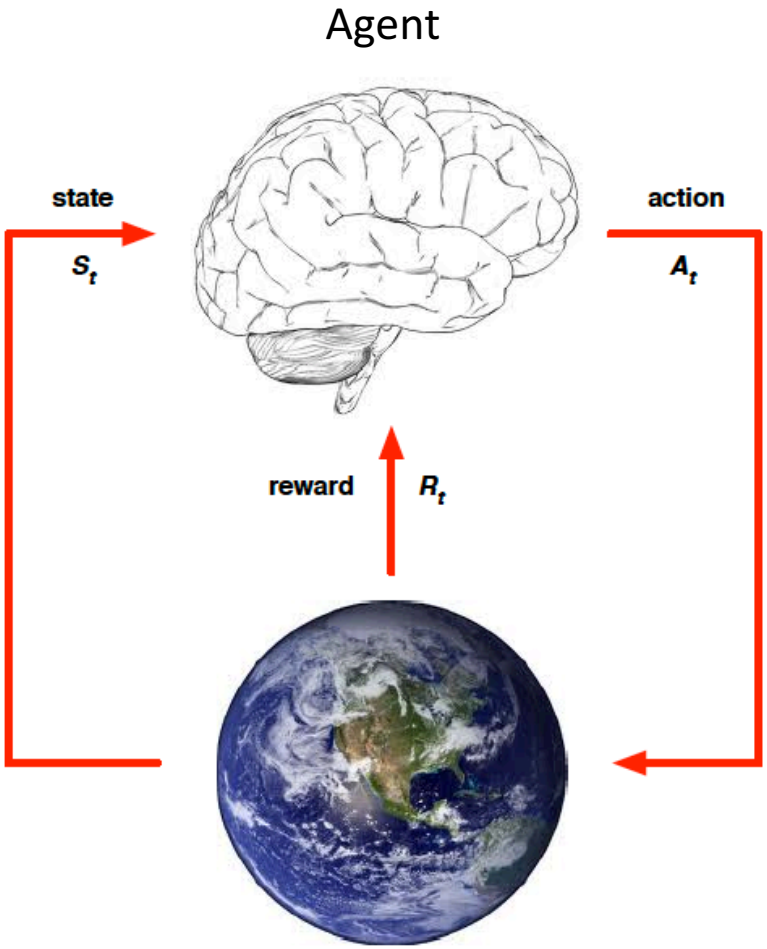
When Reinforcement Learning Meets NLP

Jun Feng

Tsinghua University

Introduction to RL

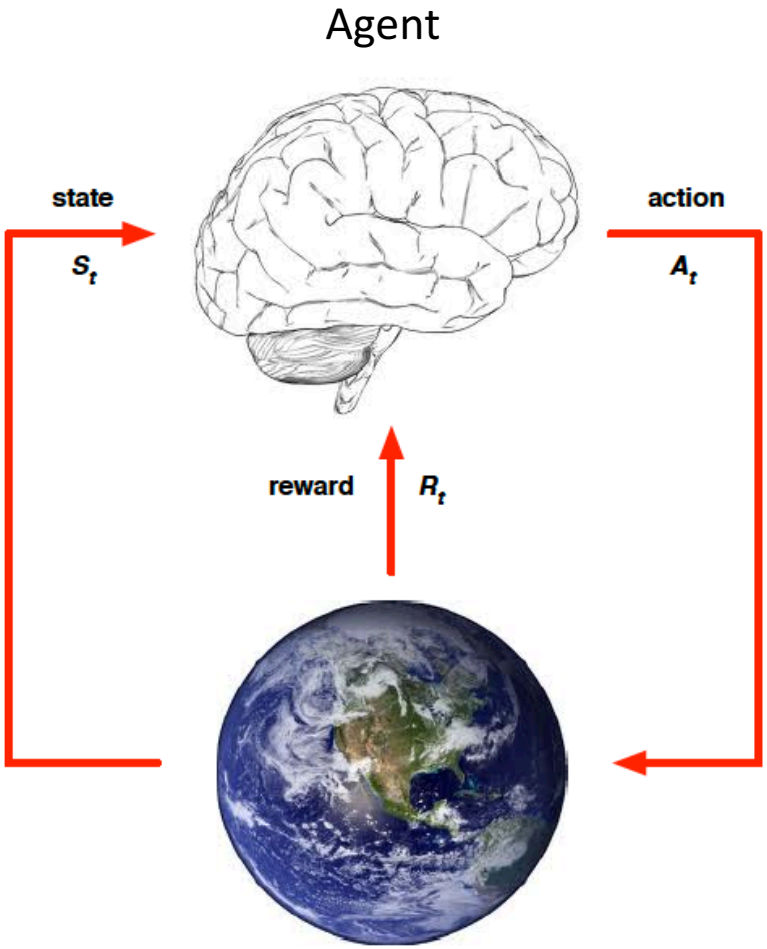


At each step t:

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

http://www0.cs.ucl.ac.uk/staff/d.silver/web/Teaching_files/intro_RL.pdf

Introduction to RL



Goal: select actions to maximize total future reward

- Trail-and-error search
- Actions may have long term consequences
- Reward may be delayed

