Deliberation Networks: Sequence Generation Beyond One-Pass Decoding

Yingce Xia[1], Fei Tian[2], Lijun Wu[3], Jianxin Lin[1], Tao Qin[2], Nenghai Yu[1], Tie-Yan Liu[2]

1University of Science and Technology of China, Hefei, China

2Microsoft Research, Beijing, China

3Sun Yat-sen University, Guangzhou, China

Speaker: AntNLP-Chenrui Li

Outline

1.Introduction

2.Motivation

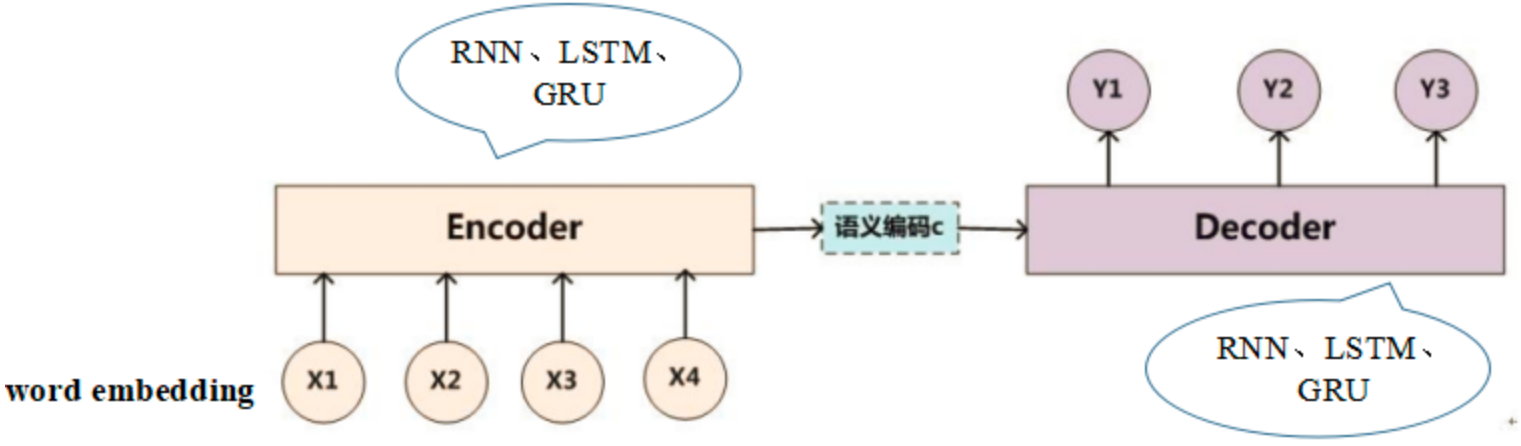
3.Framewrok

4.Application

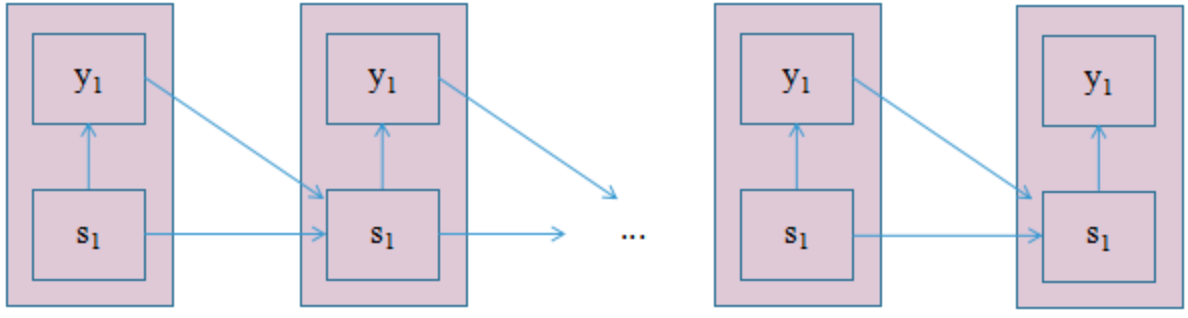
5.Conclusion

Introduction

Standard Encoder-Decoder Model



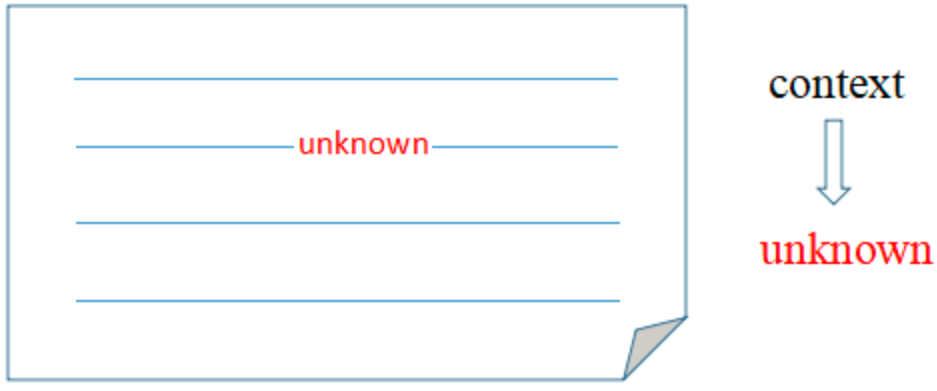
Decoder



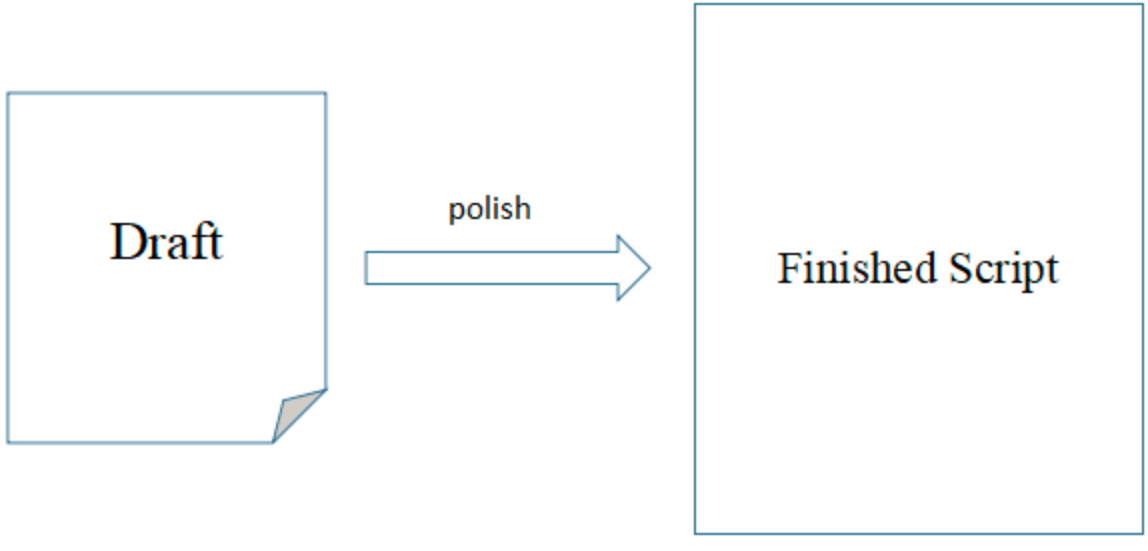
Task: NMT, Dialog System, Question Generation, Summarization Generation...

