dictionaries

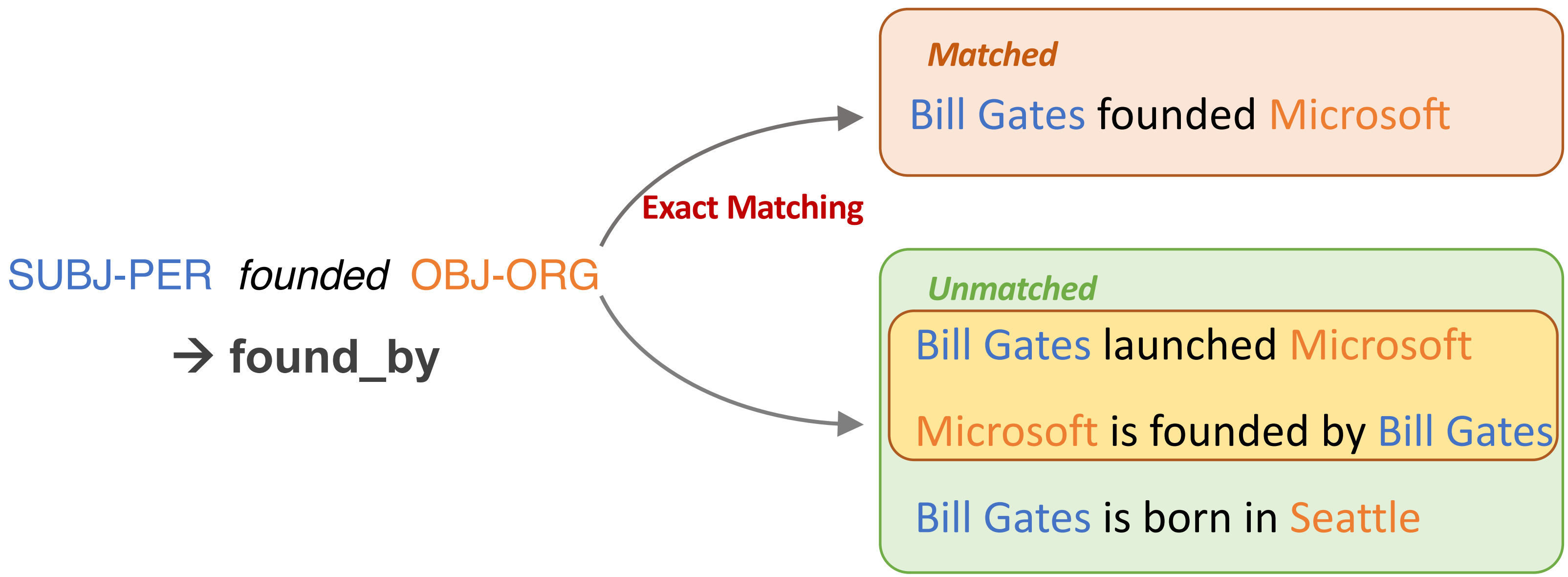


AutoNER ≈ **300** expert annotated articles on BC5CDR dataset

Neural Rule Grounding for Low-Resource Relation Extraction

Joint work with Wenxuan Zhou & Hunter Lin, *under submission*

Applying Surface Rules for Relation Extraction



Different words but semantically similar

Two Types of Methods

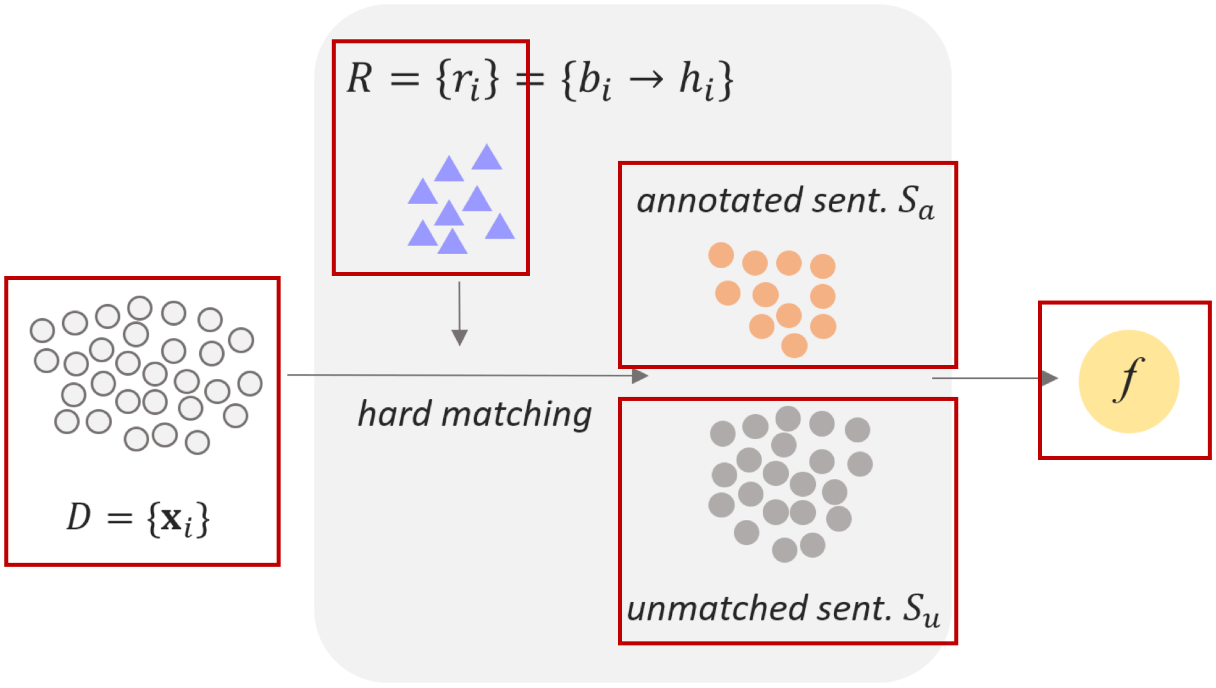
Deep learning approaches:

- Pros:
  - Latent representation
  - Good generalization
- Cons:
  - **Data hungry**
  - **Hard to interpret**

Rule-based approaches:

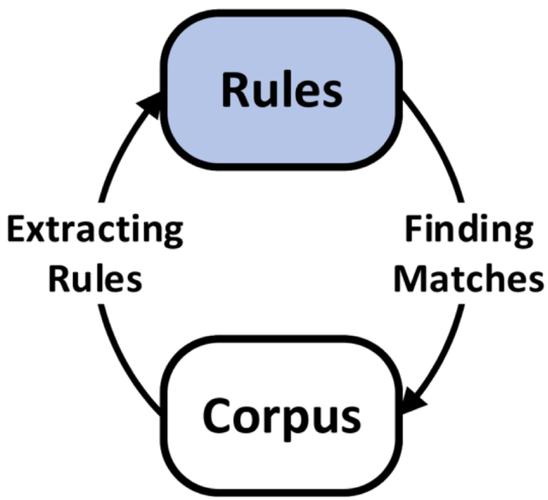
- Pros:
  - Data independent
  - Easy to interpret
  - High precision
- Cons:
  - **Low recall (Hard to generalize)**
  - Missing context information