Introduction to RL

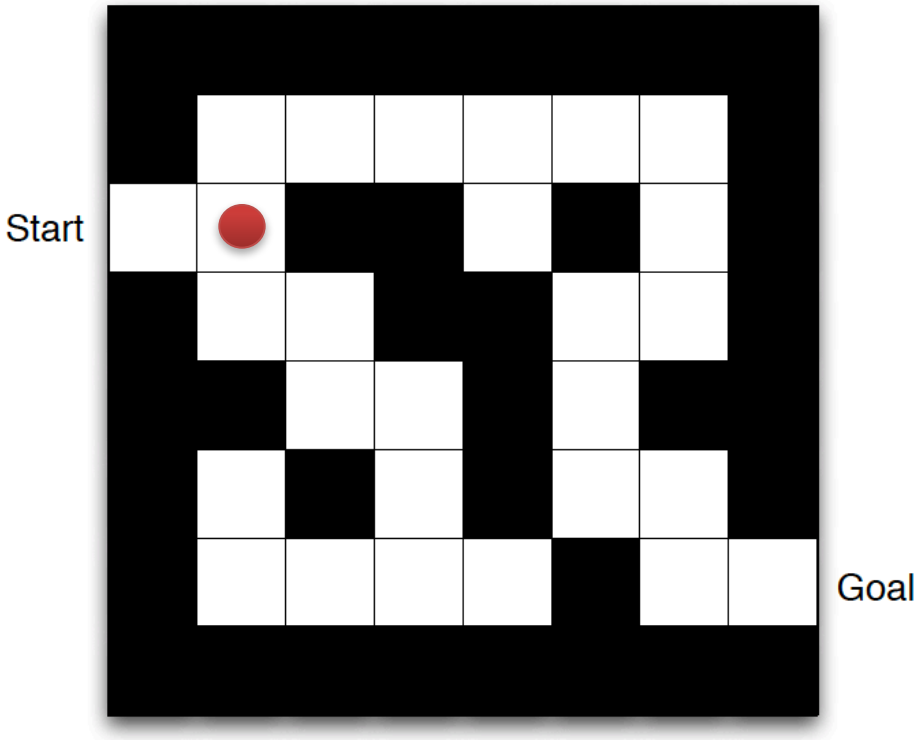
- Policy is the agent's behavior
  - maps from state to action
  - deterministic policy:

$$a = \pi(s)$$

  - stochastic policy:

$$\pi(a|s) = \mathbb{P}[A_t = a|S_t = s]$$

Maze Example

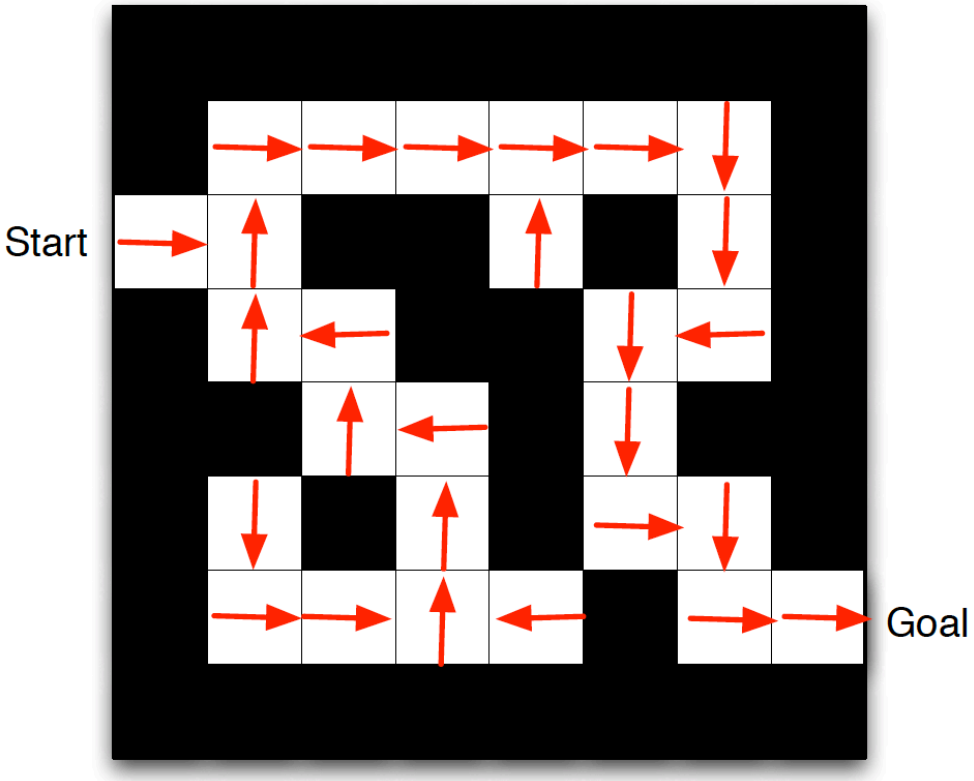


States: Agent's location
Actions: N, E, S, W
Delayed Reward:

- 100 if arrives the goal
- -100 if arrives the dead end
- -1 per time-step

http://www0.cs.ucl.ac.uk/staff/d.silver/web/Teaching_files/intro_RL.pdf

Maze Example



Arrows represent policy $\pi(s)$ for each state s

http://www0.cs.ucl.ac.uk/staff/d.silver/web/Teaching_files/intro_RL.pdf

When Reinforcement Learning Meets NLP

- Introduce two works:
  - Relation Classification(Extraction)
  - Text Classification

