Fast Learning with Explanation and Prior Knowledge

Sean (Xiang) Ren

Department of Computer Science

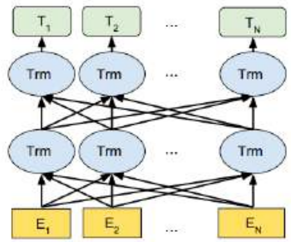
Information Science Institute

USC

http://inklab.usc.edu

Recipe for Modern NLP Applications

Model



+

Labeled Data

SWAG

SQuAD
The Stanford Question Answering Dataset

GLUE

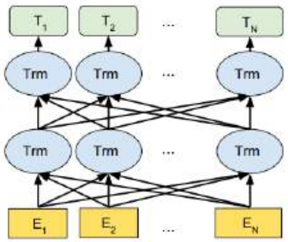
+

Computing Power

?

Recipe for Modern NLP Applications

Model

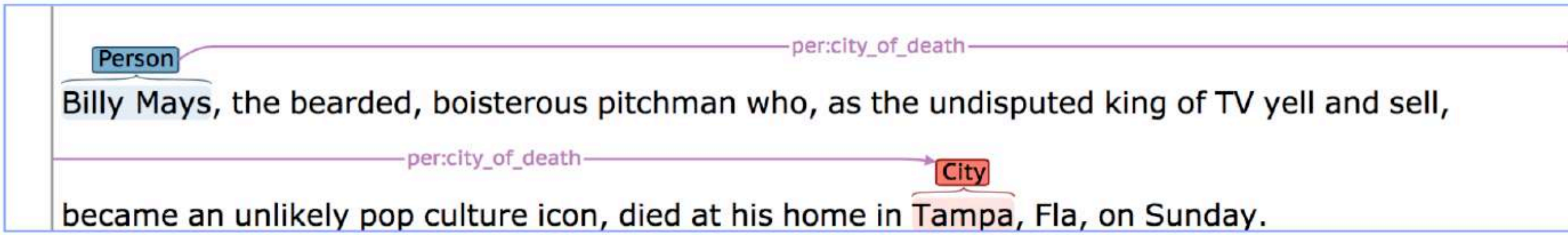


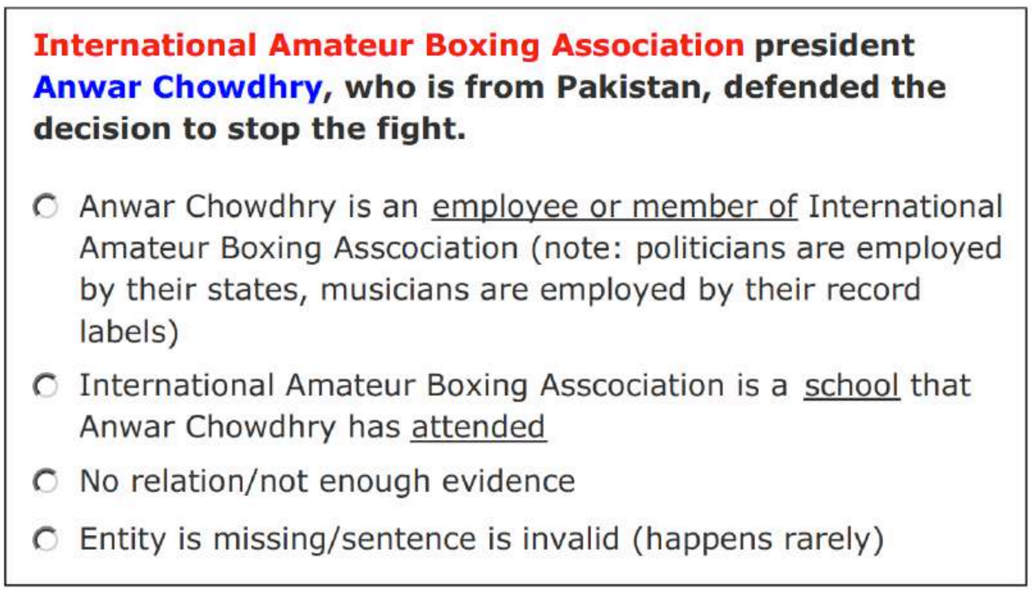
Model architectures and computing power are transferrable across applications

labeled data is not!

Computing Power

Creating Labeled Data for Relation Extraction



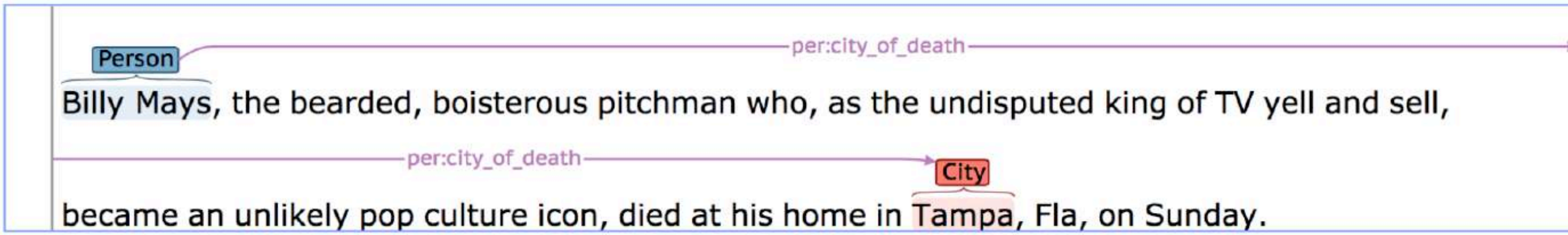

International Amateur Boxing Association president Anwar Chowdhry, who is from Pakistan, defended the decision to stop the fight.

- Anwar Chowdhry is an employee or member of International Amateur Boxing Asscociation (note: politicians are employed by their states, musicians are employed by their record labels)
- International Amateur Boxing Asscociation is a school that Anwar Chowdhry has attended
- No relation/not enough evidence
- Entity is missing/sentence is invalid (happens rarely)

TACRED dataset: 106k labeled instances for 41 relations, crowd-sourced via Amazon Mechanical Turk

(Zhang et al., 2018)

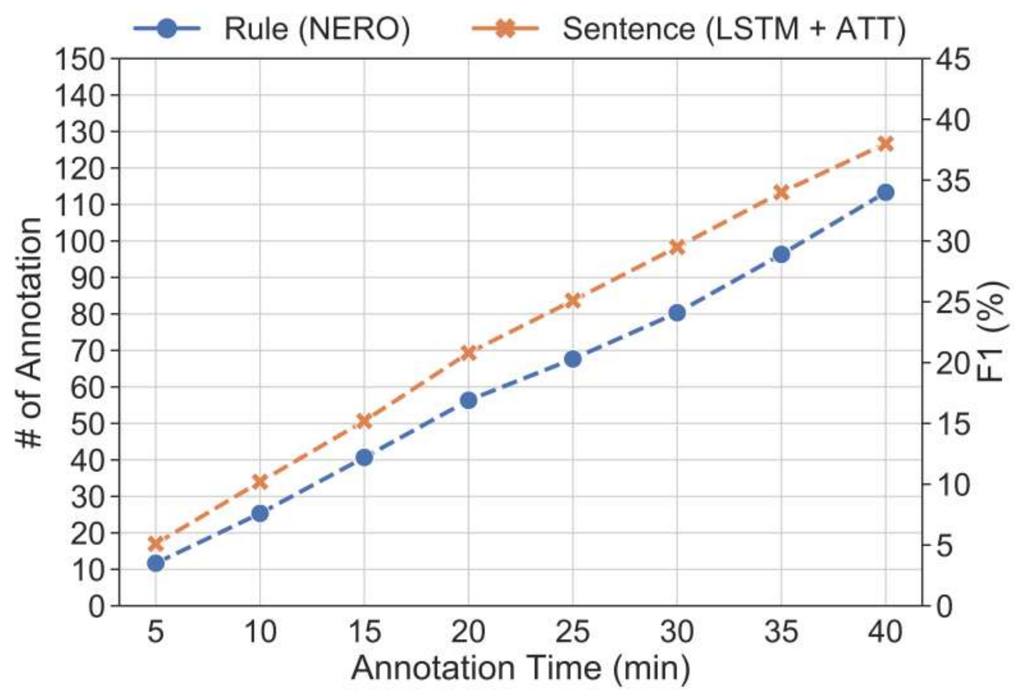
Creating Labeled Data for Relation Extraction



Cost on Amazon Mechanical Turk: $0.5 per instance → $53k!

Time cost: ~20 second per instance → 7+ days

(Zhou et al., WWW20)



Labeled data for more complex tasks

SQuAD
The Stanford Question Answering Dataset

Paragraph 1 of 43

Spend around 4 minutes on the following paragraph to ask 5 questions! If you can't ask 5 questions, ask 4 or 3 (worse), but do your best to ask 5. Select the answer from the paragraph by clicking on 'Select Answer', and then highlight the smallest segment of the paragraph that answers the question.

Oxygen is a chemical element with symbol O and atomic number 8. It is a member of the chalcogen group on the periodic table and is a highly reactive nonmetal and oxidizing agent that readily forms compounds (notably oxides) with most elements. By mass, oxygen is the third-most abundant element in the universe, after hydrogen and helium. At standard temperature and pressure, two atoms of the element bind to form dioxygen, a colorless and odorless diatomic gas with the formula O
2. Diatomic oxygen gas constitutes 20.8% of the Earth's atmosphere. However, monitoring of atmospheric oxygen levels show a global downward trend, because of fossil-fuel burning. Oxygen is the most abundant element by mass in the Earth's crust as part of oxide compounds such as silicon dioxide, making up almost half of the crust's mass.

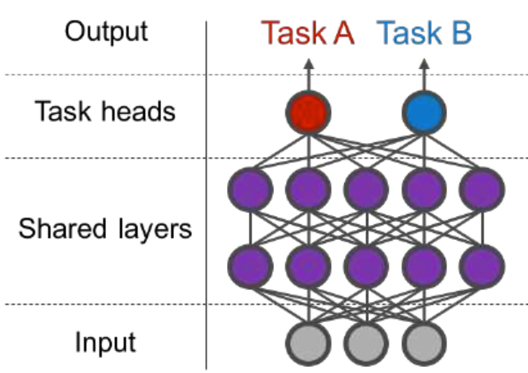
When asking questions, **avoid using** the same words/phrases as in the paragraph. Also, you are encouraged to pose **hard questions**.

Ask a question here. Try using your own words

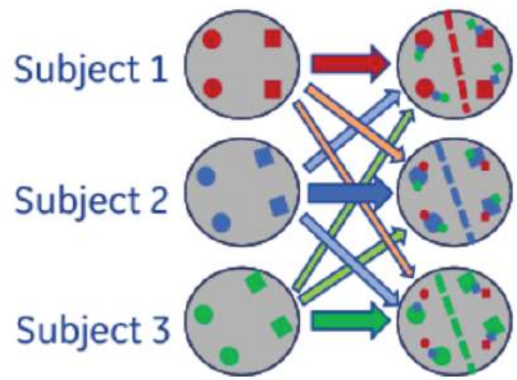
Select Answer

(Rajpurkar et al., 2018)

Towards faster learning (with less labels)

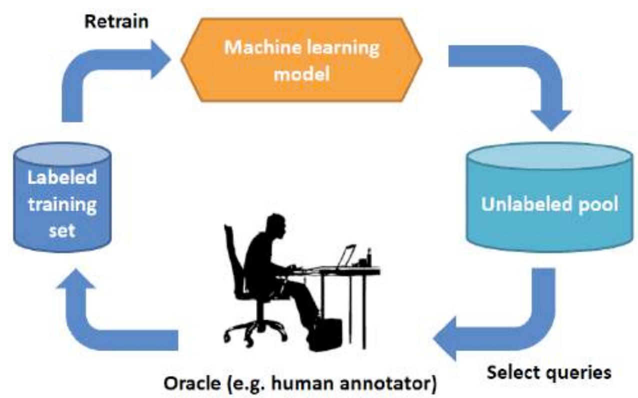


Multi-task Learning



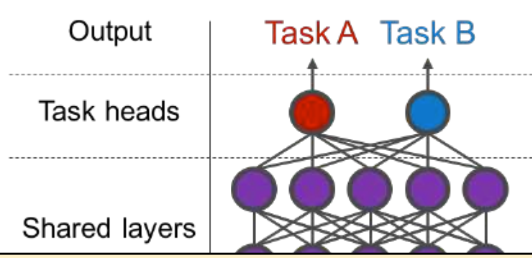
Transfer Learning

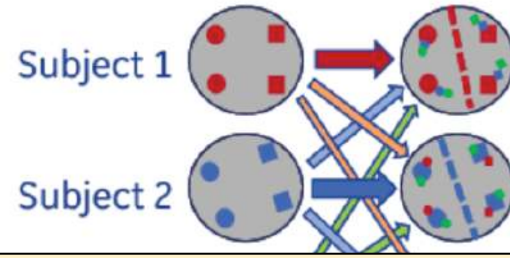
Distant Supervision



Active Learning

Towards faster learning (with less labels)





