

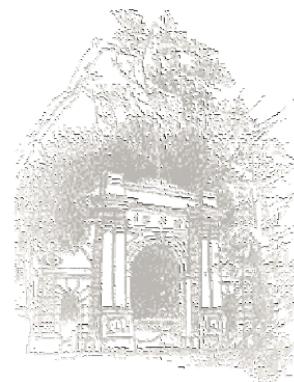
Reinforcement Learning in Natural Language Processing

Minlie Huang (黃民烈)

AI Lab., Dept. of Computer Science
Tsinghua University

aihuang@tsinghua.edu.cn

<http://coai.cs.tsinghua.edu.cn/html>

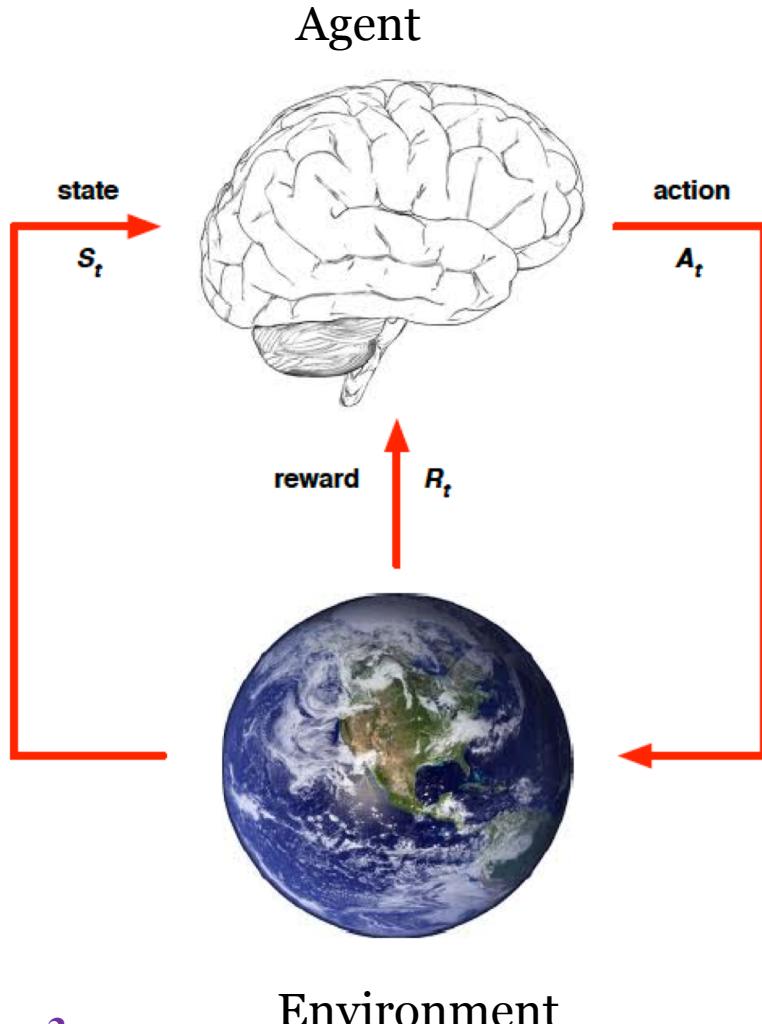


Our Recent Papers on RL

- Finding structures in text (**AAAI 2018**)
- Data Denoising in Relation Extraction (**AAAI 2018**)
 - ◆ One of **TOP 10 NLP** papers in **2017** voted by **PaperWeekly**
- Label correction in noisy labeling problems (**IJCAI-ECAI 2018**)
- Hierarchical Relation Extraction (**AAAI 2019**)
- Learning to Collaborate: Joint Ranking Optimization (**WWW 2018**)
 - ◆ Multi-agent reinforcement learning; deterministic policy; actor-critic
- Search Result Aggregation with HRL (in preparation to **SIGIR 2019**)



Reinforcement Learning



At each step t :

- The agent observes a **state S_t** from the environment
- The agent executes **action A_t** based on the observed state
- The agent receives scalar **reward R_t** from the environment
- The environment transfers into a new state **S_{t+1}**

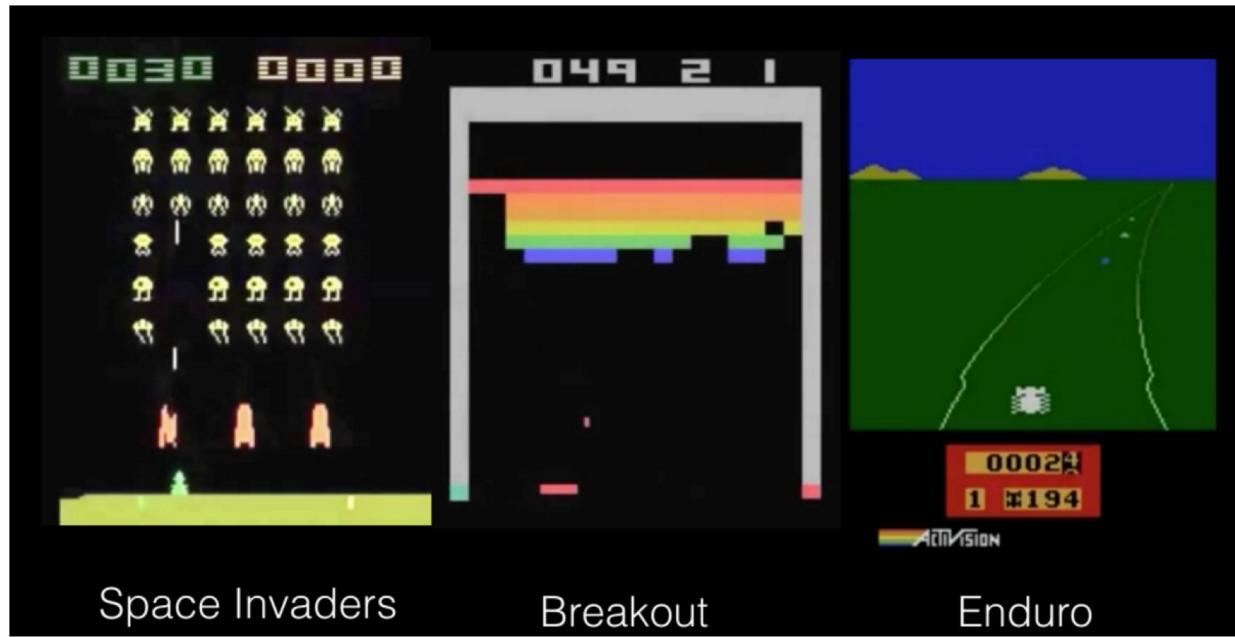
Reinforcement Learning

- **Sequential decision:** current decision affects future decision
- **Trial-and-error:** just try, do not worry about making mistakes
 - ◆ **Explore** (new possibilities)
 - ◆ **Exploit** (with the current best policy)
- **Future accumulative reward:** maximizing the future rewards instead of just the intermediate rewards at each step



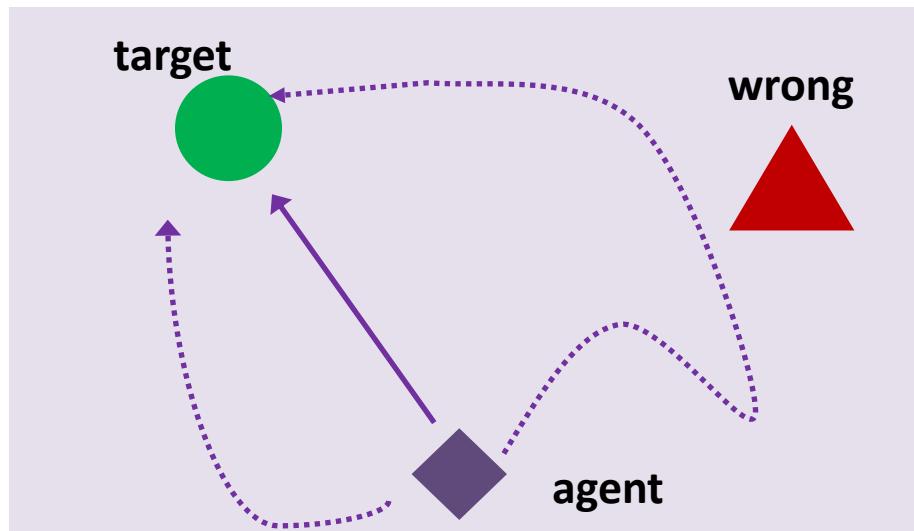
Difference to Supervised Learning

- Supervised learning: given a set of samples (x_i, y_i) , estimate $f: X \rightarrow Y$
- Given each example, the supervisor tells what is **correct**



Difference to Supervised Learning

- The agent know **what** a true goal is, but do not know **how to** achieve that goal
- Through interactions with the environment (trial-and-error)
- Many possible solutions (policies), which is optimal?



