Teaching Machine through Human Explanations

Xiang Ren

Department of Computer Science & Information Science Institute
University of Southern California

http://inklab.usc.edu
A Surprisingly “Simple” Recipe for Modern NLP

Model

+ 

Labeled Data

+ 

Computing Power
A Surprisingly “Simple” Recipe for Modern NLP

Model

+ Labeled Data

+ Computing Power

pip install transformers
from transformers import BertModel
from transformers import RobertaModel
A Surprisingly “Simple” Recipe for Modern NLP

Model + Labeled Data + Computing Power
A Surprisingly “Simple” Recipe for Modern NLP

Model + Labeled Data + Computing Power

pip install transformers
from transformers import BertModel
from transformers import RobertaModel

aws ec2 run-instances \
  --instance-type p3.2xlarge \
  --instance-type p3.16xlarge
A Surprisingly “Simple” Recipe for Modern NLP

Model architectures and computing power are transferrable across applications, but labeled data is not!

Computing Power

```
aws ec2 run-instances
  --instance-type p3.2xlarge
  --instance-type p3.16xlarge
```

```
pip install transformers
from transformers import BertModel
from transformers import RobertaModel
```
Cost of data labeling: relation extraction

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

(Zhang et al., 2018)
Cost of data labeling: relation extraction

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

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

(Zhou et al., WWW’20)
Cost of data labeling: more complex task

SQUAD dataset: 23k paragraphs
Mechanical Turk: $9 per 15 paragraphs (1 hour)

Total Cost > $13k
Time Cost > 60 days

Paragraph 1 of 43

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 O₂.

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.

(Rajpurkar et al., 2018)
Workaround for (less) data labeling?

**Multi-task/transfer/active learning** are applied to improve model adaptation and generalization to new data (distribution)

- Assumptions about source-to-target data distribution “gap”
- Annotation format: “instance-label” pairs → carries limited information
How “labels” alone could make things wrong

Models are prone to capture spurious patterns (between labels and features) in training

- There has been a rise and fall of hate against the jews
  - New York Times

reliability and robustness of the models?
From “labels” to “explanations of labels”

“One explanation generalizes to many examples”

**Input:** … but it was a little hot anyway so our **TERM** was very nice
**Label:** **Positive**
**Explanation:** the phrase “**very nice**” is within 3 words after the **TERM**.
From “labels” to “explanations of labels”

“One explanation generalizes to many examples”

Input: … but it was a little hot anyway so our TERM was very nice
Label: Positive
Explanation: the phrase “very nice” is within 3 words after the TERM.

Input: It’s such a wonderful place and the TERM here is very nice!
Get Label Automatically: Positive

Input: Oh my god! The TERM here is extraordinary!
Get Label Automatically: Positive

Input: The TERM and environment are both very nice!
Get Label Automatically: Positive

One explanation generalizes to many examples.
Learning from Human Explanation

Machine digests human rationale and learns how to make decisions

http://inklab.usc.edu/leanlife/ (Khanna et al., ACL’20 Demo)
This Talk

Learning models from labels + explanations
- An explanation-based learning framework
- Soft rule grounding for data augmentation (Zhou et al. WWW20)
- Modularized neural network for soft grounding (Wang et al. ICLR’20)
- Explanation for cross-sentence tasks (Ye et al., EMNLP’20 Findings)

Refining models with labels + explanations
- Explanation regularization (Jin et al. ACL’20)
- Explanation-based model refinement (Yao et al. In Submission)
What is an explanation?

Salient spans
Highlight important substrings in the input.

Post-hoc Explanations
Interpret a model’s prediction after it’s trained.

Natural Language
Write free-form sentences that justifies an annotation.

Q: How many touchdown passes did Culter throw in the second half?
A: 3

Question: After getting drunk people couldn’t understand him, it was because of his what?
Choices: lower standards, slurred speech, falling down
Explanation: People who are drunk have difficulty speaking.

---

Dua et al., 2020
Zaidan et al., 2007
Lei et al. 2016

Ribeiro et al., 2016
Jin et al., 2020

Camburu et al., 2018
Rajani et al., 2019
Our Focus: *Natural Language Explanations*

... targeting individual **data instances** or **features**,

**Input:** The TERM is vibrant and eye-pleasing with several semi-private booths on the right side of ...

**Label:** Positive

**Explanation:** The term is followed by "vibrant" and "eye-pleasing"

... describing **existence of concepts**, **properties of concepts**, **interactions of concepts**,

**Importance Heat-map:**

Explanation: ... “Sweden” is less than 3 dependency steps from “failure”... Adjust “Sweden” to non-hate; adjust “failure” to hate.
Our Focus: *Natural Language Explanations*

... targeting individual **data instances** or **features,**

**Input:** The TERM is vibrant and eye-pleasing with several semi-private booths on the right side of ...

**Label:** Positive

**Explanation:** The term is **followed by** "vibrant" and "eye-pleasing"

**Importance Heat-map:**

**Explanation:** ...

... "Sweden" is less than 3 dependency steps from "failure"... Adjust "Sweden" to non-hate; adjust "failure" to hate.