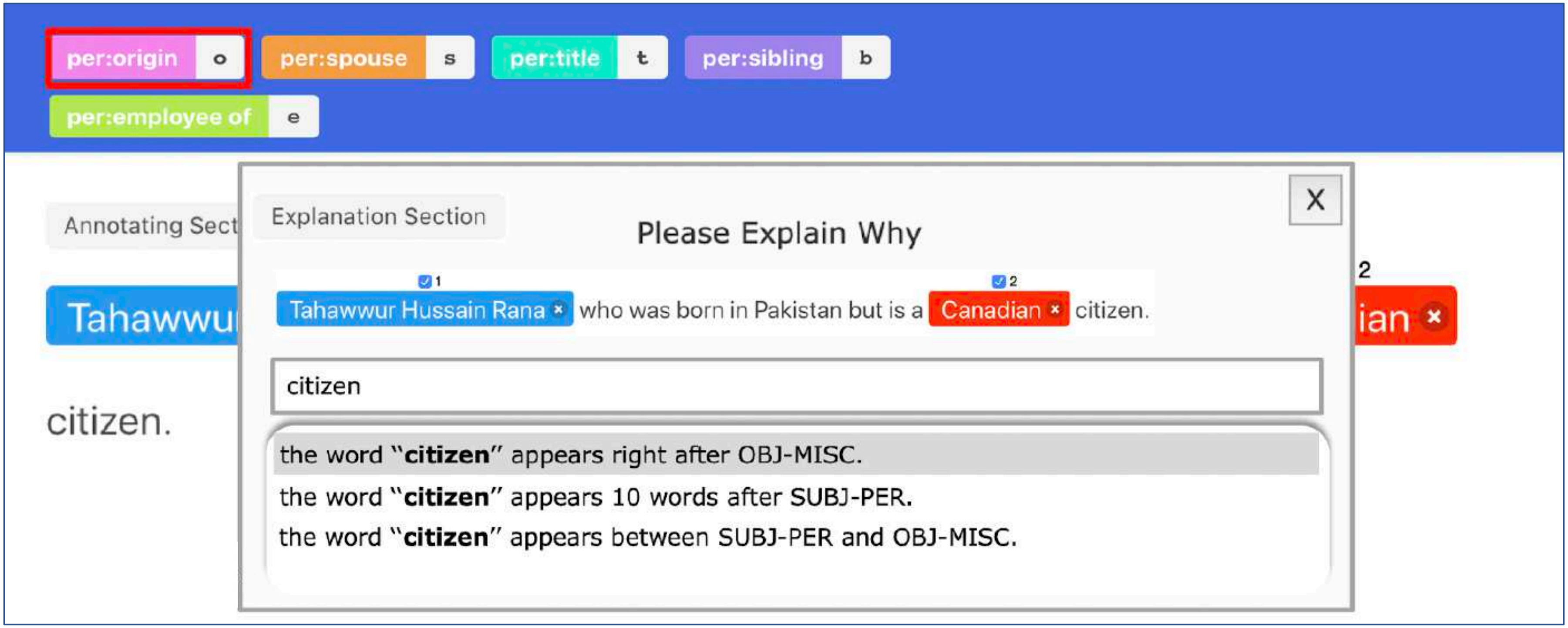
Challenges: availability of related data sources & strong assumptions on data distributions

Distant Supervision

Active Learning

Our Idea: High-level Human Supervisions

Our Idea: High-level Human Supervisions



Machine digests human rationale and learns how to make decisions

This Talk

Q1 How to augment model training with rules?

Soft rule grounding for data augmentation (Zhou et al. WWW20)

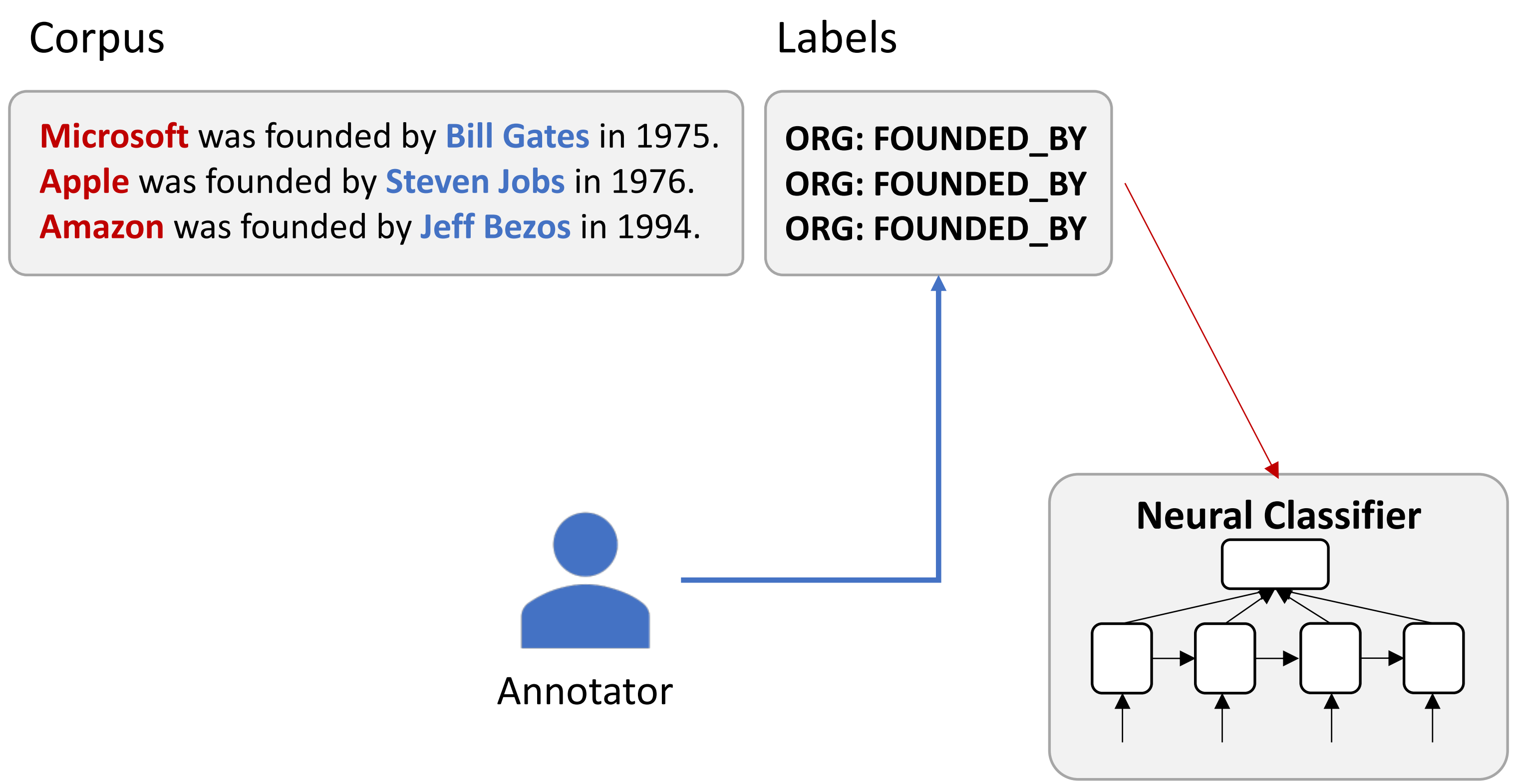
Q2 How to handle compositional natural language input?

Neural execution tree for NL explanation (Wang et al. ICLR20)

Q3 How to incorporate prior knowledge as inductive bias?

Knowledge-aware graph networks (Lin et al. EMNLP19)

Standard pipeline for data annotation



Slow, redundant annotation efforts on similar instances!

Alternative Labeling Scheme: Surface Pattern Rules

Corpus

Microsoft was founded by Bill Gates in 1975.
Apple was founded by Steven Jobs in 1976.
Amazon was founded by Jeff Bezos in 1994.

Labels

ORG: FOUNDED_BY
ORG: FOUNDED_BY
ORG: FOUNDED_BY

SUBJ-ORG was founded by **OBJ-PER** → **ORG: FOUNDED_BY**

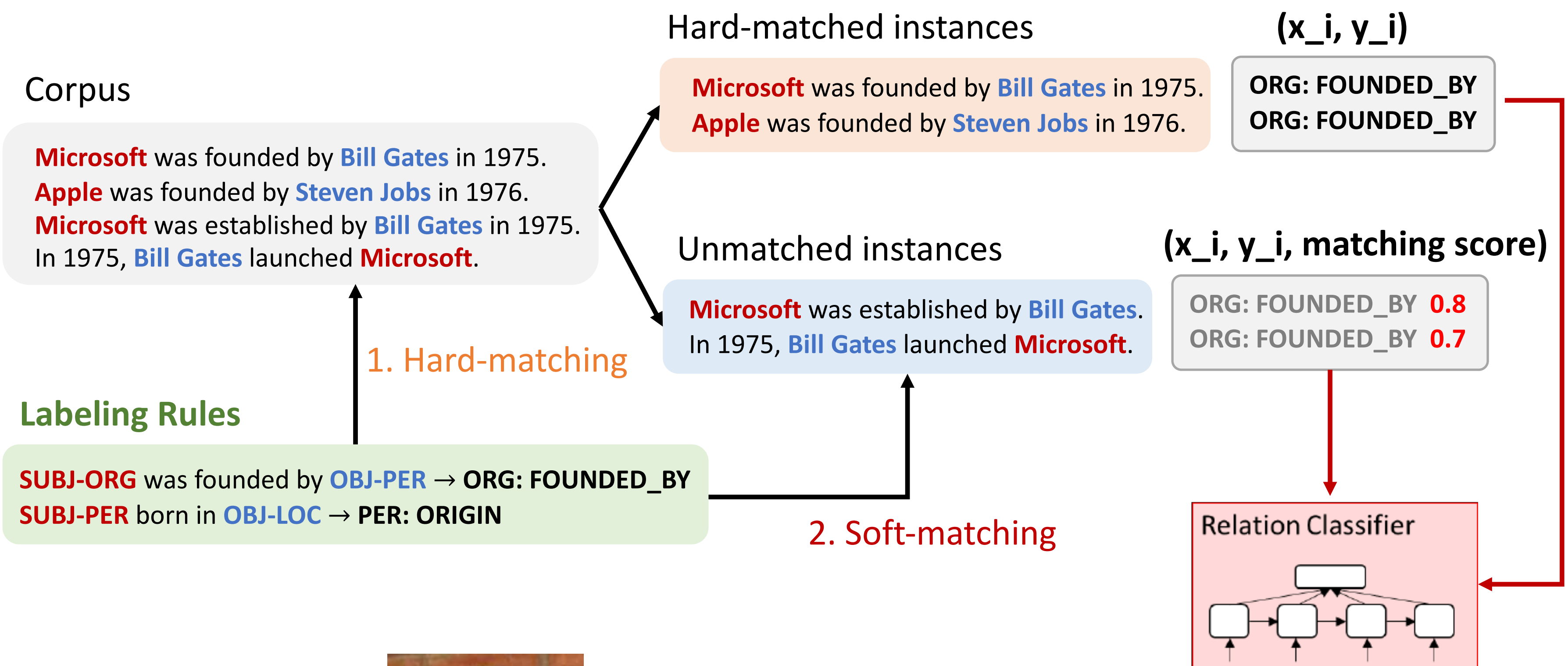
Annotator

Annotate contextually similar instances via much fewer rules!

(Hearst, 1992)

Neural Rule Grounding for Data Augmentation

Generalizing rule coverage via soft matching to instances



(Zhou et al, WWW20)

A Learnable, Soft Rule Matching Function

Unmatched instances

Microsoft was established by **Bill Gates**.
In 1975, **Bill Gates** launched **Microsoft**.