Learning a DNN with Only Rules & Unlabeled Sentences



$r = b \to h$: *X born in the town of Y* $\to$ *(X, city_of_birth, Y)*

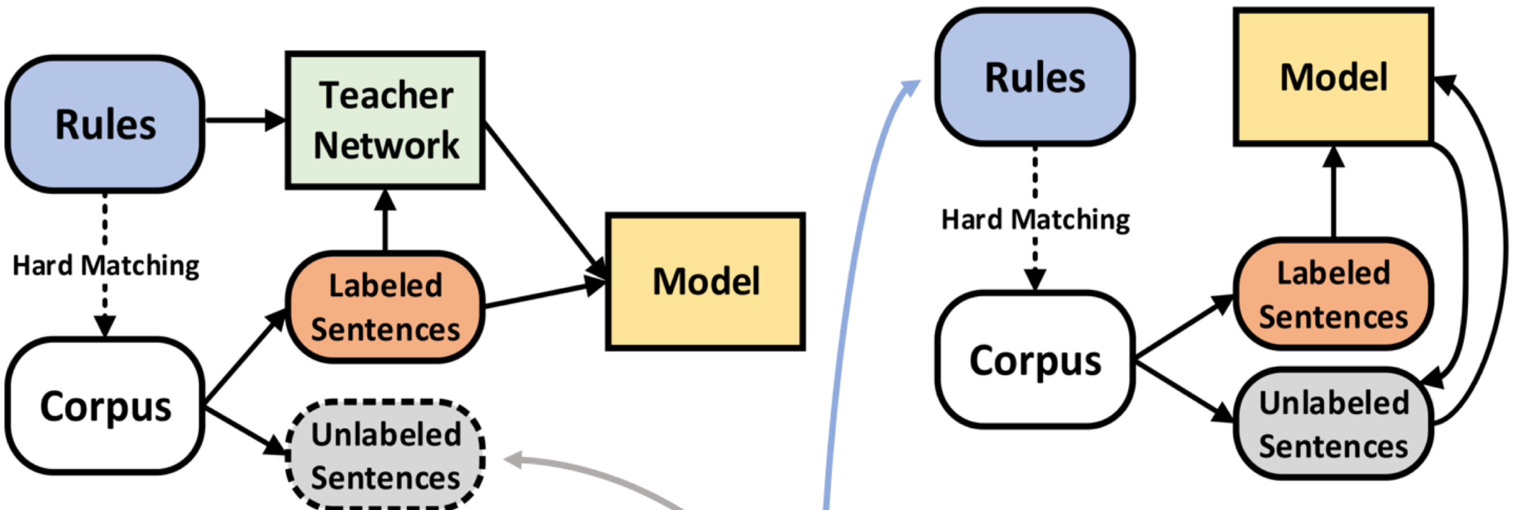
Learning from Patterns/Rules



(A) Bootstrapping

Suffer from error propagation:
The errors in model are reinforced and accumulated

Learning from Patterns/Rules

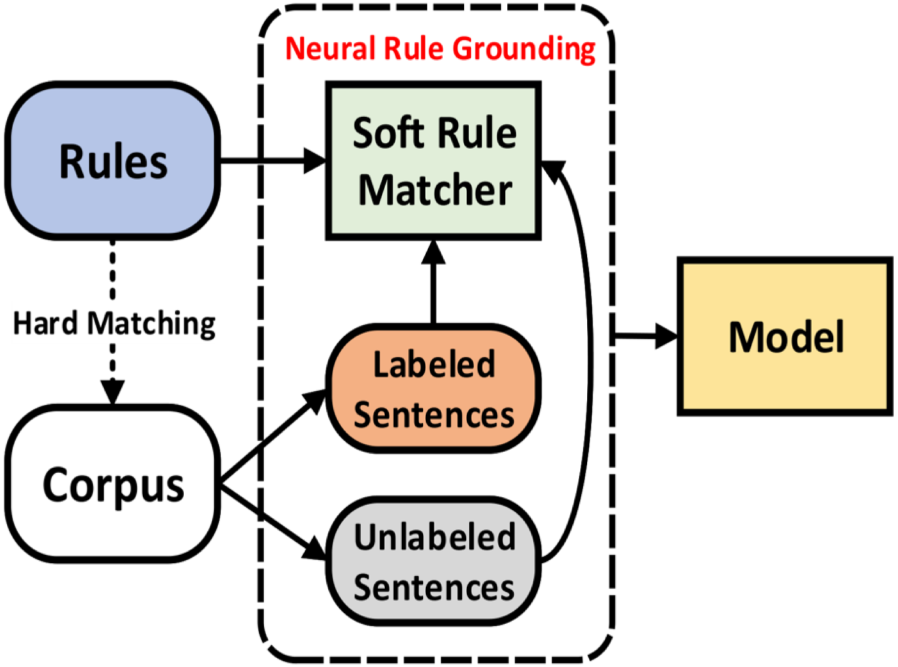


(B) Knowledge Distillation

(C) Self Learning

*No supervision from either **rules** or **unlabeled data***

Learning by Soft Rule Grounding



(D) REGD

Proposing a soft rule matcher to match rules on unlabeled sentences

Learning a *Soft Rule Matching* Function

$f_s : (S \cup P) \times P \to [-1, 1]$

SUBJ-PER *founded* OBJ-ORG

Soft grounding

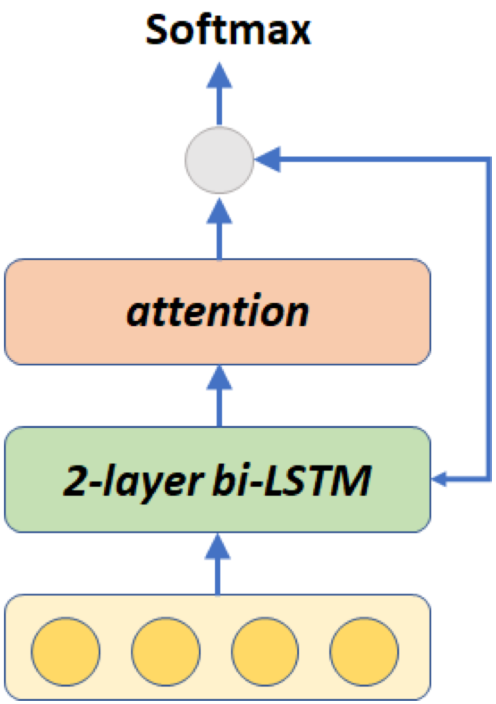
$f_s(s, p)$

Bill Gates founded Microsoft	1.0
Bill Gates launched Microsoft	0.9
Microsoft is founded by Bill Gates	0.8
Bill Gates is born in Seattle	0.3

- Perfect matching → score = 1
- Other cases → score = ?

(Zhou et al., 2019)

