JUMPER: Learning When to Make Classification Decisions in Reading

Xianggen Liu, Lili Mou, Haotian Cui, Zhengdong Lu, Sen Song

Tsinghua University & DeeplyCurious.ai & adeptmind.ai

Background

Text Classification	92.5 ACC on AG
POS Tagging	97.1 ACC on WJS
Machine Translation	40.5 BLEU on WMT EN2FR
QA	83.9 ACC on SQuAD 1.0
Document Summarization	
Chunking	
Dialogue Modeling	. . .
NER	

Background

Text Classification

92.5 ACC on AG

POS Tagging

97.1 ACC on WJS

Machine Translation

40.5 BLEU on WMT EN2FR

QA

83.9 ACC on SQuAD 1.0

Document Summarization

Chunking

. . .

Dialogue Modeling

NER

How does the model make such a decision?

Motivation: reading text like human

During reading,

people obtain information from text orderly,

look for clues, ⟶ Rationales

perform reasoning, ⟶ Symbolic and distributed reasoning

and continue obtaining new information.

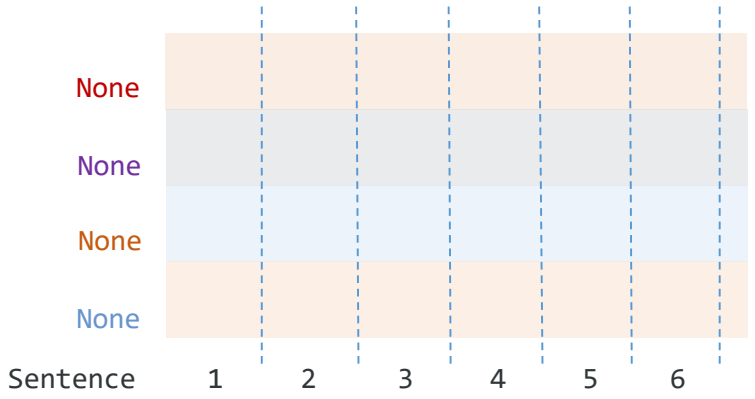
Understanding ⟶ Sequential decision process

Decision Process

Tsinghua University

Paragraph….

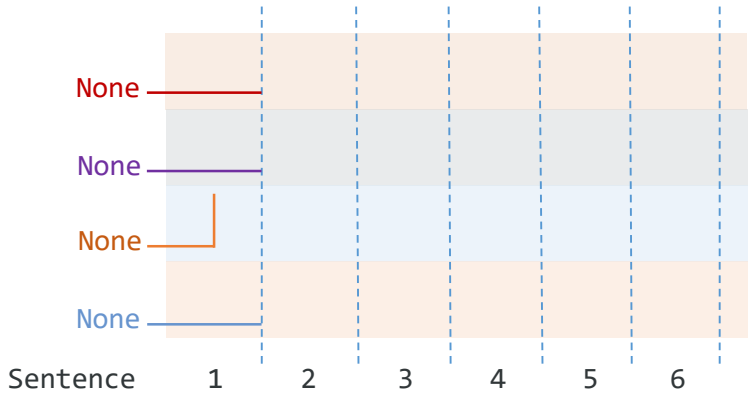
1. What is the authorized level of injury? (10)
2. Is it an occupational injury? (Yes)
3. Did the incident happen during working hours? (Yes)
4. Did the labor contract end? (unknown)



Decision Process

When I was building the scaffolding,

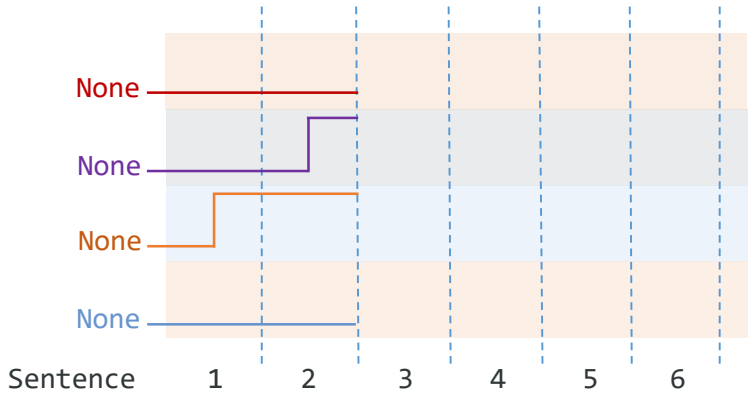
1. What is the authorized level of injury? (10)
2. Is it an occupational injury? (Yes)
3. Did the incident happen during working hours? (Yes)
4. Did the labor contract end? (unknown)