清華大學
Tsinghua University

Reinforcement Learning for Relation Classification from Noisy Data

Jun Feng, Minlie Huang, Li Zhao, Yang Yang,

Xiaoyan Zhu

AAAI-18

Background

- Relation Classification(Extraction)

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

Relation: *BornIn*

- Distant Supervision

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

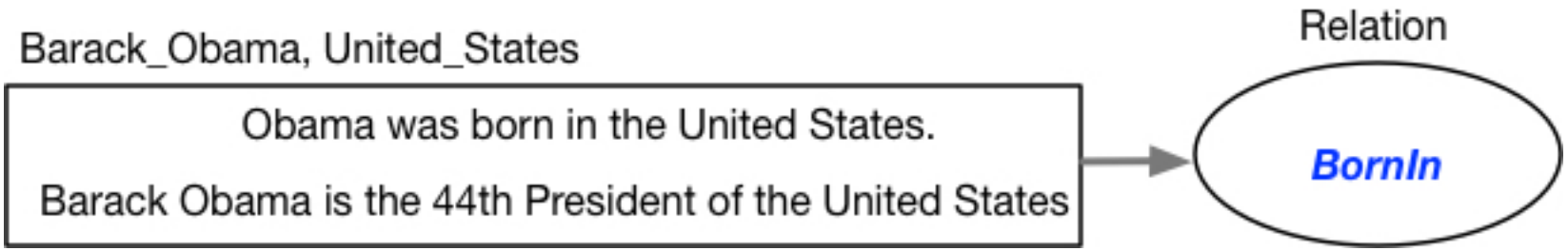
Triple in knowledge base:<Barack_Obama, *BornIn*, United_States>

Relation: *BornIn*

Suffers from the noisy labeling problem

Motivation

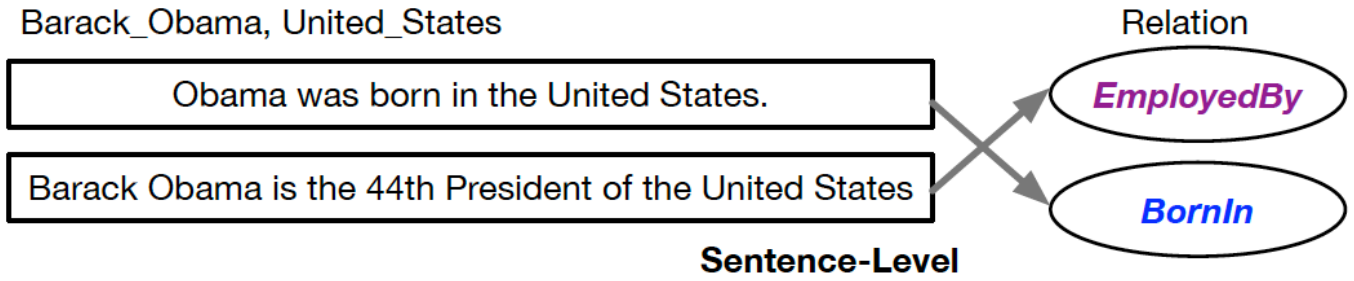
- Previous studies adopt multi-instance learning to consider the noises of instances



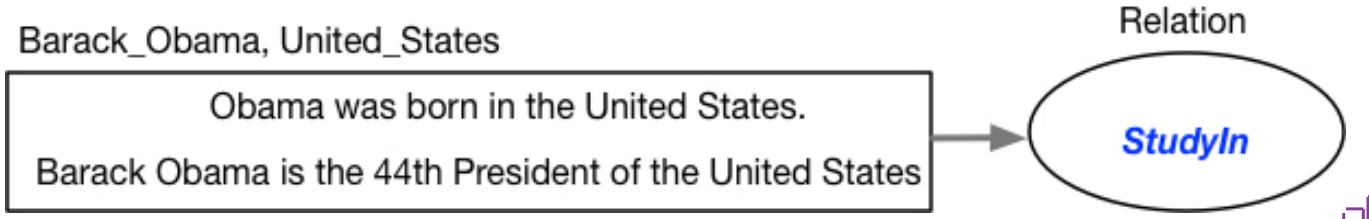
Bag-Level

Motivation

- Two limitation of previous works:
  - Unable to handle the sentence-level prediction

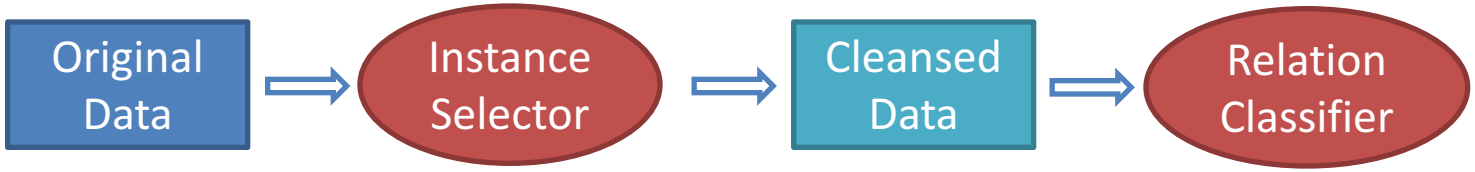


  - Sensitive to the bags with all noisy sentences which do not describe a relation at all