... describing **existence of concepts**, **properties of concepts**, **interactions of concepts,**
Our Focus: *Natural Language Explanations*

... targeting individual **data instances** or **features**, 

**Input:** The TERM is vibrant and eye-pleasing with several semi-private booths on the right side of ... 
**Label:** Positive 

Explanation: The term is followed by "vibrant" and "eye-pleasing"

**Importance Heat-map:**

Explanation: ... “Sweden” is less than 3 dependency steps from “failure”... Adjust “Sweden” to non-hate; adjust “failure” to hate.

... describing **existence of concepts**, **properties of concepts**, **interactions of concepts**, ...
Our Focus: *Natural Language Explanations*

... targeting individual **data instances** or **features**,

**Input:** The TERM is vibrant and eye-pleasing with several semi-private booths on the right side of ...  
**Label:** Positive

**Explanation:** The term is followed by "vibrant" and "eye-pleasing"

**Importance Heat-map:**

Explanation: ... “Sweden” is less than 3 dependency steps from “failure”... Adjust “Sweden” to non-hate; adjust “failure” to hate.

... describing **existence of concepts**, **properties of concepts**, **interactions of concepts**, ... and being...

**Compositional**  
Putting pieces of evidence together and applying logic.

**Self-contained**  
Clear, deterministic, closely associated to the instance or feature.

**Locally Generalizable**  
May generalize and become applicable to unseen instances.
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., ICLR’20) [http://inklab.usc.edu/project-NExT](http://inklab.usc.edu/project-NExT)
How to incorporate explanations in model learning?

**Representation Engineering**

Use explanations as feature functions, or as hidden representation directly.

Srivastava et al., 2017
Murty et al., 2020

**Auxiliary Task**

Train a decoder to generate explanations from hidden representations.

Rajani et al., 2019
Mu et al., 2020
How to incorporate explanations in model learning?

**Representation Engineering**

Use explanations as feature functions, or as hidden representation directly.

*Srivastava et al., 2017*

*Murty et al., 2020*

---

**Auxiliary Task**

Train a decoder to generate explanations from hidden representations.

*Rajani et al., 2019*

*Mu et al., 2020*

---

**Create Noisy Annotations**

Use one explanation to create multiple labeled instances.

*Hancock et al. 2018*
Explanations to “labeling rules”

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

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

Candidate logical forms

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

......

Labeling rule (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 )

Candidate scoring

$P_\theta(f|e_i) = \frac{\exp \theta^T \phi(f)}{\sum_{f':f' \in \mathcal{S}_{e_i}} \exp \theta^T \phi(f')}$

$L_{parser} = \sum_{i=1}^{|S'|} \log ( \sum_{f:f(x_i)=1 \land h(f)=y_i} P_\theta(f|e_i) )$
Matching labeling rules to create pseudo labeled data

Instance

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

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.*

Unlabeled instance

*it has delicious food with a fair price.*

**Hard Matching**

\( LF(x) \)
Data Programming & Snorkel

Annotating an unlabeled dataset with **labeling functions** collected from human experts (e.g., Snorkel)

```python
def a_cause_b(x):
    Pattern(x, "{} causes {}")
```

Collect labeling functions → Smoking causes lung diseases → Obtain noisy labels

(Ratner et al., 2017; Ratner et al., 2019)
Challenge: Language Variations

Corpus

Microsoft was founded by Bill Gates in 1975.
Apple was founded by Steven Jobs in 1976.
Microsoft was established by Bill Gates in 1975.
In 1975, Bill Gates launched Microsoft.

Labels

ORG: FOUNDED_BY
ORG: FOUNDED_BY
No Matched!
No Matched!

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

Annotator

Have to exhaust all surface patterns?
Neural Rule Grounding for rule generalization

Generalizing one rule to many instances

Corpus

Microsoft was founded by Bill Gates in 1975.
Apple was founded by Steven Jobs in 1976.
Microsoft was established by Bill Gates in 1975.
In 1975, Bill Gates launched Microsoft.