Decision Process

When I was building the scaffolding, I was hit by a steel tube,

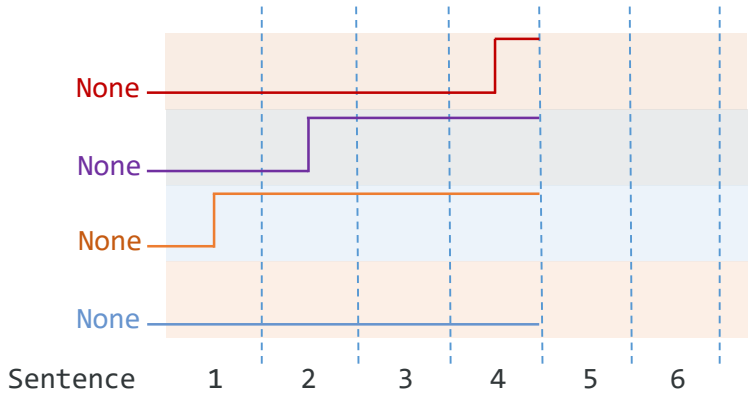
1. What is the authorized level of injury? (10)
2. Is it an occupational injury? (Yes)
3. Did the incident happen during working hours? (Yes)
4. Did the labor contract end? (unknown)



Decision Process

When I was building the scaffolding, I was hit by a steel tube, but my boss didn't buy insurance for me. I was identified as ten-level disabled,

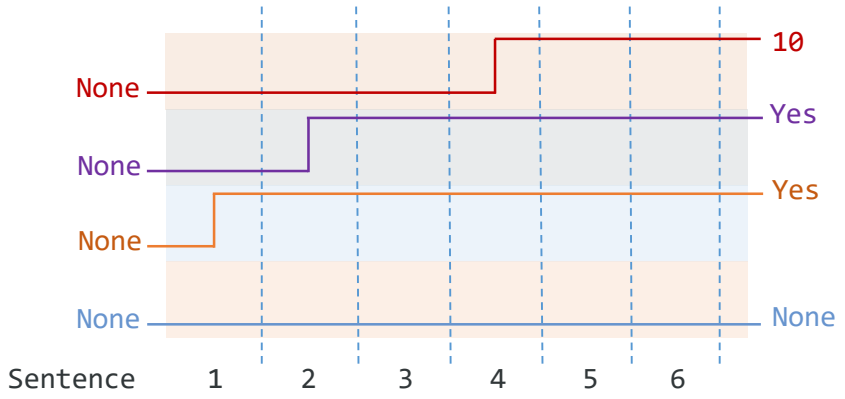
1. What is the authorized level of injury? (10)
2. Is it an occupational injury? (Yes)
3. Did the incident happen during working hours? (Yes)
4. Did the labor contract end? (unknown)



Decision Process

When I was building the scaffolding, I was hit by a steel tube, but my boss didn't buy insurance for me. I was identified as ten-level disabled, and I need to rest for 4 months. How much money can I be compensated?

1. What is the authorized level of injury? (10)
2. Is it an occupational injury? (Yes)
3. Did the incident happen during working hours? (Yes)
4. Did the labor contract end? (unknown)