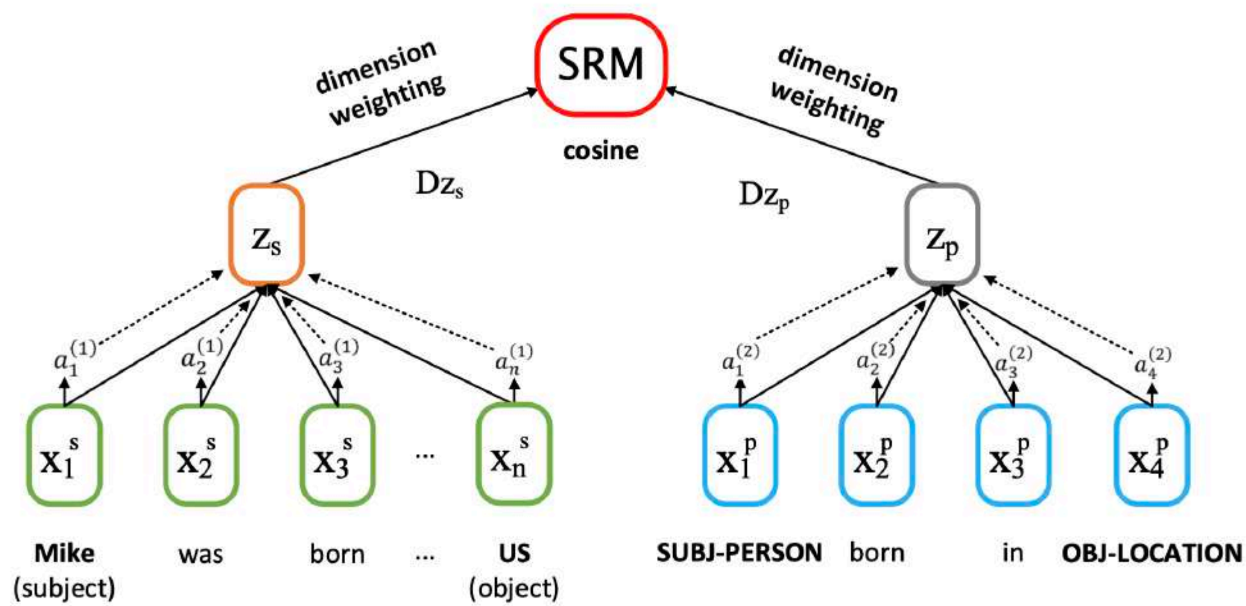
(x_i, y_i, matching score)

ORG: FOUNDED_BY **0.8**
ORG: FOUNDED_BY **0.7**

Labeling Rules

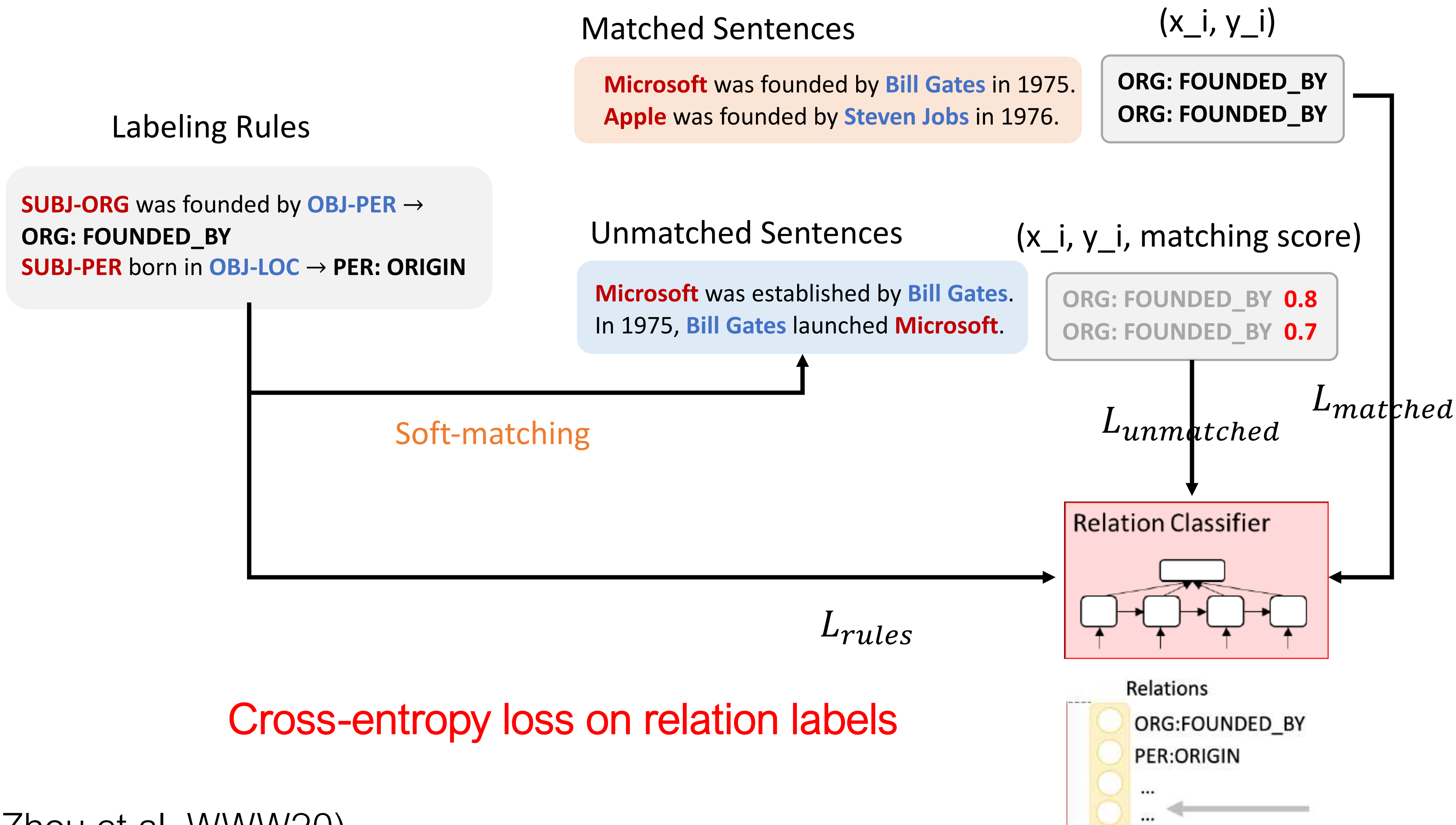
ENT1 was founded by **ENT2** → **ORG: FOUNDED_BY**
ENT1 born in **ENT2** → **PER: ORIGIN**

2. Soft-matching



(Zhou et al, WWW20)

Joint Parameter Learning: **Relation Extractor** + Soft Rule Matcher

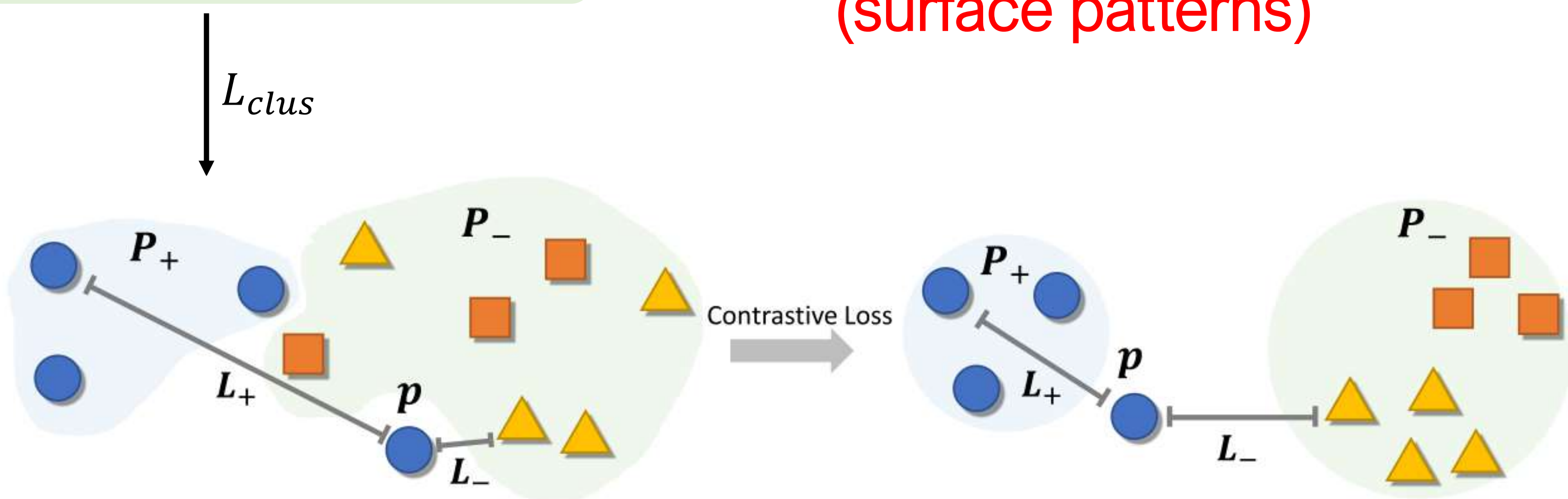


(Zhou et al, WWW20)

Joint Parameter Learning: Relation Extractor + **Soft Rule Matcher**

Labeling Rules

ENT1 was founded by **ENT2** → **ORG: FOUNDED_BY**
ENT1 born in **ENT2** → **PER: ORIGIN**

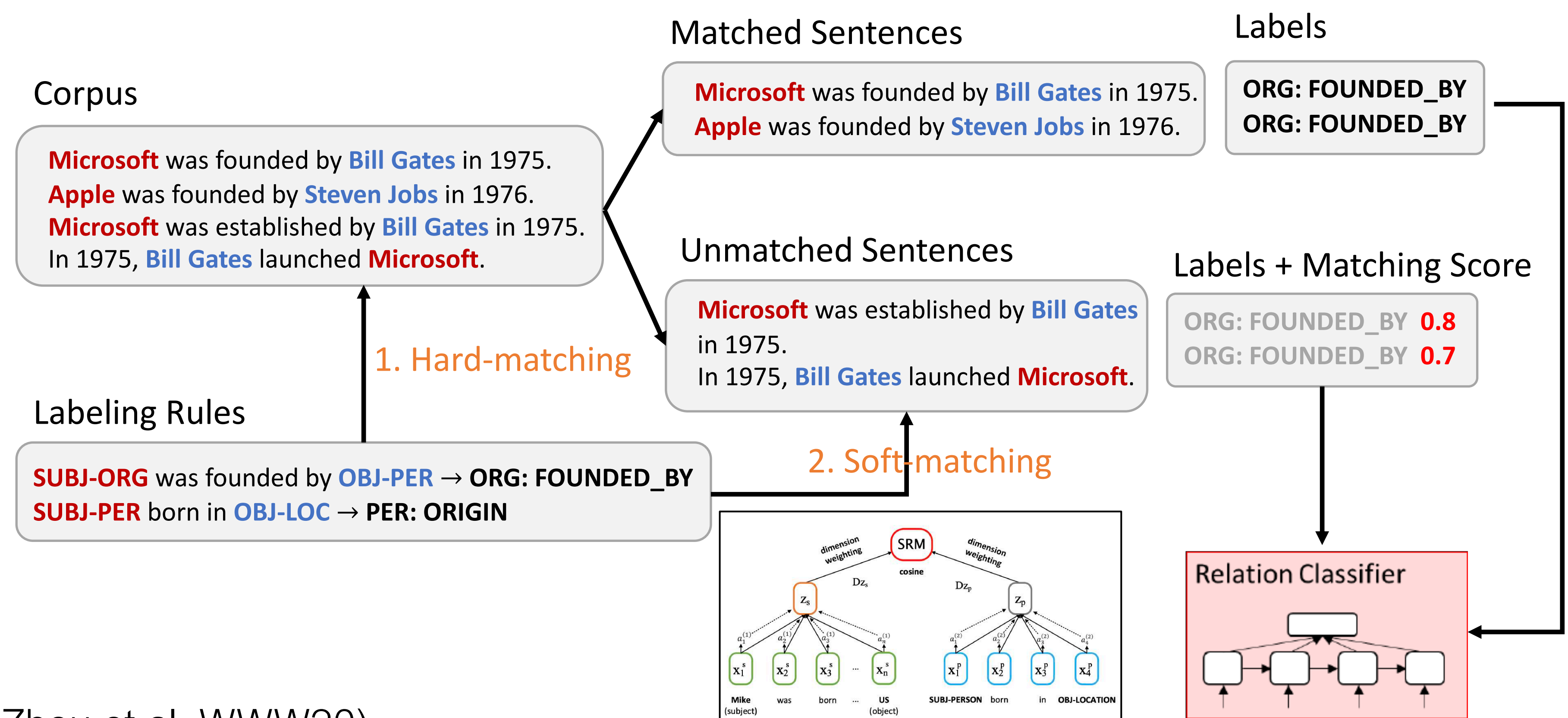


Contrastive loss for discriminating by rule bodies (surface patterns)