Labeling Rules

SUBJ-ORG was founded by OBJ-PER \rightarrow ORG: FOUNDED_BY
SUBJ-PER born in OBJ-LOC \rightarrow PER: ORIGIN

Hard-matched instances

Microsoft was founded by Bill Gates in 1975.
Apple was founded by Steven Jobs in 1976.

Unmatched instances

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

(x_i, y_i)

(x_i, y_i, matching score)

ORG: FOUNDED_BY 0.8
ORG: FOUNDED_BY 0.7

Relation Classifier

Best Paper runner-up, WWW’20

(Zhou et al, WWW20)
A Learnable, Soft Rule Matching Function

Unmatched instances

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

Labeling Rules

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

2. Soft-matching

(Zhou et al, WWW20)
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!
Learning with Hard & Soft Matching

\[ U = \{ x_i \} \]

\[ B_a = \{(x_i, y_i)\} \]

\[ B_u = \{(x_j, \hat{y}_j, \omega_j)\} \]

\[ L_a = -\frac{1}{|B_a|} \sum_{(x_i, y_i) \in B_a} \log p(y_i | x_i) \]

\[ L_u = -\sum_{(x_j \in B_u)} \omega_j \log p(\hat{y}_j | x_j) \]

**New Challenge:**

Compositional nature of the human explanations

per: nationality, because the words “is a” appear right before ENT2 and the word “citizen” is right after ENT2.
Neural Execution Tree (NExT) for Soft Matching

Labeling function:

```python
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)
```

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

Sentence examples:
- SUBJECT was murdered on OBJECT
- SUBJECT was killed in OBJECT
- SUBJECT, who died on OBJECT
- ......
Neural Execution Tree (NExT) for Soft Matching

Labeling function

```
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)
```

Explaination

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

Sentence

```
SUBJECT was murdered on OBJECT
SUBJECT was killed in OBJECT
SUBJECT, who died on OBJECT
```

(Wang et al., ICLR'20)
Neural Execution Tree (NExT) for Soft Matching

Labeling function:
```python
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)
```

Sentence:
- SUBJECT was murdered on OBJECT
- SUBJECT was killed in OBJECT
- SUBJECT, who died on OBJECT

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

(Wang et al., ICLR’20)
Neural Execution Tree (NExT) for Soft Matching

Labeling function:
```python
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)
```

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

(Wang et al., ICLR’20)
Neural Execution Tree (NExT) for Soft Matching

Labeling function

```python
def LF(x):
    return (1 if And(Word('who died'), AtMost(Left(OBJECT), Num(3, tokens)))
            Is(Word('who died'), Between(SUBJECT, OBJECT)))
    else 0)
```

Sentence

- SUBJECT was murdered on OBJECT
- SUBJECT was killed in OBJECT
- SUBJECT, who died on OBJECT

Explanation

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

(Wang et al., ICLR’20)
Module Functions in NExT

1. String matching

2. Soft counting

3. Soft logic

\[ p_1 \cap p_2 = \max(p_1 + p_2 - 1, 0), \]

\[ p_1 \lor p_2 = \min(p_1 + p_2, 1), \quad \neg p = 1 - p, \]

4. Deterministic functions

(Wang et al., ICLR’20)
Study on Label Efficiency (TACRED)

Annotation time cost:
\[
giving a label + an explanation \sim 2 \times giving a label
\]

(Wang et al., ICLR’20)
Problem: Extending to complex tasks that go beyond a single sentence?
Question: What is the atomic number for Zinc?
Context: Zinc is a chemical element with symbol Zn and atomic number 30.
Answer: 30

Explanation: X is atomic number. Y is Zinc. The question contains "number", so the answer should be a number. The answer is directly after X. "for" is directly before Y and directly after X in the question.
**Explanations for Machine Reading Comprehension**

**Question:** What is the atomic number for Zinc?
**Context:** Zinc is a chemical element with symbol Zn and atomic number 30.
**Answer:** 30

**Explanation:** X is atomic number. Y is Zinc. The question contains "number", so the answer should be a number. The answer is directly after X. "for" is directly before Y and directly after X in the question.

**Use the explanation to answer similar questions!**

**Question:** What is the phone number for CS front desk?
**Context:** You can contact CS front desk with phone number 213-000-0000.

(Ye et al., Findings EMNLP 2020) Data explorer: [http://inklab.usc.edu/mrc-explanation-project/data/](http://inklab.usc.edu/mrc-explanation-project/data/)
Use explanation to answer a similar question

A Seen Example

**Question**: What is the atomic number for Zinc?

**Context**: Zinc is a chemical element with symbol Zn and atomic number 30.

**Explanation**:

X is **atomic number**.