Model

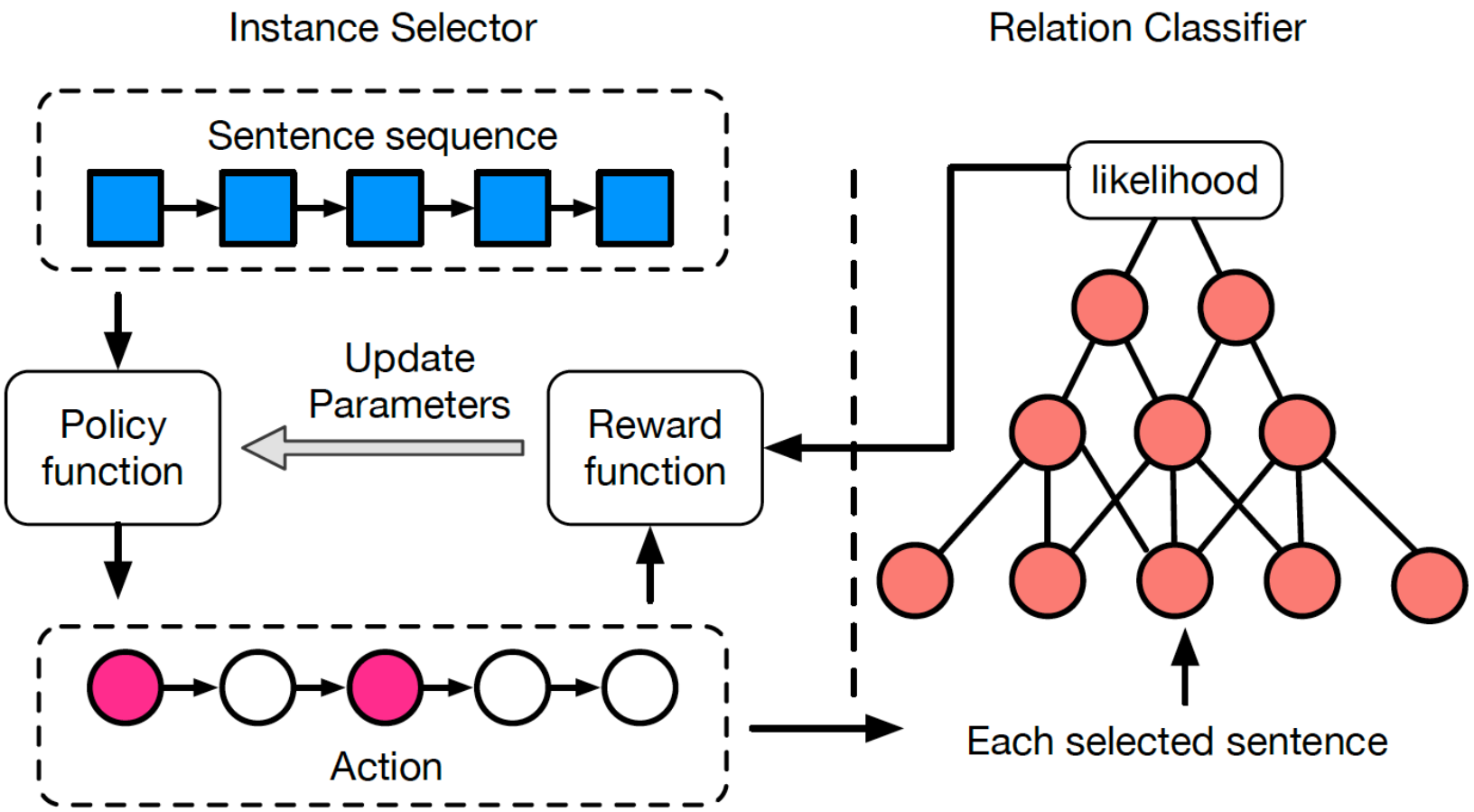
- The model consists of an **instance selector and a relation classifier**



- Challenges:
  - instance selector has no explicit knowledge about which sentences are labeled incorrectly
    - Weak supervision -> delayed reward
    - Trail-and-error search

    → Reinforcement Learning
  - how to train the two modules jointly

Model


Instance Selector
Relation Classifier
Sentence sequence
likelihood
Update
Parameters
Policy
function
Reward
function
Action
Each selected sentence

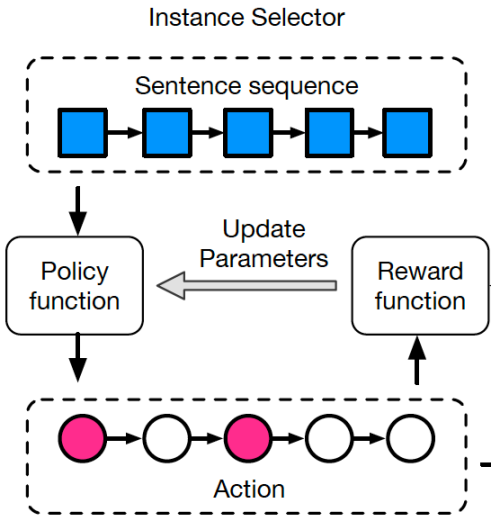
Instance Selector

- We cast instance selection as a reinforcement learning problem
  - State: $\boldsymbol{F}(s_i)$ the current sentence, the already selected sentences, and the entity pair
  - Action: select the current sentence or not

