Modeling **Coherence** for **Discourse** Neural Machine Translation

Hao Xiong, Zhongjun He, Hua Wu and Haifeng Wang
xionghao05@baidu.com
Natural Language Processing, Baidu Inc.

Contents

- Backgrounds
- Model Architecture
- Experiments
- Conclusion

Discourse Translation

Source

Sent 1: 我们加入霓虹，我们加入柔和的粉蜡色，我们使用新型材料。

Sent 2: 人们爱死这样的建筑了。

Sent 3: 我们不断地建造。

Reference

Sent 1: We add neon and we add pastels and we use new materials.

Sent 2: And you love it.

Sent 3: And we can't give you enough of it.

Discourse Neural Machine Translation

Reference

Sent 1: We add neon and we add pastels and we use new materials.

Sent 2: And you love it.

Sent 3: And we can’t give you enough of it.

Translation

Sent 1: We add the neon, we add soft, flexible crayons, and we use new materials.

Sent 2: ***[conj]$_{miss}$*** People love architecture.

Sent 3: ***[conj]$_{miss}$*** We keep building ***[coref]$_{miss}$***.

Discourse **N**eural **M**achine **T**ranslation

Reference

Sent 1: We add neon and we add pastels and we use new materials.

Sent 2: And you love it.

Sent 3: And we can’t give you enough of it.

Translation

Missing Conjunctions and Coreference

Sent 1: We add the neon, we add soft, flexible crayons, and we use new materials.

Sent 2: ***[conj]$_{miss}$*** People love architecture.

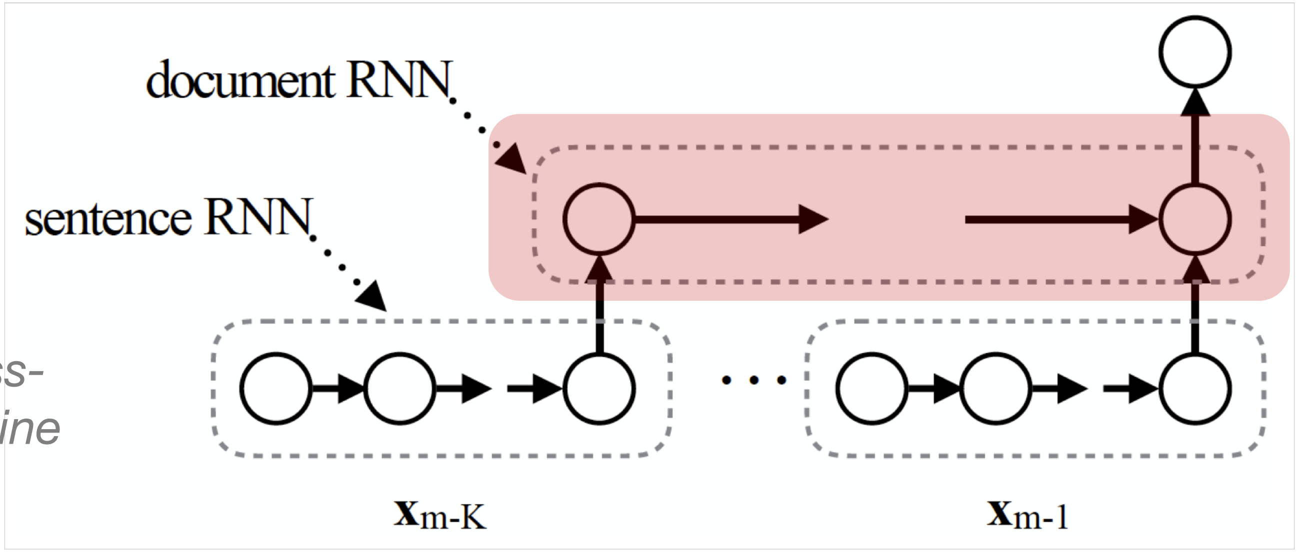
Sent 3: ***[conj]$_{miss}$*** We keep building ***[coref]$_{miss}$***.

Drawbacks of Traditional DNMT

- Translate each sentence independently
- Lack of *Discourse Coherence*
- Lack of using *Discourse Context*

Previous Solutions

Enhance the RNN with ***discourse context***



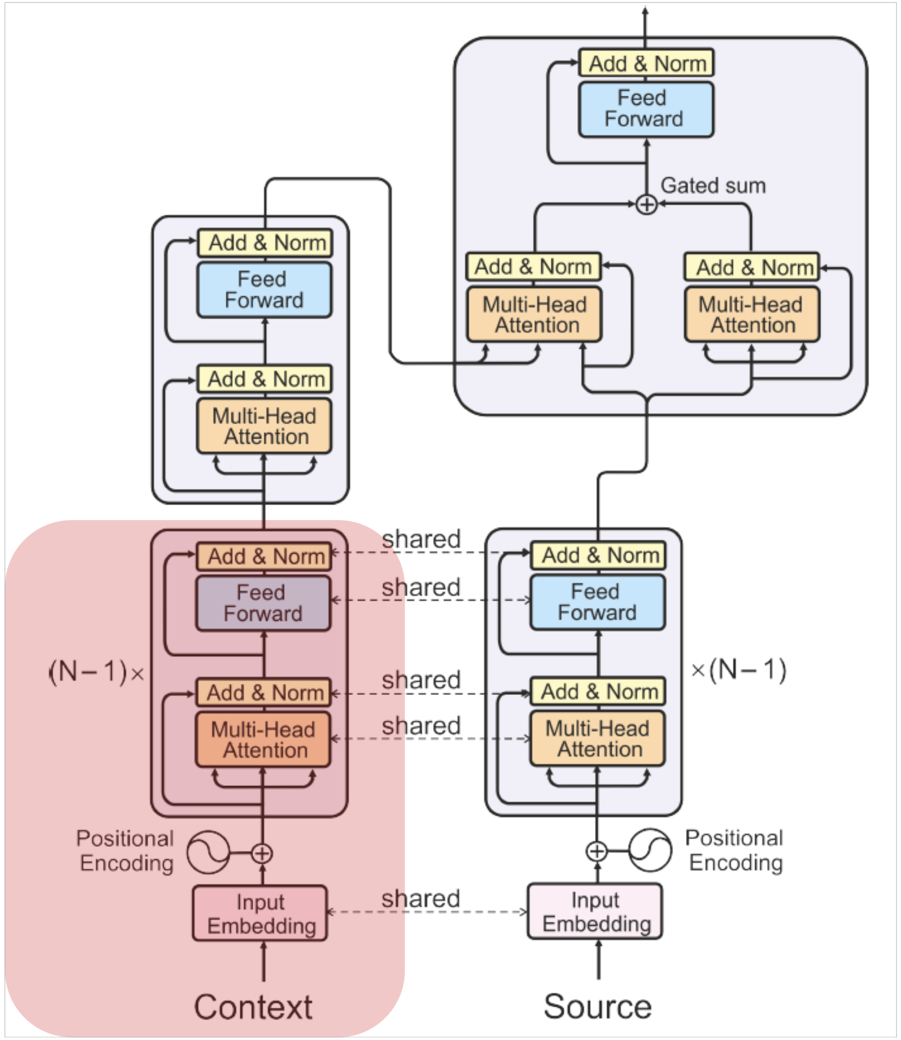
Longyue Wang et., Exploiting Cross-sentence Context for Neural Machine Translation. EMNLP 2018

Previous Solutions

Exploit discourse context

Resolve ***anaphora***

Elena Voita et., Context-aware Neural Machine Translation Learns Anaphora Resolution. ACL 2018



Previous Solutions

Focus on exploiting ***discourse context***

No work on **discourse coherence** for **DNMT**

Tiedemann, J., and Scherrer, Y. Neural Machine Translation with Extended Context. WDMT 2017
Kuang Shaohui et., Cache-based Document-level Neural Machine Translation. Arxiv 2017
Zhaopeng Tu et., Learning to Remember Translation History with a Continuous Cache. TACL 2018
Maruf, S., and Haffari, G. Document Context Neural Machine Translation with Memory Networks. ACL 2018

Our Solution

First Round: Translate each sentence independently

Sent 1: We add the neon, we add soft, flexible crayons, and we use new materials.