Modeling choice of text classification

Generally, the goal is to transform some source sequence

$$X = x_1, x_2, x_3, \cdots, x_N$$

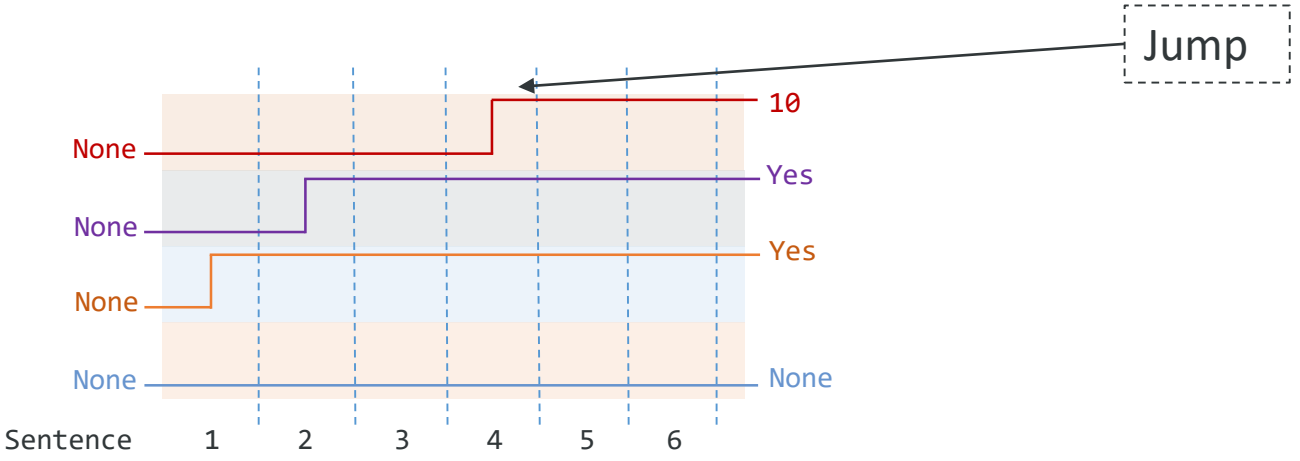
into an integer.

Here, our goal is to transform the source sequence X into a sequence of decisions S:

$$S = s_1, s_2, s_3, \cdots, s_N$$

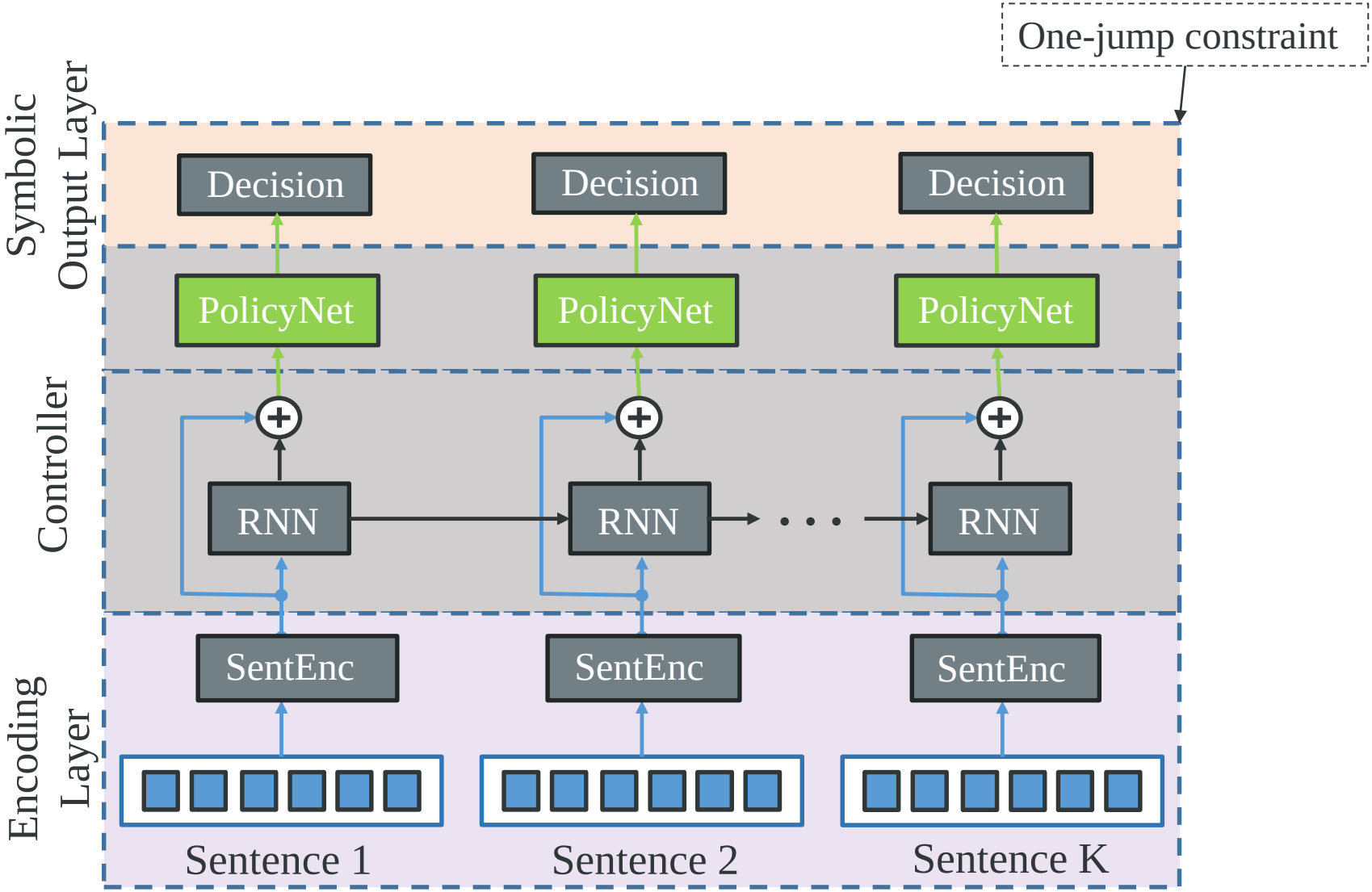
Which reflects the dynamic decision-making process in reading.