Applying RL in NLP

◎ Challenges

- ◆ **Sparse** reward (few feedback when making decisions)
- ◆ **Difficulty** in reward function design
- ◆ **High-dimensional** action space (e.g. **Language generation**)
- ◆ High **variance** in training RL algorithms
- ◆ Extremely **expensive** to involve **simulator** (e.g. **dialog**)



Applying RL in NLP

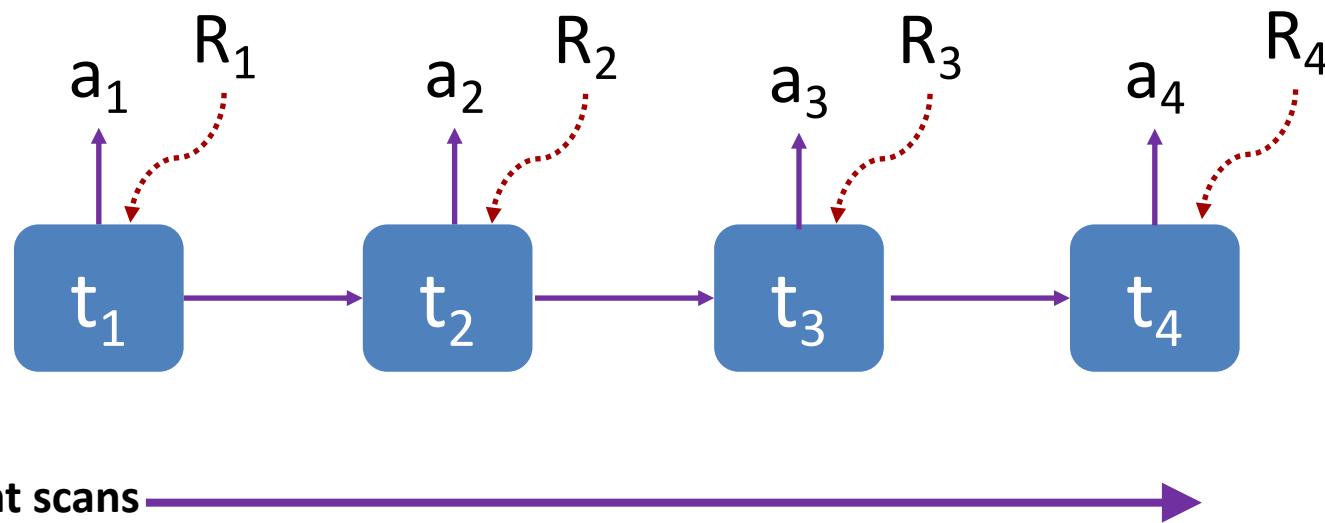
◎ Strengthens of RL

- ◆ **Weak supervision** without explicit annotations
- ◆ **Trial-and-error**: probabilistic exploring
- ◆ **Accumulative rewards**: encoding expertise/prior knowledge in reward design



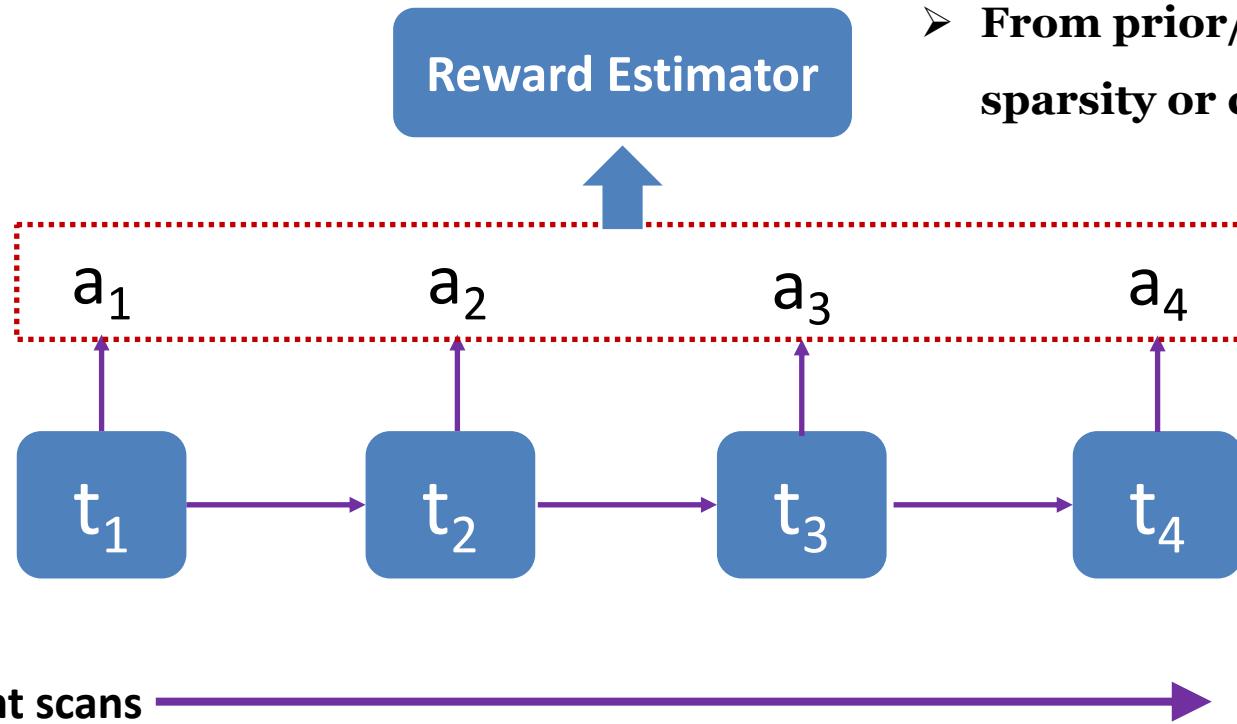
Applying RL in NLP

- Immediate rewards: t could be word/sentence, or any symbol



Applying RL in NLP

Delayed rewards



- Comparing with gold-standard:
BLEU\ACC\F1
- By classifier: likelihood
- From prior/domain expertise:
sparsity or continuity



Why RL in NLP

- Learning to **search and reason**
- Directly optimize the **end metrics** (BLEU, ROUGE, Acc, F₁)
 - ◆ Machine translation, language generation, summarization
- Make discrete operations “**BP-able**” in deep learning
 - ◆ Sampling
 - ◆ Argmax
 - ◆ Binary operations in neural networks



RL in NLP & Search

○ Search and reasoning

- ◆ Find optimal model structure (e.g. autoML)
- ◆ Search for representation structure
- ◆ Search for reasoning path

○ Instance selection

- ◆ Selecting unlabeled data in SSL or co-training
- ◆ Selecting mini-batch order in SGD
- ◆ Data denoising
- ◆ Label correction in noisy labeling

○ Strategy optimization

- ◆ Language generation, dialogue strategy, ranking systems



Learning Structured Representation for Text Classification

Tianyang Zhang, Minlie Huang, Li Zhao. Learning Structured
Representation for Text Classification via Reinforcement
Learning. **AAAI 2018.**

The Problem ...