Y is **Zinc**.

The question contains "number", so the answer should be a number.

The answer is directly after X.

"for" is directly before Y and directly after X in the question.

An Unseen Example

**Question**: What is the phone number for CS front desk?

**Context**: You can contact CS front desk with phone number 213-000-0000.

**Answer**: ? 213-000-0000

**Matching Procedure**:

X and Y are noun phrases in the question.

- X = phone number, phone, number, CS front desk, front desk
- Y = phone number, phone, number, CS front desk, front desk

ANS is a number

- ANS = 213-000-0000

List each combination

- Comb1: X = phone number, Y = CS front desk, ANS = 213-000-0000
- Comb2: X = front desk, Y = phone number, ANS = 213-000-0000
- Comb3: X = phone, Y = front desk, ANS = 213-000-0000

For each combination, see if all constraints are satisfied

- For Comb1, every constraint is satisfied
- For Comb2, X "for" is directly before Y and directly after X in the question.
- For Comb3, X The answer is directly after X.

Matching Result

- X = phone number, Y = CS front desk, ANS = 213-000-0000

(Ye et al., Findings EMNLP 2020)
How can we generalize with softened matching?

Question: What is the telephone number for CS front desk?
Context: You can contact CS front desk with phone number 213-000-0000.
Answer: 213-000-0000 (with confidence 0.8)

(Ye et al., Findings EMNLP 2020)
How can we generalize with softened matching?

Question: What is the telephone number for CS front desk?
Context: You can contact CS front desk with phone number 213-000-0000.
Answer: 213-000-0000 (with confidence 0.8)

Reference sentence
What is the telephone number for CS front desk?
Target sentence
You can contact CS front desk with phone number 213-000-0000.

(Ye et al., Findings EMNLP 2020)
How can we generalize with softened matching?

Question: What is the phone number for CS front desk?
Context: If you want to contact CS front desk, the phone number is 213-000-0000.
Answer: ? 213-000-0000 (with confidence 0.75)

(Ye et al., Findings EMNLP 2020)
How can we *generalize* with softened matching?

**Question:** What is the **phone number** for **CS front desk**?

**Context:** If you want to contact **CS front desk**, the **phone number** is 213-000-0000.

**Answer:** 213-000-0000 (with confidence 0.75)

(Ye et al., Findings EMNLP 2020)
Results on SQUAD: Label Efficiency

Collecting one answer takes 43 seconds. Collecting one answer with explanation takes 151 seconds (3.5x slower).

But if we compare performance when annotation time is held constant...

Or if we want to achieve 70% F1 on SQuAD, You need either 1,100 answers (13.1 hours) or 26 answers with explanations (1.1 hours)

12x speed-up 😊

(Ye et al., Findings EMNLP 2020)
Now, suppose you have a working model

Task: Hate Speech Detection

Input: ... Sweden has been proved to be a failure...

A Trained Classifier

Prediction: Non-hate  🙁  Wrong Prediction!

A Trained Classifier

Prediction: Non-hate  Hate

Update the model with the correct label...

Non-hate  🙁  We only have one example ...

Tell the model why it got wrong...
Can we update a model through human explanations on “why it goes wrong”?
Refining neural models through compositional explanations

1. Inspect Post-hoc Explanation Heatmaps

2. Write Compositional Explanation

Because the word “Sweden” is a country, “failure” is negative, and “Sweden” is less than 3 dependency steps from “failure”, attribution score of “Sweden” should be decreased. Attribution score of “failure” should be increased. The interaction score of “Sweden” and “failure” should be increased.

3. First-Order Logic Rule

@Is(Word1, country) ∧ @Is(Word2, negative) ∧ @LessThan(Word1, Word2) → DecreaseAttribution(Word1) ∧ IncreaseAttribution(Word2) ∧ IncreaseInteraction(Word1, Word2).

4. Rule Matching

“Another Reminder that Britain’s establishment is stupid beyond the point of saving.”

Attribution score of “Britain” should be decreased. Attribution score of “stupid” should be increased. The interaction score of “Britain” and “stupid” should be increased.