(Zhou et al, WWW20)

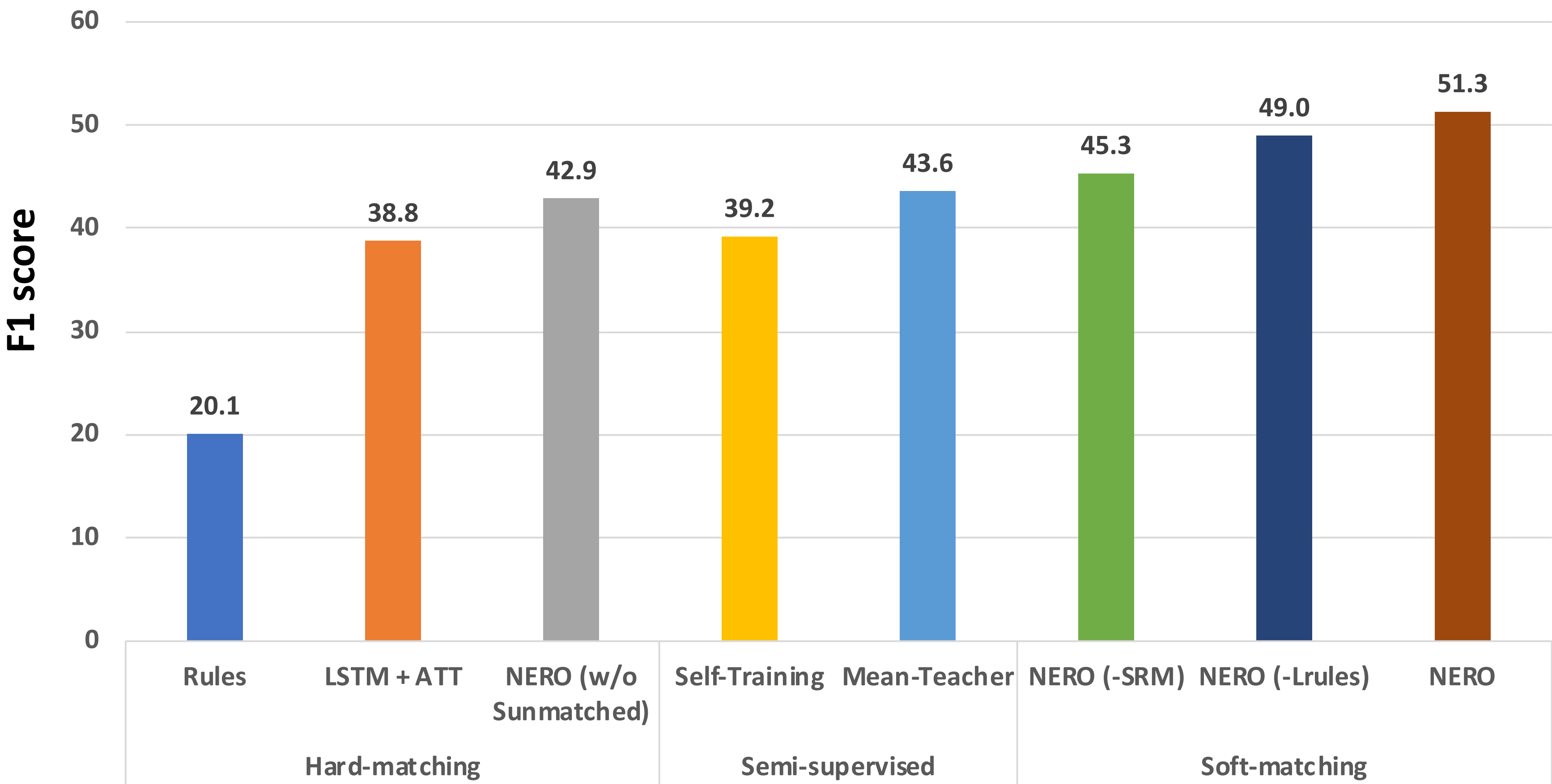
Joint Parameter Learning: **Relation Extractor + Soft Rule Matcher**

$$L = L_{matched} + \alpha \cdot L_{unmatched} + \beta \cdot L_{rules} + \gamma \cdot L_{clus}$$



(Zhou et al, WWW20)

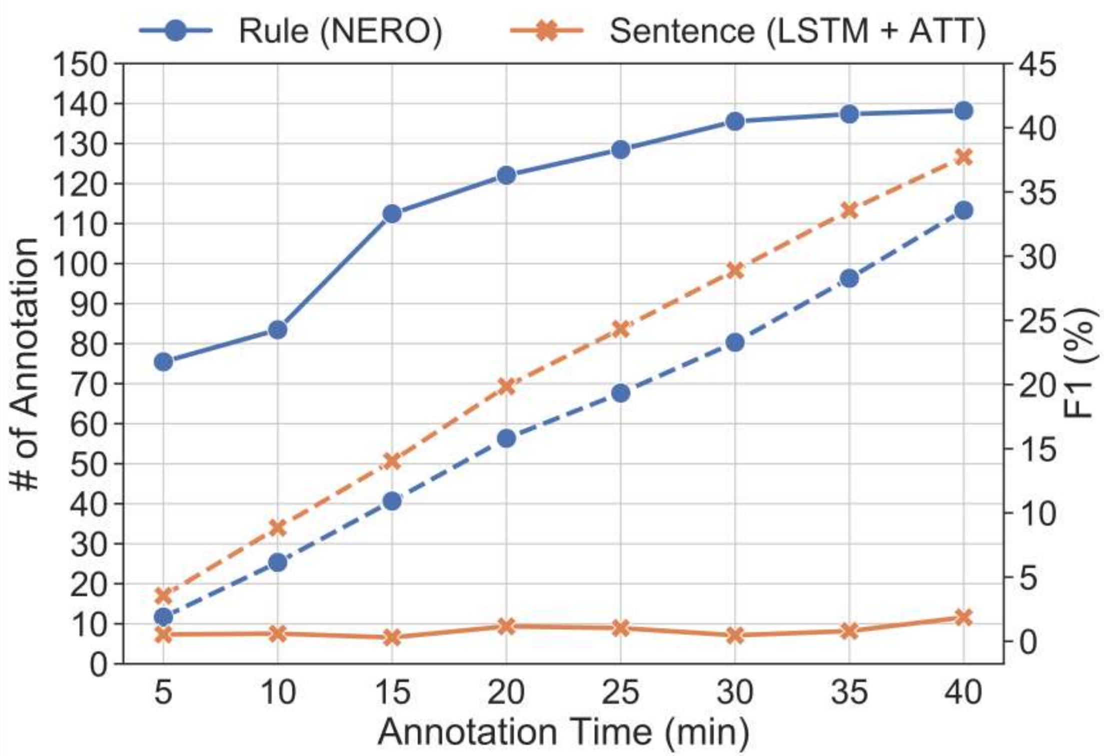
Results on Relation Extraction



Relation Extraction Performance (in F1 score) on TACRED

Study on Label Efficiency

Spent 40min on labeling instances from TACRED



Dashed: Avg # of **rules** / **sentences** labeled by annotators.

Solid: Avg **model F1** trained with corresponding annotations.

{Rules + Neural Rule Grounding} produces much more effective model with limited time!

Standard annotation pipeline

view each example → assess the example → provide a label

Rule-based annotation pipeline

Annotator → view several examples → summarize rules →

Labeling rules

SUBJ-ORG was founded by **OBJ-PER**
→ **ORG: FOUNDED_BY**

Better label efficiency

Less user-friendly, limited expressiveness

Problem: *Can users provide more* ***complex*** *clues to explain their thought process, in a* ***natural*** *way?*

Learning with Natural Language Explanations

Sentiment on ENT is positive or negative?

x_1: There was a long wait for a table outside, but it was a little too hot in the sun anyway so our ENT was very nice.

→

Users' natural language explanations

Positive, because the words "*very nice*" is within 3 words after the ENT.

Relation between ENT1 and ENT2?

x_2: Officials in Mumbai said that the two suspects, David Headley, and ENT1, who was born in Pakistan but is a ENT2 citizen, both visited Mumbai before the attacks.

→

per: nationality, because the words "*is a*" appear right before ENT2 and the word "*citizen*" is right after ENT2.

Wang et al., to appear at ICLR'20 http://inklab.usc.edu/project-NExT

Explanations to “labeling functions”

Explanation

The words “who died” precede OBJECT by no more than three words and occur between SUBJECT and OBJECT

predicate assigning

@Word @Quote(who died) @Left @OBJECT @AtMost @Num @Token @And @Is @Between @SUBJECT @And @OBJECT

CCG parsing

Candidate logical forms

@And (@Is (@Quote (‘who died’), @AtMost (@Left (@OBJECT), @Num (@Token))), @Is (@Word (‘who died’), @Between (@SUBJECT , @OBJECT)))

......

......

Candidate scoring

$$P_\theta(f|\mathbf{e}_i) = \frac{\exp \boldsymbol{\theta}^T \boldsymbol{\phi}(f)}{\sum_{f':f' \in \mathcal{Z}_{\mathbf{e}_i}} \exp \boldsymbol{\theta}^T \boldsymbol{\phi}(f')}$$

$$L_{parser} = \sum_{i=1}^{|S'|} \log \Big(\sum_{f:f(\mathbf{x}_i)=1 \wedge h(f)=y_i} P_\theta(f|\mathbf{e}_i) \Big)$$

inference

$$f_i = \arg\max_f P_{\theta^*}(f|\mathbf{e}_i)$$

function assigning

Labeling function (most plausible)

def LF (x) :
Return (1 if : And (Is (Word (‘who died’), AtMost (Left (OBJECT), Num (3, tokens))), Is (Word (‘who died’), Between (SUBJECT , OBJECT))); else 0)

(Srivastava et al., 2017; Zettlemoyer & Collins, 2012)

Hard matching for data augmentation

Instance

Sentence: quality ingredients preparation all around, and a very fair price for NYC.

Question: What is the sentiment polarity w.r.t. “price” ?

Human labeling

Label result

Label: Positive

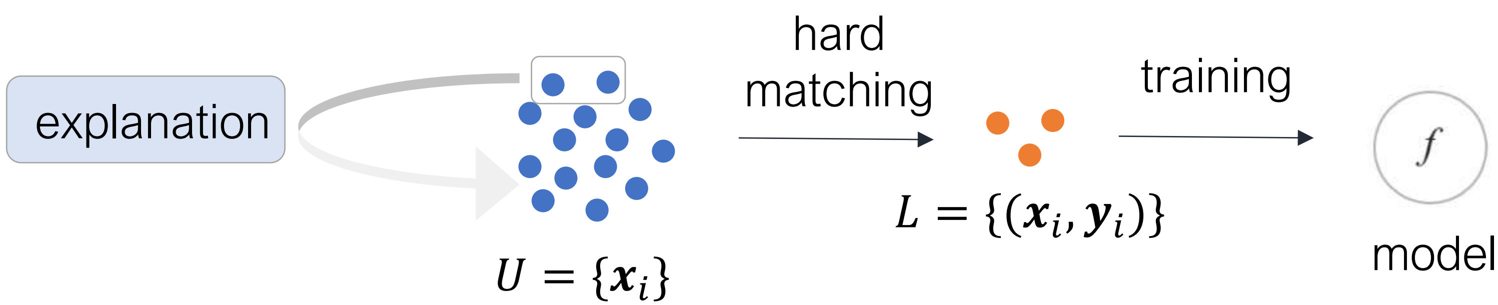
Explanation: because the word “price” is directly preceded by fair.

Hard Matching

unlabeled instance

Sentence: it has delicious food with a fair price.