Sentence Encoding



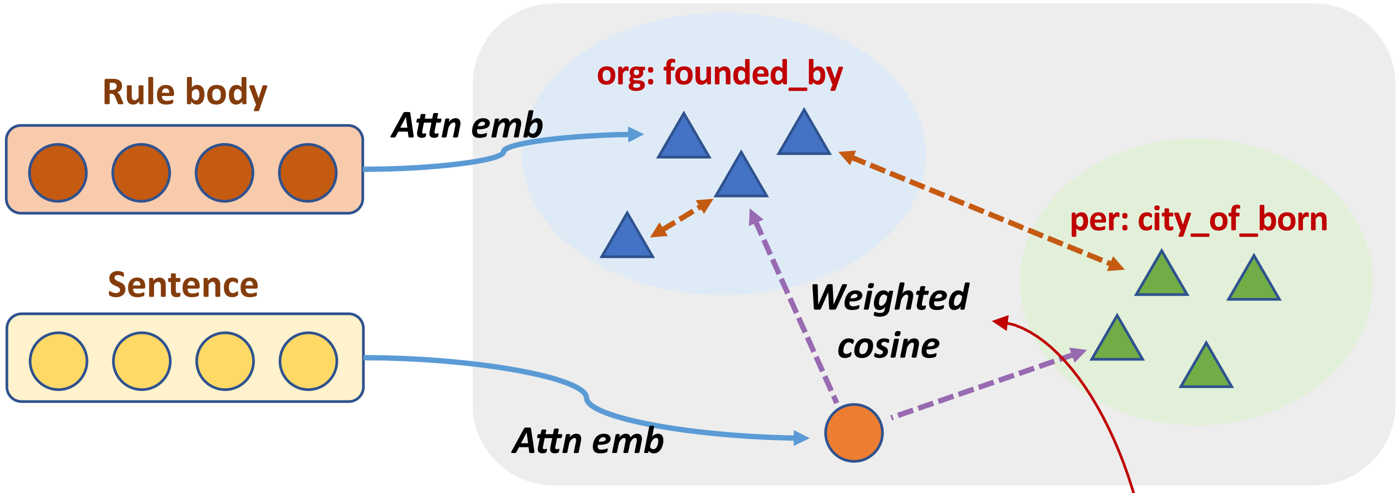
$$h_t = \mathrm{BiLSTM}(h_{t-1}, e_t)$$

$$s_t = v_h^T \tanh(W_h h_t)$$

$$a_t = \frac{\exp(s_t)}{\sum_{i=1}^{n} \exp(s_i)}$$

$$c = \sum_{t=1}^{n} a_t h_t$$

Learning a *Soft Rule Matching* Function



$$l_{sim} = \max_{p_1 \in P_+} L_+(p, p_1) + \max_{p_2 \in P_-} L_-(p, p_2)$$

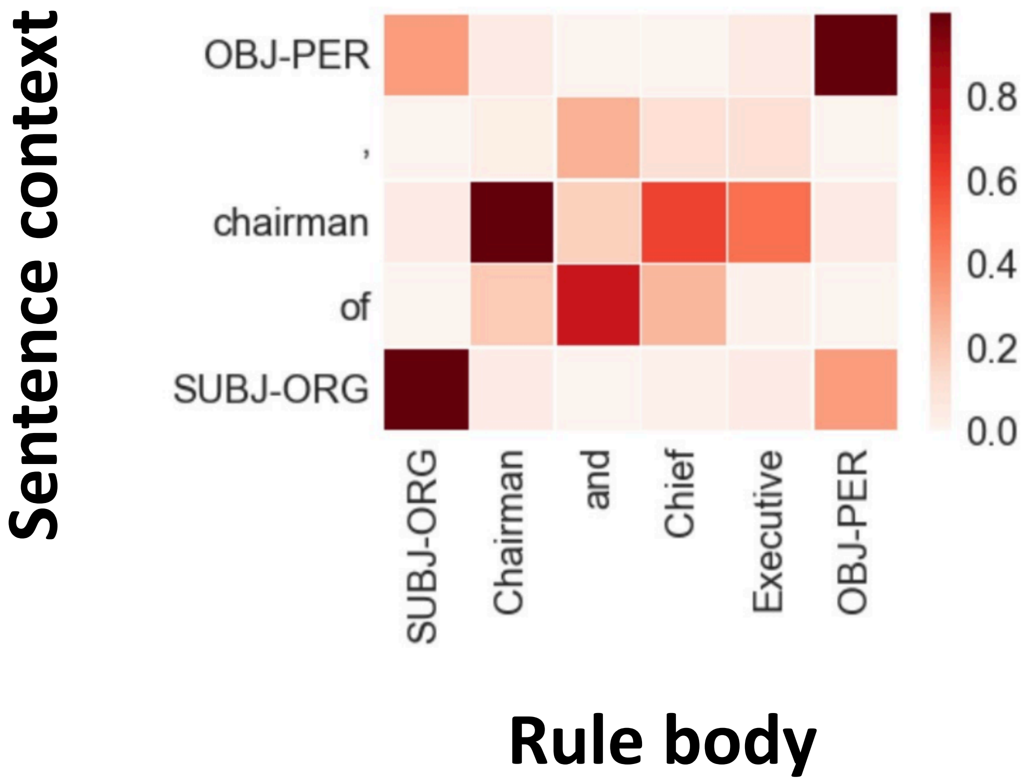
$$L_+ = \left(\tau_+ - f(p, p_1)\right)_+^2$$

$$L_- = \left(f(p, p_2) - \tau_-\right)_+^2$$

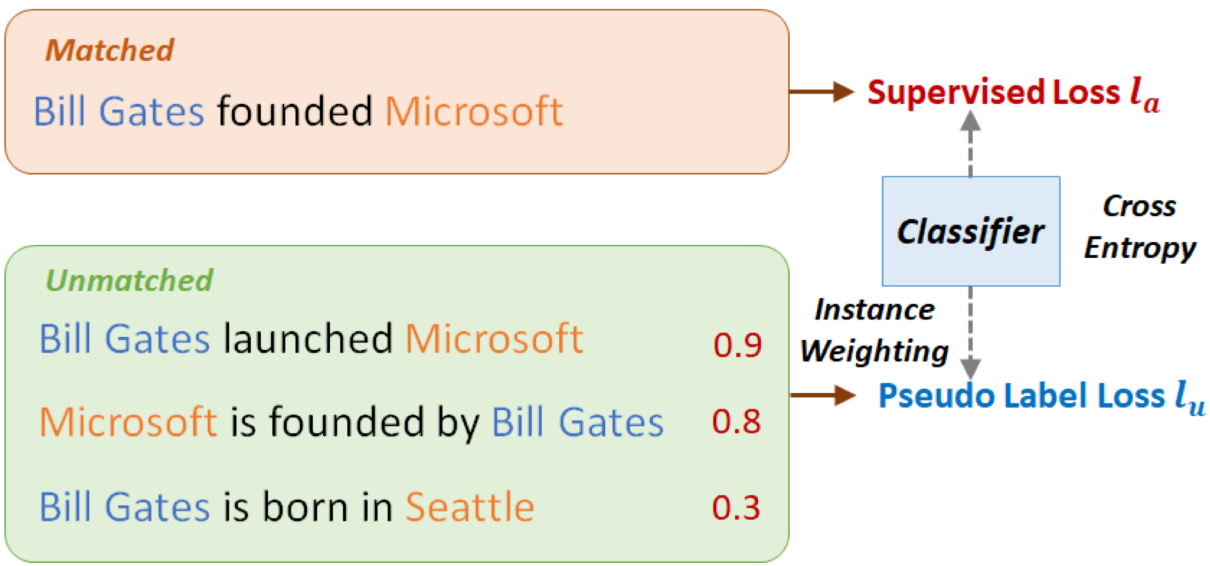
$$f_s(W_1, W_2) = \frac{z_1^T D^T D z_2}{\|z_1 D\| \|z_2 D\|}$$

(Zhou et al., 2019)

Interpretable Soft Rule Matching



REGD: Soft Rule Matching for Semi-supervised Learning



$$u_s = f(s, p^*)$$

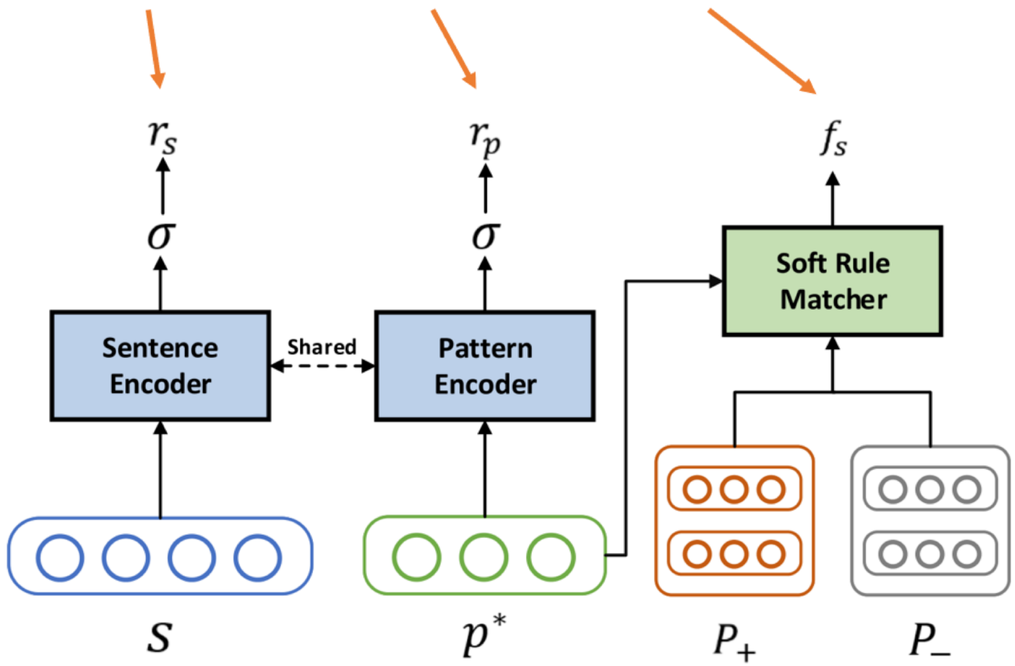
$$w_s = \frac{\exp(\theta u_s)}{\sum_{i=1}^{N_b} \exp(\theta u_i)}$$

$$l_u = -\sum_{i=1}^{n} w_i \cdot \log p(r'_s | s)$$

Assign each unmatched sentence a pseudo label and weight by soft matching.