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

  - Reward

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



Instance Selector

- Optimization:
  - we aim to maximize the expected total reward

$$J(\Theta) = V_\Theta(s_1|B)$$

$$= E_{s_1,a_1,s_2,\ldots,s_i,a_i,s_{i+1}\ldots}[\sum_{i=0}^{|B|+1} r(s_i|B)]$$

  - According to the policy gradient theorem and the REINFORCE algorithm, the gradient is:

$$\Theta \leftarrow \Theta + \alpha \sum_{i=1}^{|B|} v_i \nabla_\Theta \log \pi_\Theta(s_i, a_i)$$

Relation Classifier

- we adopt a CNN architecture to classify relations

$$\mathbf{L} = \mathrm{CNN}(\mathbf{x})$$

$$p(r|x;\mathbf{\Phi}) = softmax(\mathbf{W}_r * tanh(\boldsymbol{L}) + \mathbf{b}_r)$$

- Optimization: we define the objective function of the relation classifier using cross-entropy as follows:

$$\mathcal{J}(\Phi) = -\frac{1}{|\hat{X}|}\sum_{i=1}^{|\hat{X}|}\log p(r_i|x_i;\Phi)$$

Training

- Overall Training Procedure
  1. Pre-train the CNN model of the relation classifier
  2. Pre-train the policy network of the instance selector with the CNN model fixed
  3. Jointly train the CNN model and the policy network

Experiment

- Dataset
  - NYT and developed by (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.

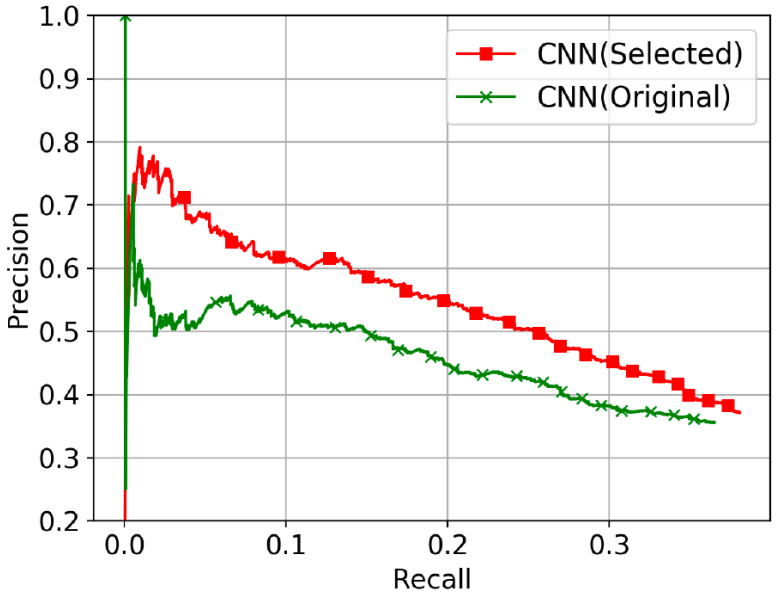
Experiment

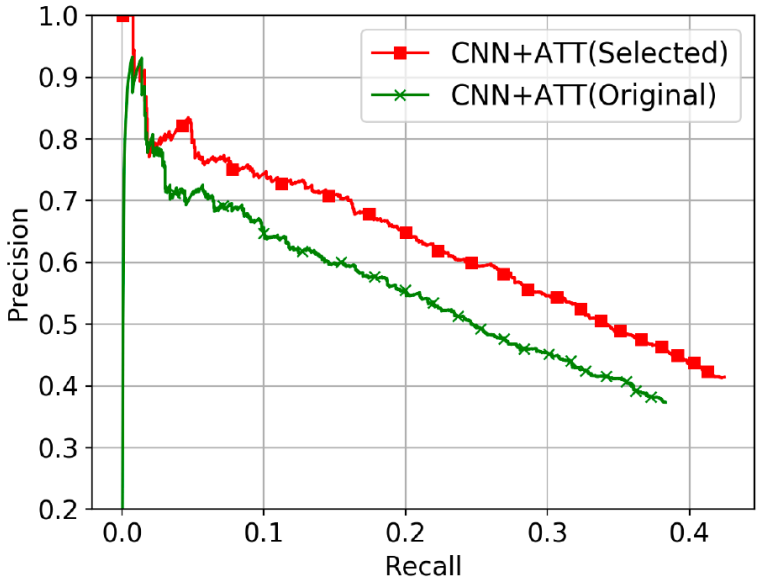
- Sentence-Level Relation Classification

Method	Macro F_1	Accuracy
CNN	0.40	0.60
CNN+Max	0.06	0.34
CNN+ATT	0.29	0.56
CNN+RL(ours)	**0.42**	**0.64**