One-Jump Constraint



$$\sum |s_t - s_{t-1}| = 1 \quad \forall t \in 1, 2, \ldots, T$$

Jumper


One-jump constraint
Symbolic Output Layer
Decision
Decision
Decision
PolicyNet
PolicyNet
PolicyNet
Controller
RNN
RNN
RNN
SentEnc
SentEnc
SentEnc
Encoding Layer
Sentence 1
Sentence 2
Sentence K

Jumper: learning

Reward:

$$R_{\text{final}}^{(j,i)} = \mathbf{1}_{\{\boldsymbol{s}_{T_j}^{(j,i)} = \boldsymbol{t}^{(j,i)}\}}$$

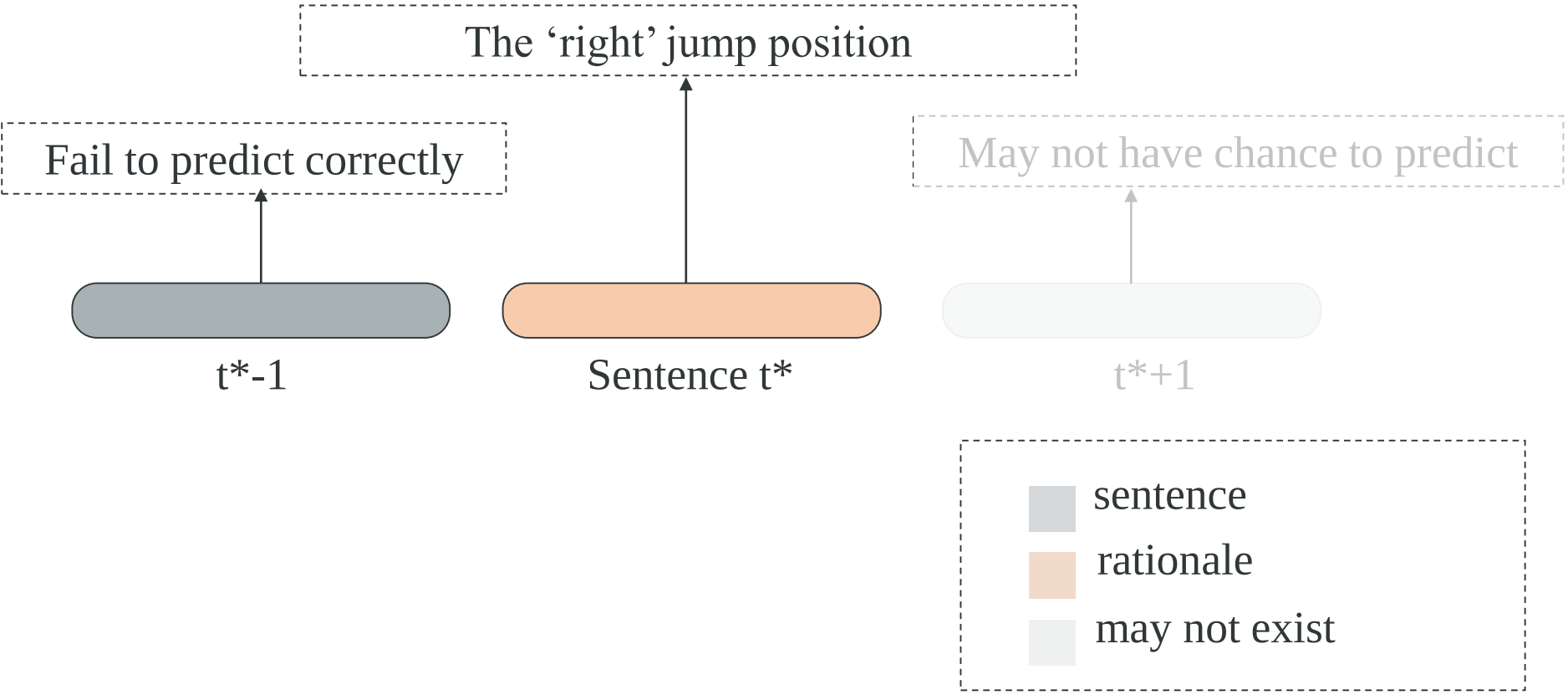
for data point j and slot i, where $\boldsymbol{s}_{T_j}^{(j,i)}$ is the symbolic state at the end of the paragraph and $\boldsymbol{t}^{(j,i)}$ is the groundtruth.

