That Doesn't Make Sense!
A Case Study in Actively Annotating Model Explanations

Sameer Singh
University of California, Irvine

NIPS 2017 Workshop on Learning with Limited Labeled Data
Relation Extraction

Given two entities, and all the sentences that mention them, Identify the relations expressed between them.

Relation Extractor

birthplace
visited

President Obama was born in Hawaii ... Obama is a native of the state of Hawaii...

The president visited Hawaii as part of his ... Barack often wears grass skirts, a common costume in Hawaii.
Can we explain predictions to help users understand and debug?
Most people are married to one person. “is native to” is same as birthplace relation.

I don’t understand. Give me labeled data.

Sigh... okay.
Barack, spouseOf, Michelle ✔
Barack, spouseOf, Ann Dunham ✗

How can we make it easy for users to inject prior knowledge?
Current Supervision Approaches

Problem 1: Each annotation takes time
Problem 2: Each annotation is a drop in the ocean
A More Intuitive Paradigm

Explanation

\[ X \text{ was at wedding with } Y \Rightarrow X \text{ husband of } Y \Rightarrow \text{spouseOf}(X, Y) \]

Annotated Explanation

\[ X \text{ was at wedding with } Y \Rightarrow X \text{ husband of } Y \checkmark \Rightarrow \text{spouseOf}(X, Y) \]

User

Prediction

\[ \text{spouseOf}(\text{Barack}, \text{AnnDunham})? \]

Learned Model

Learn

Update Model

Pick
Explaining Relation Extraction

**Explanation**

X was at wedding with Y
=> X husband of Y
=> spouseOf(X, Y)

**User**

**Annotated Explanation**

X was at wedding with Y
X => X husband of Y
✓⇒ spouseOf(X, Y)

**Prediction**

spouseOf(Barack, AnnDunham)?

**Learned Model**

**Update Model**

**Learn**

**Pick**

**Explain**
Implication Chains as Explanations

spouseOf(Barack, AnnDunham)? ➔ X was at wedding with Y

employee(Marvin Minsky, MIT)? ➔ X was a cognitive scientist at Y

\[ e(x, \theta) = \arg\max_{e \in \mathcal{E}(x)} Q(e, x, \theta) + I(e) \]

Relation prediction

All implication chains
Possible chains that are valid for the prediction x

Faithfulness of the chain
How much does the model believe in the chain?

Interpretability
Keep chains short
Explaining Relation Extraction

Prediction to explain: spouseOf(Barack, AnnDunham)

Logic Implication Chains
sequence of steps to get the prediction

\[
\begin{align*}
X \text{ was at wedding with } Y &\implies X \text{ husband of } Y \\
&\implies \text{spouseOf}(X, Y)
\end{align*}
\]

Model’s belief in the explanation

\[\mathcal{P}_{\theta}(f)\]

Original Model

Space of all possible descriptions
Logic Representation of Relations

- Relations are binary predicates
  \[ \text{bornIn}(a, b) = \top \text{ or } \bot \]
  \[ \text{was-born-in}(a, b) = \top \text{ or } \bot \]
  where \( a, b \in \{ \text{“Bernie Sanders”, “Brooklyn”, “Michelle Obama”, \ldots} \} \)

- Facts are ground atoms:
  \[ \mathcal{F} = \left\{ \begin{array}{l}
  \text{bornIn}(\text{Bernie Sanders}, \text{Brooklyn}) \\
  \text{was-born-in}(\text{Bernie Sanders}, \text{Brooklyn}) \\
  \text{spouse}(\text{Barack Obama}, \text{Michelle Obama}) \\
  \vdots
\end{array} \right\} \]

- Relation Extraction models maximize the probability of ground atoms
  \[ \theta^* = \arg \max_{\theta} \sum_{f \in \mathcal{F}} \log P_{\theta}(f) \]
Model’s belief in a formula $f$

- For facts, we know this belief:

- Otherwise, recurse...