Experiment

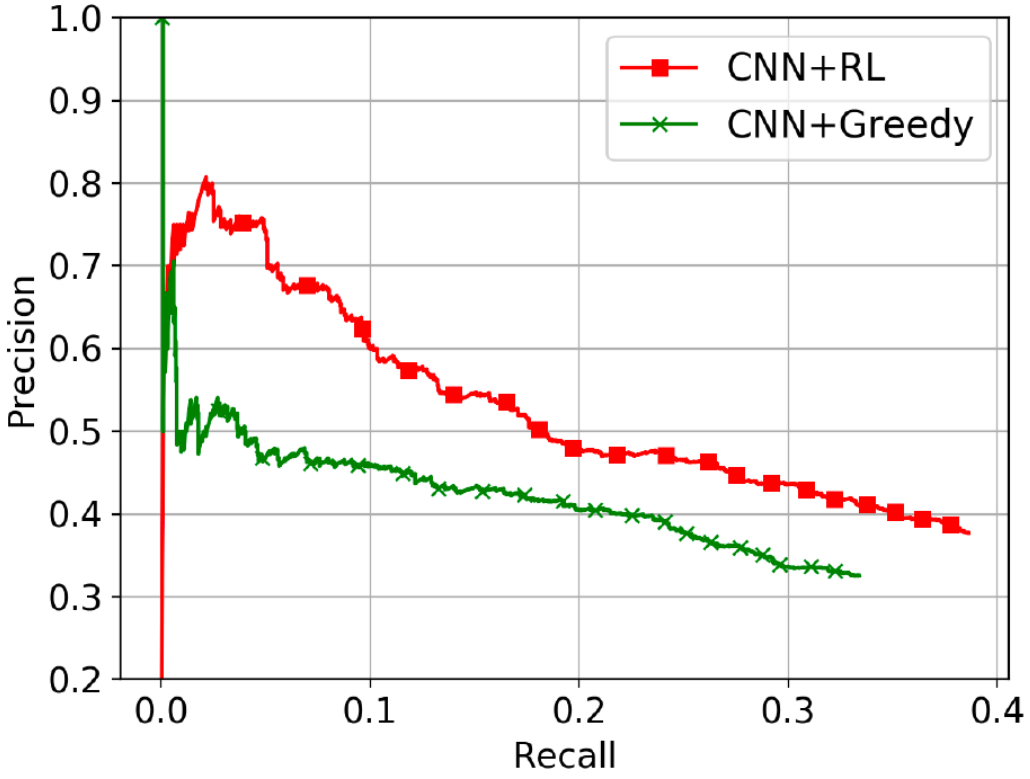
- Evaluation the performance of the instance selector





Experiment

- Evaluation the performance of the instance selector



Case Study

Bag I (**Entity Pair**: fabrice_santor, france; **Relation**:/people/person/nationality)	CNN+RL	CNN+ATT	CNN+Max
though not without some struggle, federer, the world 's top-ranked player, advanced to the fourth round with a thrilling, victory over the crafty **fabrice_santoro** of **france**, who is ranked 76th.	1	0.60	0
in his quarterfinal , nalbandian overwhelmed unseeded **fabrice_santoro** of **france**	1	0.39	1
fabrice_santoro, 33 , of **france** finally reached the quarterfinals in a major on his 54th attempt by defeating the 11th-seeded spaniard david ferrer	1	0.01	0
Bag II (**Entity Pair**: jonathan_littel, france; **Relation**:/people/person/nationality)			
jonathan_littell, a new york-born writer whose french-language novel about a murderous and degenerate officer has been the sensation of the french publishing season, on monday became the first american to win **france**'s most prestigious literary award, the prix goncourt	0	0.89	1
after a languid intercontinental auction that stretched for more than a week, the american rights to **jonathan_littell**'s novel les bienveillantes, which became a publishing sensation in **france**, have been sold to harpercollins, the publisher confirmed yesterday.	0	0.11	0

Contribution

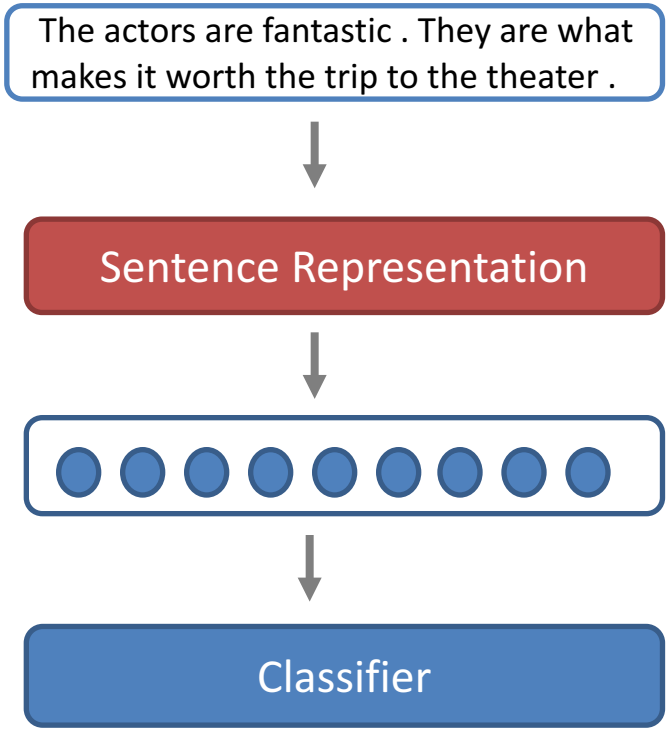
- We propose a new model to extract relations at the sentence level on the noisy data.
- We formulate instance selection as a reinforcement learning problem, just with a weak supervision signal from the relation classifier.
- Our solution for instance selection can be generalized to other tasks that employ noisy data or distant supervision.