REINFORCE Algorithm:

$$\nabla_\Theta \mathcal{J}(\Theta) = \mathbb{E}_{\pi_\Theta}\left[\sum_{t=1}^{T} \nabla_\Theta R_{t:T} \log \pi_\Theta(\boldsymbol{a}_t|\boldsymbol{c}_t)\right]$$

$$\approx \sum_{j=1}^{N}\sum_{i=1}^{I}\sum_{t=1}^{T_{\text{jump}}^{(j,i)}} \frac{1}{NT_j} R_{t:T_{\text{jump}}^{(j,i)}}^{(j,i)} \nabla_\Theta \log \pi_\Theta\left(\boldsymbol{a}_t^{(j,i)}|\boldsymbol{c}_t^{(j)}\right)$$

Jumper: learning when to jump



Experiments

Model	MR	AG	OI-Level	OI-InjIdn
CNN† [Kim, 2014]	81.00	–	–	–
fasttext† [Joulin *et al.*, 2017]	–	92.50	–	–
Bi-GRU	77.80	92.44	94.75	73.25
CNN	80.80	92.58	96.25	74.25
Self-Attentive	**82.10**	91.40	97.00	73.25
Hierarchical CNN-GRU	80.23	92.49	95.75	74.75
JUMPER	80.67	**92.62**	**97.25**	**75.50**

Table 3: Test accuracy (%) on MR, AG, and OI datasets. †Results quoted from previous papers.

Dataset	MR	AG	OI-Level	OI-InjIdn
Avg # of sub-sentences	2.17	3.46	4.88	4.88
Avg jumping position	1.46	2.04	3.23	2.87
Reduced %	32.7%	41.0%	33.8%	41.2%