Sent 2: People love architecture.

Sent 3: We keep building.

Our Solution

First Round: Translate each sentence independently

***Second Round:* Deliberate** the first round translation

Sent 1: We add the neon, we add soft, flexible crayons, and we use new materials.

Sent 2: People love ***it***.

Sent 3: We keep building ***it***.

Our Solution

First Round: Translate each sentence independently

Second Round: **Deliberate** the first round translation

Reward the coherent translation

Sent 1: We add the neon, we add soft, flexible crayons, and we use new materials.

Sent 2: ***And*** people love ***it***.

Sent 3: ***And*** we keep building ***it***.

Our Solution

First Round: Translate each sentence independently

Second Round: **Deliberate** the first round translation

Reward the coherent translation

Sent 1: We add the neon, we add soft, flexible crayons, and we use new materials.

Sent 2: ***And*** people love ***it***.

Sent 3: ***And*** we keep building ***it***.

Not very well

but acceptable 😀

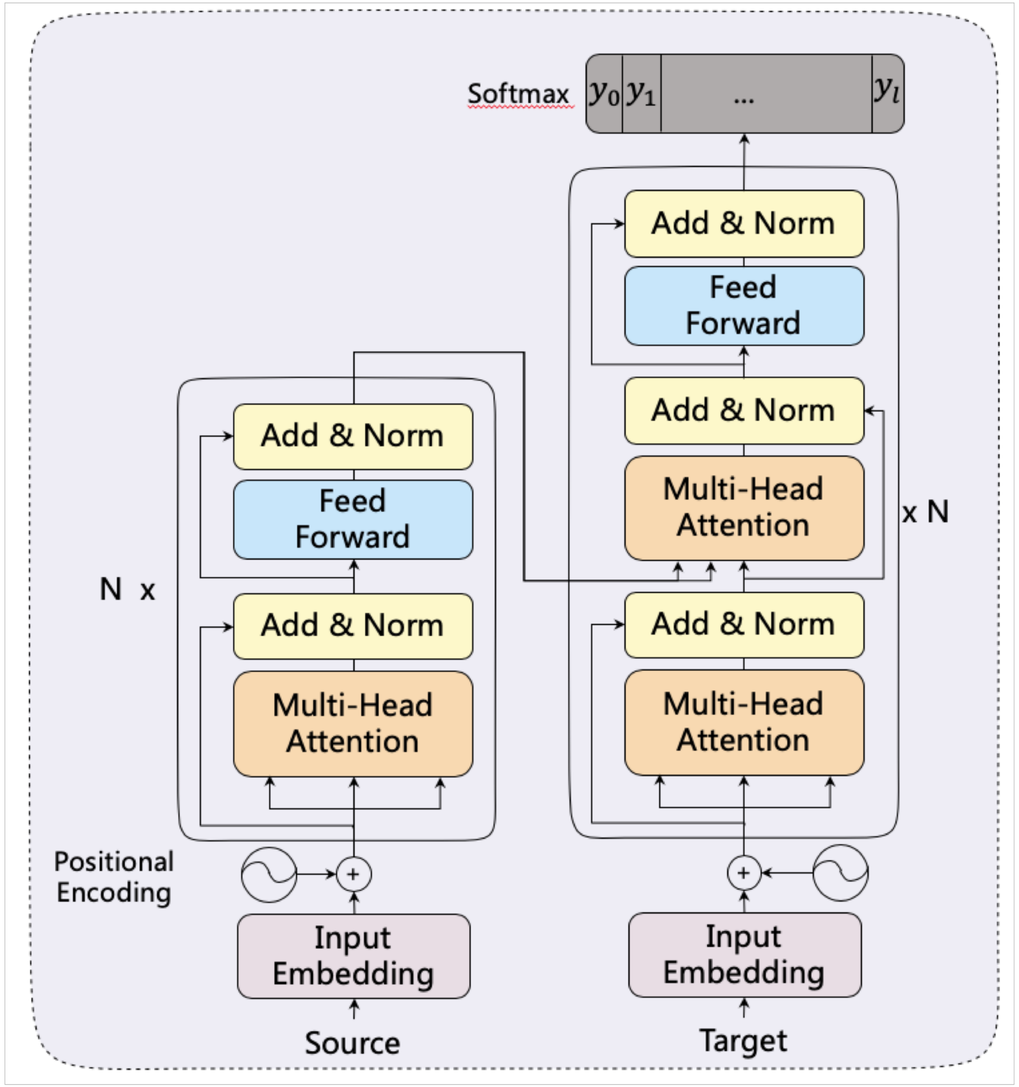
Contents

Overall Architecture

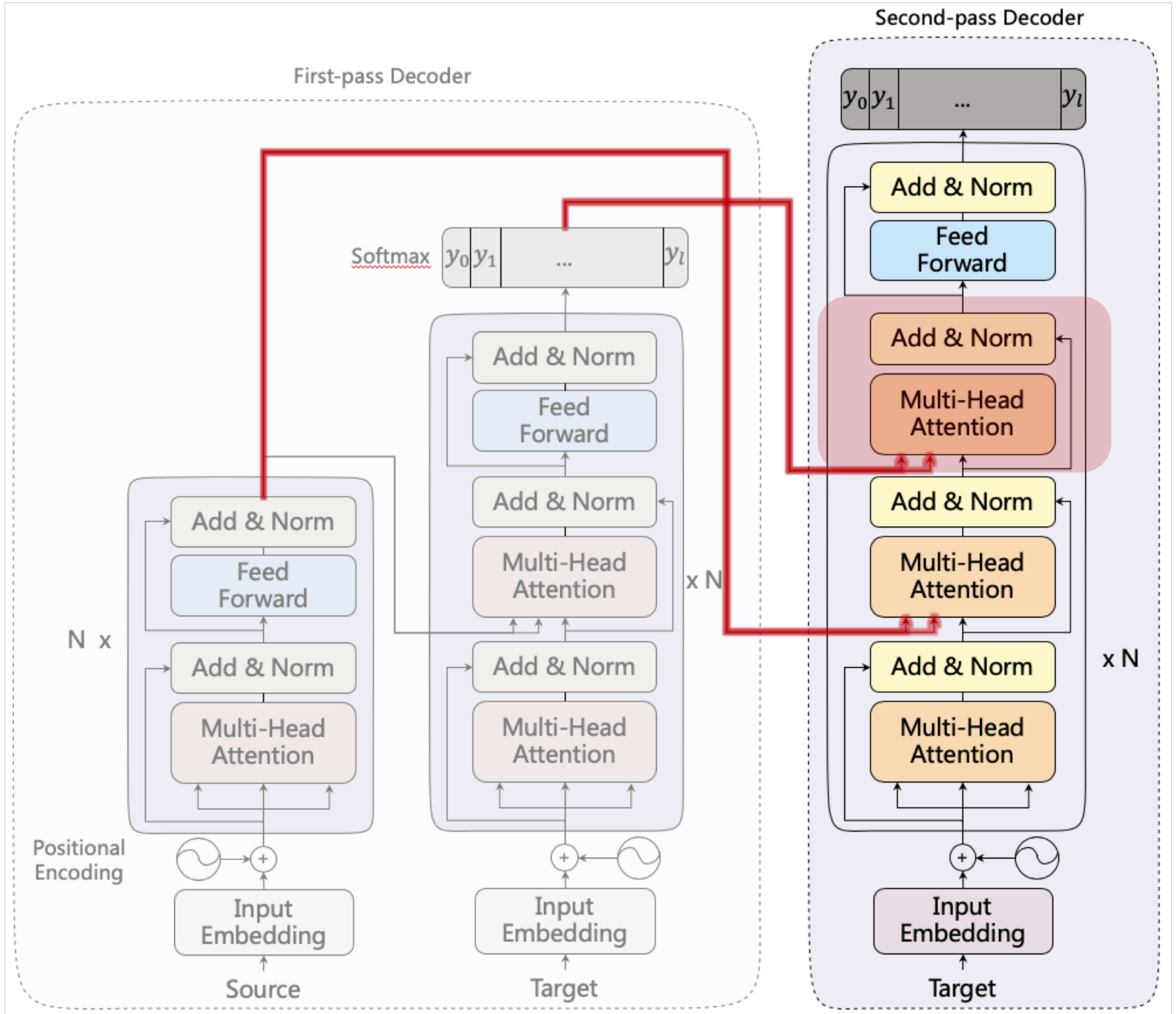
First Round:

Canonical ***Transformer***

Vaswani, Ashish et., Attention is All You Need. NIPS 2017



Overall Architecture

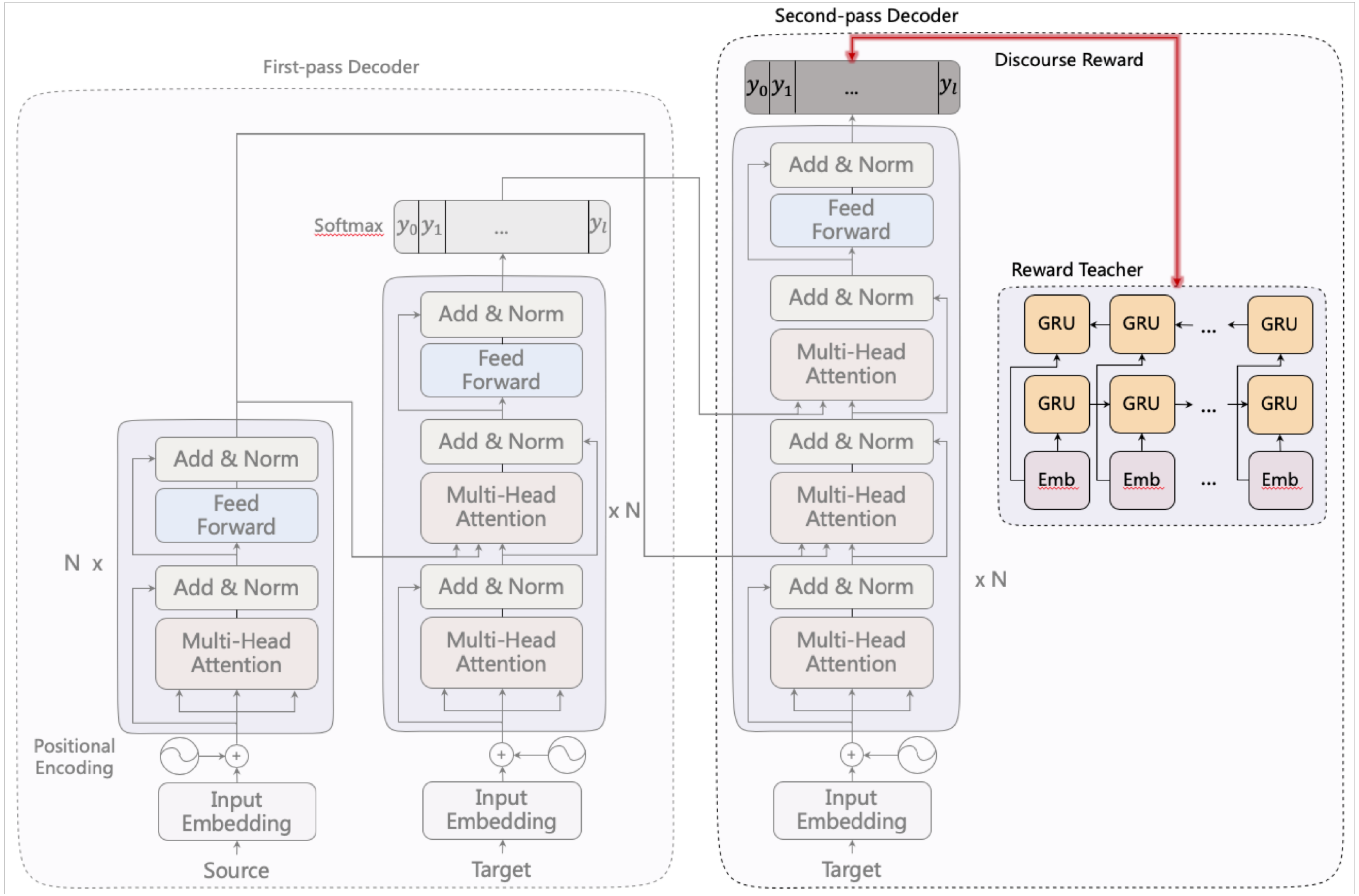


Second Round:

Deliberation Network

Yingce Xia et., Deliberation Networks: Sequence Generation Beyond One-Pass Decoding. NIPS 2017

Overall Architecture



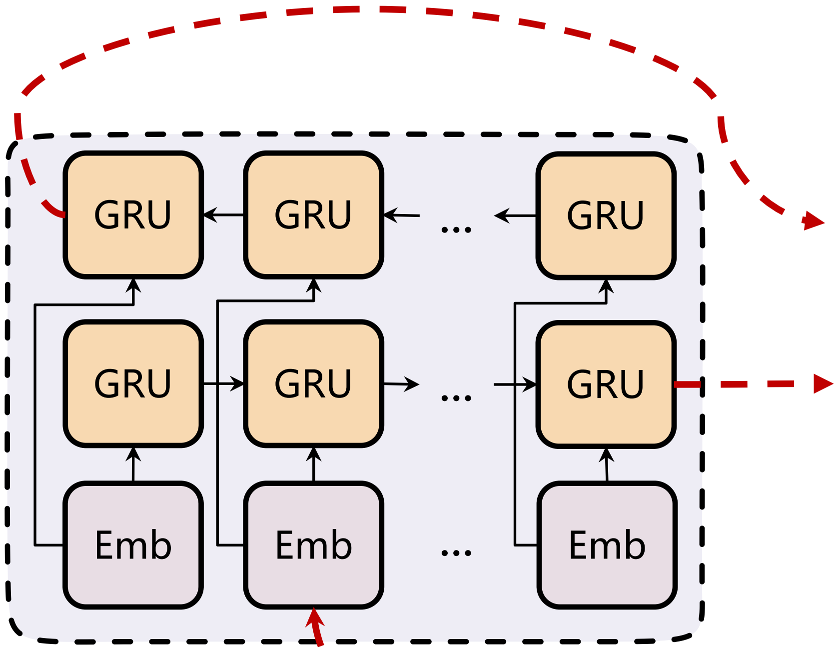
Reward

Discourse Coherent

Translation

Bosselut Antoine et., Discourse-Aware Neural Rewards for Coherent Text Generation. NAACL 2018

Reward Teacher



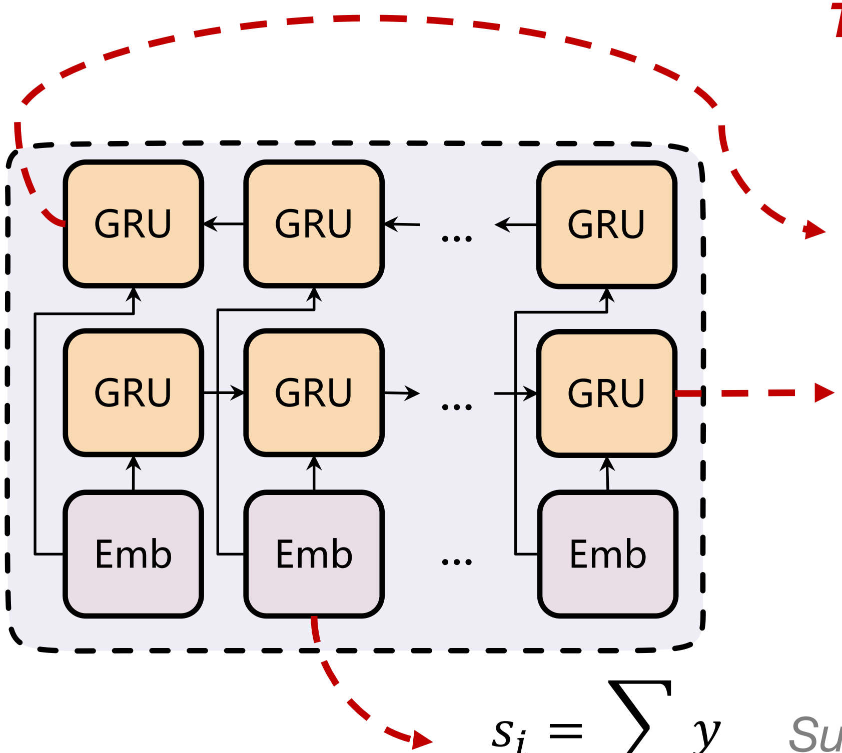
$f(\overleftarrow{s})$ *Sentences in reverse order*

$f(\vec{s})$ *Sentences in normal order*

$$s_i = \sum_{y \in s_i} y$$ *Summation of embeddings from each words*

Reward Teacher

Train to minimize $L = \frac{\langle f(\vec{s}), f(\overleftarrow{s}) \rangle}{\|f(\vec{s})\| \|f(\overleftarrow{s})\|}$



$f(\overleftarrow{s})$ *Sentences in reverse order*

$f(\vec{s})$ *Sentences in normal order*

$s_i = \sum_{y \in s_i} y$ *Summation of embeddings from each words*

Reward Teacher

$[-\mathbf{1},\mathbf{1}]$ $[-\mathbf{1},\mathbf{1}]$

$$S_1 = \frac{\left\langle f(\vec{s}) , f(\overrightarrow{\boldsymbol{s}^1}) \right\rangle}{\| f(\vec{s}) \| \left\| f(\overrightarrow{\boldsymbol{s}^1}) \right\|} - \frac{\left\langle f(\overleftarrow{s}) , f(\overrightarrow{\boldsymbol{s}^1}) \right\rangle}{\| f(\overleftarrow{s}) \| \left\| f(\overrightarrow{\boldsymbol{s}^1}) \right\|}$$

$\boldsymbol{s}$: *reference*

$\boldsymbol{s}^1$: *translation 1*

$\boldsymbol{s}^2$: *translation 2*

$$S_2 = \frac{\left\langle f(\vec{s}) , f(\overrightarrow{\boldsymbol{s}^2}) \right\rangle}{\| f(\vec{s}) \| \left\| f(\overrightarrow{\boldsymbol{s}^2}) \right\|} - \frac{\left\langle f(\overleftarrow{s}) , f(\overrightarrow{\boldsymbol{s}^2}) \right\rangle}{\| f(\overleftarrow{s}) \| \left\| f(\overrightarrow{\boldsymbol{s}^2}) \right\|}$$

If $\boldsymbol{S_1 > S_2}$ ***then***

$\boldsymbol{s}^1$ ***is more coherent than*** $\boldsymbol{s}^2$

Policy Learning

***Self-critical* Training**

***greedy search* translation** y^*

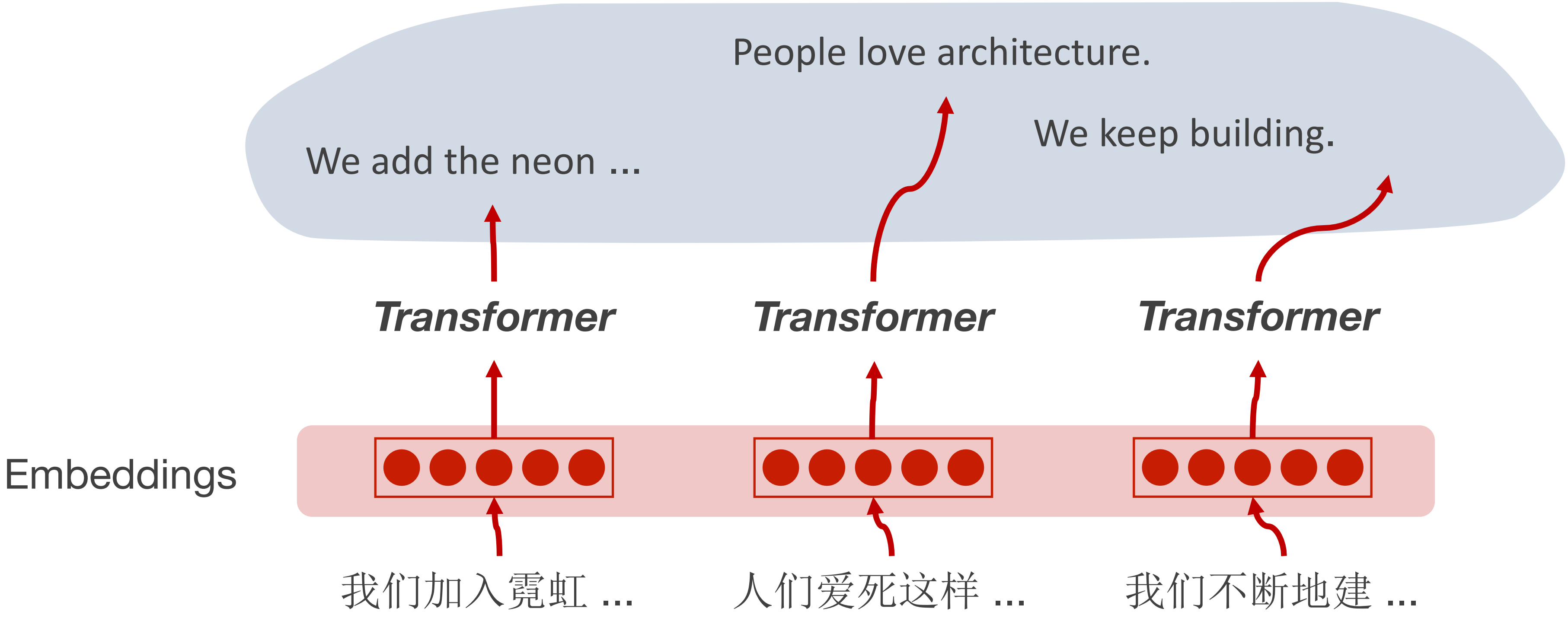
***sample* translation** $y^\wedge$

$$L_{rl} = -\sum_{i}^{n}\sum_{t}^{T_i}(r(y^\wedge) - r(y^*)) \cdot logP(y_t)$$

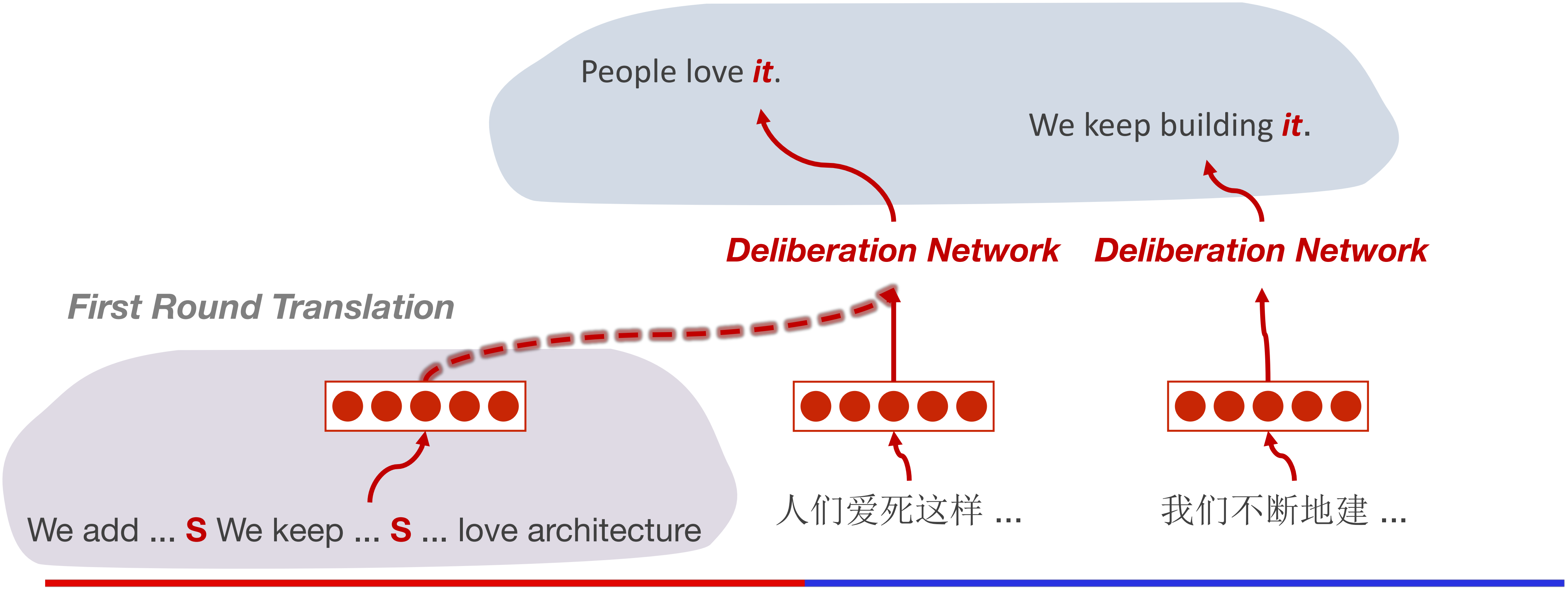
Rennie Steven J. et., Self-critical Sequence Training for Image Captioning. CVPR 2017

Running Example

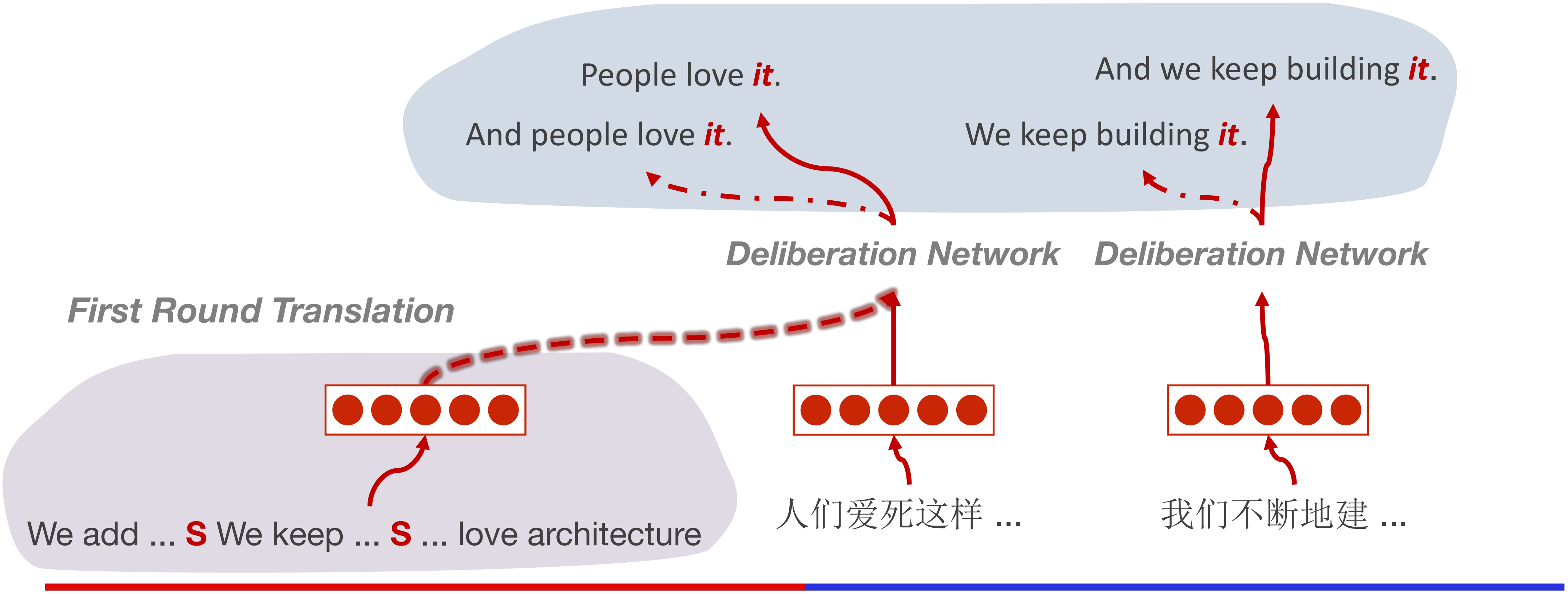
First Round Decoding



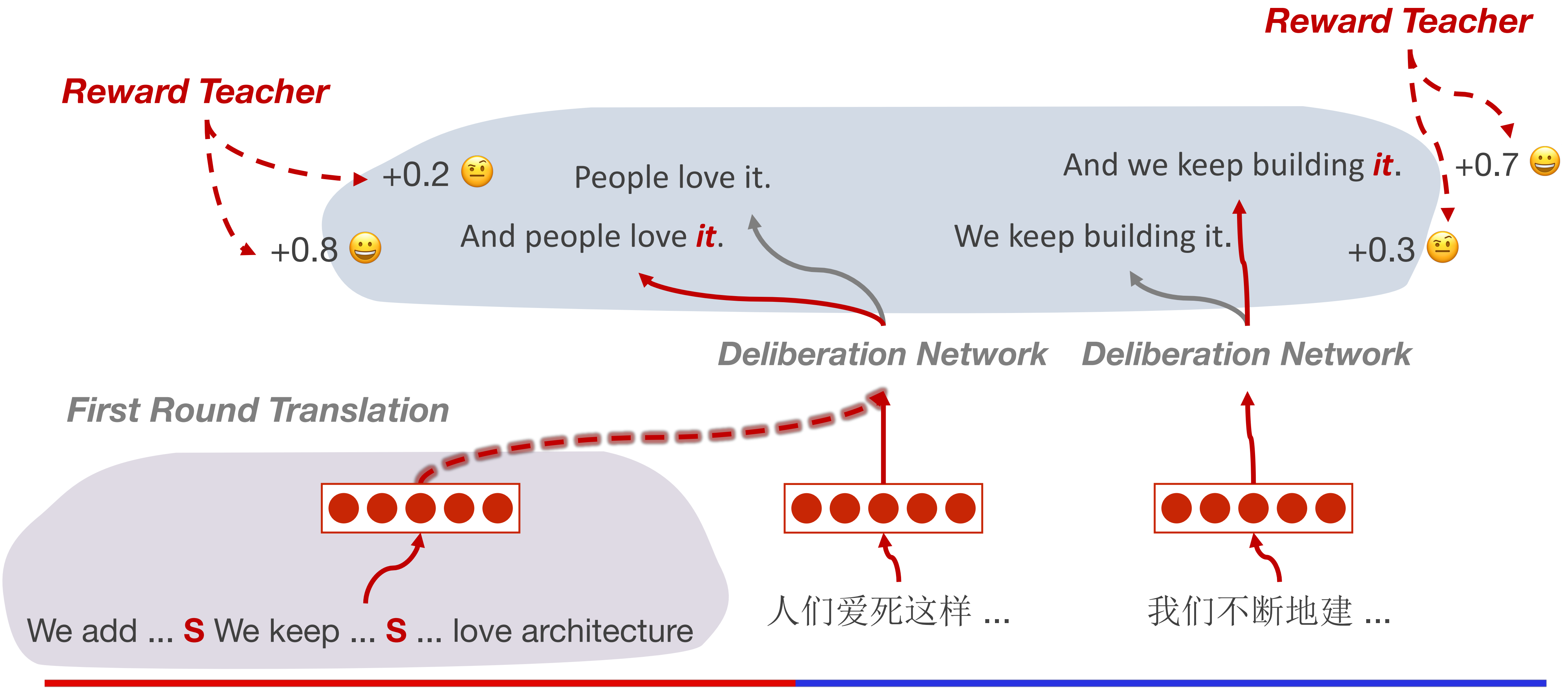
Deliberation Network



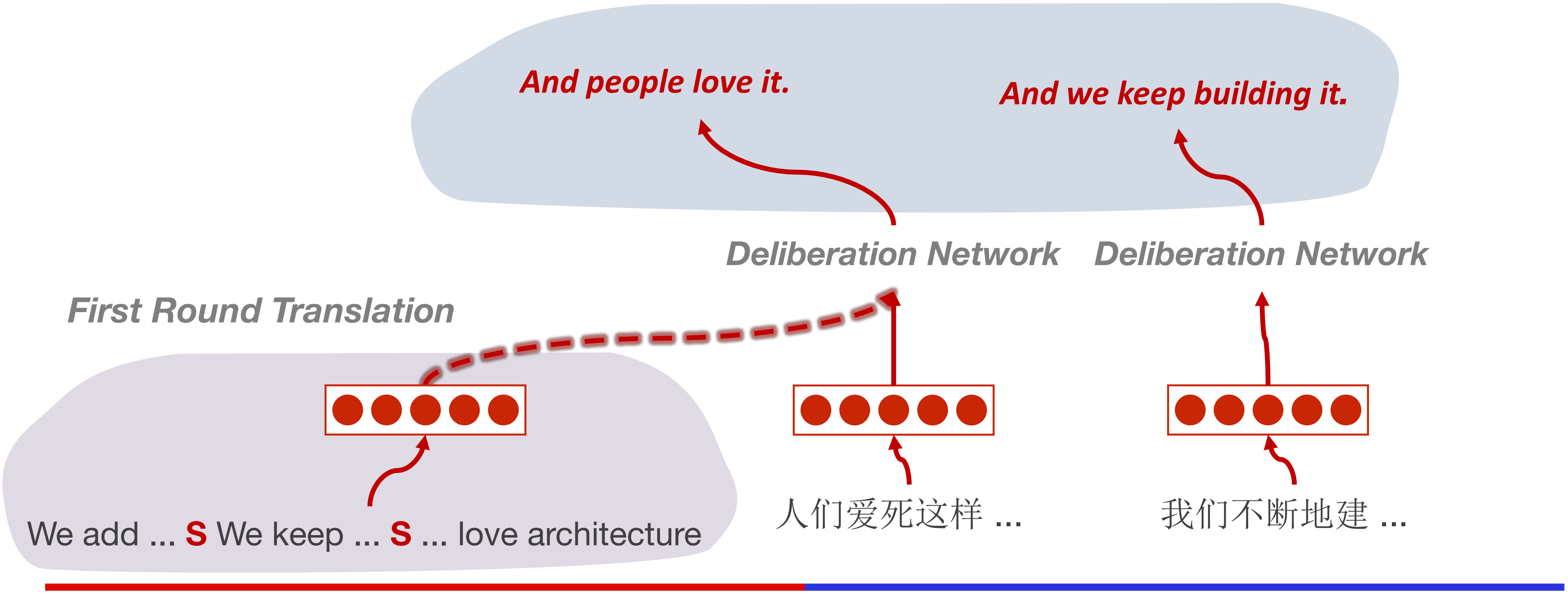
Self-critical Training



Discourse Reward



Two-pass Round Translation



Contents

- Backgrounds
- Model Architecture
- Experiments
- Conclusion

Data Preprocess

- Chinese Segmenter: [Jieba]
- English Tokenizer: [Moses Tokenizer]
- BPE size: Chinese(20K), English(18K)
- Data Size

Corpus	Talks	Sentences
Training	14,258	231,266
Dev	48	879
Test	234	3,874

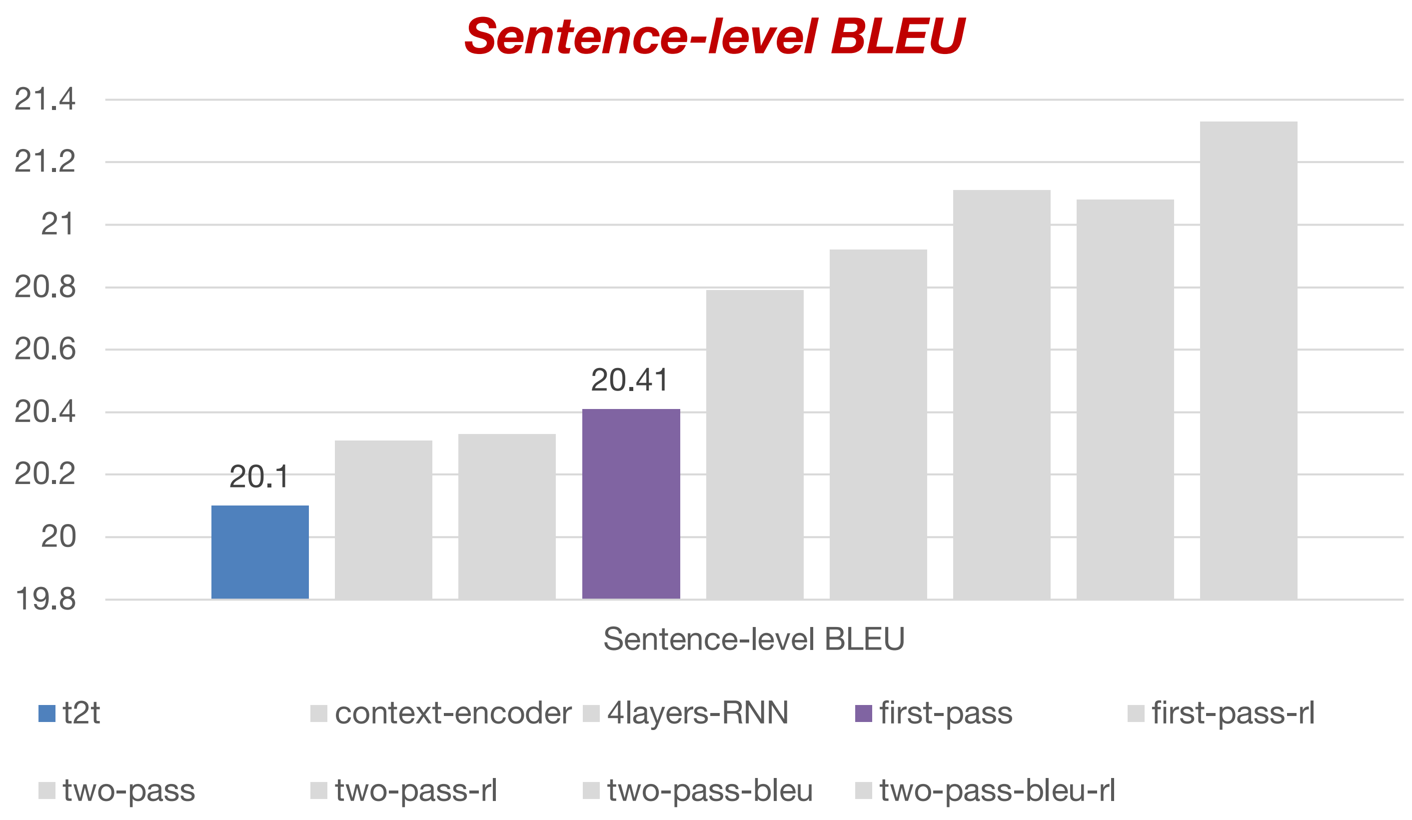
Systems

t2t	tensor2tensor V1.6.5
context-encoder	reimplementation of the work Voita et al.(2018)
first-pass	Train to minimize the first round decoding
first-pass-rl	*first-pass* with RL training
two-pass	Train to minimize the deliberation network
two-pass-rl	*two-pass* with RL training
two-pass-bleu	with BLEU as its reward
two-pass-bleu-rl	with BLEU and Reward Teacher as reward

Training Details

- *transform-base* version of hyperparameters
- *batch_size*: 320 (tokens)
- Reward Teacher
  - embedding size: 100
  - hidden size: 100
  - dropout: 0.3