Problems with hard matching

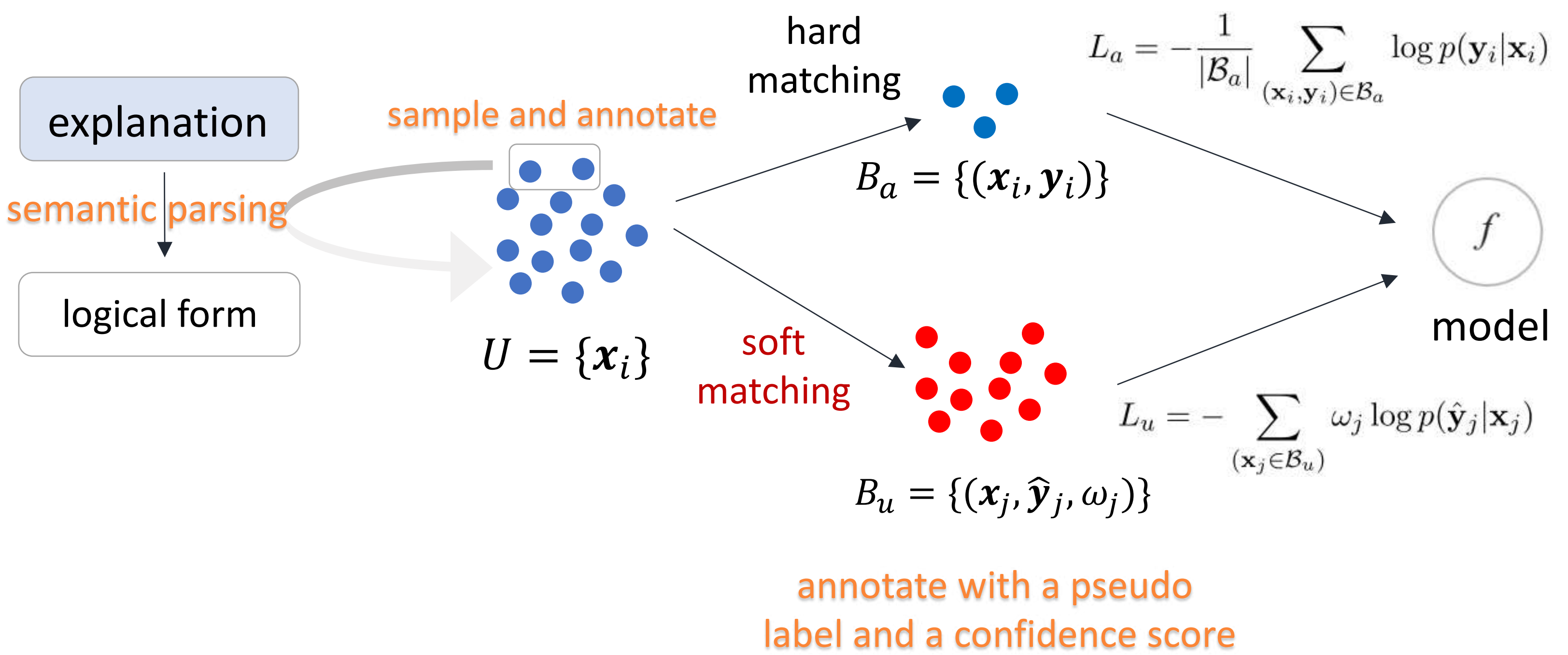


Challenge 1: *language variations* on both explanation predicates & contextual clues

per: nationality, because the words "*is a*" appear right before ENT2 **and** the word "*citizen*" is right after ENT2.

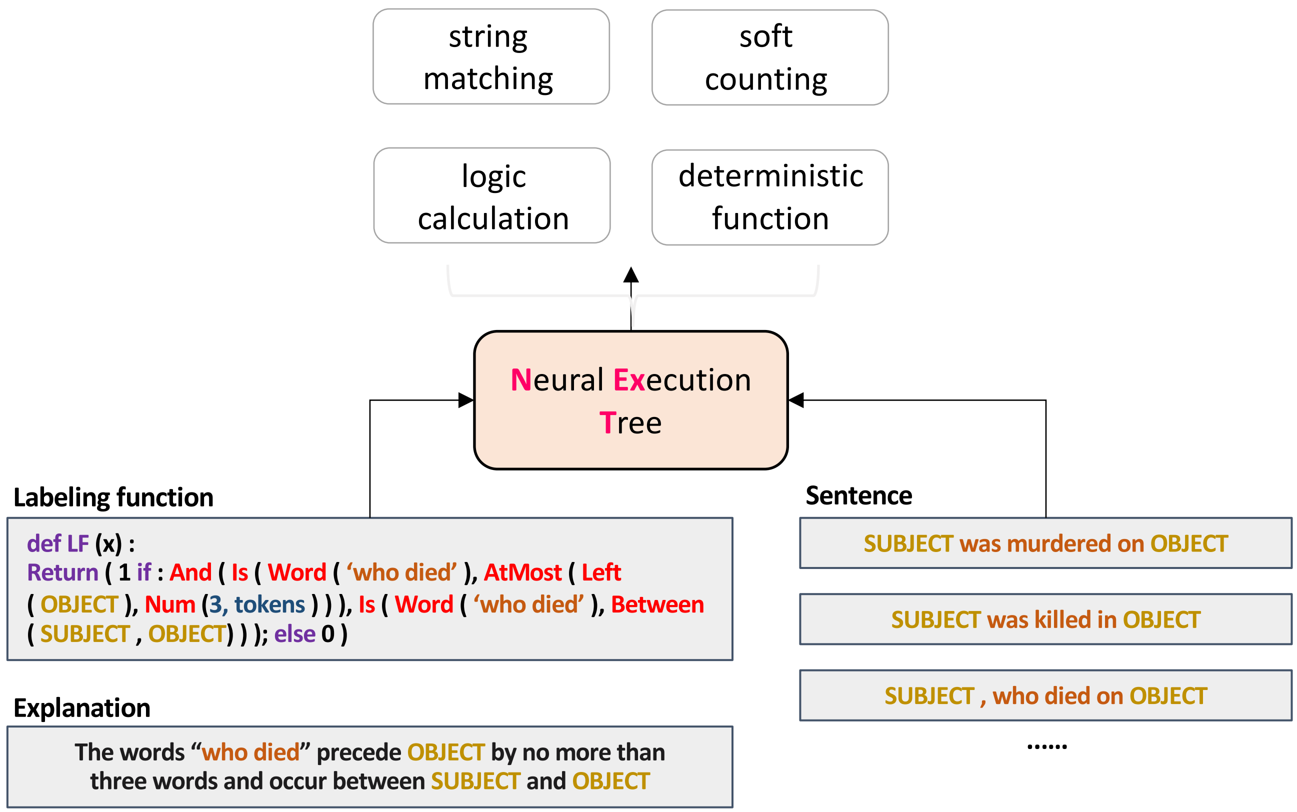
Challenge 2: *compositional nature* of the explanations

Learning with Hard & Soft Matching



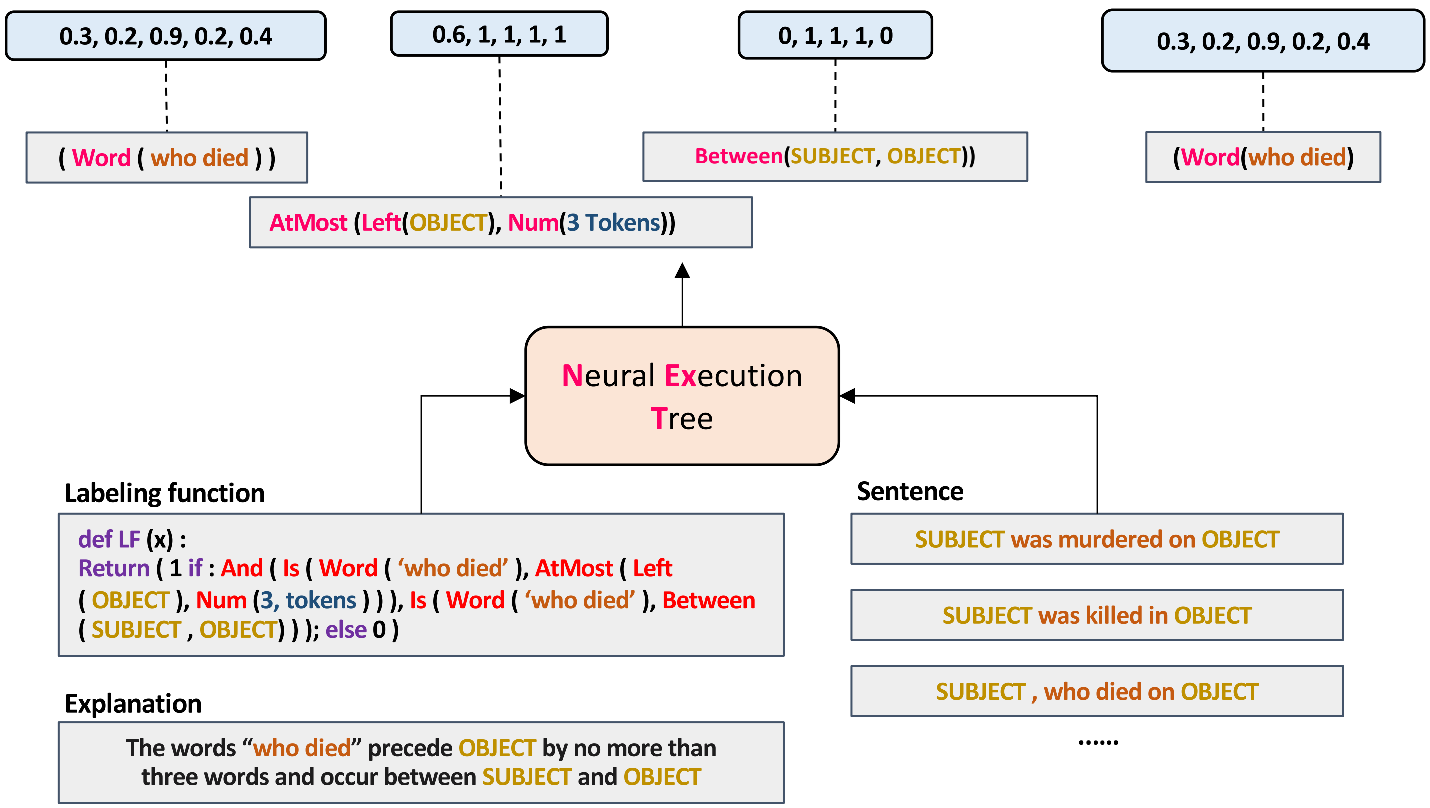
(Wang et al., ICLR20)

Neural Execution Tree (NExT) for Soft Matching



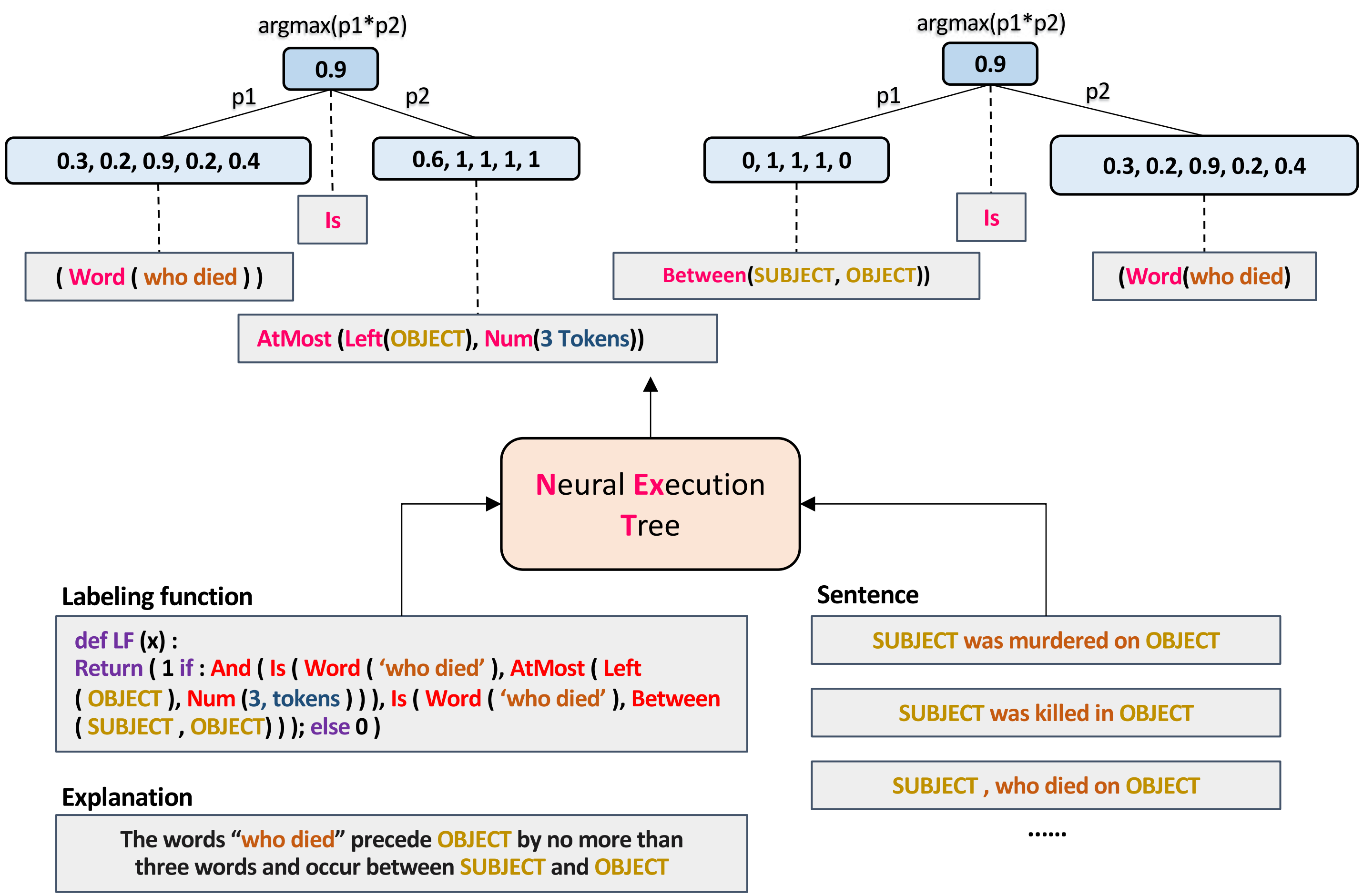
(Wang et al., ICLR20)