Table 4

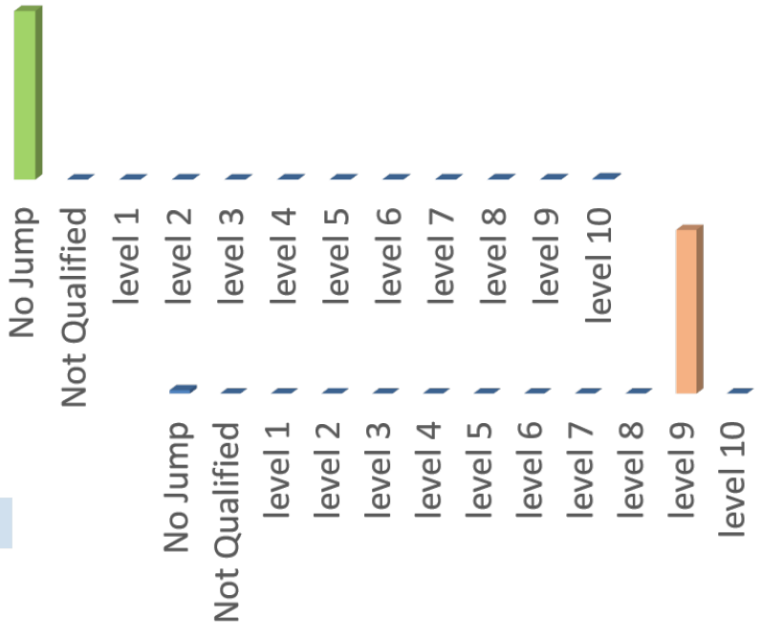
Experiments

Model	CA	JA	OA
CNN	96.25	94.81	91.25
Self-Attentive	97.00	98.45	95.50
Hierarchical CNN-RNN	96.00	98.18	94.25
JUMPER	**97.25**	**100**	**97.25**

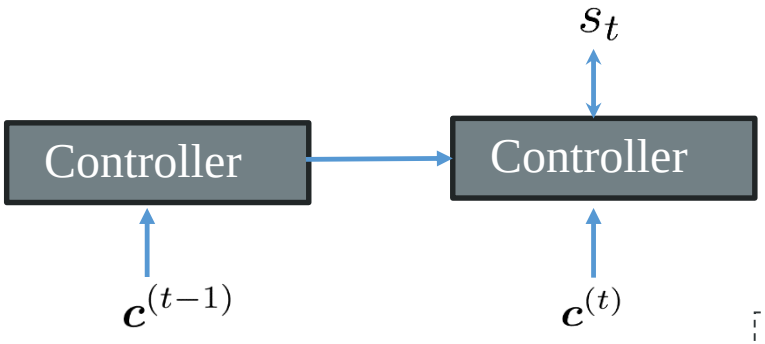
Table 5: Performance of finding the key rationale in the OI-Level dataset, where information is often local. **CA**: Classification accuracy. **JA**: Jumping accuracy. **OA**: Overall accuracy.

Translation:

Had a business trip to Suzhou on 21 Oct last year,
Riding my colleague's bicycle and crashed to fence,
Comminuted fracture in leg,
I started work on 21 Sep,
and had the incident in Oct,
I don't want to work any more.
How much money should the company compensate?
The authorized injury level is 9 !
Which department shall I appeal to for compensation?



Backtracking word-level clues



Real increase of $\boldsymbol{c}^{(t-1)}$

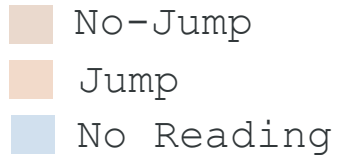
Expectation of the increase of $\boldsymbol{c}^{(t-1)}$

$$\mathscr{D} = \text{top}_D\left(\frac{\partial \log(p_t(s_t))}{\partial \boldsymbol{c}^{(t-1)}} \odot (\boldsymbol{c}^{(t)} - \boldsymbol{c}^{(t-1)})^2\right)$$

Matching rate of the two

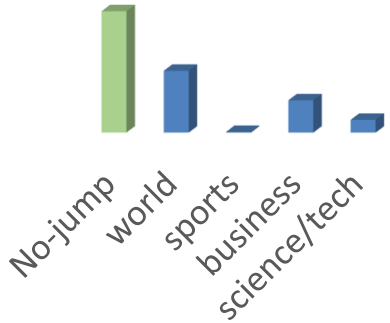
$$w_d = \text{argmax}\{c_1, \cdots, c_K\}$$

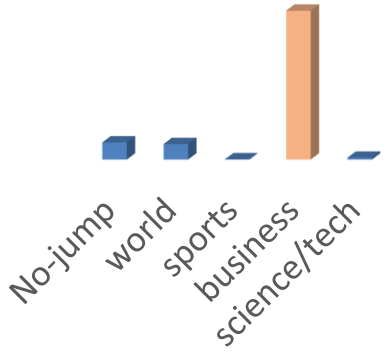
When rationales is scattered



MR:

As the characters get more depressed, the story gets more tiresome , especially as it continues to mount a conspicuous effort to be profound .





