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

+ Computing Power
Recipe for Modern NLP Applications

Model architectures and computing power are transferrable across applications

labeled data is not!
Creating Labeled Data for Relation Extraction

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

---

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

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 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.

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

citizen.
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

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**

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

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 → ORG: FOUNDED_BY
SUBJ-PER born in OBJ-LOC → PER: ORIGIN

1. Hard-matching

Hard-matched instances

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

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

Unmatched instances

(x_i, y_i)

(x_i, y_i, matching score)

Relation Classifier

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

Labeling Rules

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

2. Soft-matching

Unmatched instances

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

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

- ORG: FOUNDED_BY 0.8
- ORG: FOUNDED_BY 0.7

(Zhou et al, WWW20)
Joint Parameter Learning: Relation Extractor + Soft Rule Matcher

Labeling Rules

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

Matched Sentences

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

Unmatched Sentences

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

Soft-matching

Matched Sentences $(x_i, y_i)$

Unmatched Sentences $(x_i, y_i, matching score)$

Cross-entropy loss on relation labels $L_{rules}$

$(Zhou \ et \ al, \ WWW20)$
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

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

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** → **ORG: FOUNDED_BY**
**SUBJ-PER** born in **OBJ-LOC** → **PER: ORIGIN**

Matched Sentences

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

Unmatched Sentences

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

Labels

- ORG: FOUNDED_BY

Labels + Matching Score

- ORG: FOUNDED_BY 0.8
- ORG: FOUNDED_BY 0.7

Relation Classifier

(Zhou et al, WWW20)
Results on Relation Extraction

Relation Extraction Performance (in F1 score) on TACRED

<table>
<thead>
<tr>
<th>Method</th>
<th>F1 Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rules</td>
<td>20.1</td>
</tr>
<tr>
<td>LSTM + ATT</td>
<td>38.8</td>
</tr>
<tr>
<td>NERO (w/o Sunmatched)</td>
<td>42.9</td>
</tr>
<tr>
<td>Self-Training</td>
<td>39.2</td>
</tr>
<tr>
<td>Mean-Teacher</td>
<td>43.6</td>
</tr>
<tr>
<td>NERO (-SRM)</td>
<td>45.3</td>
</tr>
<tr>
<td>NERO (-Lrules)</td>
<td>49.0</td>
</tr>
<tr>
<td>NERO</td>
<td>51.3</td>
</tr>
</tbody>
</table>
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**?

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

→ **Positive**, because the words “very nice” is within 3 words after the ENT.

Relation between ENT1 and ENT2?

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

→ **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](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 ) ) )

......

......

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 )

Candidate scoring

\[ P(x|e_i) = \frac{\exp \theta^T \phi(f)}{ \sum_{f' : f' \in Z_{e_i}} \exp \theta^T \phi(f') } \]

\[ L_{parser} = \sum_{i=1}^{\left| S' \right|} \log \left( \sum_{f : f(x_i) = 1 \land h(f) = y_i} P_\theta(f|e_i) \right) \]

(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”? 

Label result

Label: Positive

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

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

Challenge 2: *compositional nature* of the explanations

per: nationality, because the words "is a" appear right before ENT2 and the word "citizen" is right after ENT2.
Learning with Hard & Soft Matching

\[ U = \{x_i\} \]

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

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

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

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

(Wang et al., ICLR20)
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**
- 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)
```

Explanation

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., ICLR20)
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., ICLR20)
Neural Execution Tree (NExT) for Soft Matching

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

(Wang et al., ICLR20)
Modules in NeXT

1. String matching

2. Soft counting

3. Soft logic

\[
p_1 \land p_2 = \max(p_1 + p_2 - 1, 0),
\]
\[
p_1 \lor p_2 = \min(p_1 + p_2, 1), \quad \neg p = 1 - p,
\]

4. Deterministic functions

(Wang et al., ICLR20)
Study on Label Efficiency (TACRED)

Annotation time cost: 
* giving a label + an explanation \(\approx\) 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
  - \textit{SUBJ-ORG} was founded by \textit{OBJ-PER}
  - \textit{→ 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 \textit{TERM}.  

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

Class Label: if the choice is correct or not

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

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].

Most plausible predictions are far from common truth

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

CONCEPTNET

An open, multilingual knowledge graph
Neural-Symbolic Reasoning with Commonsense KG

Symbol Space

Semantic Space

Where do adults use glue sticks?
A: classroom  B: office  C: desk drawer

Question

Answer Candidates

(Bill Yuchen Lin et al. EMNLP19)
Multi-relational Graph as Inductive Bias

KagNet

Statement

Question

Answer

Language Encoder (e.g. BERT)

Statement Vector

Graph Vector

Graph Encoder*

Plausibility score

Concept Recognition

Question Concepts

Answer Concepts

Graph Construction via Path Finding

Schema Graph

(Bill Yuchen Lin et al. 2019)
KagNet: Knowledge-aware Graph Network

\[ g = \frac{\sum_{i,j} [R_{i,j} \cdot T_{i,j}]}{|C_q| \times |C_a|} \]

KagNet: Knowledge-aware Graph Networks for Commonsense Reasoning

(Bill Yuchen Lin et al. 2019)
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](https://www.tau-nlp.org/csqa-leaderboard)
Transferability

KagNet → BERT-Large

CSQA

SWAG

WSC

59.01% vs 56.53%

53.51% vs 51.23%

No Training!

Training!

Interpretability

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

A: fountain pen ✓ (KagNet); B: printer (BERT);
C: squid D: pencil case (GPT); E: newspaper

1. select concept pairs of high att. scores

ink –PartOf– fountain pen
ink –RelatedTo– container –IsA– fountain pen
fill –HasSubEvent– ink –AtLocation– fountain pen
fill –RelatedTo– container –IsA– fountain pen
write –UsedFor– pen
write –UsedFor– pen –IsA– fountain pen
paper –RelatedTo– write –UsedFor– fountain pen

...... 2. Ranking via path-level attn.
Conclusion

(\textit{Label-efficient}) Learning from high-level human supervisions that are \textit{abstractive, compositional}, and \textit{linguistically complex}

Q1 How to augment model training with rules?

\begin{itemize}
  \item \textbf{Soft rule grounding for data augmentation} (Zhou et al. WWW20)
\end{itemize}

Q2 How to handle compositional natural language input?

\begin{itemize}
  \item \textbf{Neural execution tree for NL explanation} (Wang et al. ICLR20)
\end{itemize}

Q3 How to incorporate prior knowledge as inductive bias?

\begin{itemize}
  \item \textbf{Knowledge-aware graph networks} (Lin et al. EMNLP19)
\end{itemize}
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

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
 Kenneth 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

Research Partnership

Funding

Microsoft Academic
SPARK AT STANFORD

Microsoft

J.P. Morgan
Google
Schmidt Family Foundation
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