REGD: Soft Rule Matching for Semi-supervised Learning

$$l = \overbrace{l_a + \alpha \cdot l_{pat} + \beta \cdot l_{sim}}^{\textit{supervised learning}} + \gamma \cdot l_u$$

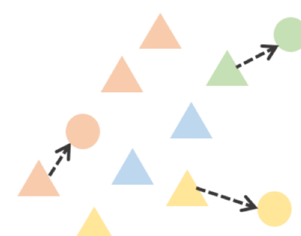


(Zhou et al., 2019)

REGD: Soft Rule Matching for Semi-supervised Learning

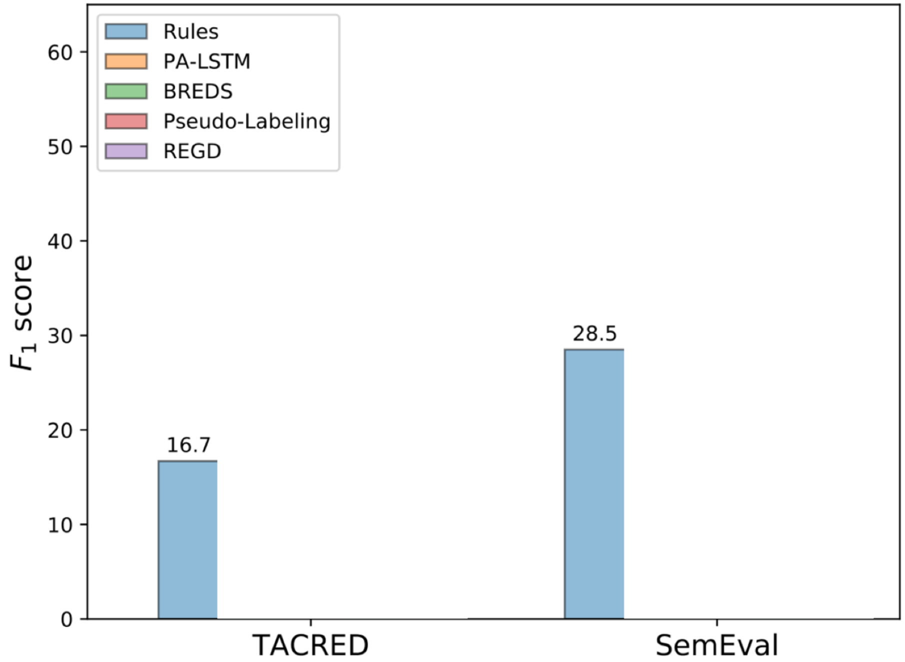
$$l = \overbrace{l_a + \alpha \cdot l_{pat} + \beta \cdot l_{sim}}^{\textit{supervised learning}} + \gamma \cdot l_u$$

trained on S_u: pseudo-labeling



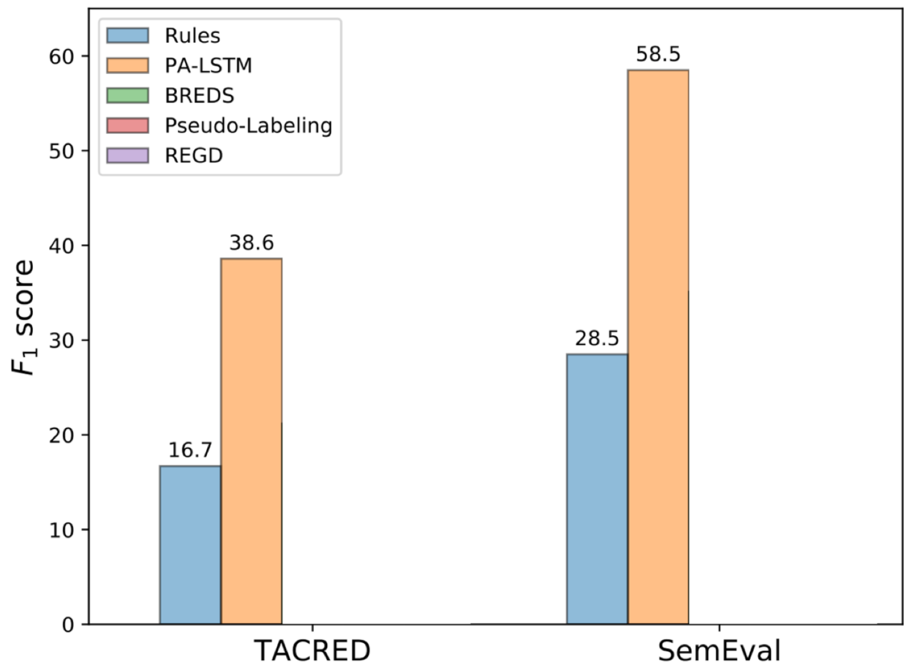
(Zhou et al., 2019)

Performance Comparison



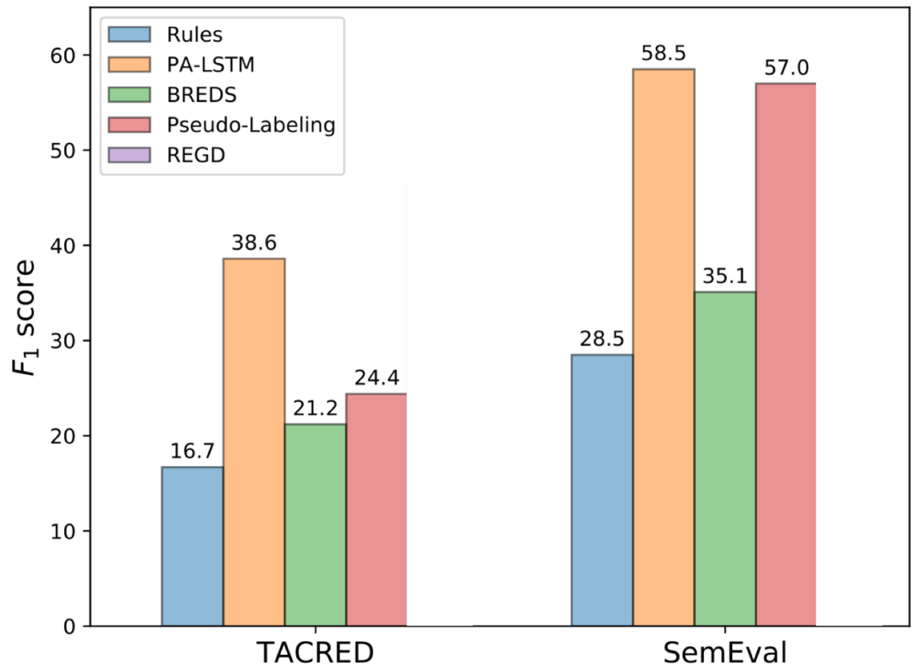
Rules have the highest precision (>80%) but lowest F1

Performance Comparison



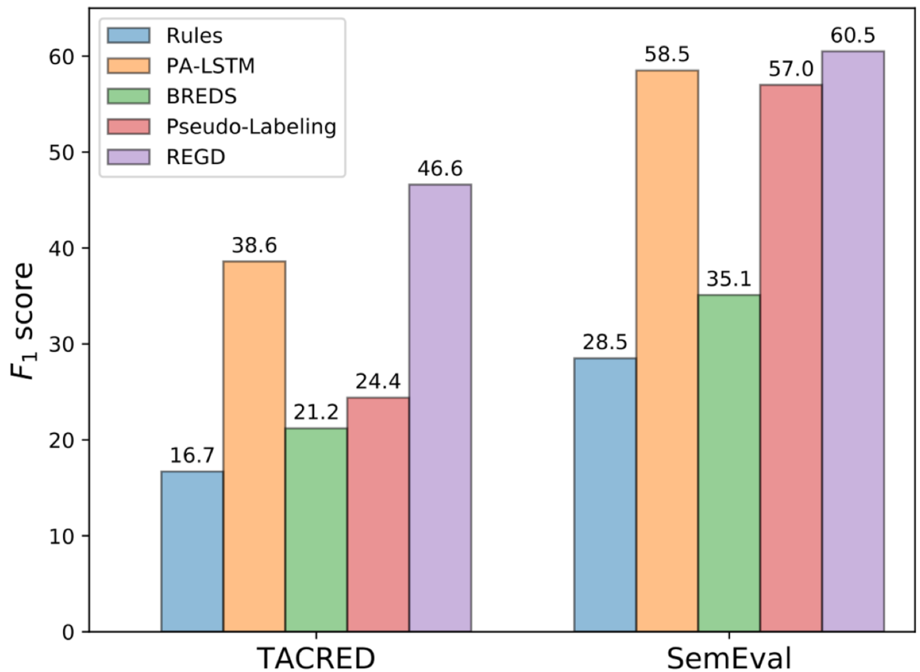
Supervised DL models generalize better than rules