AG:

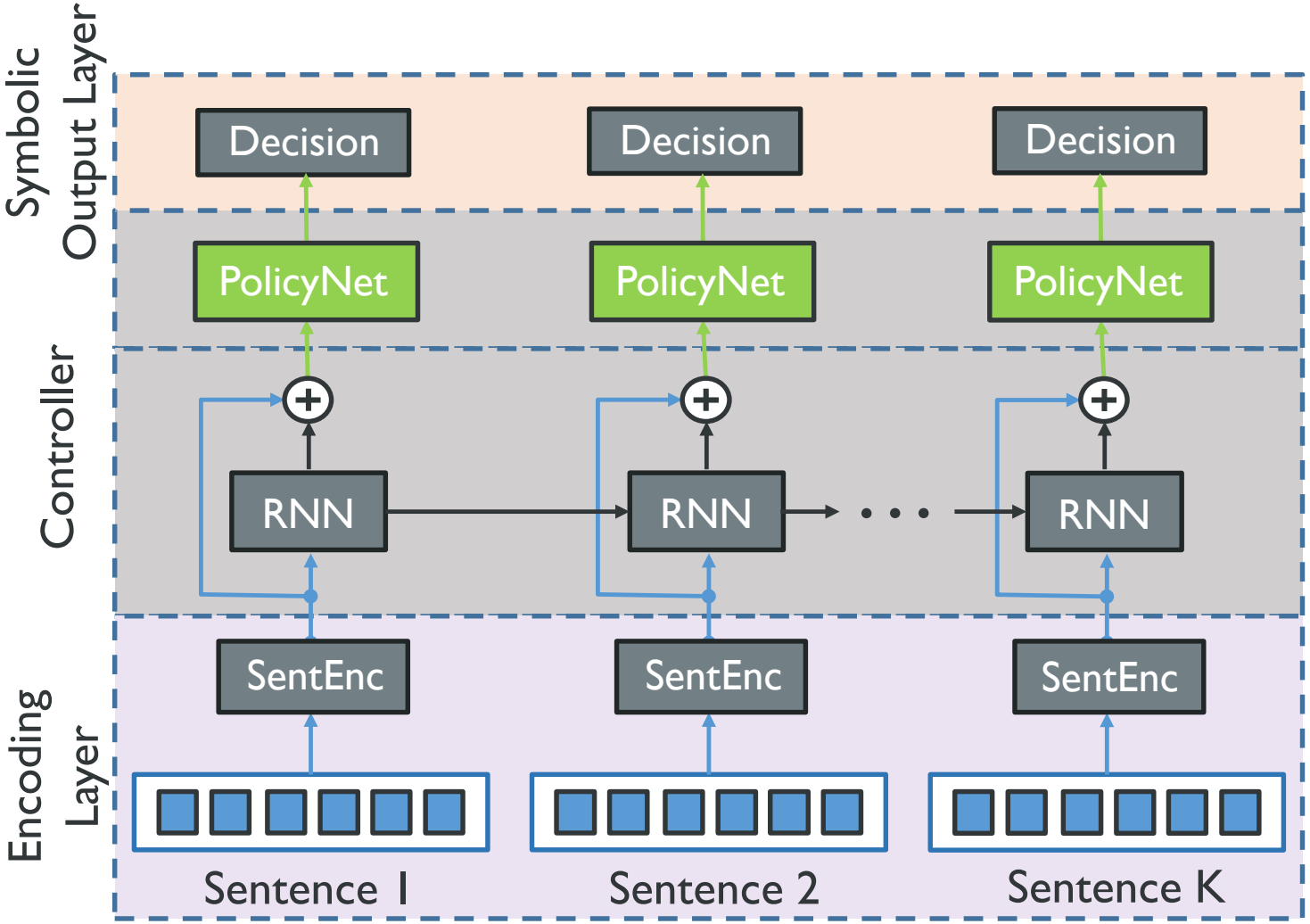
The US will offer an olive branch to Peter Mandelson , the European Union's new trade commissioner , next week by delaying any escalation of the dispute over subsidies to Airbus and Boeing , a US trade official said on Thursday .