Neural Execution Tree (NExT) for Soft Matching



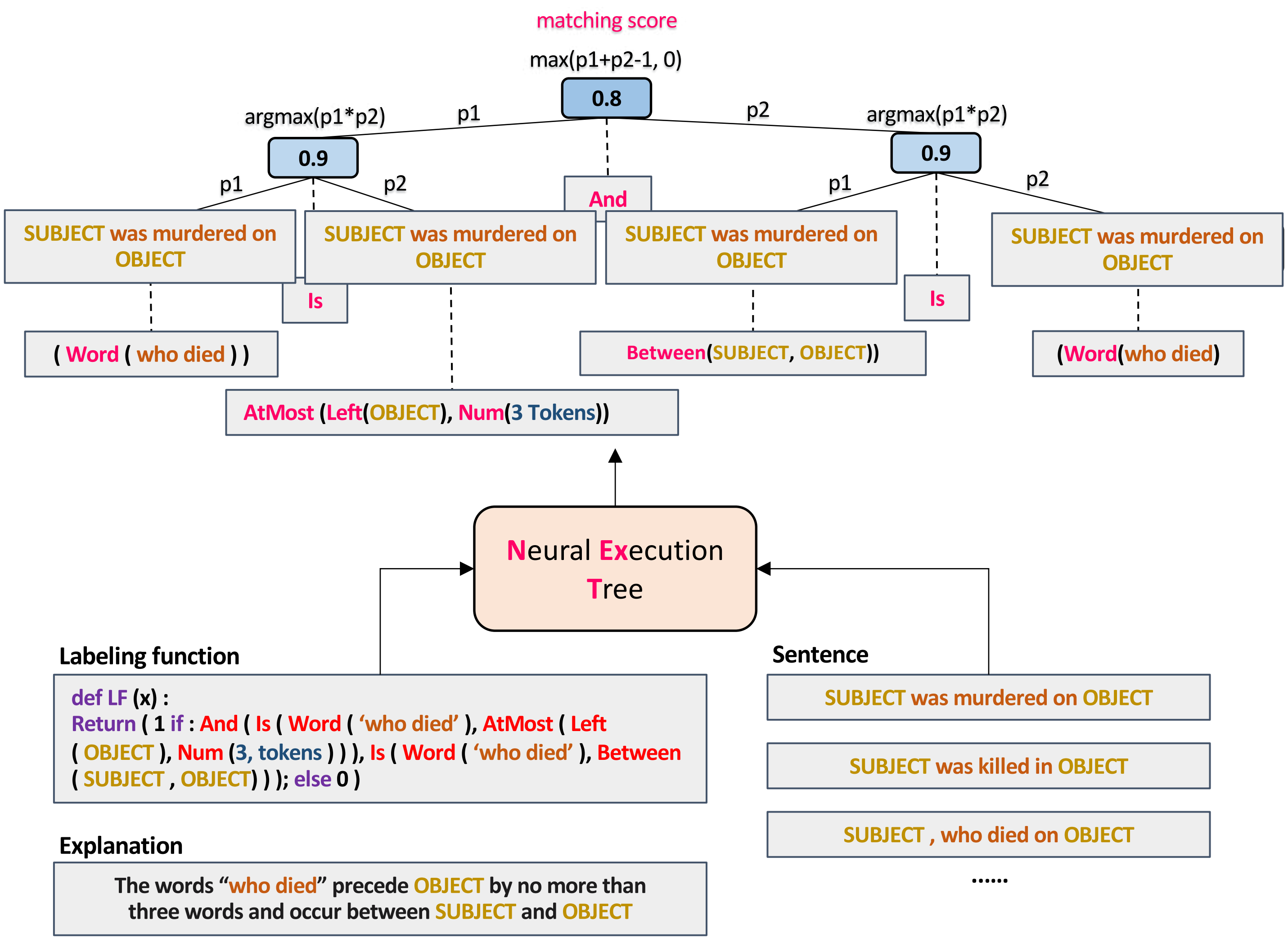
(Wang et al., ICLR20)

Neural Execution Tree (NExT) for Soft Matching



(Wang et al., ICLR20)

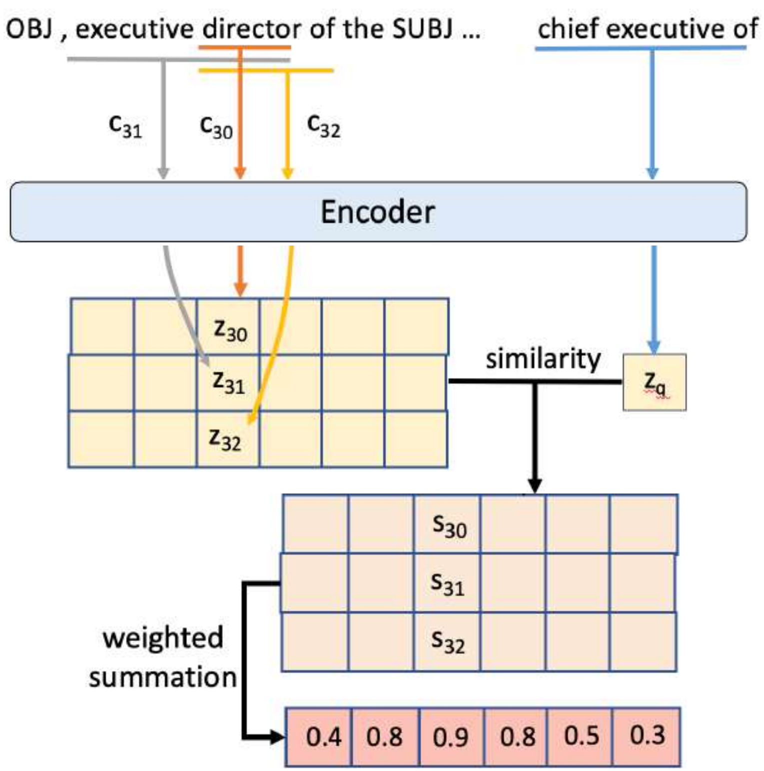
Neural Execution Tree (NExT) for Soft Matching



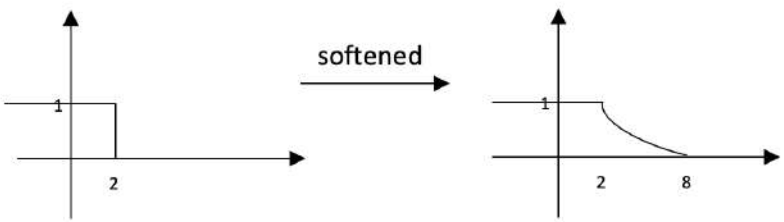
(Wang et al., ICLR20)

Modules in NeXT

1. String matching



2. Soft counting



3. Soft logic

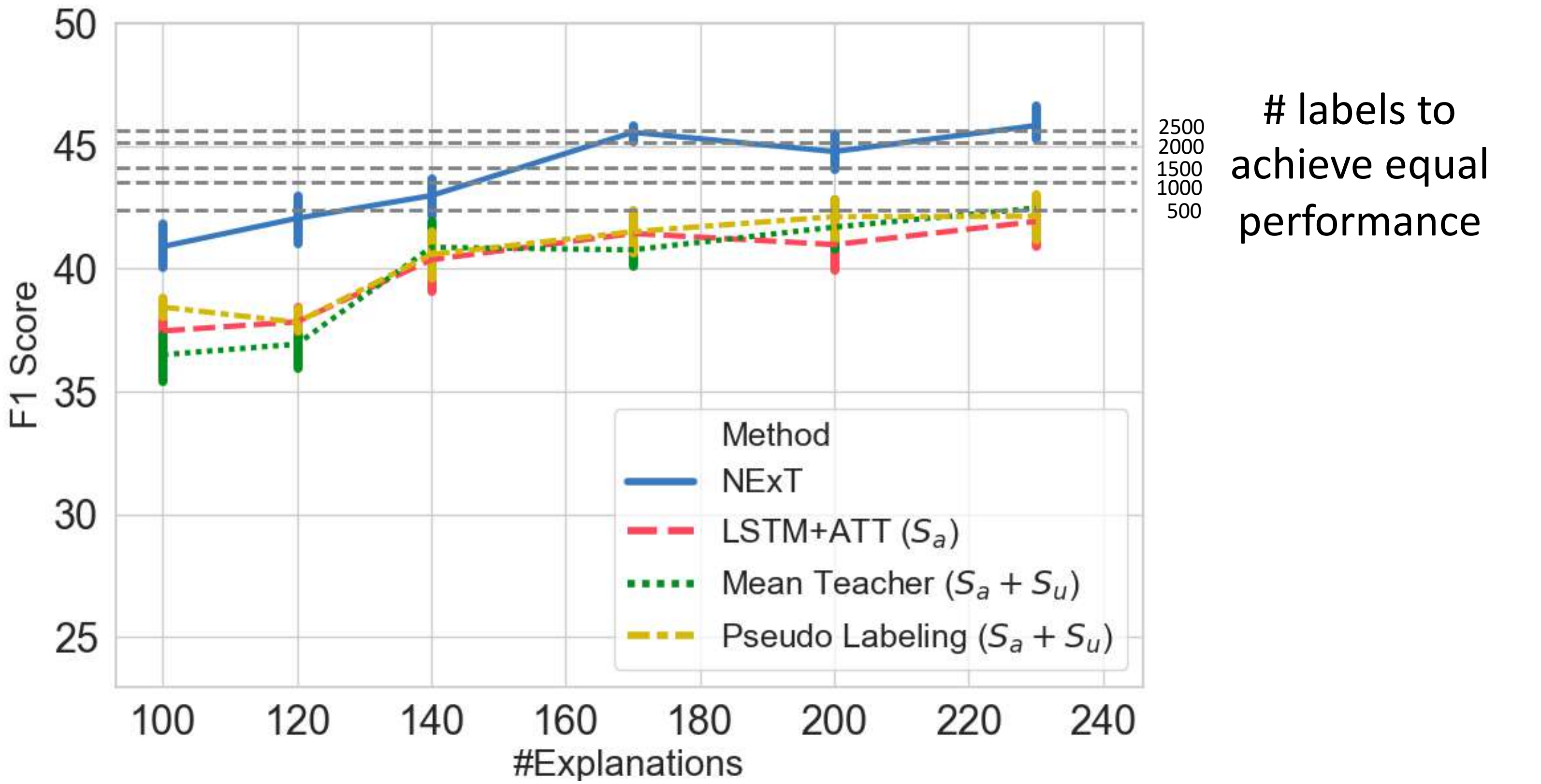
$$p_1 \wedge p_2 = \max(p_1 + p_2 - 1, 0),$$

$$p_1 \vee p_2 = \min(p_1 + p_2, 1), \quad \neg p = 1 - p,$$

4. Deterministic functions

(Wang et al., ICLR20)

Study on Label Efficiency (TACRED)



Number of explanations

Annotation time cost:

giving a label + an explanation ~= 2x *giving a label*

Standard annotation pipeline

view each example → assess the example → provide a label

Rule-based annotation pipeline

Annotator → view several examples → summarize rules →

Labeling rules

SUBJ-ORG was founded by **OBJ-PER**
→ **ORG: FOUNDED_BY**

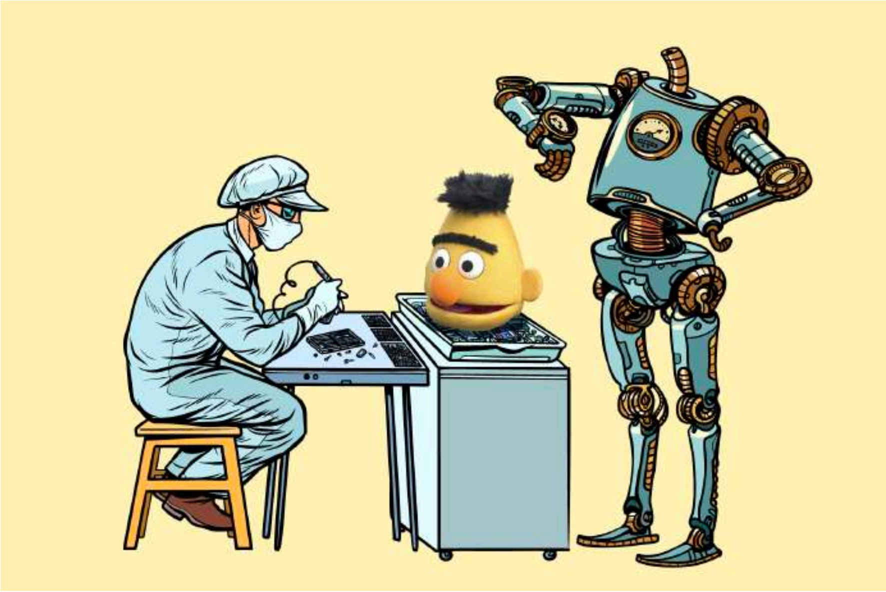
NL explanation-based annotation pipeline

Annotator → view an example → provide rationale →

NL explanations

Positive, because the words "very nice" is within 3 words after the TERM.