$$\mathcal{P}_\theta(f) = \begin{cases} 
R(a,b) & \text{then compute directly} \\
\neg f' & \text{then } 1 - \mathcal{P}_\theta(f') \\
f_1 \land f_2 & \text{then } \mathcal{P}_\theta(f_1) \mathcal{P}_\theta(f_2) \\
\forall e \mathcal{P}_\theta(f(e)) & \text{then } \prod_e \mathcal{P}_\theta(f(e)) 
\end{cases}$$

$$\mathcal{P}_\theta(\forall a,b \text{ was-born-in}(a,b) \Rightarrow \text{bornIn}(a,b)) =$$

$$\prod_{a,b} 1 - \mathcal{P}_\theta(\text{was-born-in}(a,b)) (1 - \mathcal{P}_\theta(\text{bornIn}(a,b)))$$

\[ \forall a,b \text{ was-born-in}(a,b) \Rightarrow \text{bornIn}(a,b) \]
Explaining Relation Extraction

Explanation

X was at wedding with Y
=> X husband of Y
=> spouseOf(X, Y)

Prediction

spouseOf(Barack, AnnDunham)?

User

Annotated Explanation

X was at wedding with Y
X => X husband of Y
✓ => spouseOf(X, Y)

Learn

Update Model

Learned Model

Pick

12
Explaining Relation Extraction

User

Explanation
X was at wedding with Y
=> X husband of Y
=> spouseOf(X,Y)

Annotated Explanation
X was at wedding with Y
X => X husband of Y
✓ => spouseOf(X,Y)

Learn

Prediction
spouseOf(Barack, AnnDunham)?

Pick

Learned Model

Update Model
Learning from Logical Knowledge

Explanation

X was at wedding with Y
=> X husband of Y
=> spouseOf(X, Y)

Prediction

spouseOf(Barack, AnnDunham)?

Learned Model

Learn

Annotated Explanation

X was at wedding with Y
X => X husband of Y
✓ => spouseOf(X, Y)

User

Update Model

Many different options
- Generalized Expectation
- Posterior Regularization
- Labeling functions in SNORKEL
- Andrew’s and Sebastian’s talks

[ Rocktaschel et al, NAACL 2015 ]
Logical Statements as Supervision

- If you see “was a native of”, it means birthplace
  \[ X \text{ was native of } Y \Rightarrow \text{birthplace}(X,Y) \]

- If a founder of the company is employed by the company, he’s the CEO
  \[ X \text{ is the founder of } Y \land \text{employee}(X,Y) \Rightarrow \text{ceoOf}(X,Y) \]

- Everyone is married to at most one person
  \[ \text{spouse}(X,Y) \Rightarrow \forall Y' \neg \text{spouse}(X,Y') \]
Improving the model

- Our model is maximizing probability of ground atoms
  \[ \theta^* = \arg\max_{\theta} \sum_{f \in \mathcal{F}} \log \mathcal{P}_\theta(f) \]

- But now we have a set of formulae, ground or otherwise
  \[ \mathcal{F} = \left\{ \begin{array}{l}
  \text{bornIn}(\text{Bernie Sanders, Brooklyn}) \\
  \text{was-born-in}(\text{Bernie Sanders, Brooklyn}) \\
  \text{spouse}(\text{Barack Obama, Michelle Obama}) \\
  \forall a, b \quad \text{was-born-in}(a, b) \Rightarrow \text{bornIn}(a, b) \\
  \vdots 
\end{array} \right\} \]

- Still maximizing the probability:
  \[ \theta^* = \arg\max_{\theta} \sum_{f \in \mathcal{F}} \log \mathcal{P}_\theta(f) \]

- Optimized using gradient descent
  - works for most models!
We’re evaluating whether formulae can be used instead of labeled data.

- ~2 million docs
- ~400,000 Entity pairs
- ~4000 columns
- ~50 Relations of interest
- 30 Logical Implications
Zero-Shot Learning

- Only Data (Random): 3
- Only Rules: 10
- Augment Data w/ Logic: 21
- Maximum Likelihood: 38
Learning from Logical Knowledge

X was at wedding with Y
=> X husband of Y
=> spouseOf(X,Y)

spouseOf(Barack, AnnDunham)?

X was at wedding with Y
X => X husband of Y
✓ => spouseOf(X,Y)

User

Explanation

Prediction

Learned Model

Update Model

Learn

Pick

Explain

Annotated Explanation
Learning from Logical Knowledge

Explanation:

- X was at wedding with Y
- => X husband of Y
- => spouseOf(X, Y)

Prediction:

spouseOf(Barack, AnnDunham)?