Conclusion

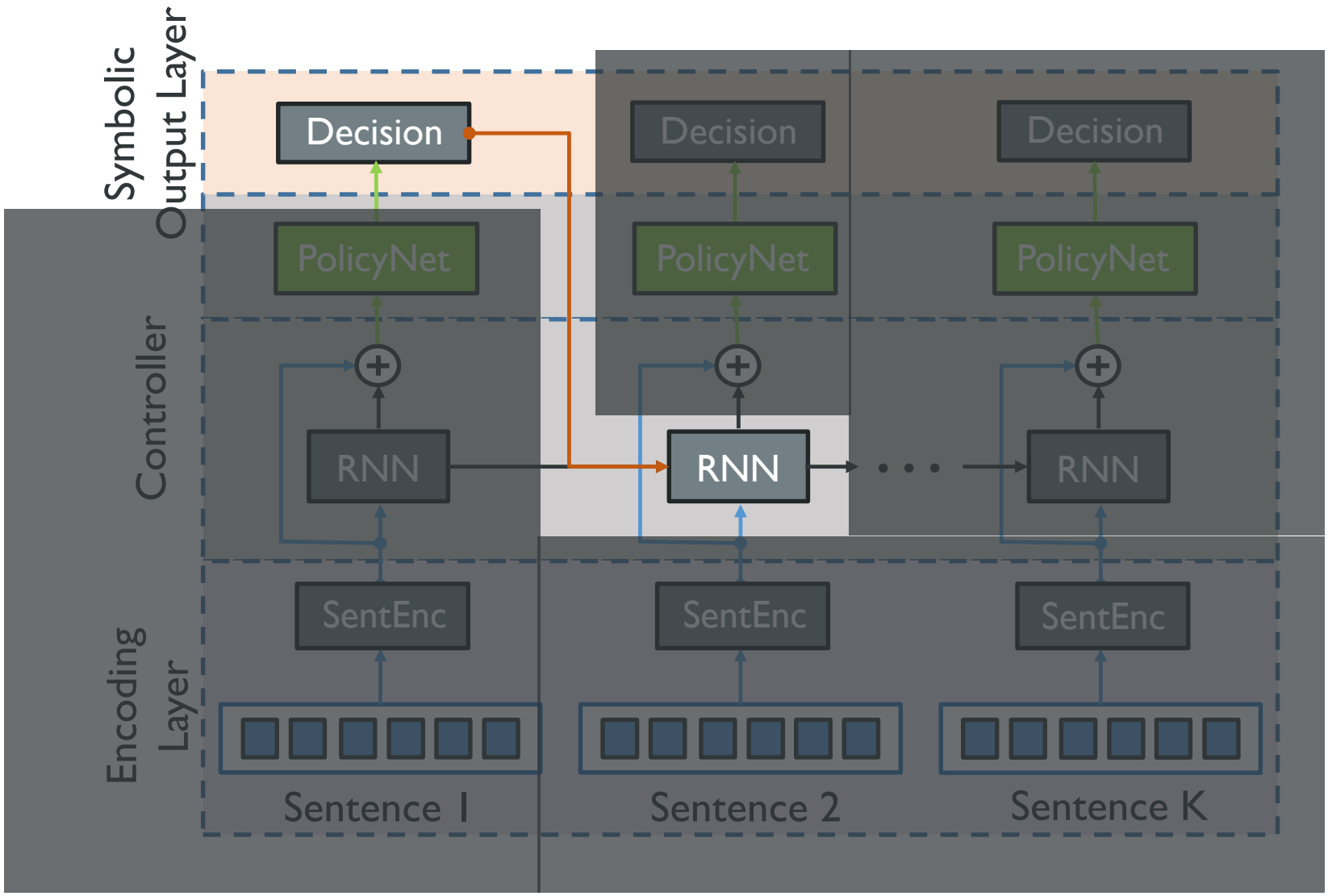
- 1) Always take decisions at the ‘appropriate’ time.
- 2) Reducing total text reading by 30–40% and often finding the key rationale of prediction.
- 3) It achieves classification accuracy better than or comparable to state-of-the-art models in several benchmark and industrial datasets.

Question?

Decision-sharing mechanism



Decision-sharing mechanism



Decision-sharing mechanism

Model	Accuracy		F_1	
	dev	test	dev	test
Bi-GRU	90.62	90.18	20.02	20.20
CNN	92.41	91.64	**30.99**	29.05
Self-Attentive	92.12	91.85	21.26	22.61
Hierarchical CNN-GRU	91.56	91.30	24.80	24.44
JUMPER	92.43	92.42	26.57	29.60
JUMPER-sharing	**92.71**	**92.65**	27.57	**30.52**

Table 6: The average accuracy and F_1 on the OI dataset using the decision-sharing mechanism.