- How can we identify task-relevant structures without explicit annotations on structure?

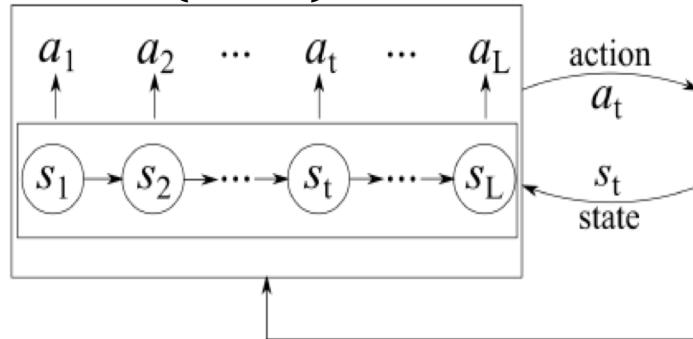
Origin text	Cho continues her exploration of the outer limits of raunch with considerable brio .
ID-LSTM	Cho continues her exploration of the outer limits of raunch with considerable brio .
HS-LSTM	Cho continues her exploration of the outer limits of raunch with considerable brio .
Origin text	Much smarter and more attentive than it first sets out to be .
ID-LSTM	Much smarter and more attentive than it first sets out to be .
HS-LSTM	Much smarter and more attentive than it first sets out to be .
Origin text	Offers an interesting look at the rapidly changing face of Beijing .
ID-LSTM	Offers an interesting look at the rapidly changing face of Beijing .
HS-LSTM	Offers an interesting look at the rapidly changing face of Beijing .

- Challenges
 - ◆ NO explicit annotations on structure-weak supervision
 - ◆ Trial-and-error, measured by delayed rewards

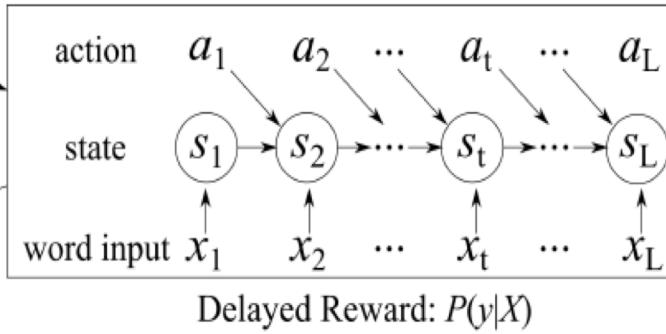


Model Structure

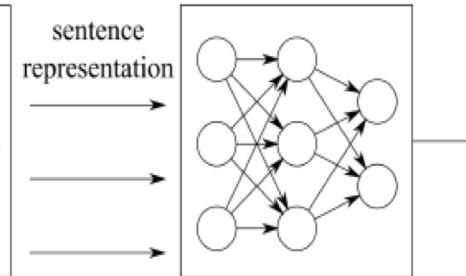
Policy Network (PNet)



Structured Representation Model



Classification Network (CNet)



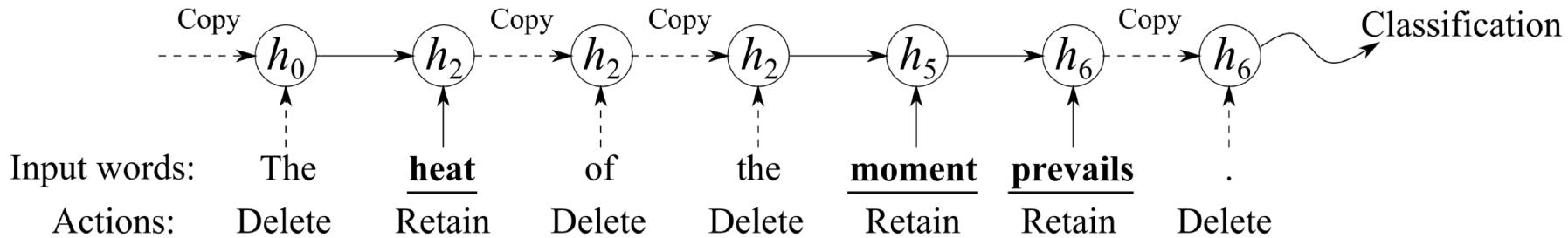
- **Policy Network** samples an action at each state when scanning the word sequence
- **Structured Representation Model** transfers action sequence to representation
 - ◆ Two models: **Information Distilled LSTM**, **Hierarchically Structured LSTM**
- **Classification Network** computes the likelihood as reward signal



Information Distilled LSTM (ID-LSTM)

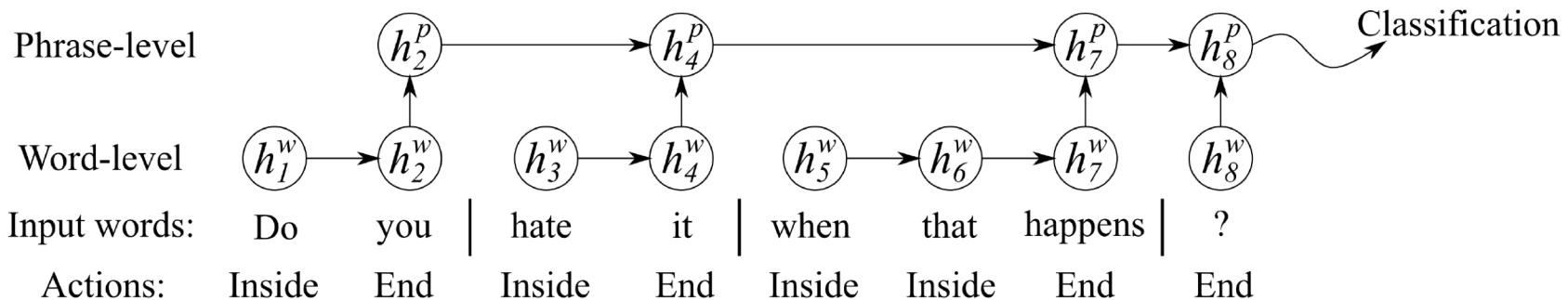
- Distill the most **important** words (or remove **irrelevant** words) to the task
- Reward signal: the classification likelihood

$$P(y|X) = \text{softmax}(\mathbf{W}_s \mathbf{h}_L + \mathbf{b}_s)$$



Hierarchically Structured LSTM(HS-LSTM)

- A structured representation by discovering hierarchical structures in a sentence
- Two-level structure:
 - ◆ Word-level LSTM + phrase-level LSTM
 - ◆ Sentence representation: the last hidden state of phrase-level LSTM



Experiment

Dataset

- ◆ **MR**: movie reviews (Pang and Lee 2005)
- ◆ **SST**: Stanford Sentiment Treebank, a public sentiment analysis dataset with five classes (Socher et al. 2013)
- ◆ **Subj**: subjective or objective sentence for subjectivity classification (Pang and Lee 2004)
- ◆ **AG**: AG's news corpus, a large topic classification dataset constructed by (Zhang, Zhao, and LeCun 2015)



Experiment

○ Classification Results

Models	MR	SST	Subj	AG
LSTM	77.4*	46.4*	92.2	90.9
biLSTM	79.7*	49.1*	92.8	91.6
CNN	81.5*	48.0*	93.4*	91.6
RAE	76.2*	47.8	92.8	90.3
Tree-LSTM	80.7*	50.1	93.2	91.8
Self-Attentive	80.1	47.2	92.5	91.1
ID-LSTM	81.6	50.0	93.5	92.2
HS-LSTM	82.1	49.8	93.7	92.5

○ Examples by ID-LSTM/HS-LSTM

Origin text	Cho continues her exploration of the outer limits of raunch with considerable brio .
ID-LSTM	Cho continues her exploration of the outer limits of raunch with considerable brio .
HS-LSTM	Cho continues her exploration of the outer limits of raunch with considerable brio .
Origin text	Much smarter and more attentive than it first sets out to be .
ID-LSTM	Much smarter and more attentive than it first sets out to be .
HS-LSTM	Much smarter and more attentive than it first sets out to be .
Origin text	Offers an interesting look at the rapidly changing face of Beijing .
ID-LSTM	Offers an interesting look at the rapidly changing face of Beijing .
HS-LSTM	Offers an interesting look at the rapidly changing face of Beijing .