Performance Comparison



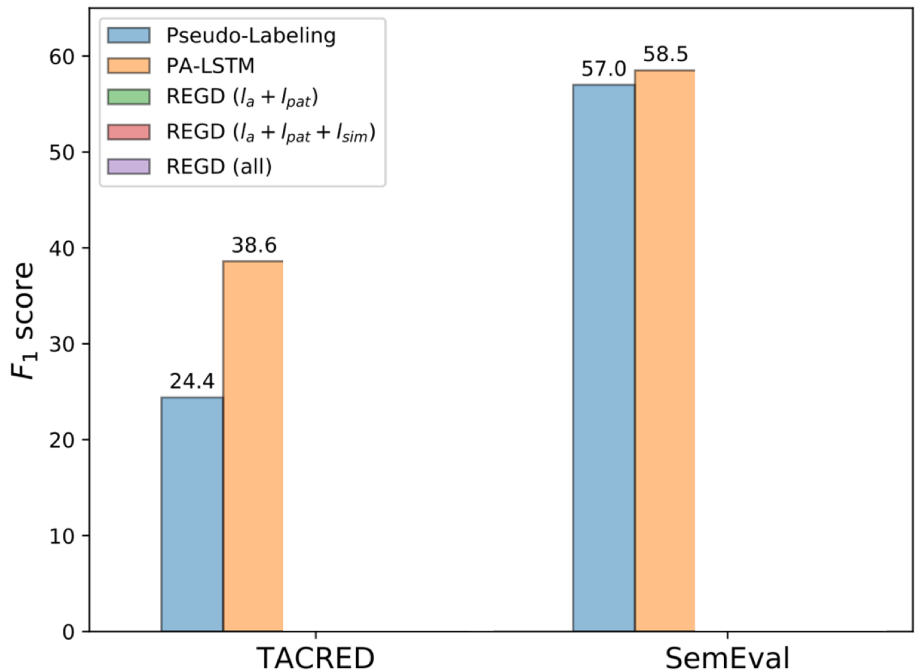
Semi-supervised models perform extremely bad since labeled data are scarce

Performance Comparison



REGD outperforms the competing baselines

Ablation on Components



Base models: PA-LSTM is equivalent to REGD with l_a only; Pseudo-Labeling is similar to adding l_u to supervised model.

Predicting on New Relations

- Apply soft rule matching to new relations with *unseen rules*

	TACRED			SemEval		
Method	P	R	F_1	P	R	F_1
Rule (exact match)	100	6.1	10.8	83.2	17.7	28.2
CBOW-GloVe	52.4	86.3	64.7	40.3	45.5	34.7
BERT	66.2	76.8	**69.5**	37.8	33.2	35.3
REGD	61.4	80.5	68.9	43.0	54.1	**45.5**

KagNet: Learning to Answer Commonsense Questions with *Knowledge-aware Graph Networks*

Joint work with Bill Lin & Jamin Chen, under submission

What is Commonsense Reasoning?

- Naïve Physics
  - Humans' **natural understanding of the physical world**
  - The ***trophy*** would not fit in the brown ***suitcase*** because **it** was too **big**. What was too **big**?
- Folk Psychology
  - Humans' **innate ability to reason about people's behavior and intentions**
  - *Person A puts his trust in Person B*, because _____ ? . (A and B are friends.)

- How can we evaluate the commonsense reasoning capacity of an NLU model?
  - Recent textual multi-choice QA datasets:
    - `CommonsenseQA` (Talmor et al. NAACL 2019)
    - `CommonsenseNLI(SWAG & HellaSwag,` Zellers et al. 2018, 2019)
    - `SocialIQA` (Sap et al. 2019)

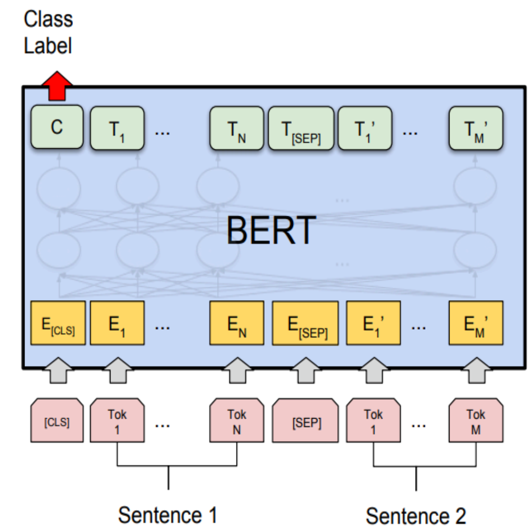
CommonsenseQA dataset (Talmor et al. 2019)

Where would I not want a fox?
hen house, england, mountains, english hunt, california

Why do people read gossip magazines?
entertained, get information, learn, improve know how, lawyer told to

What do all humans want to experience in their own home?
feel comfortable, work hard, fall in love, lay eggs, live forever

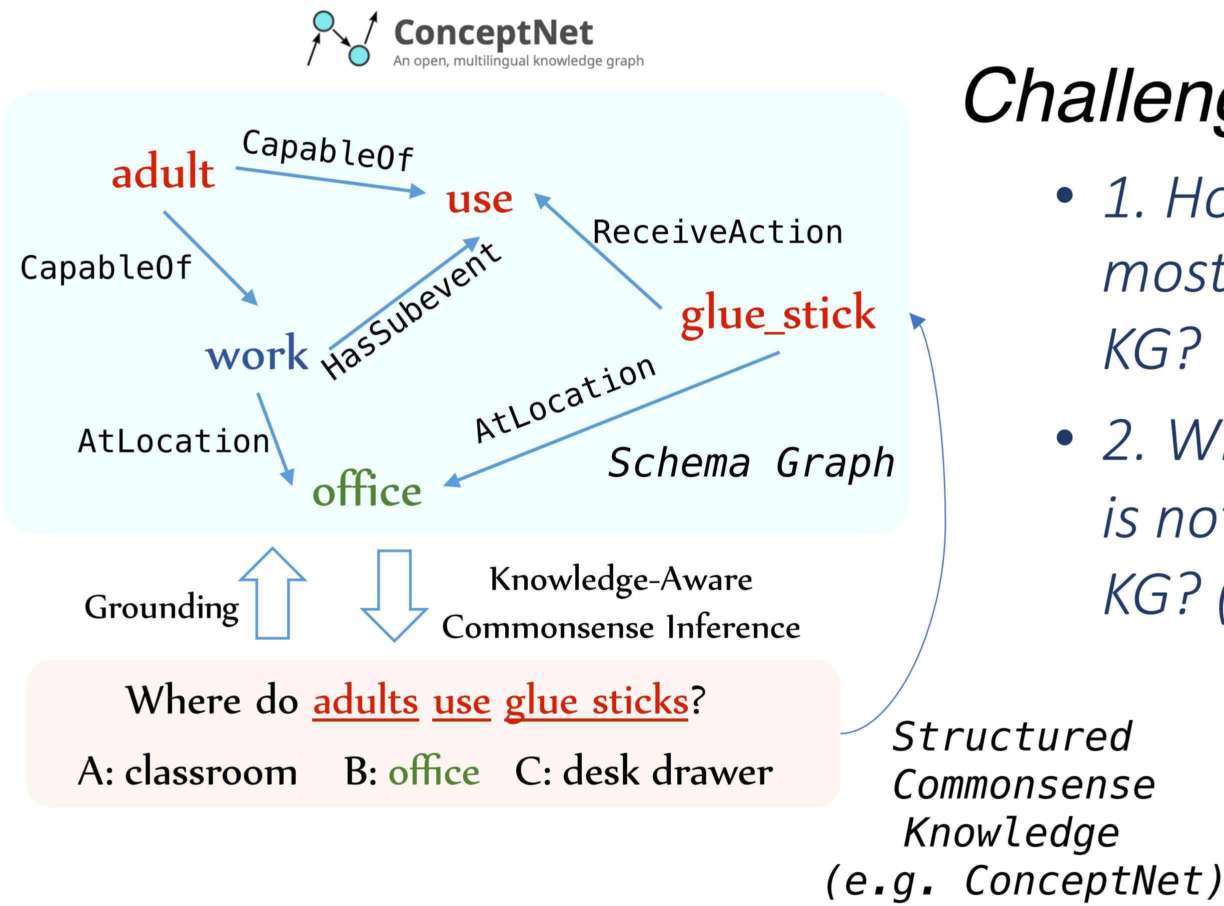
State-of-the-art Model: Fine-tuning BERT-based classifiers