Motivation

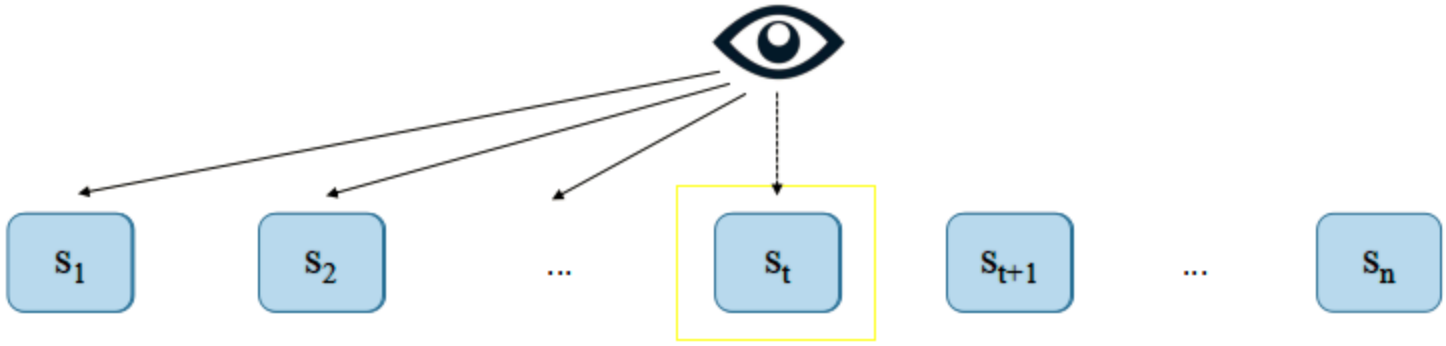
reading behavior



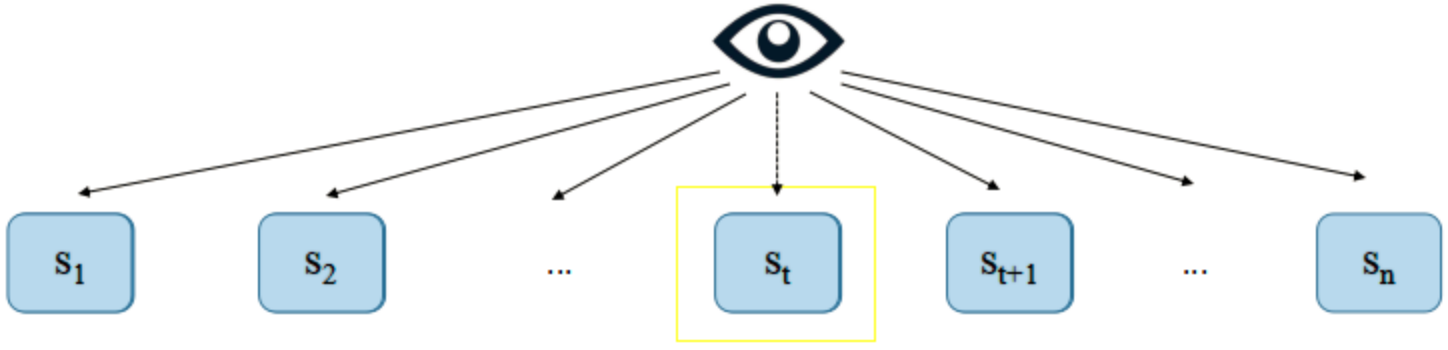
writing behavior



looking forward



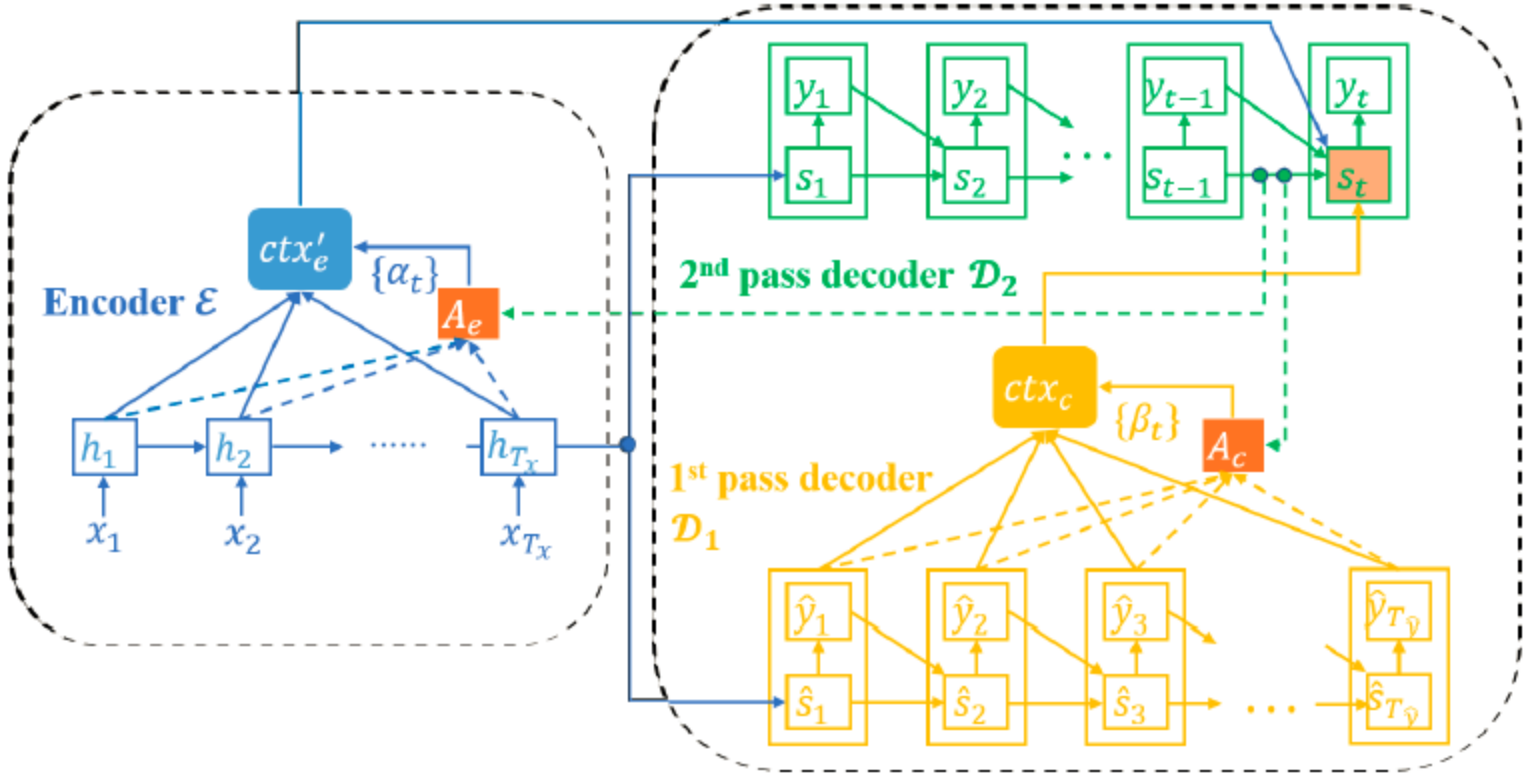
looking forward and back



Deliberation: leveraging the global information with both looking forward and back in sequence decoding.

Framework

Deliberation Network



Encoder

$$h_i = RNN(x_i, h_{i-1})$$

First-Pass Decoder

$$\hat{s}_j = RNN([\hat{y}_{j-1}; ctx_e], \hat{s}_{j-1})$$
$$ctx_e = \sum_{i=1}^{T_x} \alpha_i h_i$$
$$\alpha_i \propto exp(v_\alpha^T tanh(W^c_{att,h} h_i + W^c_{att,\hat{s}} \hat{s}_{j-1})) \quad \forall i \in [T_x]$$

$$P = sofmtmax(Linear([\hat{s}_j; cte_x; y_{j-1}]))$$

$\hat{y}_j$ is sampled out from P

Second-Pass Decoder

$$s_t = RNN([y_{t-1}; ctx_e^{'}; ctx_c], s_{t-1})$$
$$ctx_c = \sum_{j=1}^{T_{\hat{y}}} \beta_j [\hat{s}_j; \hat{y}_j]$$
$$\beta_j \propto exp(v_\beta^T tanh(W_{att,\hat{s}y}^d [\hat{s}_j; \hat{y}_j] + W_{att,s}^d s_{t-1})) \quad \forall j \in [T_{\hat{y}}]$$