Learning Structured Representation for Text Classification via Reinforcement Learning

Tianyang Zhang, Minlie Huang, Li Zhao

AAAI-18

Background



Sentence Representation

- Bag-of-words
- Convolutional neural network
- Recurrent neural network
- Attention-based methods
- Tree-structured LSTM

No structure / Using given structure

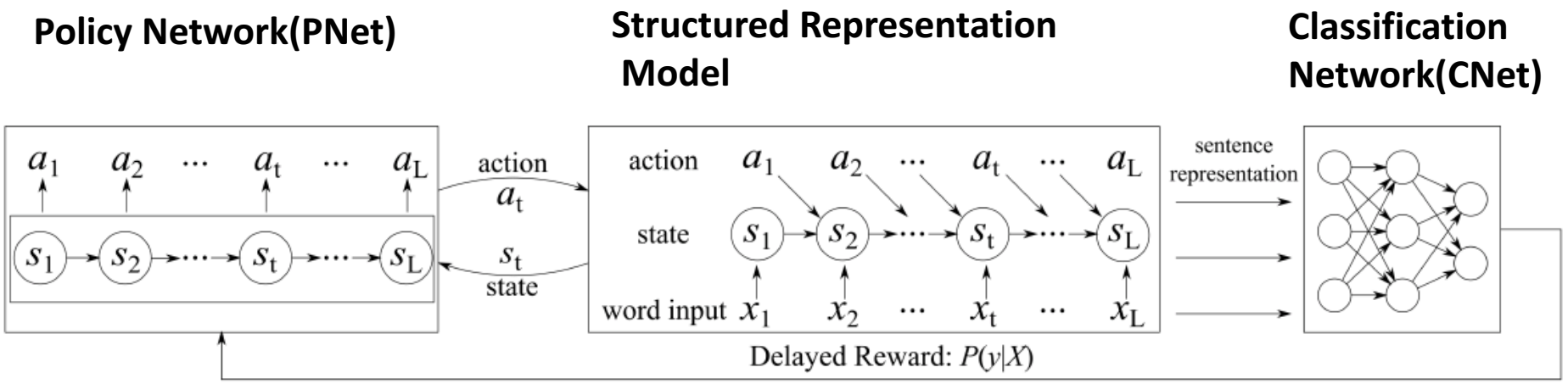
What We Want To Do …

- Identify task-relevant structures
- Build structured sentence representations over the structures
- Challenges
  - Do not have explicit structure annotations

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

→ Reinforcement Learning

Model



- Policy Network:
  - samples an action at each state
  - Two models: **Information Distilled LSTM**, **Hierarchically Structured LSTM**
- Structured Representation Model: offers state representation
- Classification Network: performs text classification, provide reward

Policy Network (PNet)

- State s_t
  - Encodes the current input and previous contexts
  - Provided by different representation models
- Action a_t
  - {Retain, Delete} in **Information Distilled LSTM**
  - {Inside, End} in **Hierarchically Structured LSTM**
  - $\pi(a_t|s_t;\Theta) = \sigma(W * s_t + b)$
- Reward r_t
  - Be calculated from the classification network
  - A factor considering the tendency of structure selection

Policy Network (PNet)

- Maximize the expected reward:

$$
\begin{aligned}
J(\Theta) &= \mathbb{E}_{(\mathbf{s_t}, a_t) \sim P_\Theta(\mathbf{s_t}, a_t)} r(\mathbf{s_1} a_1 \cdots \mathbf{s_L} a_L) \\
&= \sum_{\mathbf{s_1} a_1 \cdots \mathbf{s_L} a_L} P_\Theta(\mathbf{s_1} a_1 \cdots \mathbf{s_L} a_L) R_L \\
&= \sum_{\mathbf{s_1} a_1 \cdots \mathbf{s_L} a_L} p(\mathbf{s_1}) \prod_t \pi_\Theta(a_t | \mathbf{s_t}) p(\mathbf{s_{t+1}} | \mathbf{s_t}, a_t) R_L \\
&= \sum_{\mathbf{s_1} a_1 \cdots \mathbf{s_L} a_L} \prod_t \pi_\Theta(a_t | \mathbf{s_t}) R_L .
\end{aligned}
$$

- Update the policy network with gradient:

$$
\nabla_\Theta J(\Theta) = \sum_{t=1}^{L} R_L \nabla_\Theta \log \pi_\Theta(a_t | \mathbf{s_t})
$$

Classification Network (CNet)

- In order to train CNet, we adopt cross entropy as loss function:

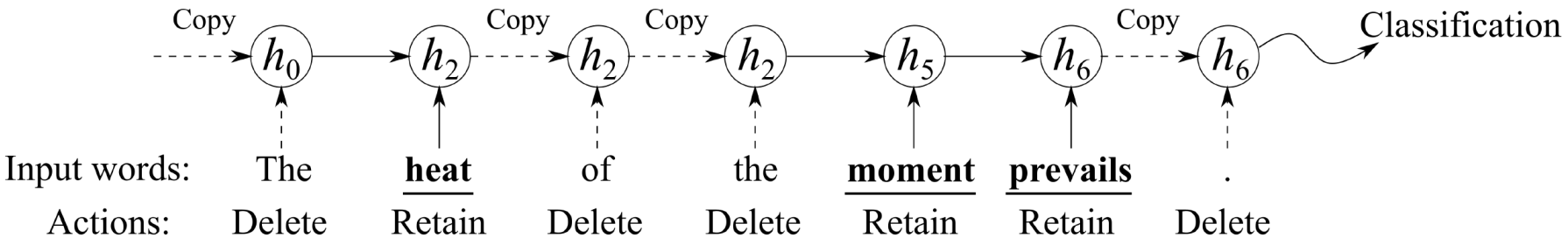
$$P(y|X) = softmax(\mathbf{W_s h_L} + \mathbf{b_s}),$$

$$\mathcal{L} = \sum_{X \in \mathcal{D}} -\sum_{y=1}^{K} \hat{p}(y, X) \log P(y|X)$$