(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG

https://www.tau-nlp.org/commonsenseqa

Our Idea: Imposing External Knowledge

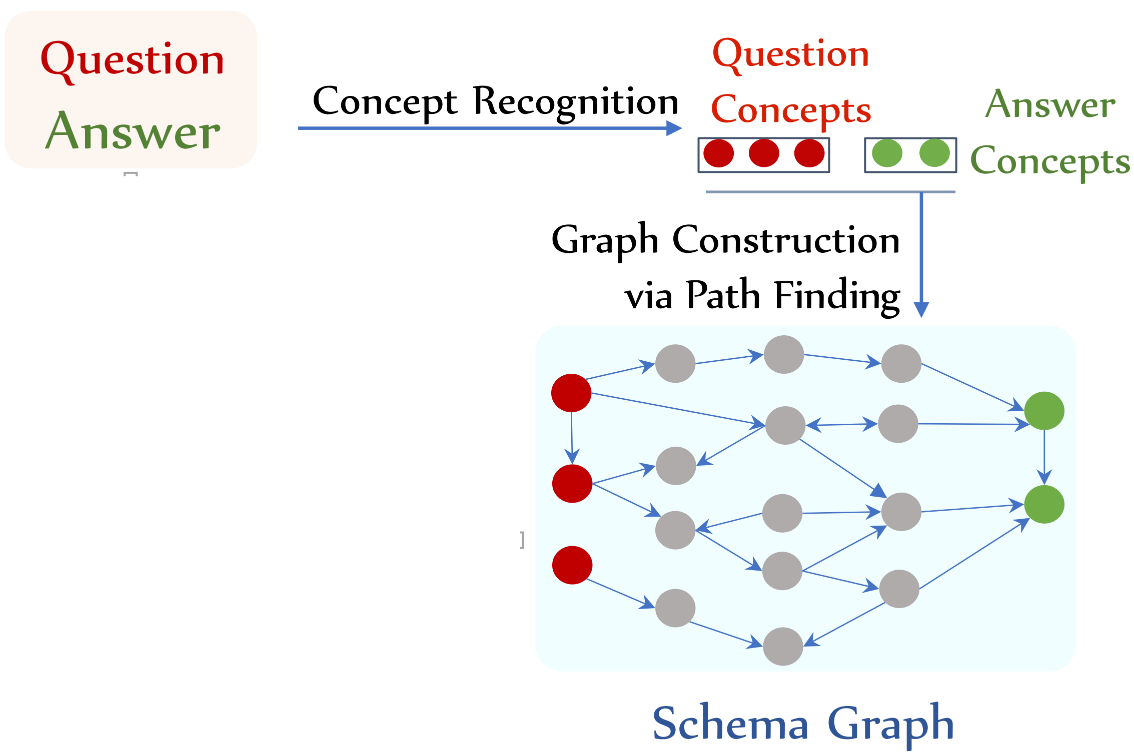


Challenges:

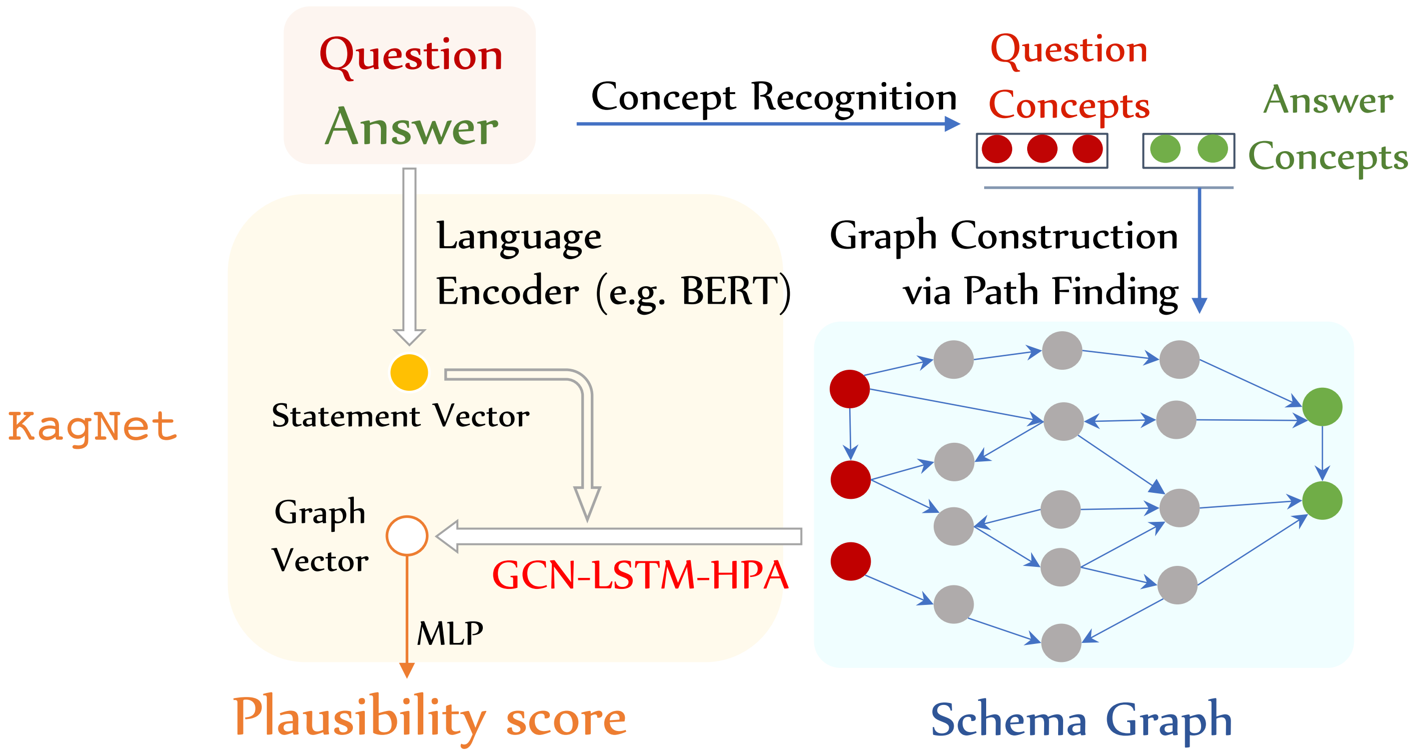
- *1. How can we find the most relevant paths in KG? (**noisy**)*
- *2. What if the best path is not existent in the KG? (**incomplete**)*