Results of ID-LSTM (Word Deletion)

Dataset	Length	Distilled Length	Removed
MR	21.25	11.57	9.68
SST	19.16	11.71	7.45
Subj	24.73	9.17	15.56
AG	35.12	13.05	22.07

Table 4: The original average length and distilled average length by ID-LSTM in the test set of each dataset.

Word	Count	Deleted	Percentage
of	1,074	947	88.18%
by	161	140	86.96%
the	1,846	1558	84.40%
's	649	538	82.90%

but	320	25	7.81%
not	146	0	0.00%
no	73	0	0.00%
good	70	0	0.00%
interesting	25	0	0.00%



Results of HS-LSTM (Phrase Discovery)

Models	SST-binary	AG's News
RAE	85.7	90.3
Tree-LSTM	87.0	91.8
Com-Tree-LSTM	86.5*	—
Par-HLSTM	86.5	91.7
HS-LSTM	87.8	92.5

Table 8: Classification accuracy from structured models. The result marked with * is re-printed from (Yogatama et al. 2017).

Dataset	Length	#Phrases	#Words per phrase
MR	21.25	4.59	4.63
SST	19.16	4.76	4.03
Subj	24.73	4.42	5.60
AG	35.12	8.58	4.09

Table 9: Statistics of structures discovered by HS-LSTM in the test set of each dataset.



Data Denoising for Relation Classification

Jun Feng, Minlie Huang, Li Zhao, Yang Yang,
Xiaoyan Zhu. Reinforcement Learning for Relation
Classification from Noisy Data. **AAAI 2018**

Noisy Labeling in Distant Supervision

- **Relation Classification:** given two entities and a sentence, identify relation labels

[Obama]_{e1} was born in the [United States]_{e2}.



Relation: *BornIn*

- Distant Supervision (**noisy labeling problem**)

Triple in knowledge base:<Barack_Obama, *BornIn*, United_States>

[Barack Obama]_{e1} is ~~the 44th President of~~ the [United States]_{e2}.



Relation: *BornIn*



Noisy Labeling in Distant Supervision

- Previous studies adopt multi-instance learning to consider the instance noises

Barack_Obama, United_States

Obama was born in the United States.

Barack Obama is the 44th President of the United States

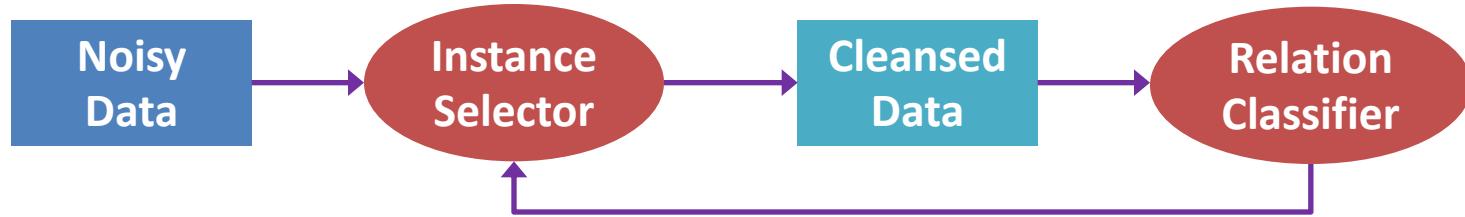
Relation

BornIn

How can we remove noisy data to improve relation extraction without explicit annotations?

Model Structure

- The model consists of an **instance selector** and a **relation classifier**



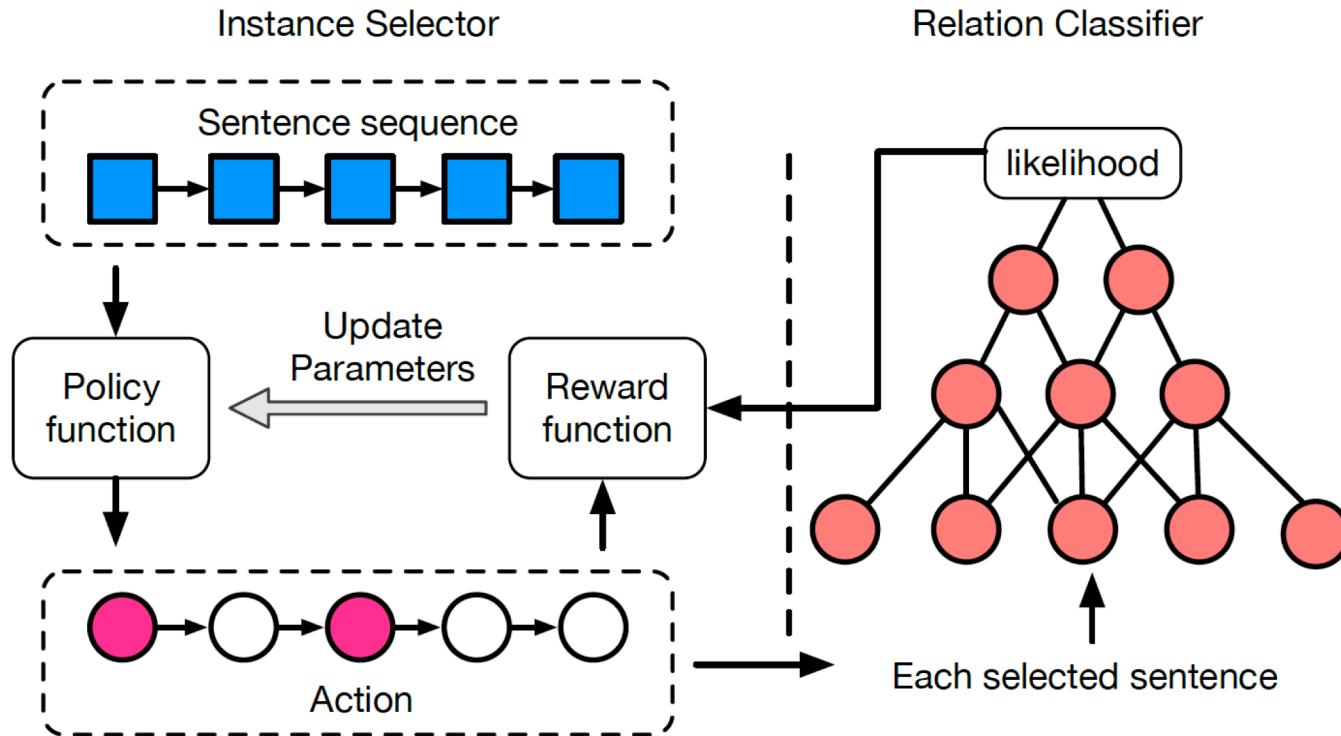
- Challenges:

- ◆ Instance selector has no explicit annotation on which sentences are labeled incorrectly

- **Weak supervision** -> delayed reward
- Trail-and-error search

Model Structure

$$r(s_i|B) = \begin{cases} 0 & i < |B| + 1 \\ \frac{1}{|\hat{B}|} \sum_{x_j \in \hat{B}} \log p(r|x_j) & i = |B| + 1 \end{cases}$$



$$\begin{aligned}\pi_\Theta(s_i, a_i) &= P_\Theta(a_i|s_i) \\ &= a_i \sigma(\mathbf{W} * \mathbf{F}(s_i) + \mathbf{b}) \\ &\quad + (1 - a_i)(1 - \sigma(\mathbf{W} * \mathbf{F}(s_i) + \mathbf{b}))\end{aligned}$$