$$P_2 = softmax(Linear([s_t; ctx_e^{'}; ctx_c; y_{t-1}]))$$

y_t is sampled out from P_2

Algorithm

$$maximize \frac{1}{n} \sum \mathcal{J}(x, y; \theta_e, \theta_1; \theta_2)$$

$$\mathcal{J}(x, y; \theta_e, \theta_1, \theta_2) = \log \sum_{y' \in \mathcal{Y}} P(y|y', E(x; \theta_e); \theta_2) P(y'|E(x; \theta_e); \theta_1)$$

$$\nabla_{\theta_1} \mathcal{J}(x, y; \theta_e, \theta_1, \theta_2) = \frac{\sum_{y' \in \mathcal{Y}} P(y|y', E(x; \theta_e); \theta_2) \nabla_{\theta_1} P(y'|E(x; \theta_e); \theta_1)}{\sum_{y' \in \mathcal{Y}} P(y|y', E(x; \theta_e); \theta_2) P(y'|E(x; \theta_e); \theta_1)},$$

According to Jensen's inequality: $f(E[x]) > E[f(x)]$, concavity

$$\tilde{\mathcal{J}}(x, y; \theta_e, \theta_1, \theta_2) = \sum_{y' \in \mathcal{Y}} P(y'|E(x; \theta_e); \theta_1) \log P(y|y', E(x; \theta_e); \theta_2)$$

$$\mathcal{J}(x, y; \theta_e, \theta_1, \theta_2) \geq \tilde{\mathcal{J}}(x, y; \theta_e, \theta_1, \theta_2)$$