User

Learned Model

Annotated Explanation:

- X was at wedding with Y
- X => X husband of Y
- ✓ => spouseOf(X, Y)

Update Model
Picking What to Annotate

**Explanation**
- X was at wedding with Y
  - => X husband of Y
  - => spouseOf(X, Y)

**User**

**Annotated Explanation**
- X was at wedding with Y
  - X => X husband of Y
  - ✔ => spouseOf(X, Y)

**Learn**

**Pick**

**Prediction**
- spouseOf(Barack, AnnDunham)?

**Learned Model**

**Update Model**
Picking the Constraint

- **Active Learning**: Annotation that effects the model the most
  - Most uncertain example, since both true and false lead to change
- Should we pick the **most uncertain constraint**?
  - \( X \text{ was born in } Y \Rightarrow X \text{ died in } Y \) \( \times \)
  - If model didn’t believe it anyway, nothing changes
- Should we pick the **most certain constraint**?
  - Likely to be correct!
  - \( X \text{ was born in } Y \Rightarrow X \text{ birthPlace } Y \) \( \checkmark \)
- Pick most confident constraint that is likely to be wrong
  - **What we do**: Most confident explanation of most uncertain example
Picking What to Annotate

Prediction
spouseOf(Barack, AnnDunham)?

Prepare Explanation
X was at wedding with Y
=> X husband of Y
=> spouseOf(X,Y)

Picking What to Annotate

Learn

User
Annotated Explanation
X was at wedding with Y
X => X husband of Y
✓ => spouseOf(X,Y)

Learned Model
Update Model

Pick
Closing the Loop

**Explanation**

X was at wedding with Y
=> X husband of Y
=> spouseOf(X, Y)

**User**

**Annotated Explanation**

X was at wedding with Y
X => X husband of Y
✓ => spouseOf(X, Y)

**Prediction**

spouseOf(Barack, AnnDunham)?

**Learn**

**Pick**

**Learned Model**

**Update Model**
Crowd Sourcing Annotations

- Generate textual phrases from dependency paths
- Annotate individual implications, 5 labels each ($0.05 per label)

Think about the facts that the following phrases suggest:

1. "X, son of Y", and
2. "X is a child of Y".

Do you think something in the first statement might imply the second?

- YES, the first phrase strongly conveys the second.
- Yes it does, but only weakly.
- I can't tell, not sure.
- No, the implication is quite weak.
- Not at all, there is very little connection between the two.
Effort of Annotations

![Average Annotation Time Graph]

- **Labels**: Average time (ms) ~ 9000
- **Explanations**: Average time (ms) ~ 7000

The graph shows a comparison between the average annotation times for labels and explanations.
Quality of Annotations

![Interannotator Agreement Graph]

- Random: Low agreement
- Labels: Moderate agreement
- Explanations: High agreement

The graph shows the level of interannotator agreement for different types of annotations.
Closing the Loop on the Trained Model

Single round of annotating explanations, and incorporating them

- 150 total implications, 5 annotators each
Interactive Relation Extraction

Real-world, large-scale application of ML and NLP
But suffers from the need for a large amount of labeled data

Actively Annotating Model Explanations

Labeled Data: *is expensive, noisy, and time-consuming to obtain*
Explanations: *are simple chains of logical implications*
Feedback on Explanations: *much easier for users to annotate*
Open Questions

• **Evaluation:**
  - How much does labeling explanations help over instances?
  - How much does it help to be “active”?

• **Pick:**
  - What is the optimal explanation to show user?
  - What is a good approximation of that?

• **Explain:**
  - Can the explanations always be black-box?
  - Can we surface latent spaces for annotation directly?

• **Learn:**
  - How do we balance higher-level supervision with observed data?
spouseOf(Barack, AnnDunham)?

X was at wedding with Y => X husband of Y => spouseOf(X, Y)

User

Explanation

Annotated Explanation

X was at wedding with Y
X => X husband of Y
✔ => spouseOf(X, Y)

Learn

Prediction

Pick

Learned Model

Update Model

Thank you!

In collaboration with Carlos Guestrin, Sebastian Riedel, Luke Zettlemoyer, Tim Rocktaschel