5. Explanation regularization

(Jin et al., ACL’20; Yao et al., In Submission)
**Explanation Regularization**

Adjust Attribution Scores

Attribution of \( p \) ("Sweden") in the sentence \( x \) ("Sweden has been proved to be a failure") towards the prediction \( c \) (Non-hate)

\[
\mathcal{L}^{\text{attr}} = \sum_{c} \sum_{p \in \mathcal{R}} \left( \phi^c(p; x) - t^c_p \right)^2;
\]

Adjust Interactions

\[
\mathcal{L}^{\text{inter}} = \sum_{c} \sum_{\{p, q\} \in \mathcal{R}} \left( \phi^c(p, q; x) - t^c_{p,q} \right)^2.
\]

Final Loss Term

\[
\mathcal{L} = \mathcal{L}' + \alpha(\mathcal{L}^{\text{attr}} + \mathcal{L}^{\text{inter}}),
\]

(Jin et al., ACL’20; Yao et al., *In Submission*)
Explanation

Regularization

Adjust Attribution Scores

\[ \mathcal{L}_{attr} = \sum_{c} \sum_{p \in \mathcal{R}} \left( \phi^c(p; x) - t^c_p \right)^2; \]

“Decrease”, adjust to zero

Adjust Interactions

\[ \mathcal{L}_{inter} = \sum_{c} \sum_{\{p, q\} \in \mathcal{R}} \left( \varphi^c(p, q; x) - \tau^c_{p, q} \right)^2. \]

Final Loss Term

\[ \mathcal{L} = \mathcal{L}' + \alpha(\mathcal{L}_{attr} + \mathcal{L}_{inter}), \]

(Jin et al., ACL’20; Yao et al., In Submission)
Explanation
Regularization

Adjust Attribution Scores

\[ \mathcal{L}^{\text{attr}} = \sum_{c} \sum_{p \in \mathcal{R}} (\phi^c(p; \mathbf{x}) - t^c_p)^2; \]

Adjust Interactions

Interaction between p("Sweden") and q("failure") towards the prediction c (Non-hate)

\[ \mathcal{L}^{\text{inter}} = \sum_{c} \sum_{\{p, q\} \in \mathcal{R}} (\varphi^c(p, q; \mathbf{x}) - \tau^c_{p, q})^2. \]

Final Loss Term

\[ \mathcal{L} = \mathcal{L}' + \alpha (\mathcal{L}^{\text{attr}} + \mathcal{L}^{\text{inter}}), \]

(Jin et al., ACL’20; Yao et al., In Submission)
Explanation

Regularization

Adjust Attribution Scores

\[
L^{\text{attr}} = \sum_c \sum_{p \in \mathcal{R}} (\phi^c(p; \mathbf{x}) - t^c_p)^2;
\]

Adjust Interactions

\[
L^{\text{inter}} = \sum_c \sum_{\{p, q\} \in \mathcal{R}} (\varphi^c(p, q; \mathbf{x}) - r^c_{p, q})^2.
\]

“Increase”, adjust to one.

Final Loss Term

\[
L = L' + \alpha(L^{\text{attr}} + L^{\text{inter}}),
\]

(Jin et al., ACL’20; Yao et al., In Submission)
Results: Hate Speech (Binary) Classification

Source dataset: HatEval → “source model”
Target dataset: Gap Hate Corpus (HGC)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>HatEval → GHC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metrics</td>
<td>Source F1 (↑)</td>
</tr>
<tr>
<td>Source model</td>
<td>64.2±0.3</td>
</tr>
</tbody>
</table>

*With only reg.*

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>- Hard reg. with IG</td>
<td>63.2±0.6</td>
<td>34.4±1.4</td>
<td>197.2</td>
</tr>
<tr>
<td>- Hard reg. with SOC</td>
<td>63.1±0.4</td>
<td>37.6±2.6</td>
<td>73.6</td>
</tr>
<tr>
<td>- Soft reg. with IG</td>
<td>63.2±0.3</td>
<td>33.2±0.8</td>
<td>204.9</td>
</tr>
<tr>
<td>- Soft reg. with SOC</td>
<td>63.2±1.1</td>
<td>39.5±1.5</td>
<td>19.4</td>
</tr>
</tbody>
</table>

**Source vs. Target F1:** model’s performance on source vs. target dataset

**FPRD:** false-positive rate difference → metric of model fairness
Take-aways

• “One explanation generalizes to many examples” --- better label efficiency vs. conventional supervision

• “Explanation carries more information than label” --- learning reliable & robust models

• Model updates via attribution/interaction on features & their compositions

• A new paradigm for constructing & maintaining NLP models?
Thank you!

USC Intelligence and Knowledge Discovery (INK) Lab

http://inklab.usc.edu/

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

xiangren@usc.edu

@xiangrenNLP