KagNet: Knowledge-Aware Graph Networks

KagNet: Knowledge-Aware Graph Networks

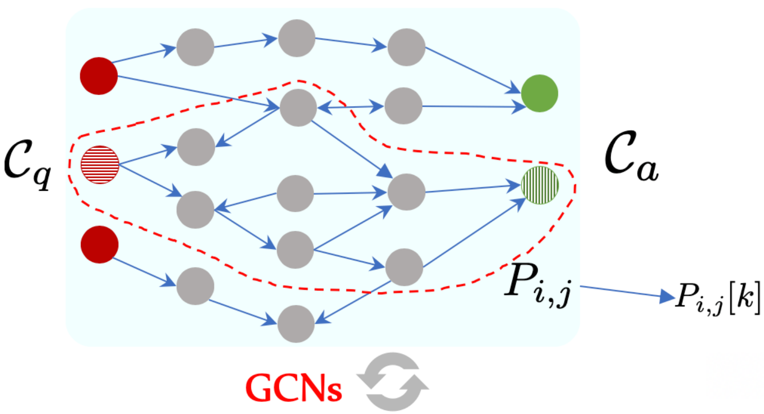


KagNet: Knowledge-Aware Graph Networks

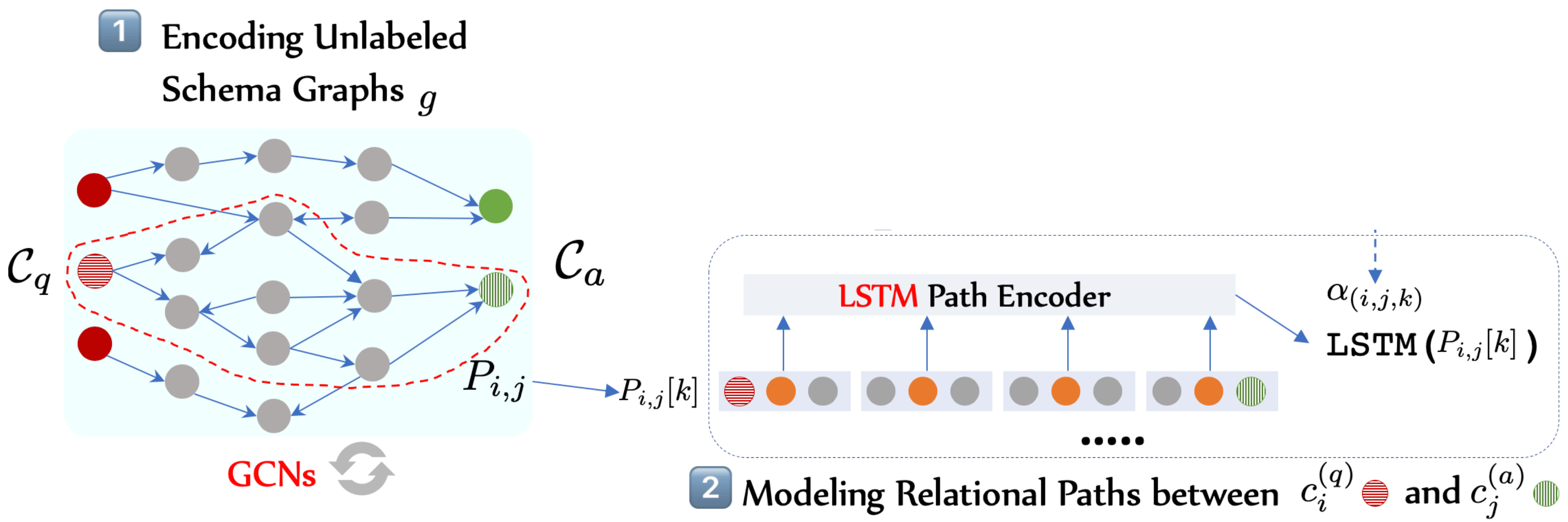


The GCN-LSTM-HPA Architecture

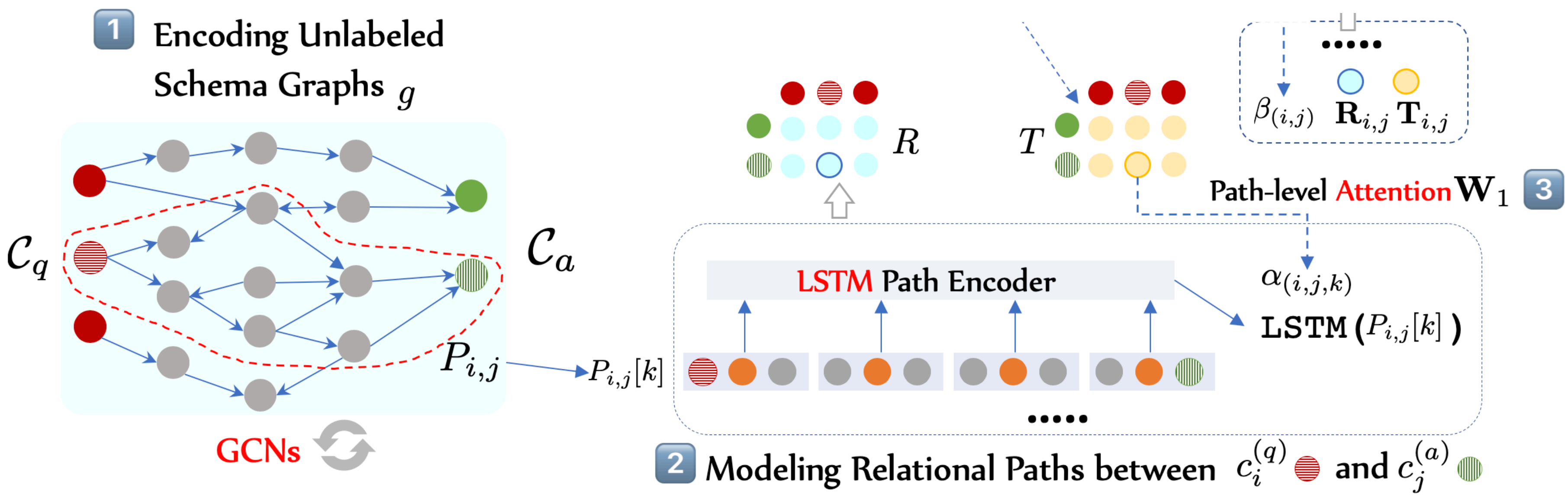
1 Encoding Unlabeled Schema Graphs g



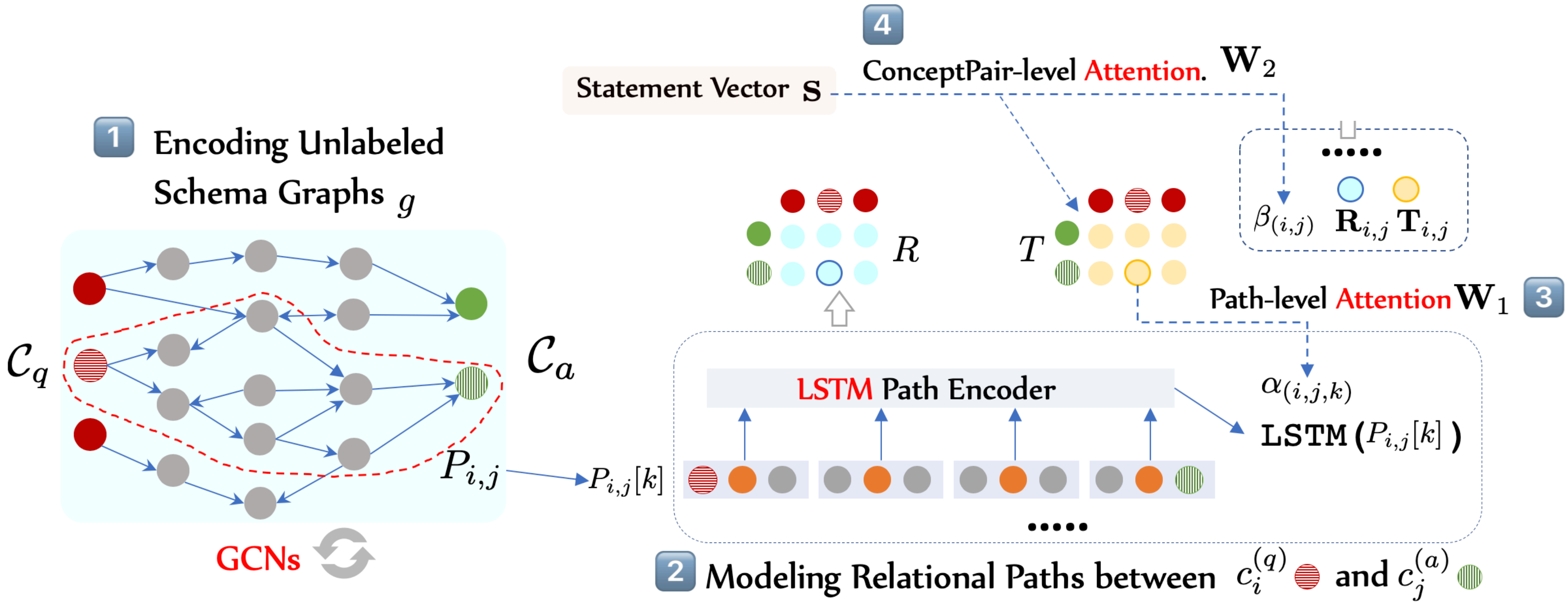
The GCN-LSTM-HPA Architecture



The GCN-LSTM-HPA Architecture



The GCN-LSTM-HPA Architecture



KagNet with Different Base Models & Trained on Varying Amounts of Data

	10(%) of IHtrain		**50**(%) of IHtrain		**100**(%) of IHtrain	
Model	IHdev-Acc.(%)	IHtest-Acc.(%)	IHdev-Acc.(%)	IHtest-Acc.(%)	IHdev-Acc.(%)	IHtest-Acc.(%)
Random guess	20.0	20.0	20.0	20.0	20.0	20.0
GPT-FINETUNING	27.55	26.51	32.46	31.28	47.35	45.58
GPT-**KAGNET**	28.13	**26.98**	33.72	**32.33**	48.95	**46.79**
BERT-BASE-FINETUNING	30.11	29.78	38.66	36.83	53.48	53.26
BERT-BASE-**KAGNET**	31.05	**30.94**	40.32	**39.01**	55.57	**56.19**
BERT-LARGE-FINETUNING	35.71	32.88	55.45	49.88	60.61	55.84
BERT-LARGE-**KAGNET**	36.82	**33.91**	58.73	**51.13**	62.35	**57.16**
Human Performance	-	*88.9*	-	*88.9*	-	*88.9*