Problem: *Can we make use of prior knowledge to constrain the model learning?*



Commonsense Reasoning in QA

Where do adults usually use glue sticks?

A: classroom **B:** office **C:** desk drawer

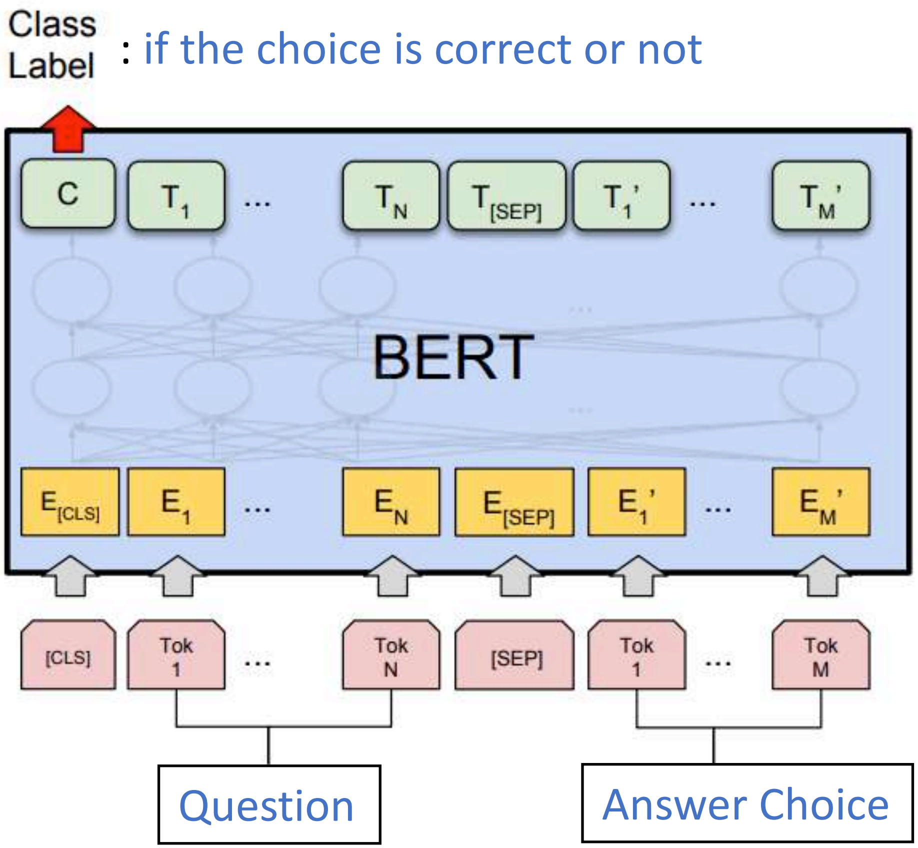
What do you need to fill with ink to write notes on an A4 paper?

A: fountain pen **B:** printer **C:** pencil

Can you choose the most plausible answer based on daily life commonsense knowledge?

(CommonsenseQA, Talmor et al., 2018)

Pre-trained LMs doesn't get it for free



Fine-tuning BERT for CommonsenseQA (12k QA pairs).

Accuracy will drop 15+% if labeled data are reduced for 10%

Limitations of Fine-tuned LMs

1. Not capturing commonsense

Most plausible predictions
are far from common truth

Masked Language Modeling

Enter text with one or more "[MASK]" tokens and BERT will generate the most likely token to substitute for each "[MASK]".

Sentence:

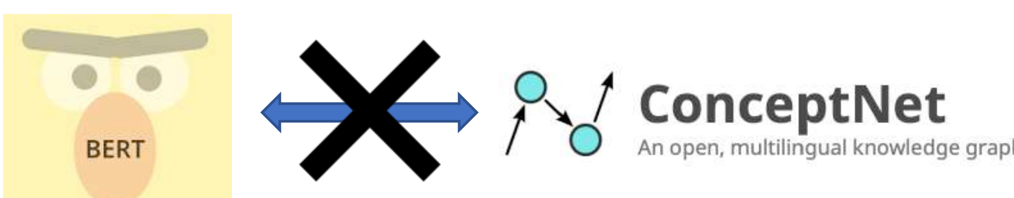
Adults usually use glue sticks at their [MASK].

Mask 1 Predictions:

- 16.4% **feet**
- 14.8% **disposal**
- 5.4% **backs**
- 3.5% **fingertips**

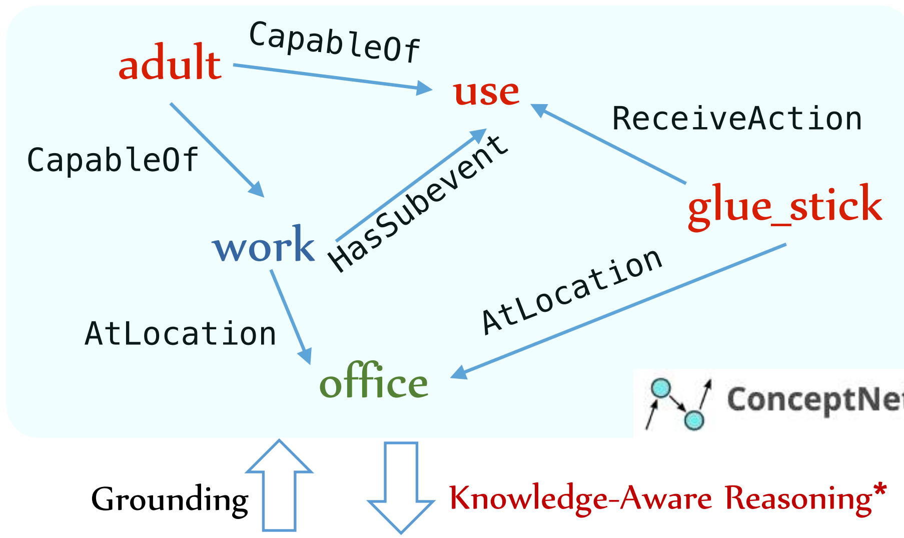
Online demo of BERT's Masked-LM https://demo.allennlp.org/masked-lm

2. Not Interpretable w/ Knowledge



Neural-Symbolic Reasoning with Commonsense KG

Symbol Space

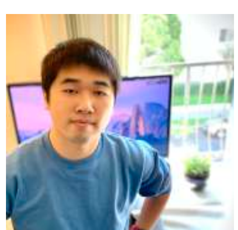


A **Schema Graph** for the choice **B**

Semantic Space

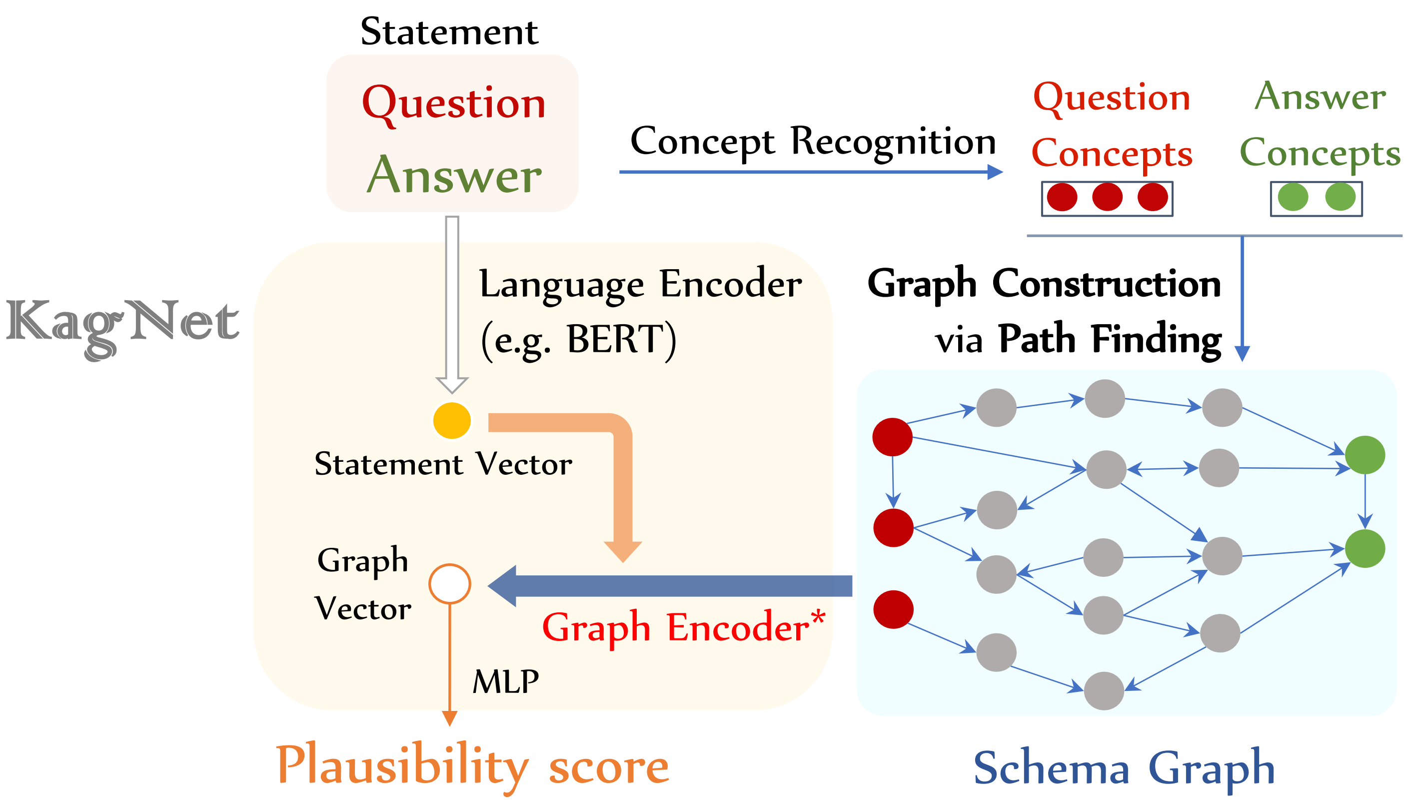
Where do adults use glue sticks? — *Question*