Denote $\tilde{\mathcal{J}}(x, y; \theta_e, \theta_1, \theta_2)$ as $\tilde{J}$. The gradients of $\tilde{J}$ w.r.t its parameters are:

$$\nabla_{\theta_1} \tilde{\mathcal{J}} = \sum_{y' \in \mathcal{Y}} P(y'|E(x;\theta_e);\theta_1) \underbrace{\log P(y|y', E(x;\theta_e);\theta_2) \nabla_{\theta_1} \log P(y'|E(x;\theta_e);\theta_1)}_{G_1};$$

$$\nabla_{\theta_2} \tilde{\mathcal{J}} = \sum_{y' \in \mathcal{Y}} P(y'|E(x;\theta_e);\theta_1) \underbrace{\nabla_{\theta_2} \log P(y|y', E(x;\theta_e);\theta_2)}_{G_2};$$

$$\nabla_{\theta_e} \tilde{\mathcal{J}} = \sum_{y' \in \mathcal{Y}} P(y'|E(x;\theta_e);\theta_1) G_e(x, y, y'; \theta_e, \theta_1, \theta_2), \text{ where } G_e \text{ is defined as follows:}$$

$$G_e = \nabla_{\theta_e} \log P(y|y', E(x;\theta_e);\theta_2) + \log P(y|y', E(x;\theta_e);\theta_2) \nabla_{\theta_e} \log P(y'|E(x;\theta_e);\theta_1)$$

Application

NMT

Shallowd Models

setting:(1)encoder,decoders are all GRU with one hidden layer (2)deliberation's encoder and decoder are initialized by the pre-trained NMT model.

Table 1: BLEU scores of En→Fr translation

Algorithm	$\mathcal{M}_{\text{base}}$	$\mathcal{M}_{\text{dec}\times 2}$	$\mathcal{M}_{\text{reviewer}\times 4}$	$\mathcal{M}_{\text{delib}}$
BLEU	29.97	30.40	30.76	**31.67**

dataset: WMT'14

Table 2: BLEU scores of Zh→En translation

Algorithm	NIST04	NIST05	NIST06	NIST08
$\mathcal{M}_{\text{base}}$	34.96	34.57	32.74	26.21
$\mathcal{M}_{\text{delib}}$	36.90	35.57	33.90	27.13