Training Procedure

Overall Training Procedure

1. Pre-train the relation classifier (CNN)
2. Pre-train the policy network of the instance selector with the relation classifier frozen
3. Jointly train the relation classifier and the policy network



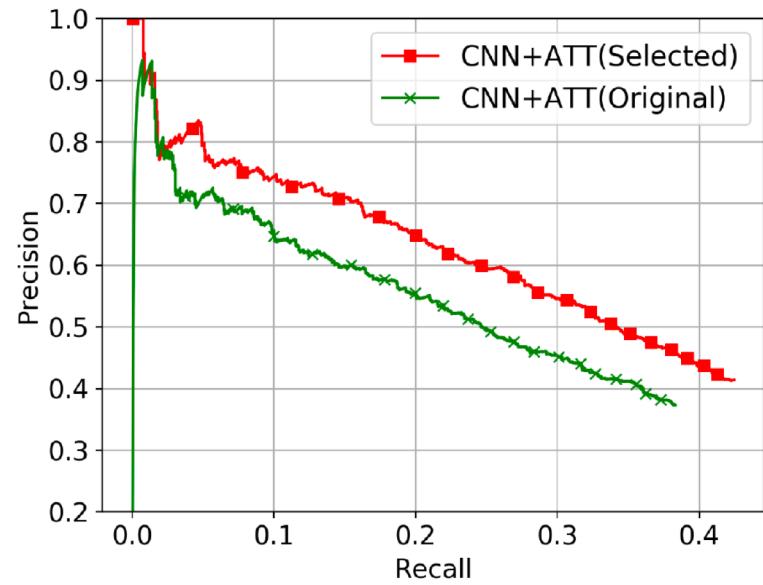
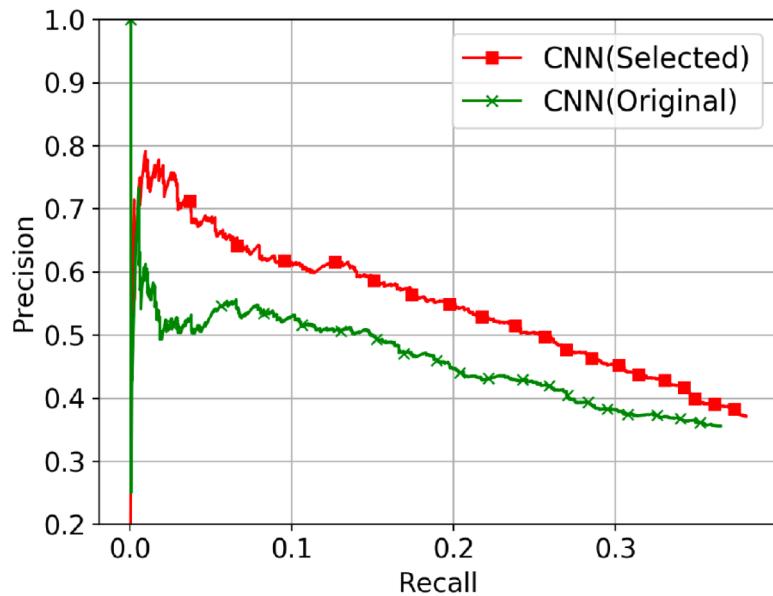
Experiment

- Dataset
 - ◆ **NYT** (Riedel, Yao, and McCallum 2010)
- Baselines
 - ◆ **CNN**: is a sentence-level classification model. It does not consider the noisy labeling problem.
 - ◆ **CNN+Max**: assumes that there is one sentence describing the relation in a bag and chooses the most correct sentence in each bag.
 - ◆ **CNN+ATT**: adopts a sentence-level attention over the sentences in a bag and thus can down weight noisy sentences in a bag.



Experiment

- Extraction performance (non-cleansed vs. cleansed)



Label Correction in Noisy Labeling for Topic Labeling

Ryuichi Takanobu, Minlie Huang, et al. A Weakly Supervised Method for Topic Segmentation and Labeling in Goal-oriented Dialogues via Reinforcement Learning. **IJCAI-ECAI 2018.**

The problem ...

Product-info	A:	The release date of < MODEL >???
	B:	< MODEL > will be available for pre-order on 19 April and launch on 26.
	A:	How long can the battery last?
	B:	It's equipped with a 4,000 mAh battery up to 8 hours of HD video playing or 10 hours of web browsing.
PaymentPromotion	A:	Can I use a coupon?
	B:	When entering your payment on the checkout page, click <i>Redeem a coupon</i> below your payment method.
	B:	You can check here for more details: < URL >.
	A:	OK. Support payment by installments?
PaymentPromotion	B:	Sure. We provide an interest-free installment option for up to 6 months.

Table 1: An example of customer service dialogues, translated from Chinese. Utterances in the same color are of the same topic.



The Challenge is NO Annotation!

- Too many data
- Too expensive

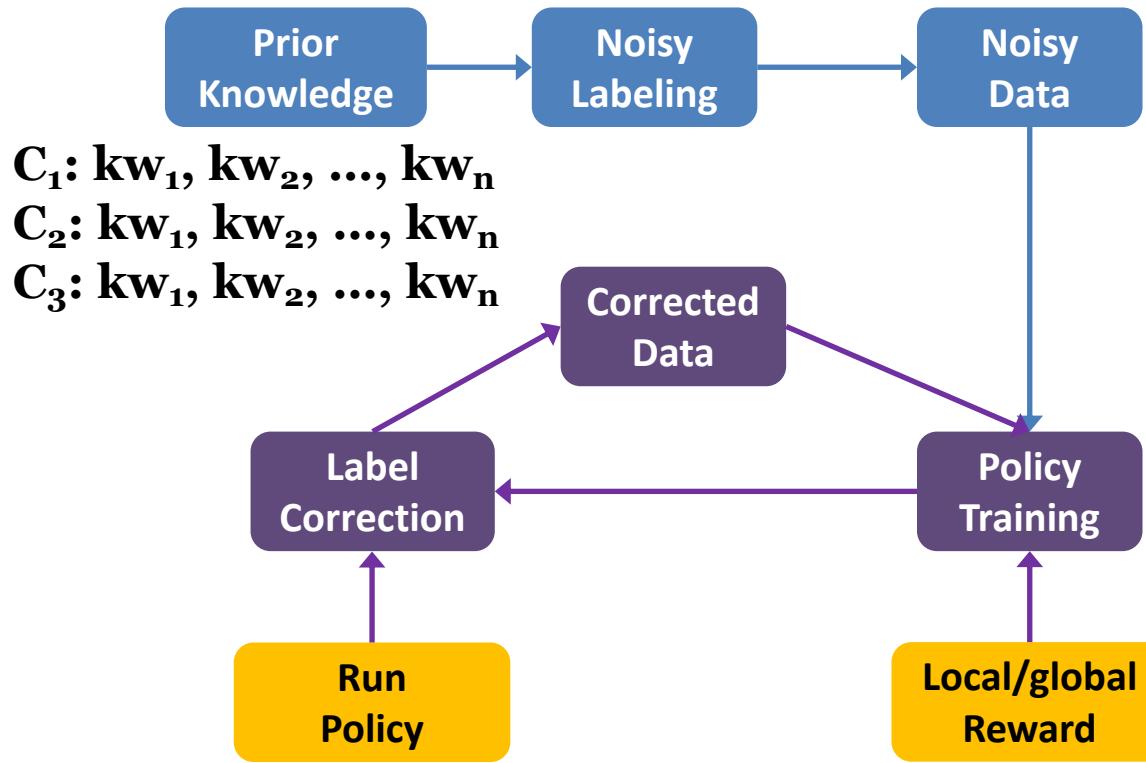
Datasets	SmartPhone	Clothing
# Topic category	7	10
# Training session	12,315	10,000
# Training utterance	430,462	338,534
# Gold-standard session	300	315
# Gold-standard utterance	10,888	10,962

Table 2: Statistics of the corpus.

How can we do topic labeling on these large-scale dialogues without much annotation efforts?

Central Idea

- Start from noisy data → correct data → refine policy



Model Structure

- State Representation Network
- Policy Network

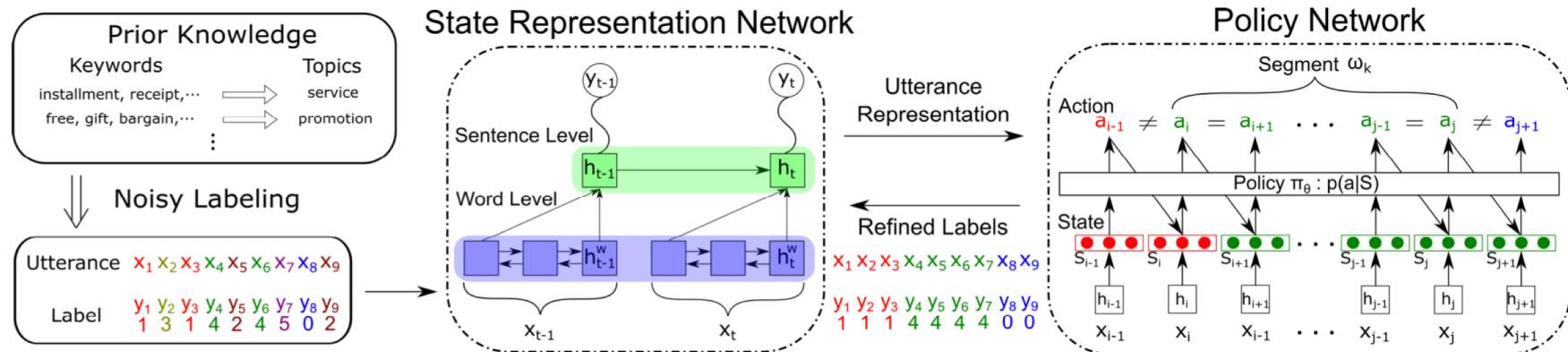


Figure 1: Illustration of the model. SRN adopts a hierarchical LSTM to represent utterances and provides state representations to PN. Data labels are refined to retrain SRN and PN to learn better state representations and policies. The label y and the action a are in the same space.



Central Idea

- ◎ **Local topic continuity:** the same topic will continue in a few dialogue turns

$$r_{int} = \frac{1}{L-1} \operatorname{sign}(a_{t-1} = a_t) \cos(\mathbf{h}_{t-1}, \mathbf{h}_t)$$

- ◎ **Global topic structure:** high content similarity within segments but low between segments

$$\begin{aligned} r_{delayed} &= \frac{1}{N} \sum_{\omega \in X} \frac{1}{|\omega|} \sum_{X_t \in \omega} \cos(\mathbf{h}_t, \boldsymbol{\omega}) \\ &\quad - \frac{1}{N-1} \sum_{(\omega_{k-1}, \omega_k) \in X} \cos(\boldsymbol{\omega}_{k-1}, \boldsymbol{\omega}_k) \end{aligned}$$



The Scanning Process

Sequence scan ↓

- C₁ **A:** The release date of < MODEL >???
- C₁ **B:** < MODEL > will be available for pre-order on 19 April and launch on 26.
-
- C₁ **A:** How long can the battery last?
- C₁ **B:** It's equipped with a 4,000 mAh battery up to 8 hours of HD video playing or 10 hours of web browsing.
-
- C₂ **A:** Can I use a coupon?
- C₂ **B:** When entering your payment on the checkout page, click *Redeem a coupon* below your payment method.
-
- C₂ **B:** You can check here for more details: < URL >.
-
- C₃ **A:** OK. Support payment by installments?
- C₃ **B:** Sure. We provide an interest-free installment option for up to 6 months.



Experiment

(a) Topic Segmentation (MAE and WD)

Model	SmartPhone		Clothing	
	MAE	WD	MAE	WD
TextTiling(TT)	13.09	.802	16.32	.948
TT+Embedding	3.59	.564	3.17	.567
STM	4.37	.505	8.85	.669
NL+HLSTM	8.25	.632	16.26	.925
Our method	2.69	.415	2.74	.446

(b) Topic Labeling (Accuracy)

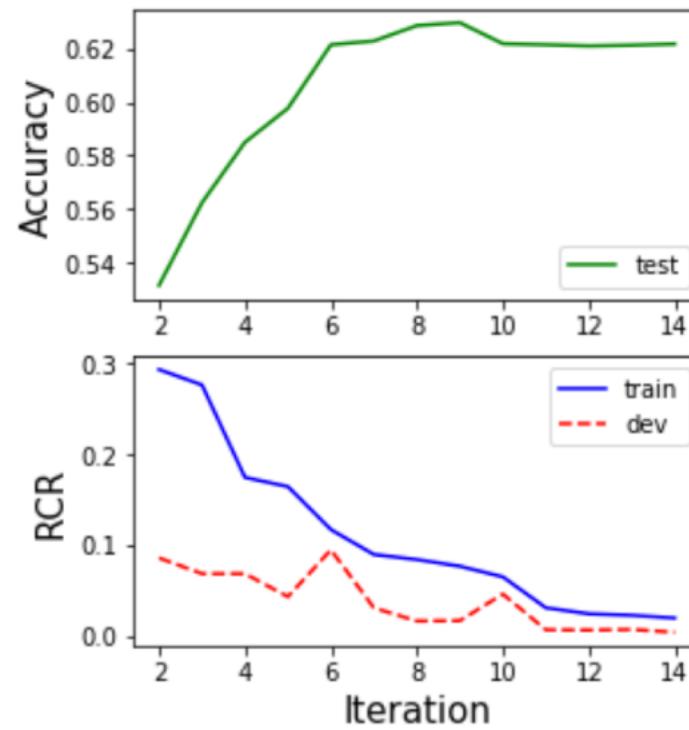
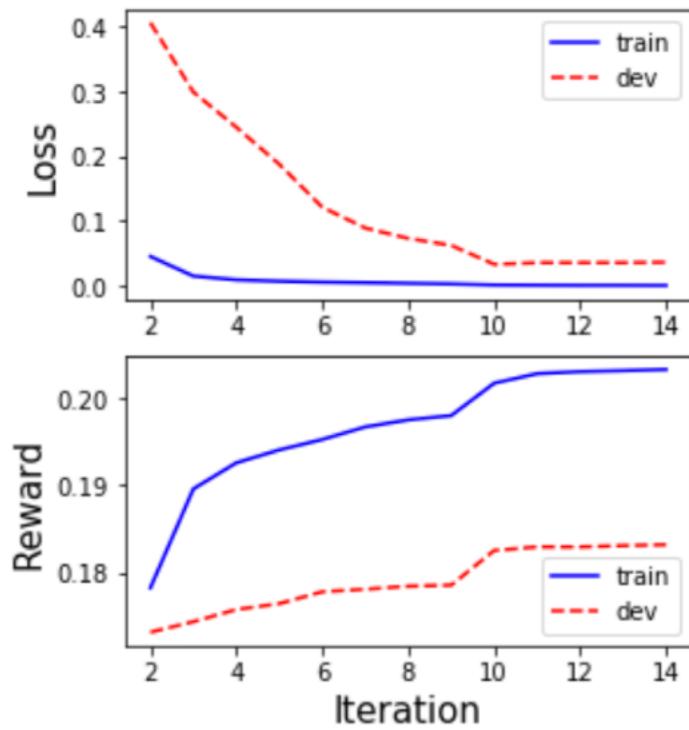
Model	SmartPhone	Clothing
Keyword Matching	39.8	31.8
NL	51.4	39.0
NL+LSTM	49.6	35.5
NL+HLSTM	52.6	40.1
Our method	62.2	48.0

Only with noisy labeling

With label correction

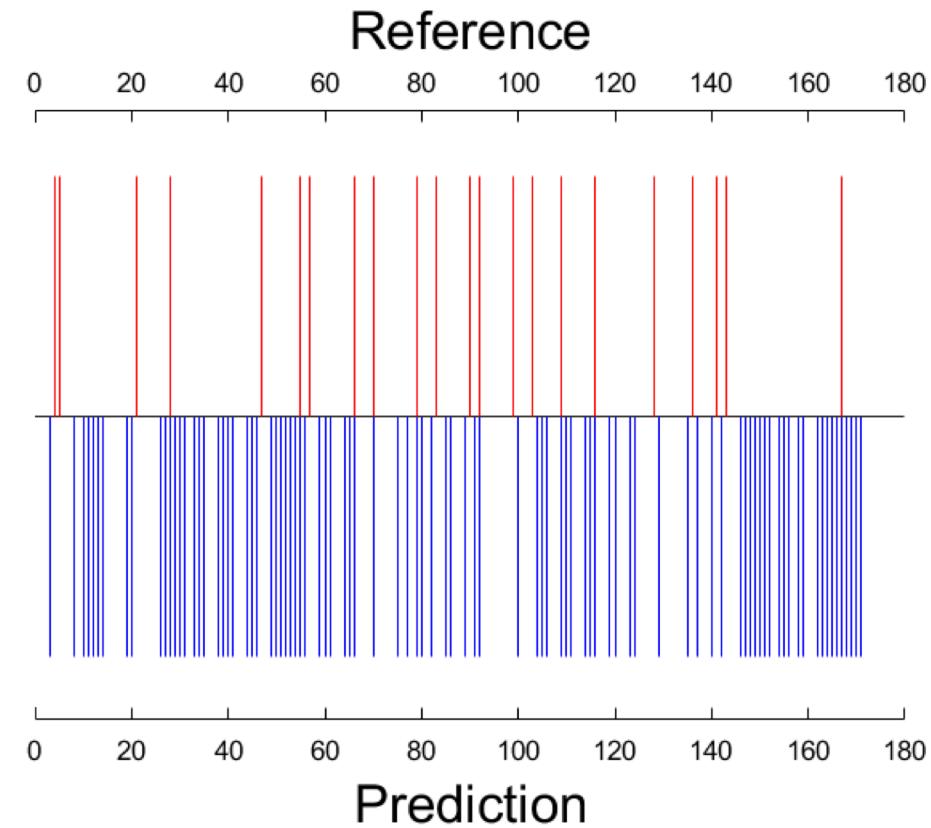
Experiment

- Training converges well (loss, reward, accuracy, RCR~relative change of data label)

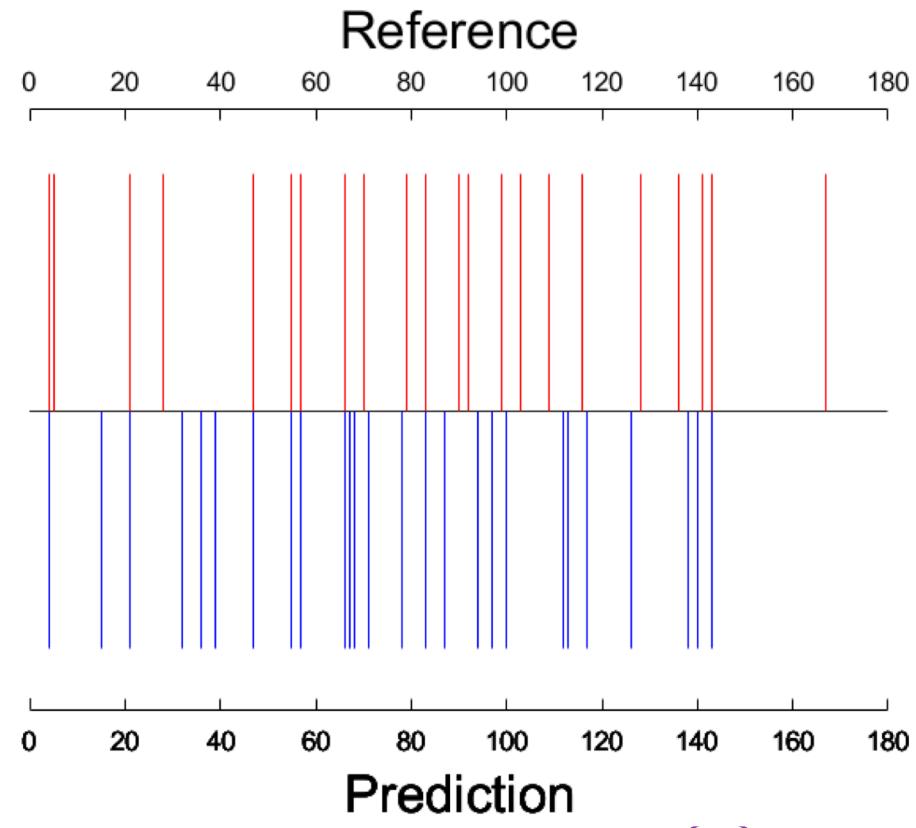


Visualization Examples

By Noisy Labeling



By Our RL Models



Hierarchical Reinforcement Learning for Relation Extraction

Ryuichi Takanobu, Tianyang Zhang, Jiexi Liu, Minlie Huang.

A Hierarchical Framework for Relation Extraction with
Reinforcement Learning. **AAAI 2019**

Relation Extraction

◎ Relation Extraction

Obama was born in the United States.



Relation Triple: ($[\text{Obama}]_{\text{es}}$, BornIn , $[\text{United States}]_{\text{et}}$)

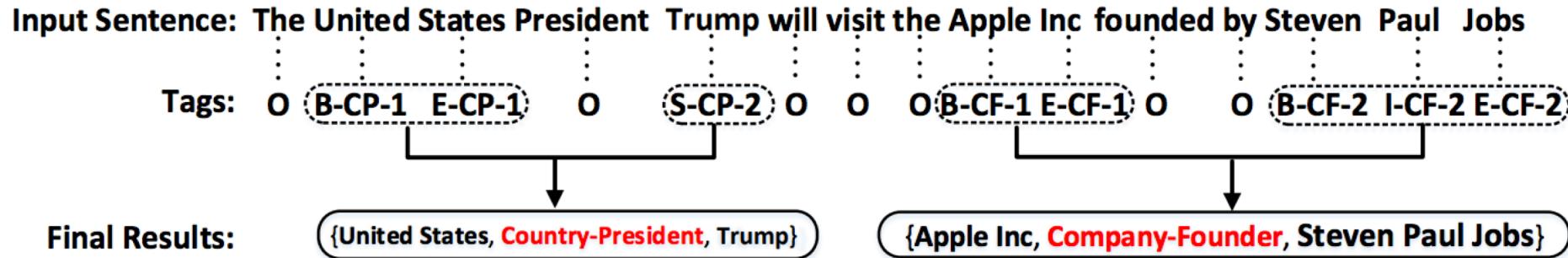
Source entity	Relation type	Target entity
---------------	---------------	---------------

- Joint extraction of **entity mentions** and **relation types**.



Existing Solutions

- Sequential Labeling (Tagging, Zheng et al. 2017)
 - ◆ Overlapping relations?



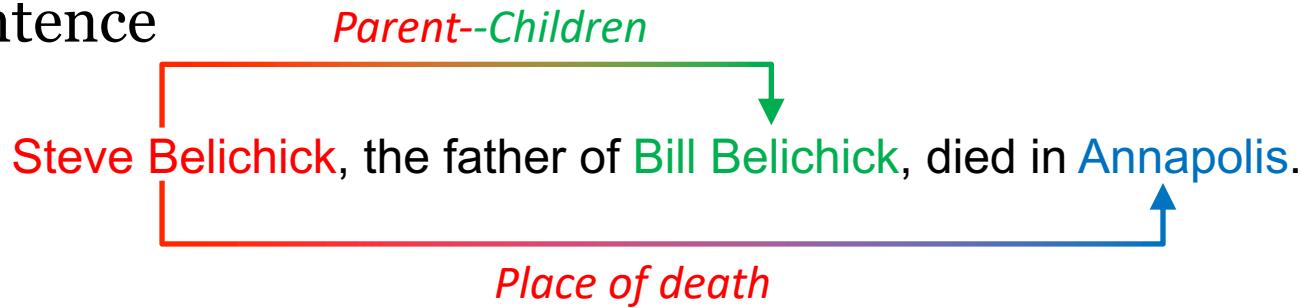
- Pair-wise relation prediction (SPTree, Miwa and Bansal 2016)
 - ◆ Enumerate all combinations



Motivation

- Complex **overlapping relations**.

- One entity participate in multiple relations in the same sentence

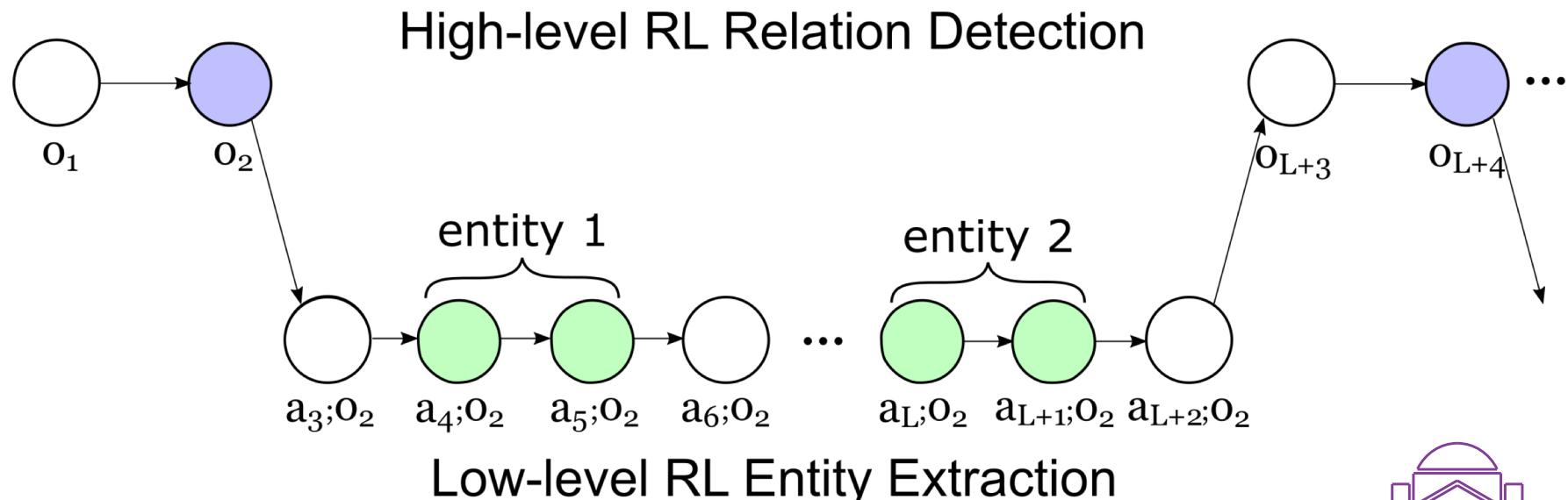


- Same entity pair in a sentence is associated with different relations



Framework

- Decomposing relation extraction into
 - ◆ **Relation indicator detection** (as option)
 - ◆ **Entity mention detection** (as primitive action, treating entity mention as argument of a relation)



An Illustration Example

Steve Belichick, the father of New England Patriots coach Bill Belichick, died of heart failure in Annapolis, at the age of 86 .

parent-children

Steve Belichick , the father of New England Patriots coach Bill Belichick ,

I

Steve Belichick , the father of New England Patriots coach Bill Belichick , died of heart failure in Annapolis , at the age of 86 .

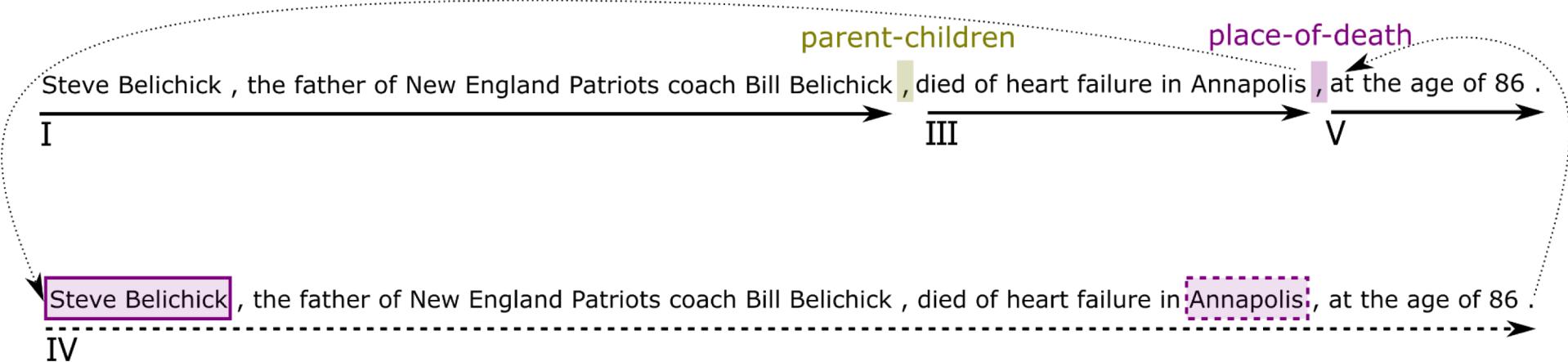
II





An Illustration Example

(Steve Belichick), the father of (New England Patriots) coach (Bill Belichick), died of heart failure in (Annapolis), at the age of 86 .



Experiment

Model	NYT10			NYT11		
	Prec	Rec	F_1	Prec	Rec	F_1
FCM	–	–	–	.432	.294	.350
MultiR	–	–	–	.328	.306	.317
CoType	–	–	–	.486	.386	.430
SPTree	.492	.557	.522	.522	.541	.531
Tagging	.593	.381	.464	.469	.489	.479
CopyR	.569	.452	.504	.347	.534	.421
HRL	.714	.586	.644	.538	.538	.538

Table 2: Main results on relation extraction.

Model	NYT10-sub			NYT11-plus		
	Prec	Rec	F_1	Prec	Rec	F_1
FCM	–	–	–	.234	.199	.219
MultiR	–	–	–	.241	.214	.227
CoType	–	–	–	.291	.254	.271
SPTree	.272	.315	.292	.466	.229	.307
Tagging	.256	.237	.246	.292	.220	.250
CopyR	.392	.263	.315	.329	.224	.264
HRL	.815	.475	.600	.441	.321	.372

47 Table 3: Performance comparison on extracting overlapping relations.

SPTree is a strong baseline
using dependency parsing trees



Summary

- In **weakly supervised** settings

- ◆ Finding text structures
- ◆ De-noising low-quality instances
- ◆ Re-assigning data labels
- ◆ Decomposing complex tasks to simple ones



Messages and Lessons

⦿ Keys to the success of RL in NLP

- ◆ Formulate a task as a **natural sequential decision** problem where current decisions affect future ones!
- ◆ Remember the **nature** of **trial-and-error** when we have no access to *full, strong supervision*.
- ◆ Encode the **expertise** or **prior knowledge** of the task in rewards.
- ◆ Applicable in many **weak supervision** settings.



Thanks for Your Attention

◎ Acknowledgements

- ◆ Prof. Xiaoyan Zhu, Dr. Li Zhao
- ◆ Jun Feng, Tianyang Zhang, Ryuichi Takanobu

◎ Contact

- ◆ Minlie Huang, Tsinghua University
- ◆ Email: aihuang@tsinghua.edu.cn
- ◆ Homepage: <http://coai.cs.tsinghua.edu.cn/hml>