Result on CommonsenseQA Leaderboard (as of 5/14)

Version 1.11 Random Split Leaderboard

(12,102 examples with 5 answer choices)

Model	Affiliation	Date	Accuracy
Human		03/10/2019	88.9
KagNet	Anonymous	05/14/2019	58.9
CoS-E	Anonymous	04/12/2019	58.2
SGN-lite	Anonymous	04/20/2019	57.1
BERTLarge	Tel-Aviv University	03/10/2019	56.7
BERTBase	University College London	03/13/2019	53.0
BERTBase	University of Melbourne	04/22/2019	52.6
GPT	Tel-Aviv University	03/10/2019	45.5
ESIM+GLOVE	Tel-Aviv University	03/10/2019	34.1
ESIM+ELMO	Tel-Aviv University	03/10/2019	32.8

https://www.tau-nlp.org/csqa-leaderboard

Knowledge-Injection Baseline Methods

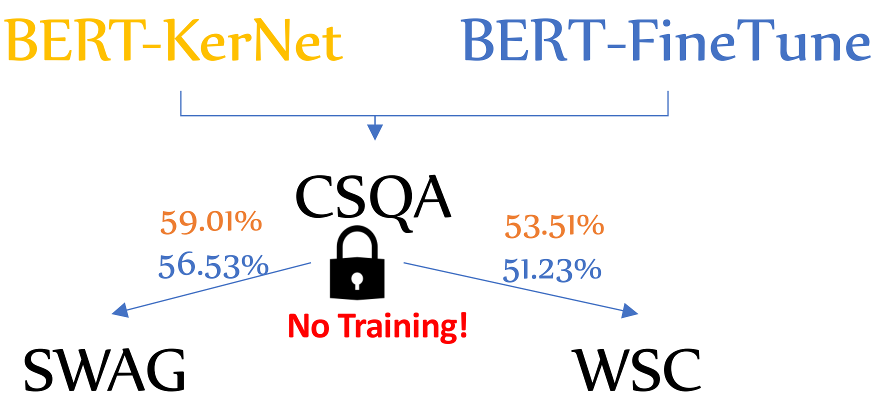
	Easy Mode		Hard Mode	
Model	IHdev.(%)	IHtest.(%)	IHdev.(%)	IHtest.(%)
Random guess	33.3	33.3	20.0	20.0
BLSTMs	80.15	78.01	34.79	32.12
+ KV-MN	81.71	79.63	35.70	33.43
+ CSPT	81.79	80.01	35.31	33.61
+ TEXTGRAPHCAT	82.68	81.03	34.72	33.15
+ TRIPLESTRING	79.11	76.02	33.19	31.02
+ **KAGNET**	83.26	**82.15**	36.38	**34.57**
Human Performance	-	*99.5*	-	*88.9*

Table 3: Comparisons with knowledge-aware baseline methods using the in-house split (both easy and hard mode) on top of BLSTM as the sentence encoder.

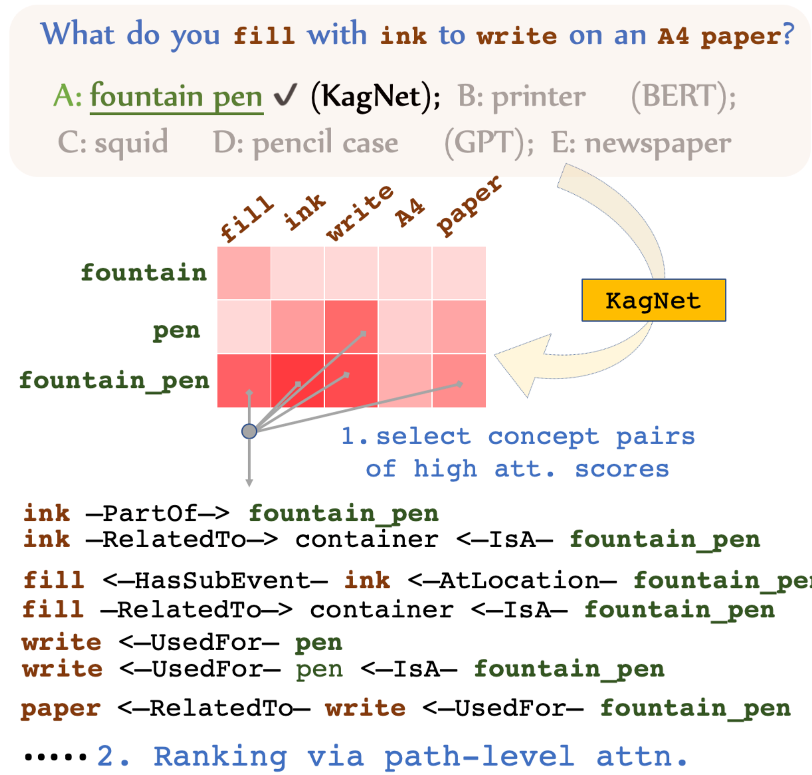
Model	IHdev.(%)	IHtest.(%)
KAGNET (STANDARD)	62.35	57.16
: replace GCN-HPA-LSTM w/ R-GCN	60.01	55.08
: w/o GCN	61.84	56.11
: #GCN Layers = 1	62.05	57.03
: w/o Path-level Attention	60.12	56.05
: w/o QAPair-level Attention	60.39	56.13
: using all paths (w/o pruning)	59.96	55.27

Table 4: Ablation study on the KAGNET framework.

Transferability



Interpretability



Summary

- Learnings
  - Where to solicit complex rules?
  - Coverage of KG grounding; completeness of KG
  - Scalability
- Some open problems
  - Inducing transferrable, latent structures from pre-trained models
  - Modular network for modeling compositional rules
  - Modeling “human efforts” in the objective

Community

- Deep Learning for Low-resource NLP (DeepLo): ACL 2018, EMNLP 2019
- Learning on Limited Data (LLD) Workshop: NeurIPS 2018, ICLR 2019
- Automated Knowledge Base Construction (AKBC)
- Open-source tools
  - **DS-RelationExtraction**: a suite of base models for relation extraction & distantly-supervised learning techniques https://github.com/INK-USC/DS-RelationExtraction
  - **AutoNER toolkit**: multiple training options (distant training, LM-augmentation, etc.) for building sequence taggers https://github.com/shangjingbo1226/AutoNER
- PubMed literature search powered by an auto-constructed, open knowledge graph

http://usc.edu/life-inet

Life-iNet

Students

Bill Lin

Priya Irukulapati

Woojeong Jin

Wenxuan Zhou

Collaborators

Jure Leskovec, Computer Science, Stanford University
Dan MacFarland, Sociology, Stanford University
Dan Jurafsky, Computer Science, Stanford University
Jiawei Han, Computer Science, UIUC
Kennth Yates, Clinical Education, USC
Craig Knoblock, USC ISI
Curt Langlotz, Bioinformatics, Stanford University
Heng Ji, Computer Science, UIUC
Kuansan Wang, Microsoft Academic
Xiaolin Shi, Snapchat
Mark Musen, Bioinformatics, Stanford University

Research Partnerships

Funding

Thank You!

- Injecting structured prior knowledge into various knowledge extraction tasks: input level vs. model level
- Aim to lower the reliance on traditional human-annotated data
- Learnings:
  - Where to solicit complex rules?
  - Coverage of KG grounding; completeness of KG
  - Scalability of computational models
- Technology Transfer: