

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 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

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

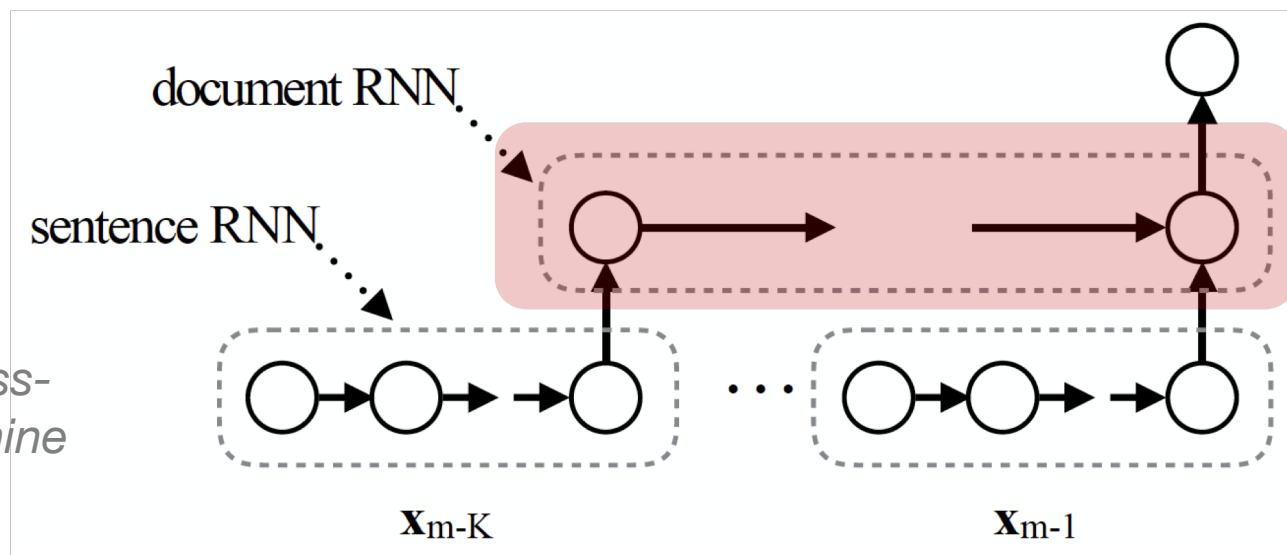
- Translate each sentence independently
- Lack of *Discourse Coherence*
- Lack of using *Discourse Context*



Previous Solutions

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

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

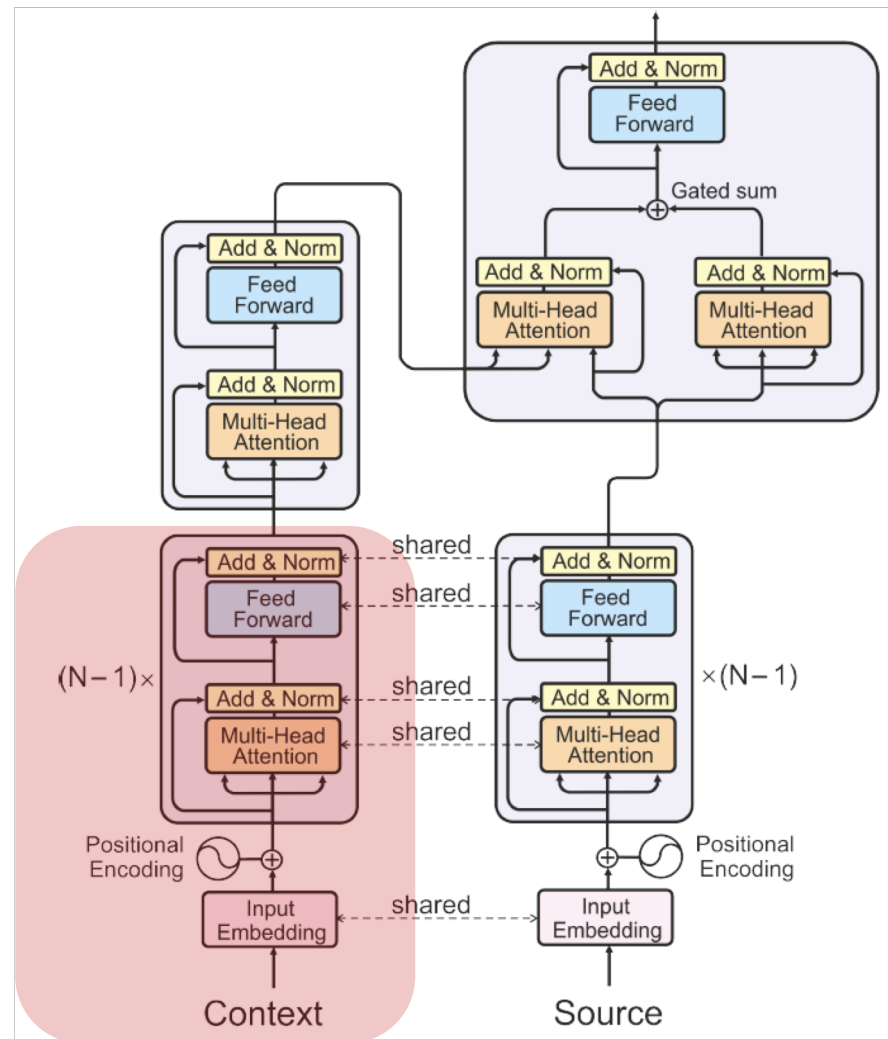


Previous Solutions

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

- Backgrounds
- **Model Architecture**
- Experiments
- Conclusion

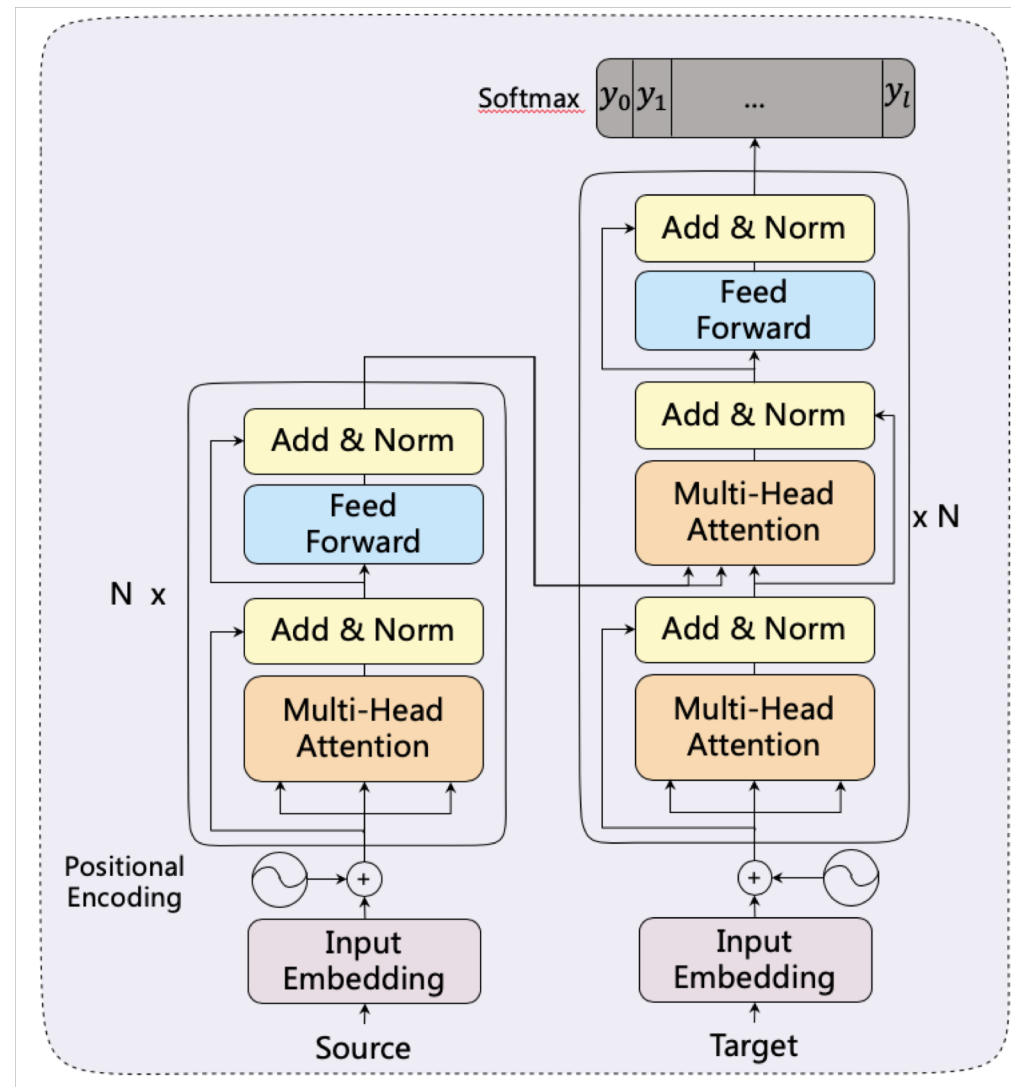


Overall Architecture

First Round:

Canonical **Transformer**

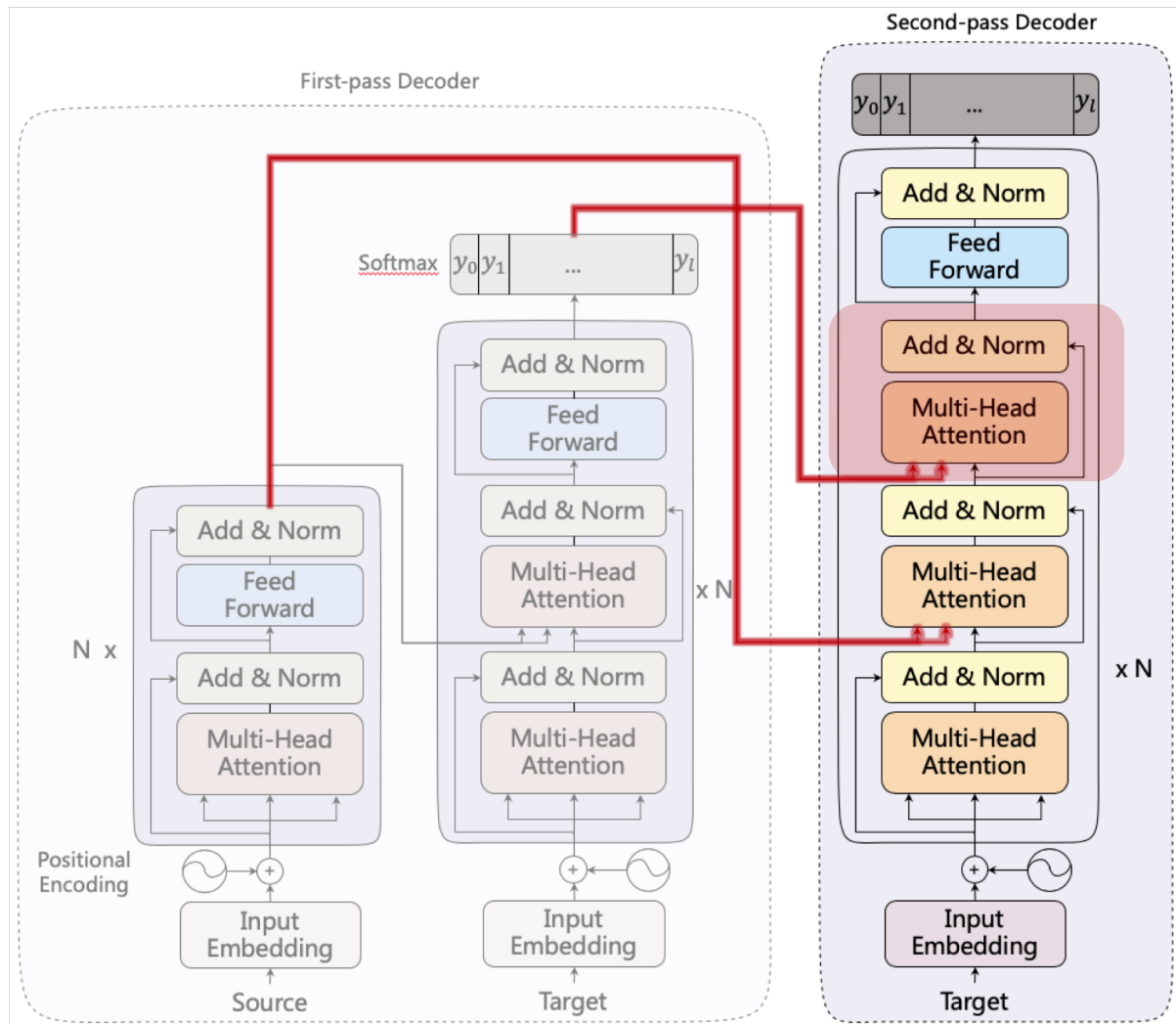
Vaswani, Ashish et., Attention is All You Need. NIPS 2017



Overall Architecture

Second Round: Deliberation Network

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



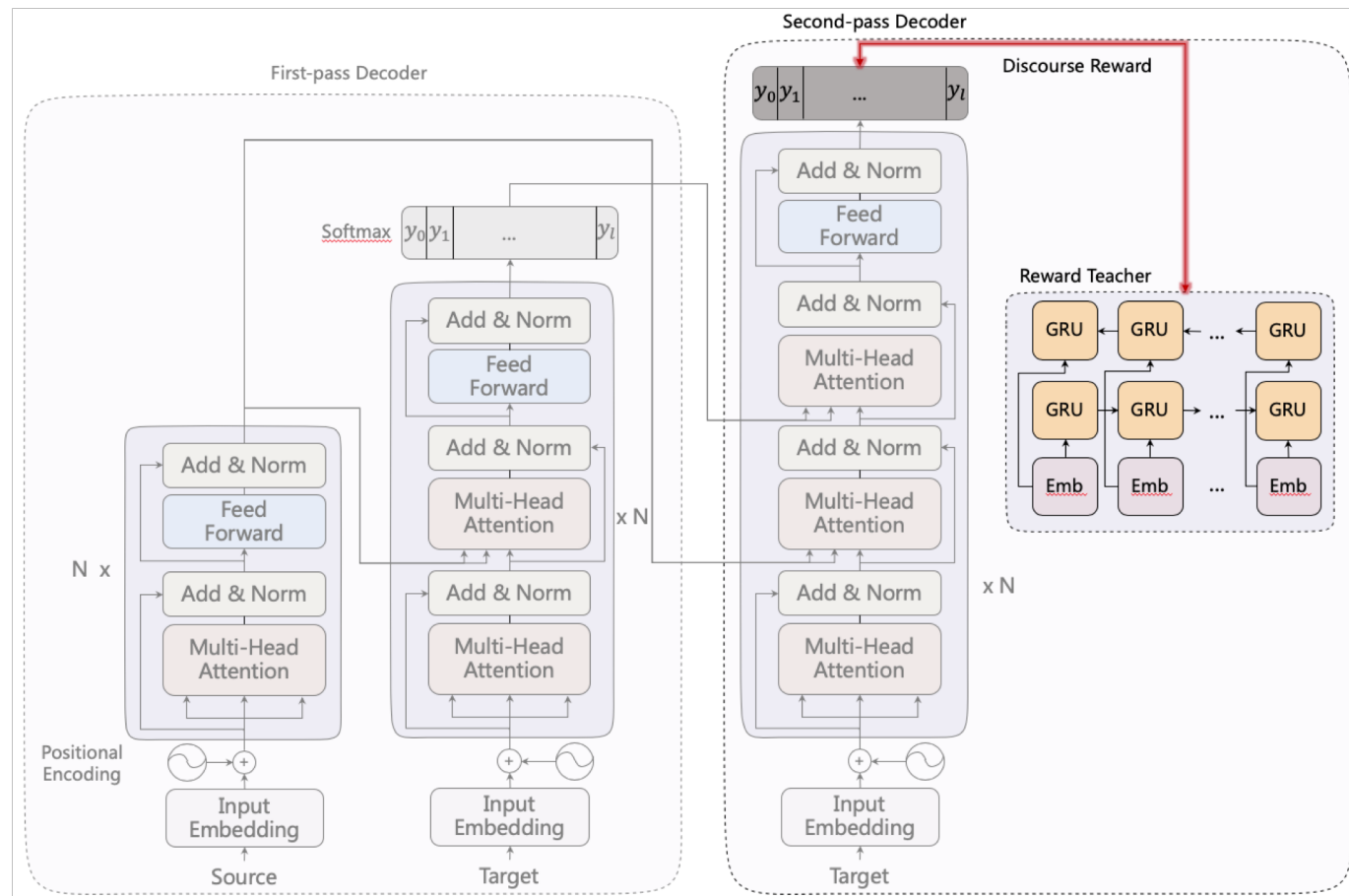
Overall Architecture

Reward

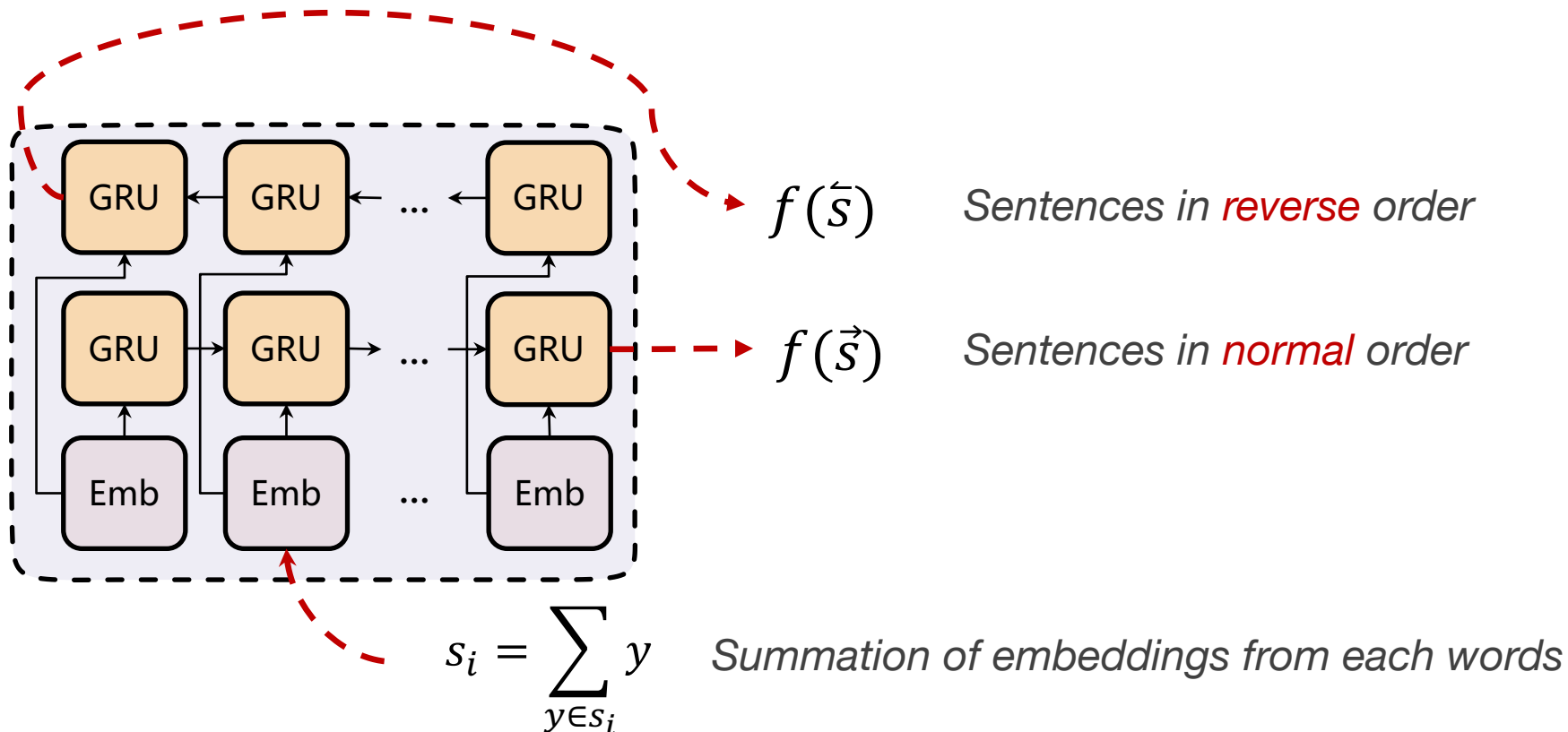
Discourse Coherent

Translation

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

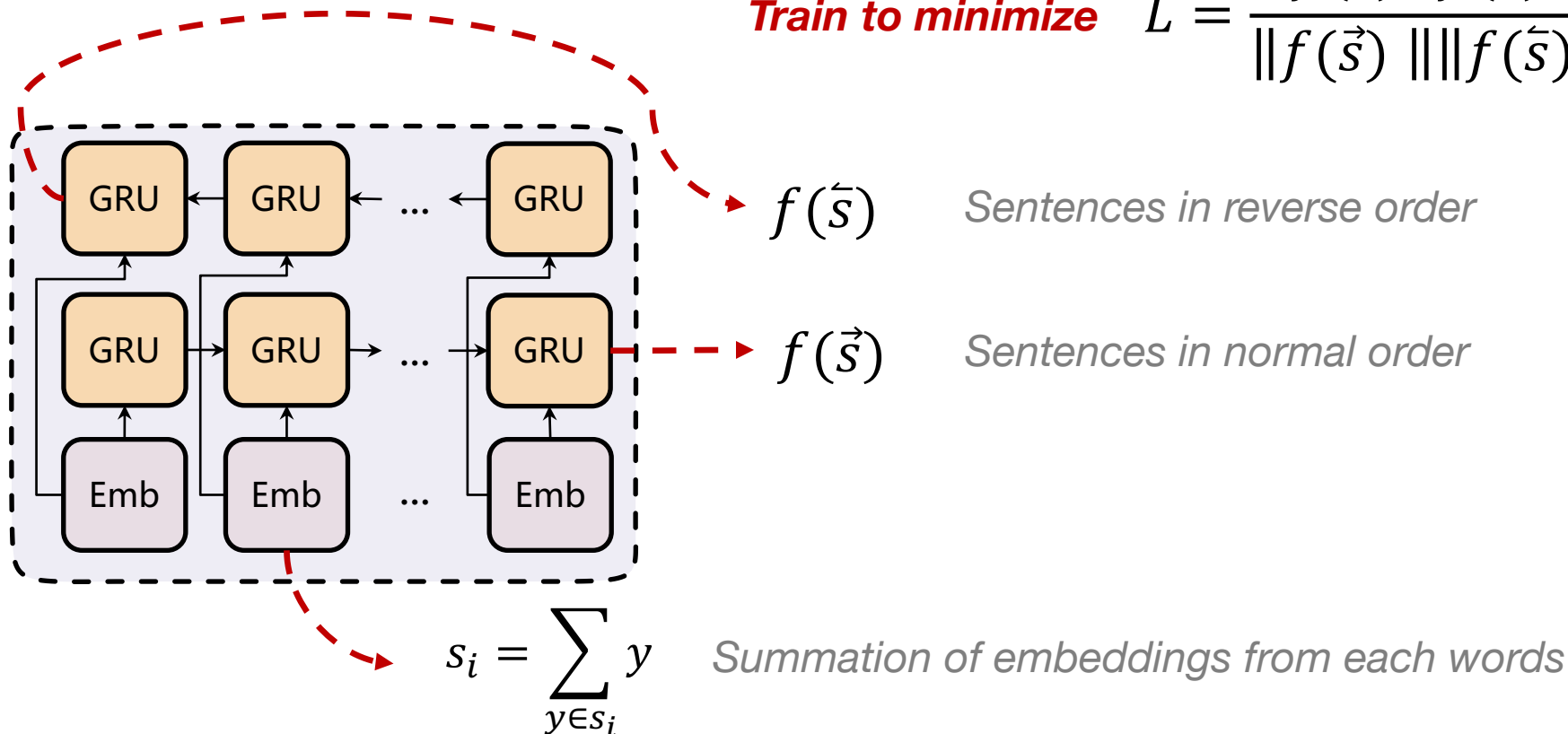


Reward Teacher



Reward Teacher

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



Reward Teacher

$[-1, 1]$

$[-1, 1]$

$$S_1 = \frac{\langle f(\vec{s}) , f(\vec{s^1}) \rangle}{\|f(\vec{s})\| \|f(\vec{s^1})\|} - \frac{\langle f(\vec{\hat{s}}) , f(\vec{s^1}) \rangle}{\|f(\vec{\hat{s}})\| \|f(\vec{s^1})\|}$$

$$S_2 = \frac{\langle f(\vec{s}) , f(\vec{s^2}) \rangle}{\|f(\vec{s})\| \|f(\vec{s^2})\|} - \frac{\langle f(\vec{\hat{s}}) , f(\vec{s^2}) \rangle}{\|f(\vec{\hat{s}})\| \|f(\vec{s^2})\|}$$

s : reference

s^1 : translation 1

s^2 : translation 2

If $S_1 > S_2$ then

s^1 is more coherent than s^2

Policy Learning

Self-critical Training

greedy search translation \mathbf{y}^*
sample translation \mathbf{y}^\wedge

$$L_{rl} = - \sum_i^n \sum_t^{T_i} (r(\mathbf{y}^\wedge) - r(\mathbf{y}^*)) \cdot \log P(y_t)$$

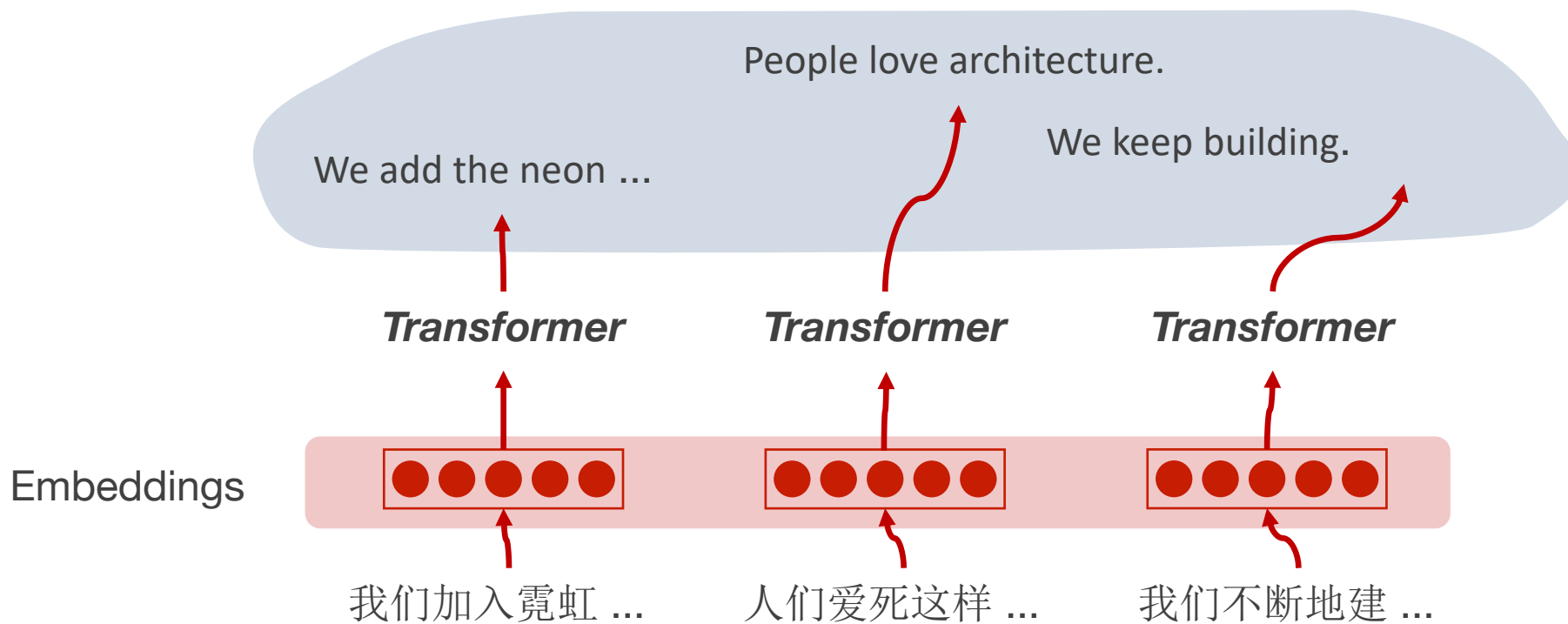
Rennie Steven J. et., Self-critical Sequence
Training for Image Captioning. CVPR 2017



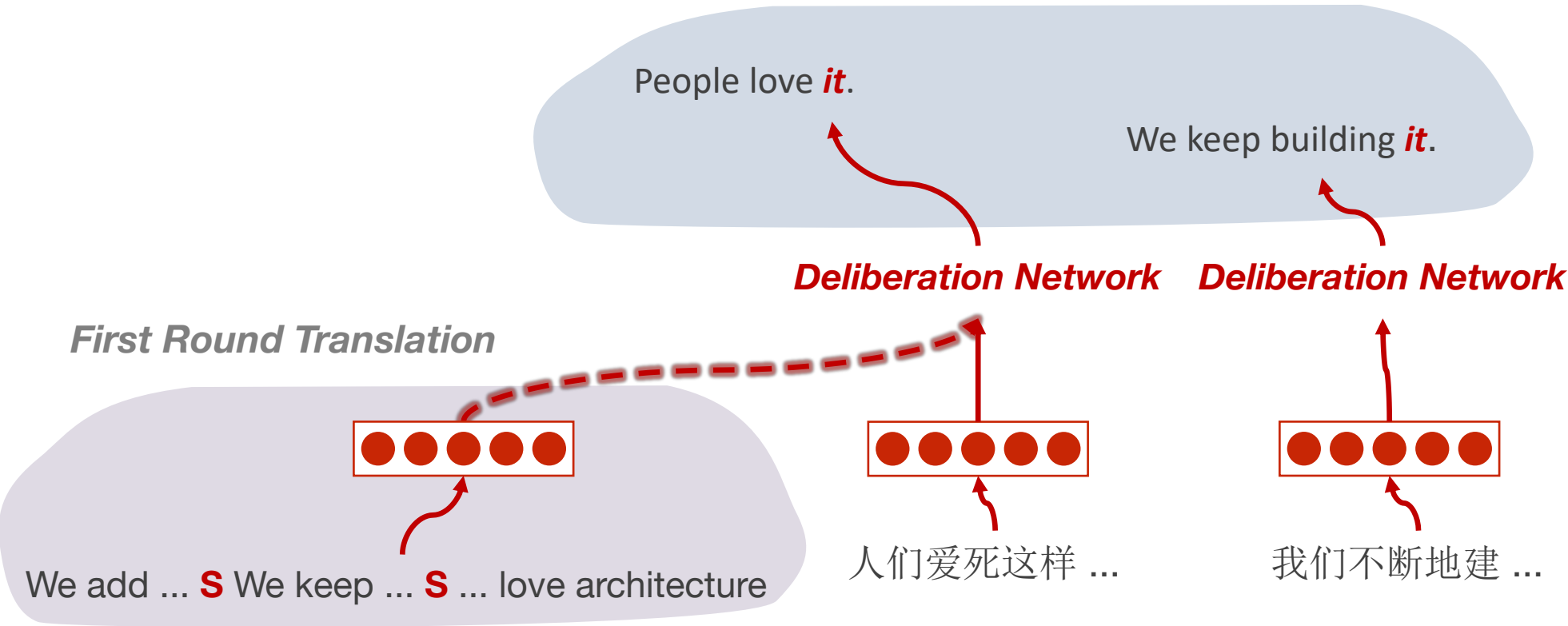
Running Example



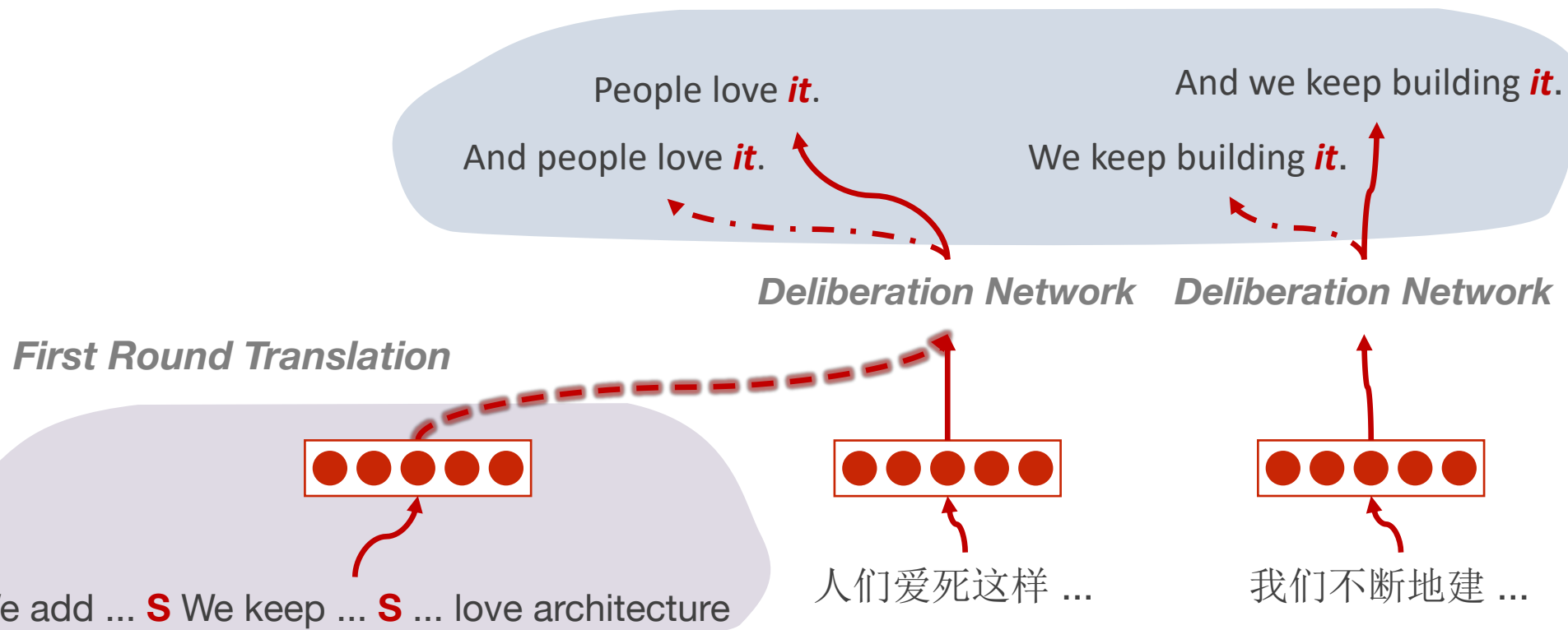
First Round Decoding



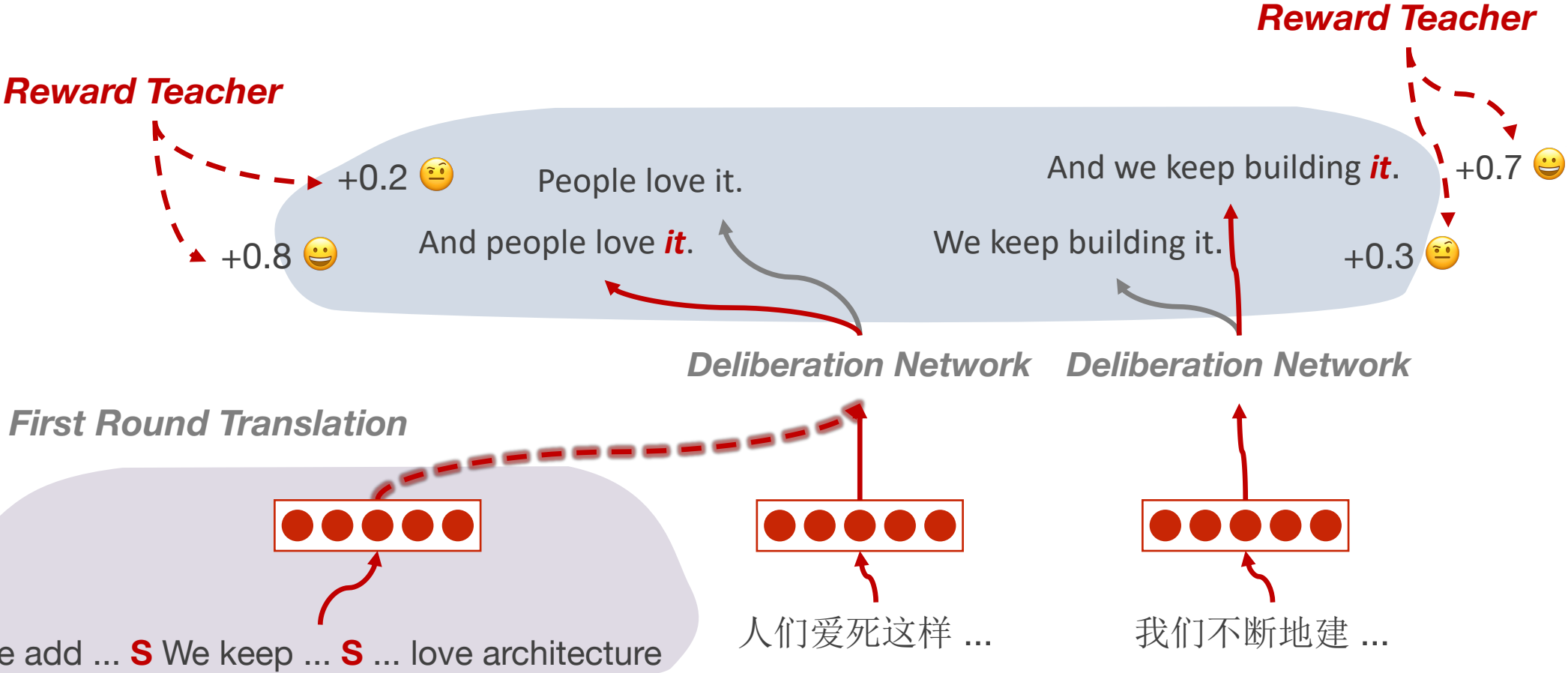
Deliberation Network



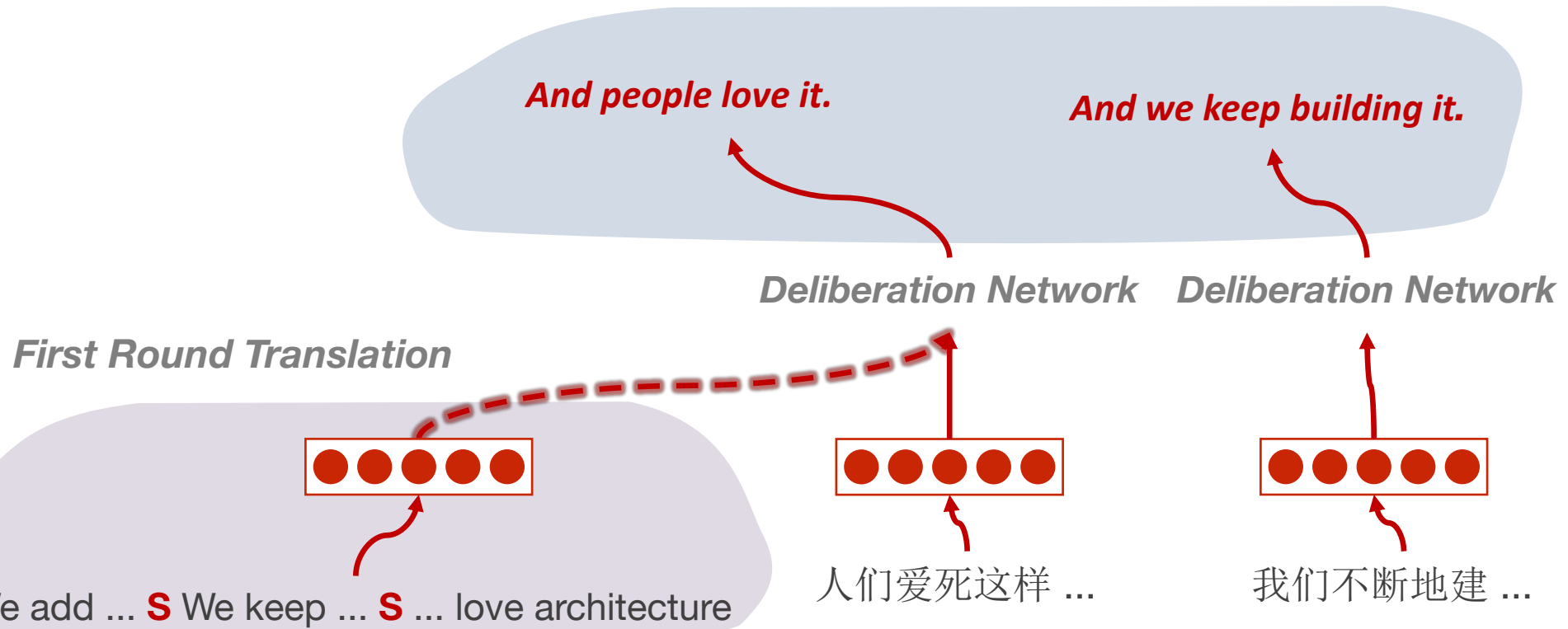
Self-critical Training



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

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

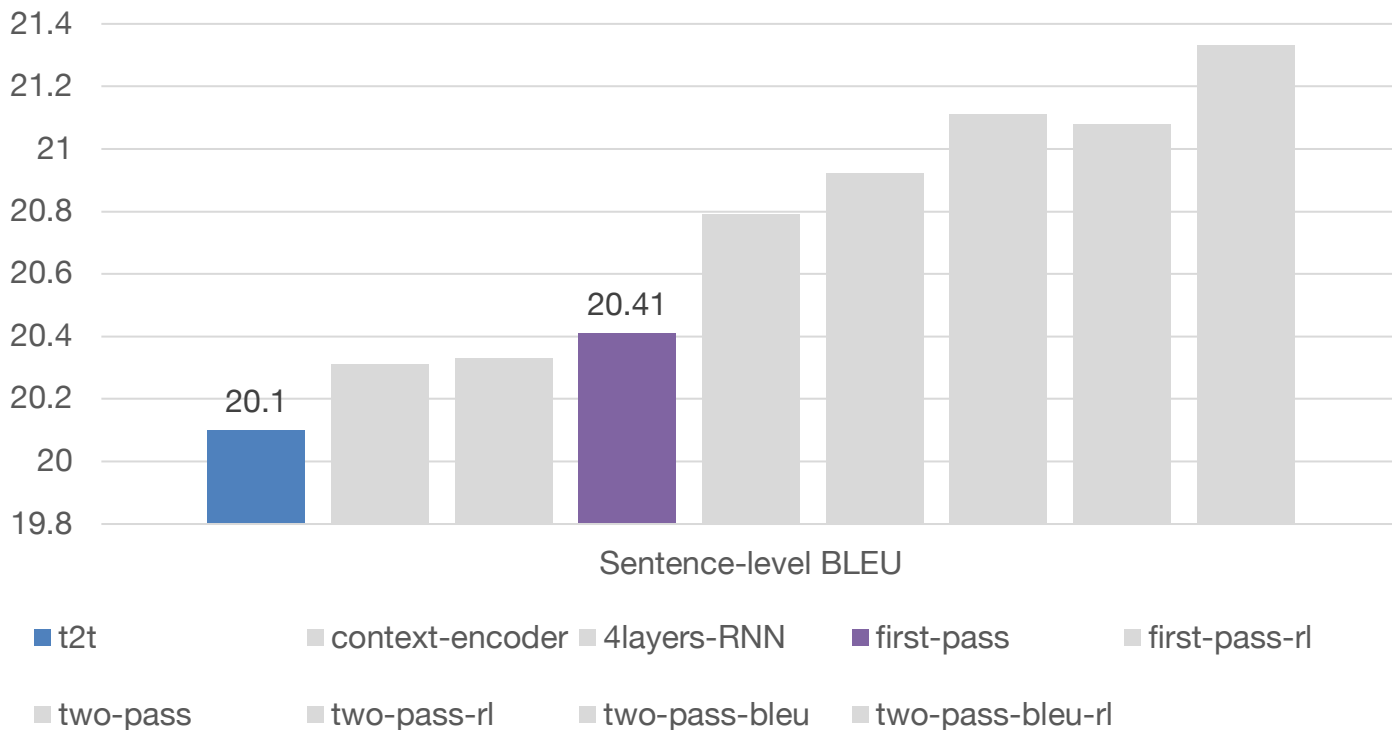
Training Details

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



Experimental Conclusion 1

Sentence-level BLEU

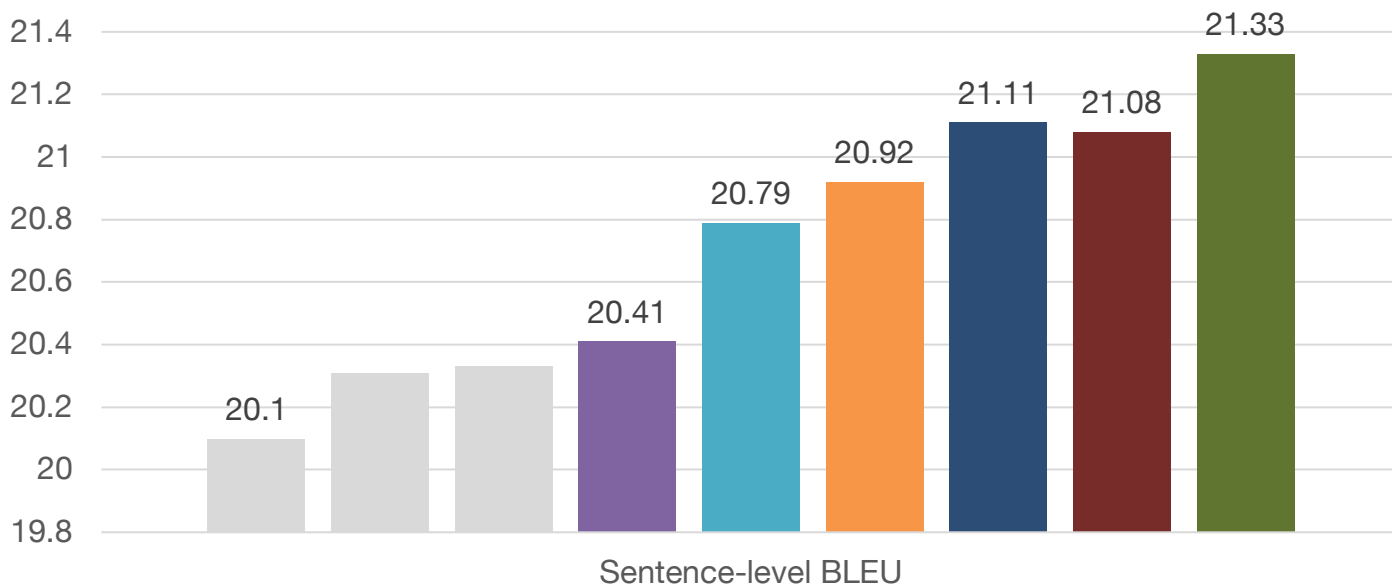


shuffle by *talk* is better than *sentence*, can be viewed as well-designed *curriculum learning*

Bengio Yoshua et., Curriculum Learning. ICML 2009

Experimental Conclusion 2

Sentence-level BLEU



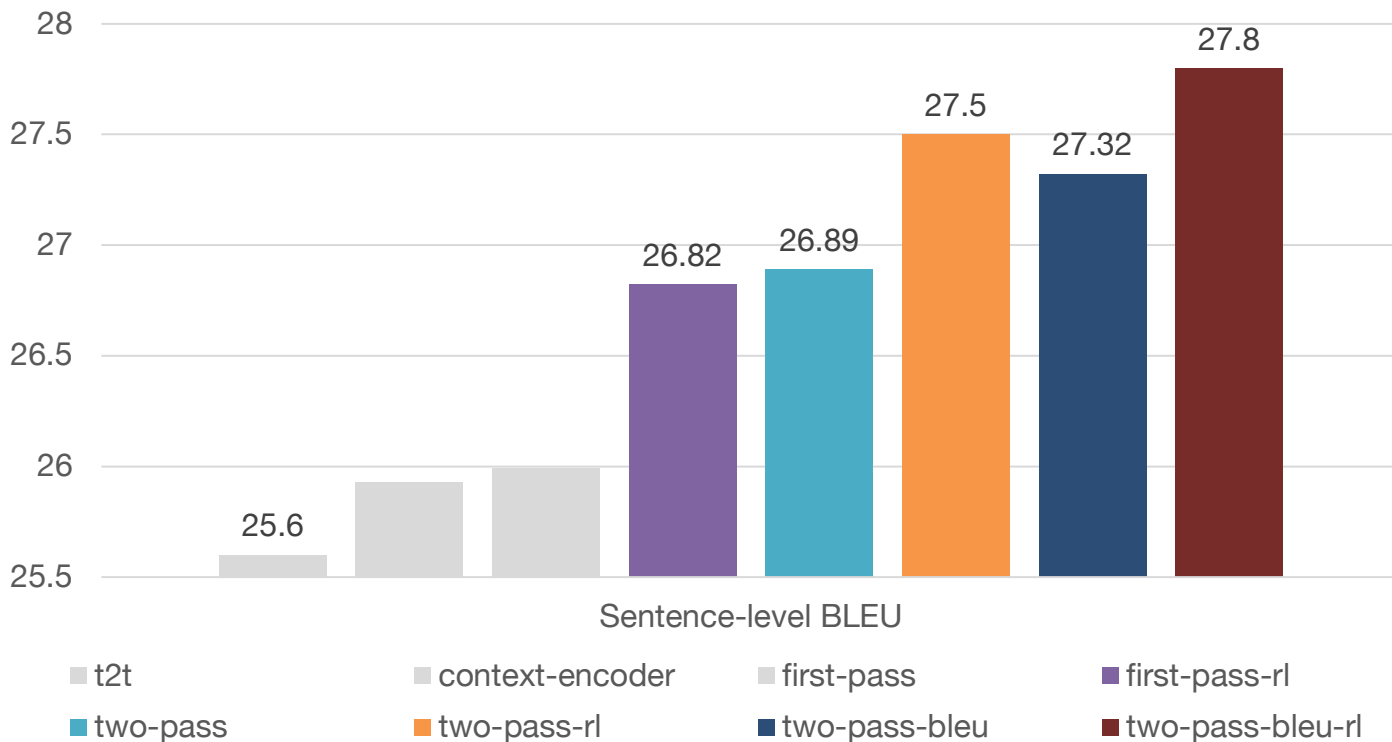
t2t context-encoder 4layers-RNN first-pass first-pass-rl
two-pass two-pass-rl two-pass-bleu two-pass-bleu-rl

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

+1.2 BLEU

Experimental Conclusion 3

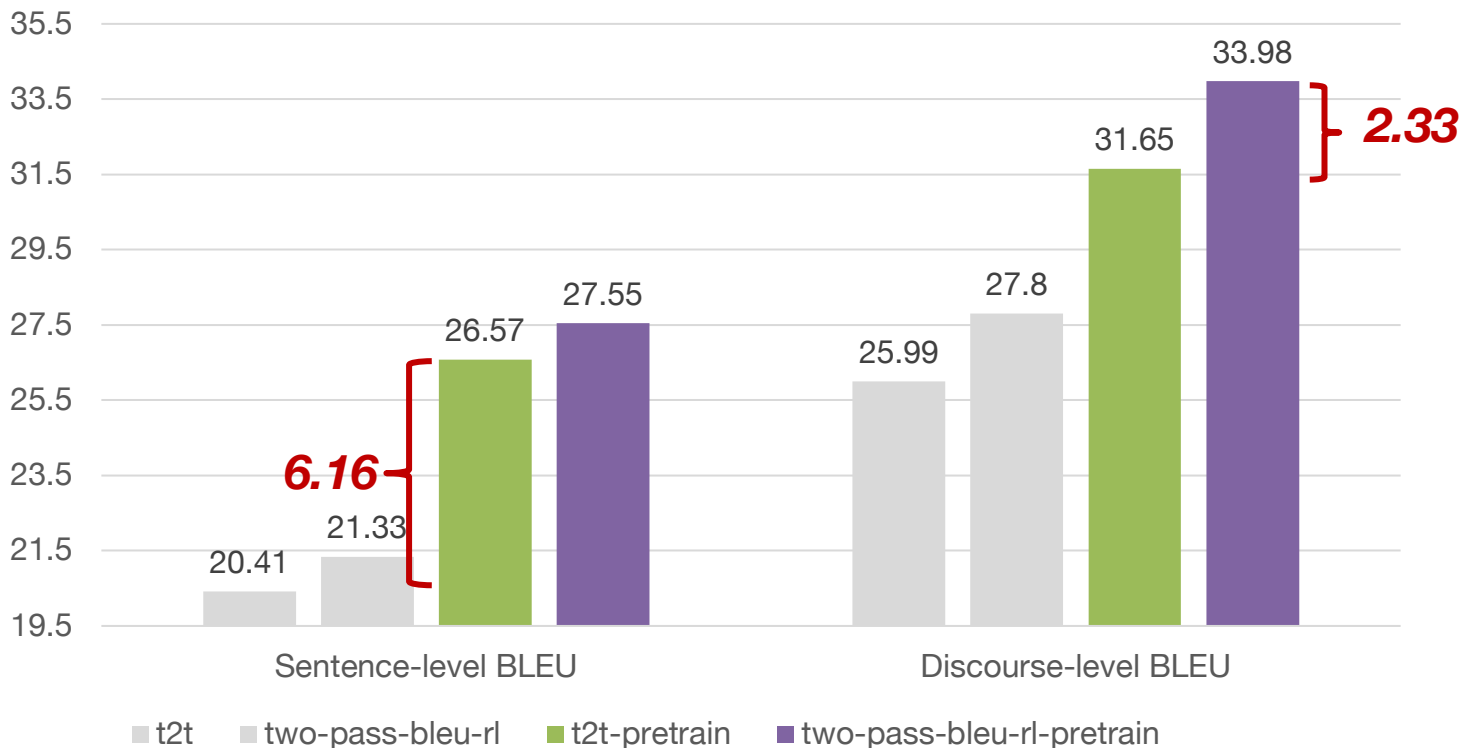
Discourse-level BLEU



RL and *second-pass decoding* improve *discourse* quality
+2.2 BLEU

Experimental Conclusion 4

Pretrain on 25M WMT Corpus

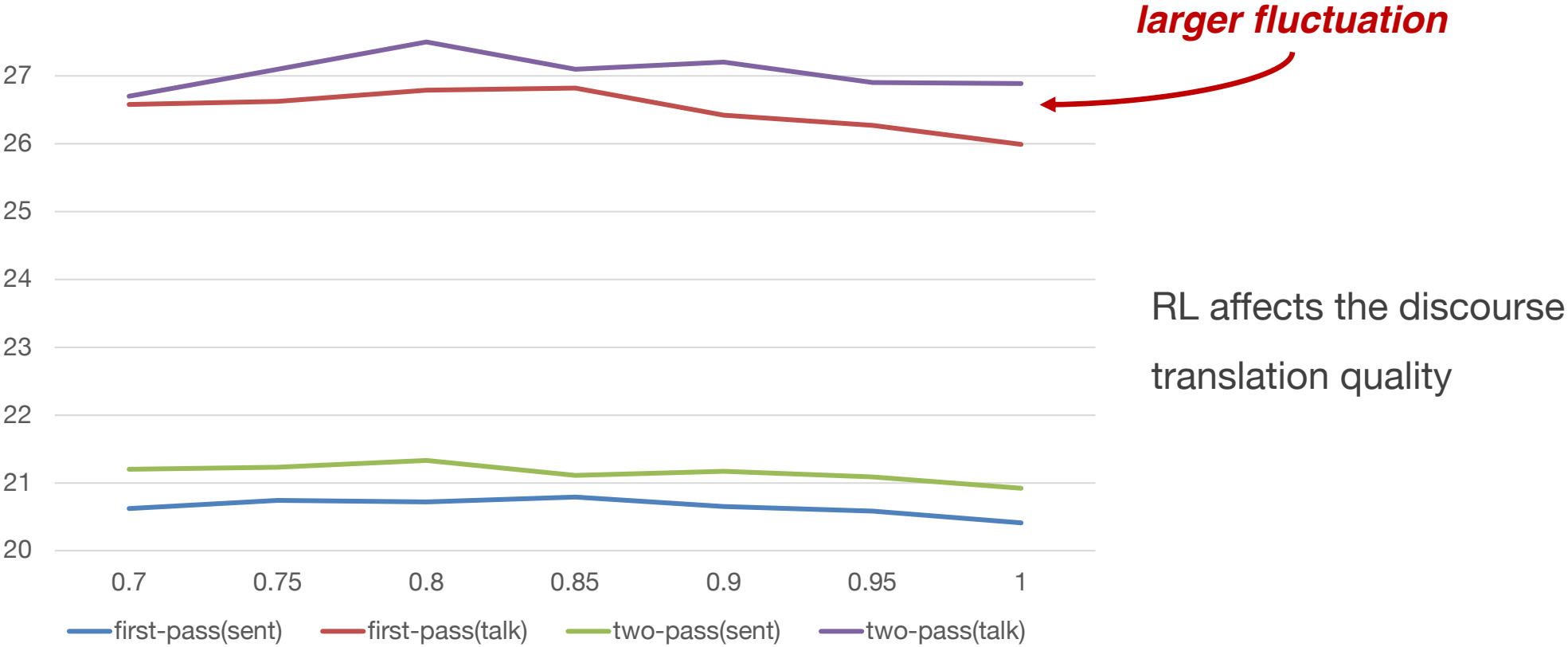


Pretraining significantly improves the performance by **+6 points**.

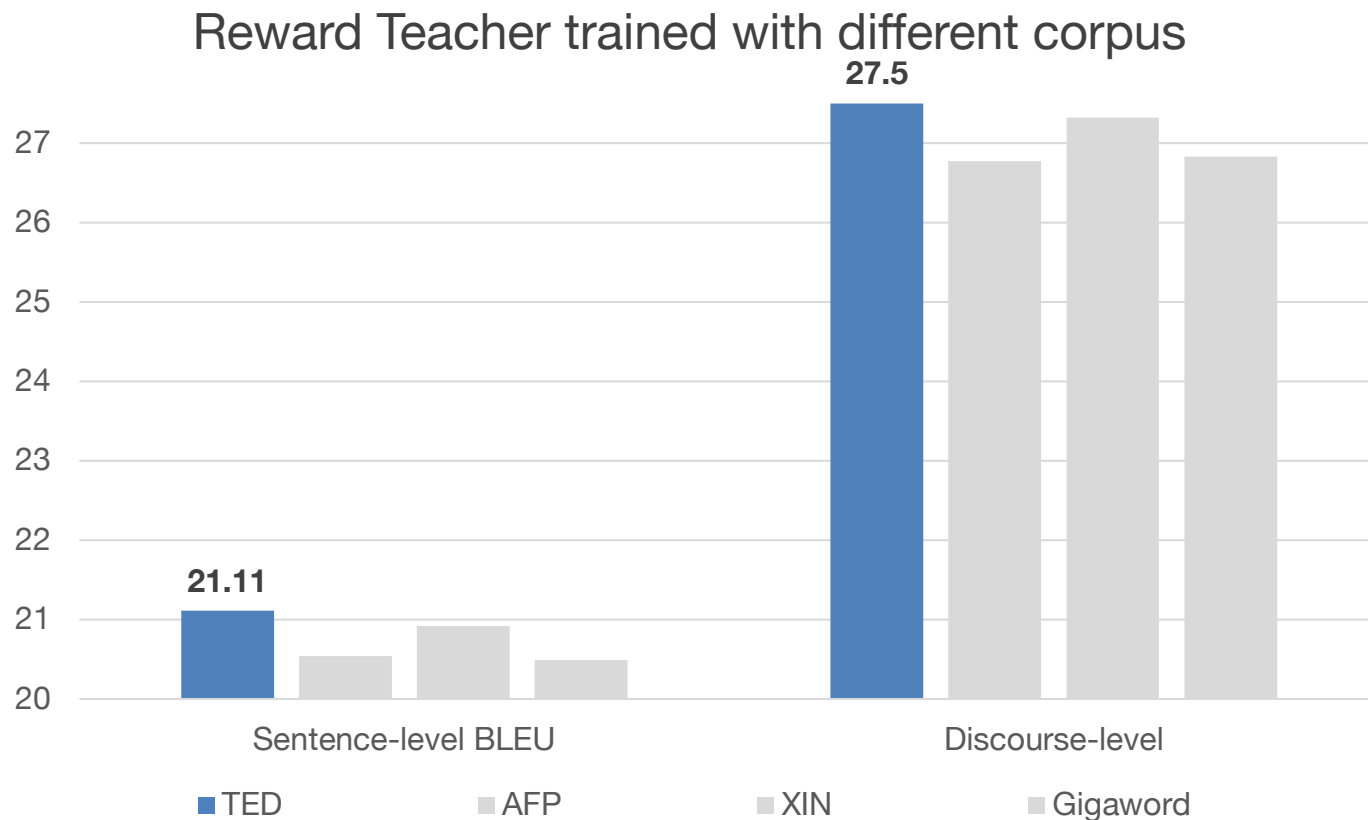
Our approach is **robust** and **practical**.

Experimental Conclusion 5

Effect of reward



Experimental Conclusion 6



Reward Teacher is better trained with in-domain corpus

Experimental Conclusion 7

Our models significantly improve the *discourse coherence*

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

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

Experimental Conclusion 8

Our model tends to using more *conjunctions*

<i>systems</i>	<i>t2t</i>	<i>two-pass-bleu-rl</i>
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



Conclusion

- **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