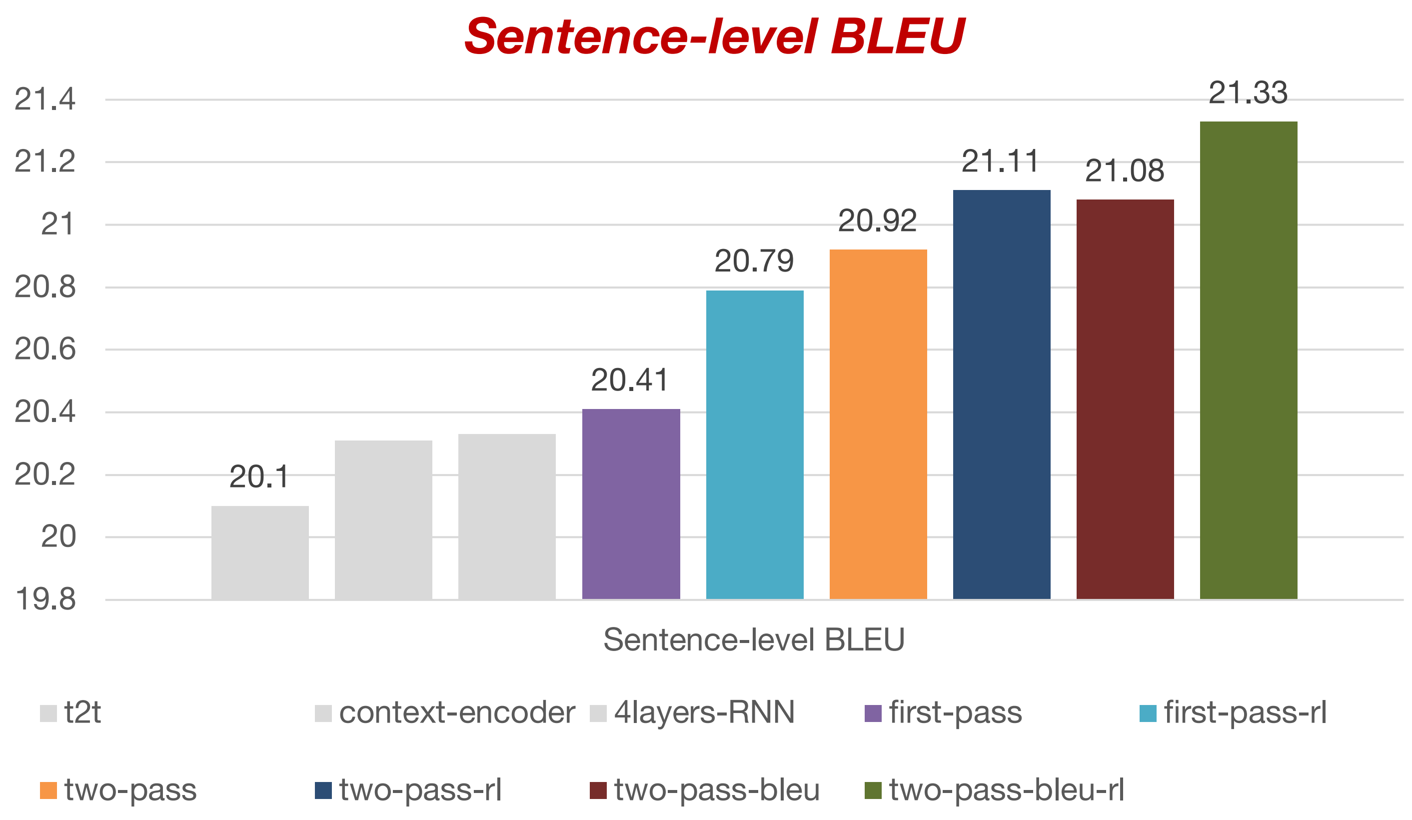
Experimental Conclusion 1



shuffle by ***talk*** is better than ***sentence***, can be viewed as well-designed ***curriculum learning***