training set: LDC

validation set: NIST2003

test set: NIST2004, NIST2005, NIST2006, NIST2008

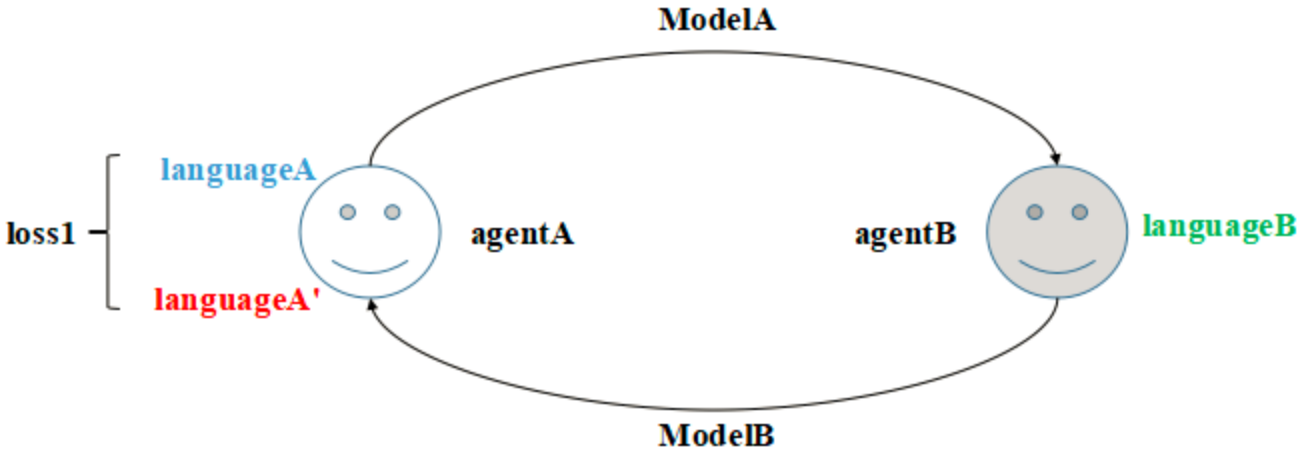
Deep Model

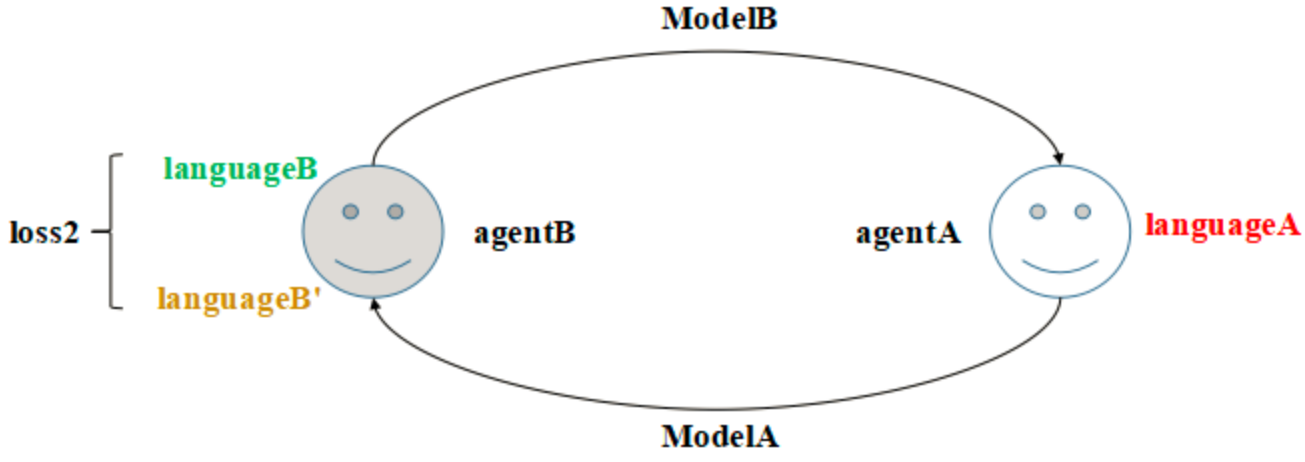
setting: encoder,decoders are all 4-layer LSTMs with residual connections.

Table 4: Comparison between deliberation network and different deep NMT systems (En→Fr).

System	Configurations	BLEU
GNMT [31]	Stacked LSTM (8-layer encoder + 8 layer decoder) + RL finetune	39.92
FairSeq [4]	Convolution (15-layer) encoder and (15-layer) decoder	40.51
Transformer [26]	Self-Attention + 6-layer encoder + 6-layer decoder	41.0
this work	Stack LSTM (4-layer encoder and 4-layer decoder)	39.51
	Stack 4-layer NMT + Dual Learning	40.53
	Stack 4-layer NMT + Dual Learning + Deliberation Network	**41.50**

Dual Learning





Instance Analysis

[Source] ***Aiji shuo, zhongdong heping*** *xieyi yuqi jiang you yige xinde jiagou .*
[Reference] ***Egypt says*** *a new framework is expected to come into being for the* ***Middle East*** *peace agreement .*
[Base] ***egypt 's middle east*** *peace agreement is expected to have a new framework , he said .*
[First-pass] ***egypt 's middle east*** *peace agreement is expected to have a new framework , egypt said .*
[Second-pass] ***egypt says the middle*** *east peace agreement is expected to have a new framework .*

[Source] *Nuowei dashiguan zhichu, "shuangfang jiang taolun ruhe gaijin luoshi tinghuo xieyi, zhe yeshi* ***san nian lai*** *shuangfang shouci* ***zai ruci gao de cengji shang jinxing mianduimian tanpan"***
[Reference] *The Norwegian embassy pointed out that , " Both sides will discuss how to improve the implementation of the cease-fire agreement , which is the first time for both sides to* ***have face-to-face negotiations at such a high level .*** *"*
[Base] *" ..... , which is the first time for the two countries to* ***conduct face-to-face talks on the basis of a high level of three years*** *, " it said .*
[First-pass] *" ..... , which is the first time for the two countries to* ***conduct face-to-face talks on the basis of a high level of three years*** *, " the norwegian embassy said in a statement .*
[Second-pass] *" ..... , which is the first time* ***in three years*** *for the two countries to* ***conduct face-to-face talks at such high level*** *, " the norwegian embassy said .*

Text Summarization

Table 5: ROUGE-{1, 2, L} scores of text summarization

Algorithm	ROUGE-1	ROUGE-2	ROUGE-L
$\mathcal{M}_{\text{base}}$	27.45	10.51	26.07
$\mathcal{M}_{\text{dec}\times 2}$	27.93	11.09	26.50
$\mathcal{M}_{\text{reviewer}\times 4}$	28.26	11.25	27.28
$\mathcal{M}_{\text{delib}}$	30.90	12.21	29.09

dataset: Gigaword Corpus

Conclusion

This work proposed deliberation networks with a second-pass decoder.