A: classroom B: office C: desk drawer — *Answer Candidates*



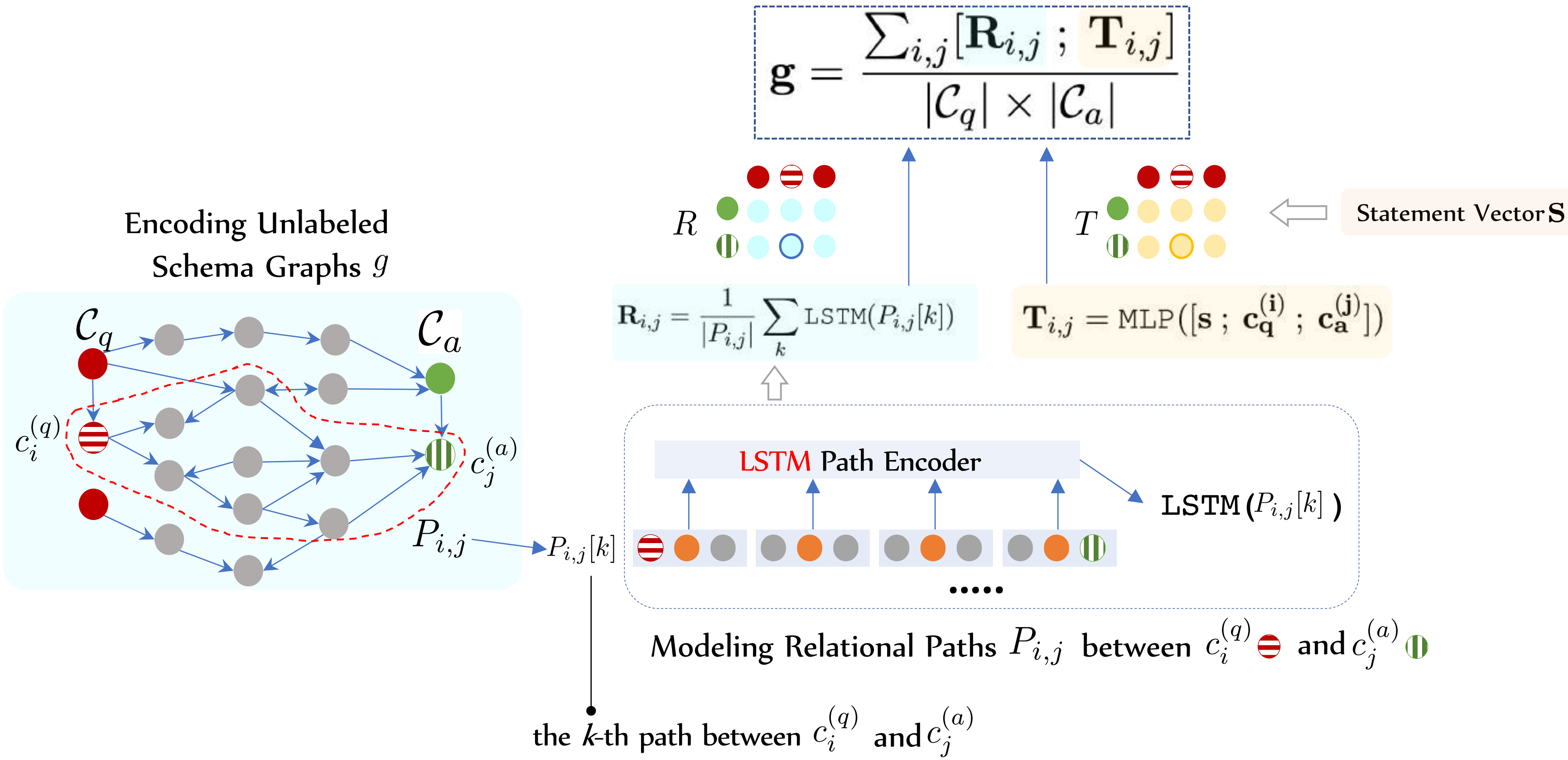
(Bill Yuchen Lin et al. EMNLP19)

Multi-relational Graph as Inductive Bias



(Bill Yuchen Lin et al. 2019)

KagNet: Knowledge-aware Graph Network



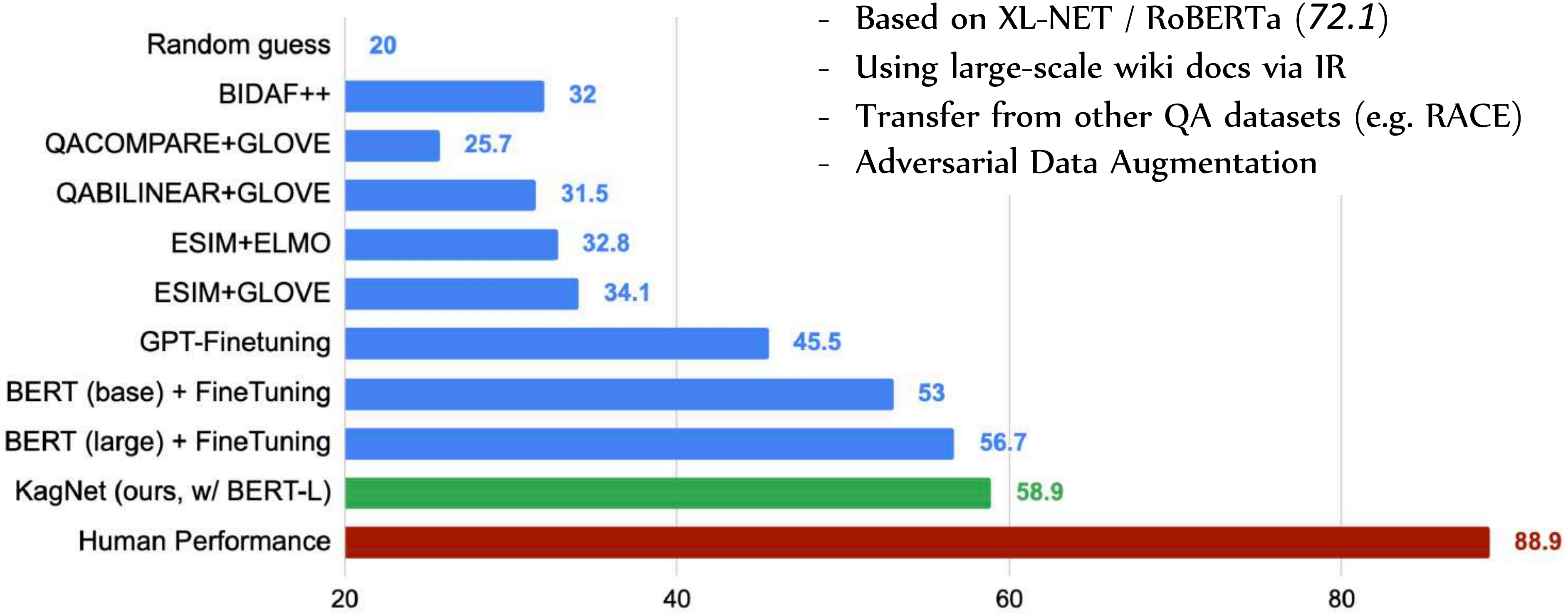
(Bill Yuchen Lin et al. 2019)

KagNet: Knowledge-Aware Graph Networks
for Commonsense Reasoning

Experiments

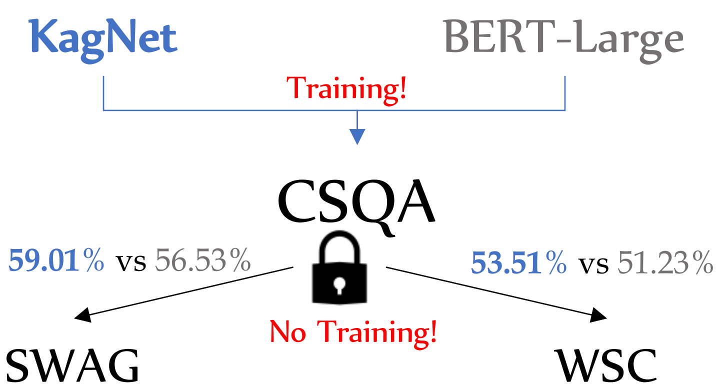
Recent follow-up submissions:

- Based on XL-NET / RoBERTa (*72.1*)
- Using large-scale wiki docs via IR
- Transfer from other QA datasets (e.g. RACE)
- Adversarial Data Augmentation

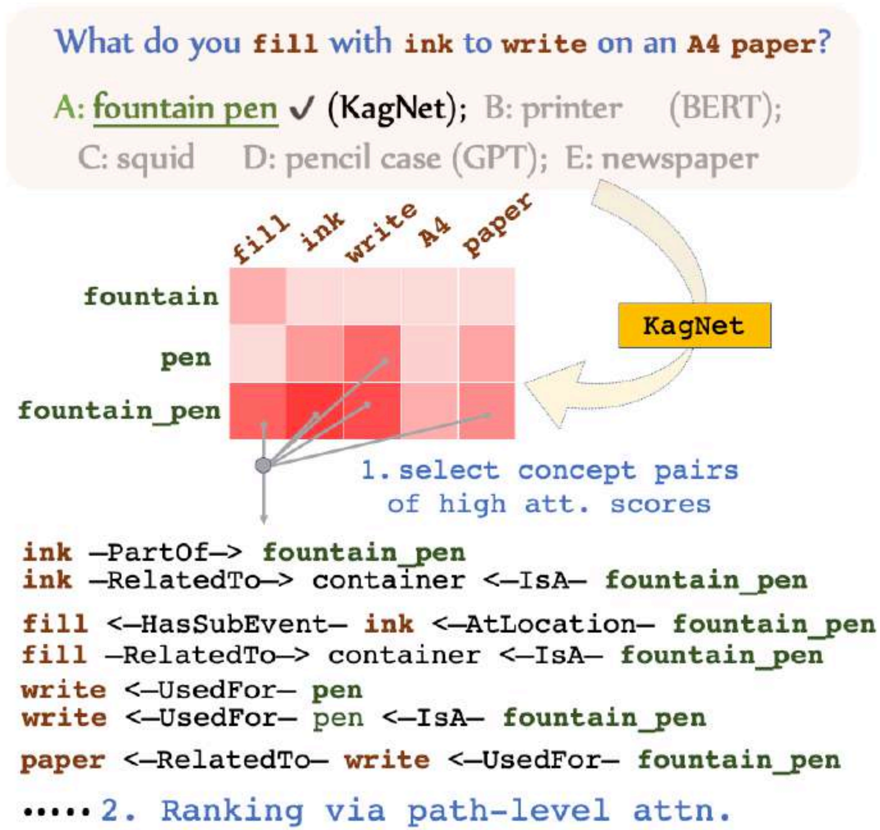


More Performance on Official Test Set: https://www.tau-nlp.org/csqa-leaderboard

Transferability



Interpretability



(Bill Yuchen Lin et al. 2019)

Conclusion

(*Label-efficient*) **Learning from high-level human supervisions** that are *abstractive*, *compositional*, and *linguistically complex*

Q1 How to augment model training with rules?

Soft rule grounding for data augmentation (Zhou et al. WWW20)

Q2 How to handle compositional natural language input?

Neural execution tree for NL explanation (Wang et al. ICLR20)

Q3 How to incorporate prior knowledge as inductive bias?

Knowledge-aware graph networks (Lin et al. EMNLP19)

Other related efforts

Q1 How to augment model training with rules?

Soft rule grounding for data augmentation (Zhou et al. WWW20)

Q2 How to handle compositional natural language input?

Neural execution tree for NL explanation (Wang et al. ICLR20)

Q3 How to incorporate background knowledge?

Knowledge-aware graph networks (Lin et al. EMNLP19)

Learning from Distant Supervision: [Ye et al., EMNLP19], [Zhang et al., NAACL19], [Shang et al., EMNLP18], [Liu et al., EMNLP17]

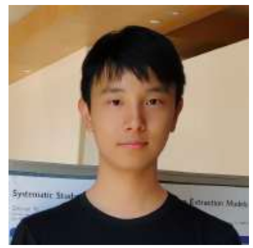
Reasoning over Heterogeneous Data: [Fu et al., EMNLP18], [Jin et al., ICLR-GRLM19], [Ying et al., NeurIPS18], [Ying et al., ICML18]

Students







Research Partnership

Microsoft Academic

SPARK AT STANFORD

Semantic Scholar

BioPortal

Collaborators

Dan MacFarland, Sociology, Stanford University
Jure Leskovec, Computer Science, Stanford University
Dan Jurafsky, Computer Science, Stanford University
Jiawei Han, Computer Science, UIUC
Morteza Dehghani, Psychology, USC
Kennth Yates, Clinical Education, USC
Craig Knoblock, USC ISI
Curt Langlotz, Bioinformatics, Stanford University
Kuansan Wang, Microsoft Academic
Leonardo Neves, Snap Research
Mark Musen, Bioinformatics, Stanford University

Funding

NSF

DARPA

IARPA
BE THE FUTURE

J.P.Morgan

SCHMIDT FAMILY FOUNDATION

Google

amazon

Adobe

Thank you!

USC Intelligence and Knowledge Discovery (INK) Lab

http://inklab.usc.edu/

Code: https://github.com/INK-USC

xiangren@usc.edu

@xiangrenNLP