Bengio Yoshua et., Curriculum Learning. ICML 2009

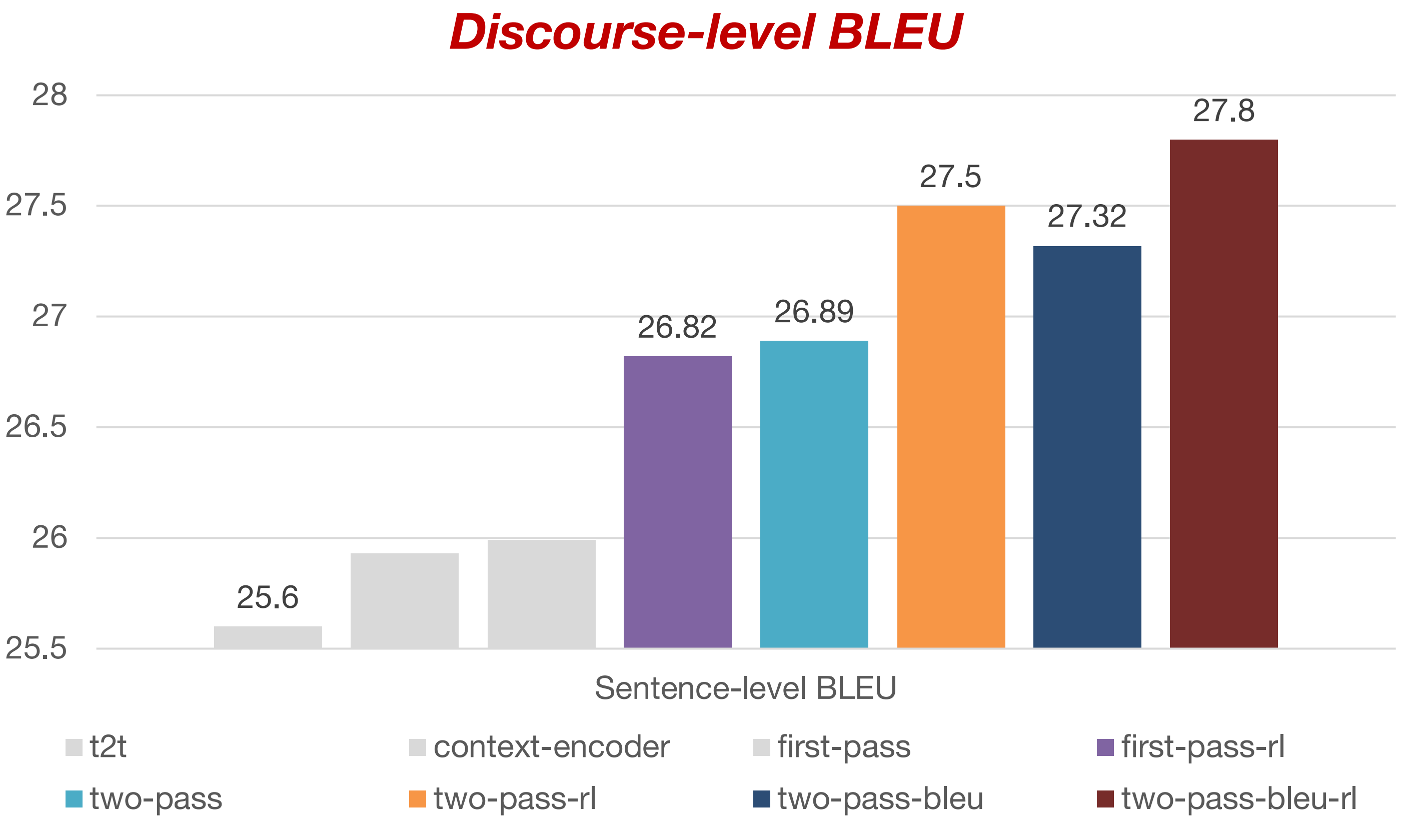
Experimental Conclusion 2



RL and ***second-pass decoding*** improve individual sentence quality

+1.2 BLEU

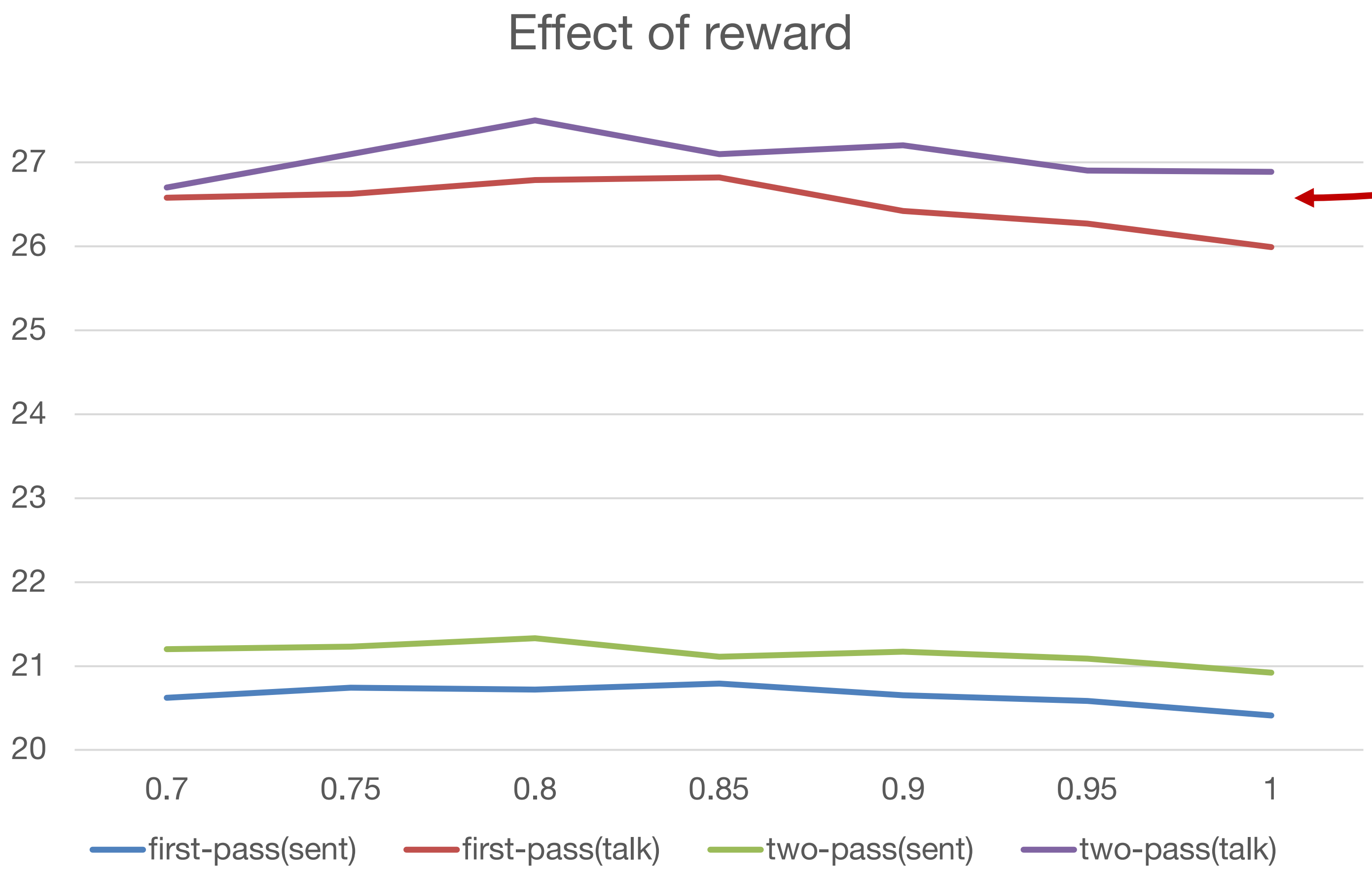
Experimental Conclusion 3



RL and ***second-pass decoding*** improve ***discourse*** quality

+2.2 BLEU

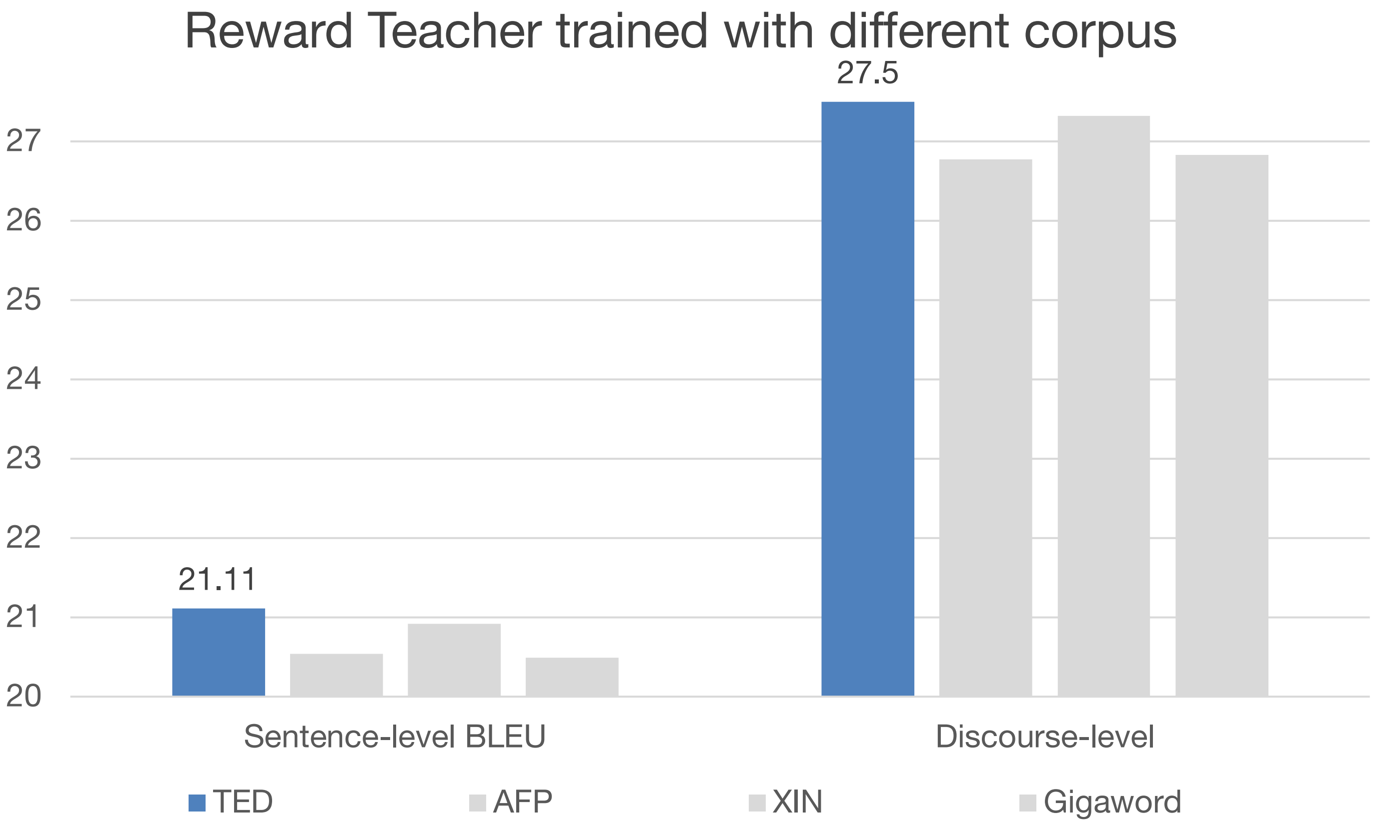
Experimental Conclusion 4



larger fluctuation

RL affects the discourse translation quality

Experimental Conclusion 5



Reward Teacher is better trained with in-domain corpus

Experimental Conclusion 6

Our model significantly improve the *discourse coherence*

systems	***tst-2013***	***tst-2014***	***tst-2015***
t2t	0.5991	0.5838	0.5939
first-pass	0.5999	0.5845	0.5943
first-pass-rl	0.6008	0.5861	0.5952
two-pass	0.6011	0.5880	0.5962
two-pass-rl	0.6032	0.5913	0.6008
two-pass-bleu-rl	0.6041	0.5938	0.6014
Human translation	0.6066	0.5910	0.6013

Lapata and Barzilay, Automatic Evaluation of Text Coherence: Models and Representations. IJCAI 2005

Experimental Conclusion 7

Our model tends to using more ***conjunctions***

systems	***t2t***	***two-pass-bleu-rl***
And	519	540
But	186	183
In	114	129
So	174	178
What	55	73

Statistics of top five frequent conjunctions in two systems.

Contents

- Backgrounds
- Model Architecture
- Experiments
- Conclusion

Conclusions

- First work on generating discourse coherent translations
- Two-pass round decoding strategy with Deliberation Network
- RL to encourage generating discourse coherent translations
- Experimental results confirm the effectiveness of our models
- Analysis reveals the contribution of our model to generate discourse coherent translations

THANKS