Information Distilled LSTM (ID-LSTM)

- Distill the most important words and remove irrelevant words
- Sentence Representation: the last hidden state of ID-LSTM

$$P(y|X) = softmax(\mathbf{W_s h_L} + \mathbf{b_s})$$



Information Distilled LSTM (ID-LSTM)

- Action: {Retain, Delete}
- States:

$$\mathbf{s_t} = \mathbf{c_{t-1}} \oplus \mathbf{h_{t-1}} \oplus \mathbf{x_t},$$

$$\mathbf{c_t}, \mathbf{h_t} = \begin{cases} \mathbf{c_{t-1}}, \mathbf{h_{t-1}}, & a_t = Delete \\ \Phi(\mathbf{c_{t-1}}, \mathbf{h_{t-1}}, \mathbf{x_t}), & a_t = Retain \end{cases}$$

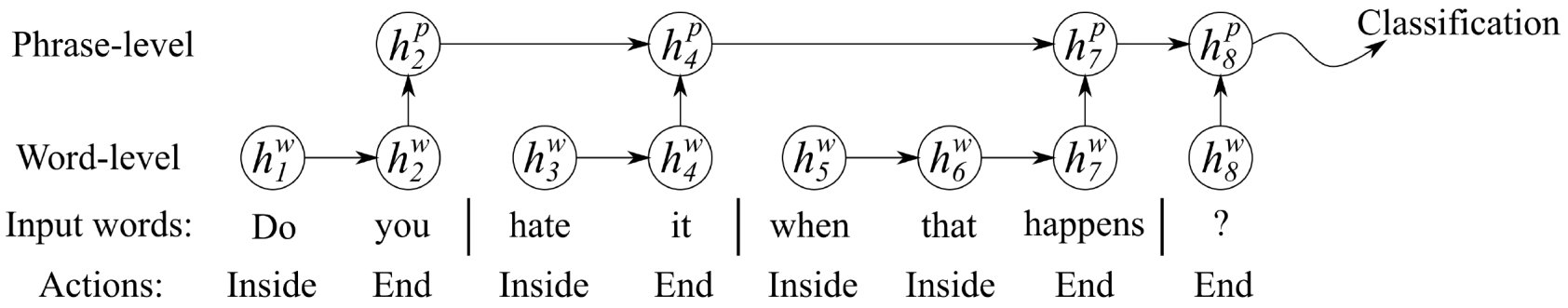
- Rewards:

$$R_L = \log P(c_g|X) + \gamma L'/L$$

the proportion of the number of deleted words to the sentence length

Hierarchically Structured LSTM(HS-LSTM)

- Build 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



Hierarchically Structured LSTM(HS-LSTM)

- Action: {Inside, End}

a_{t-1}	a_t	Structure Selection
Inside	Inside	A phrase continues at x_t.
Inside	End	A old phrase ends at x_t.
End	Inside	A new phrase begins at x_t.
End	End	x_t is a single-word phrase.

- States: $\mathbf{s_t} = \mathbf{c^P_{t-1}} \oplus \mathbf{h^P_{t-1}} \oplus \mathbf{c^w_t} \oplus \mathbf{h^w_t}$

Word-level LSTM $$\mathbf{c^w_t}, \mathbf{h^w_t} = \begin{cases} \Phi^w(\mathbf{0}, \mathbf{0}, \mathbf{x_t}), & a_{t-1} = End \\ \Phi^w(\mathbf{c^w_{t-1}}, \mathbf{h^w_{t-1}}, \mathbf{x_t}), & a_{t-1} = Inside \end{cases}$$

Phrase-level LSTM $$\mathbf{c^P_t}, \mathbf{h^P_t} = \begin{cases} \Phi^p(\mathbf{c^P_{t-1}}, \mathbf{h^P_{t-1}}, \mathbf{h^w_t}), & a_t = End \\ \mathbf{c^P_{t-1}}, \mathbf{h^P_{t-1}}, & a_t = Inside \end{cases}$$

- Rewards: $$R_L = \log P(c_g|X) - \gamma(L'/L + 0.1L/L')$$

a unimodal function of the number of phrases (a good phrase structure should contain neither too many nor too few phrases)

Experiment

- Dataset
  - MR: This dataset contains positive/negative reviews (Pang and Lee 2005)
  - SST: Stanford Sentiment Treebank, a public sentiment analysis dataset with five classes (Socher et al. 2013)
  - Subj: Subjectivity dataset. The task is to classify a sentence as subjective or objective (Pang and Lee 2004)
  - AG: AG's news corpus3, 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

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 | .

Analyze ID-LSTM

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.

Analyze ID-LSTM

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%

Table 5: The most/least deleted words in the test set of SST.

Analyze HS-LSTM

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).

Analyze HS-LSTM

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.

Conclusion

- A reinforcement learning method which learns sentence representation by discovering task-relevant structure.
- Two representation models: ID-LSTM and HS-LSTM
- state-of-the-art performance & interesting task-relevant structures

Thank You!