Content Preserving Text Generation with Attribute Controls

NIPS 2018

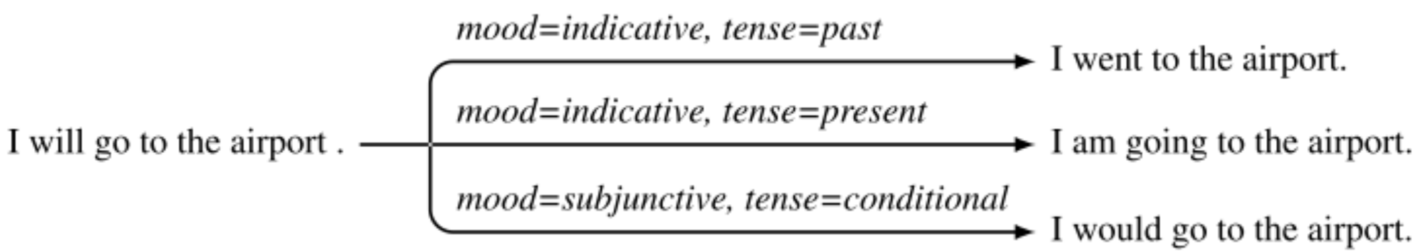
Lee et al.

발표자: 박정수

Data Mining & Information Systems Lab.
Department of Computer Science and Engineering,
College of Informatics, Korea University

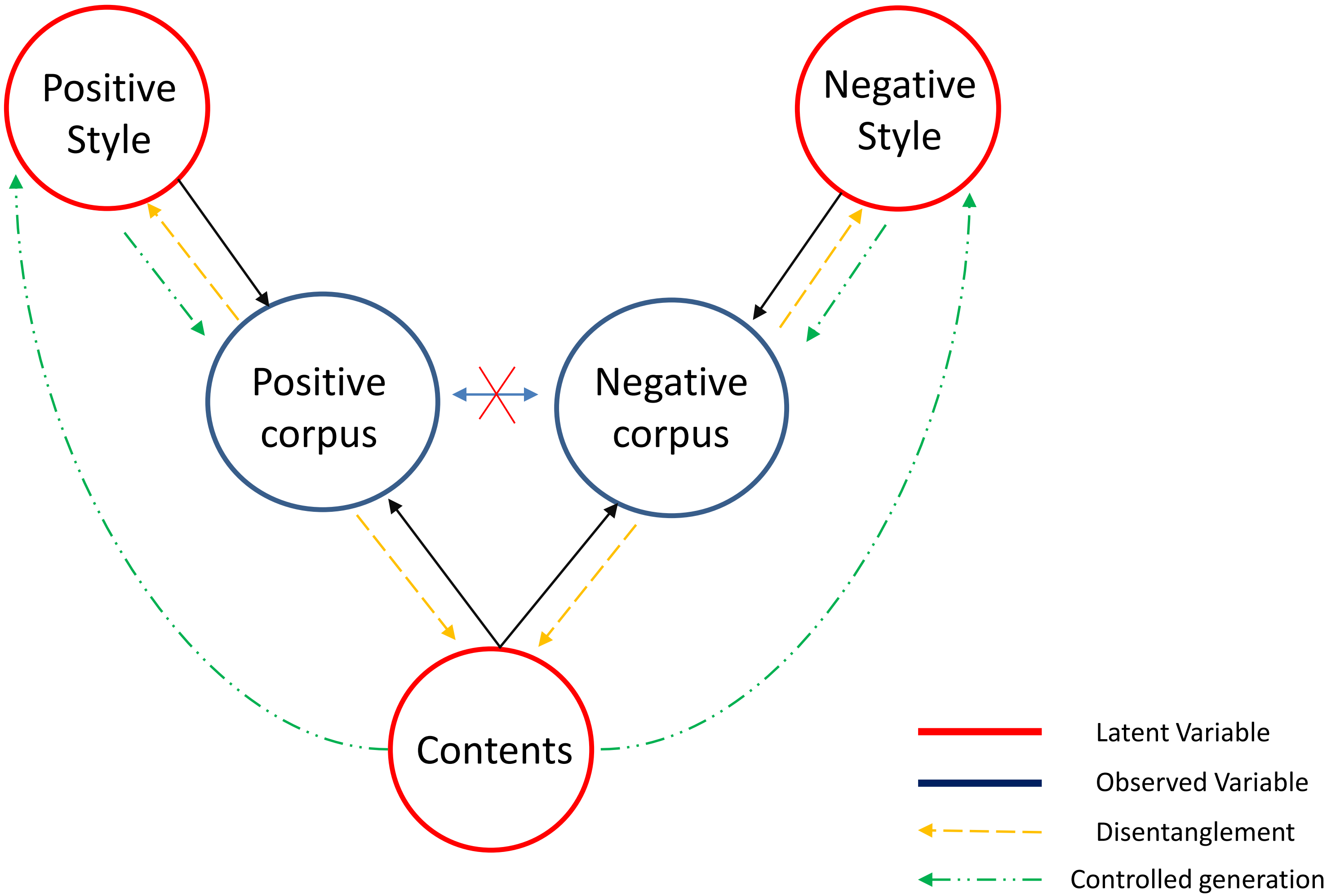
1 Introduction

How can we generate text in a controlled way?



1 Recap

Disentanglement

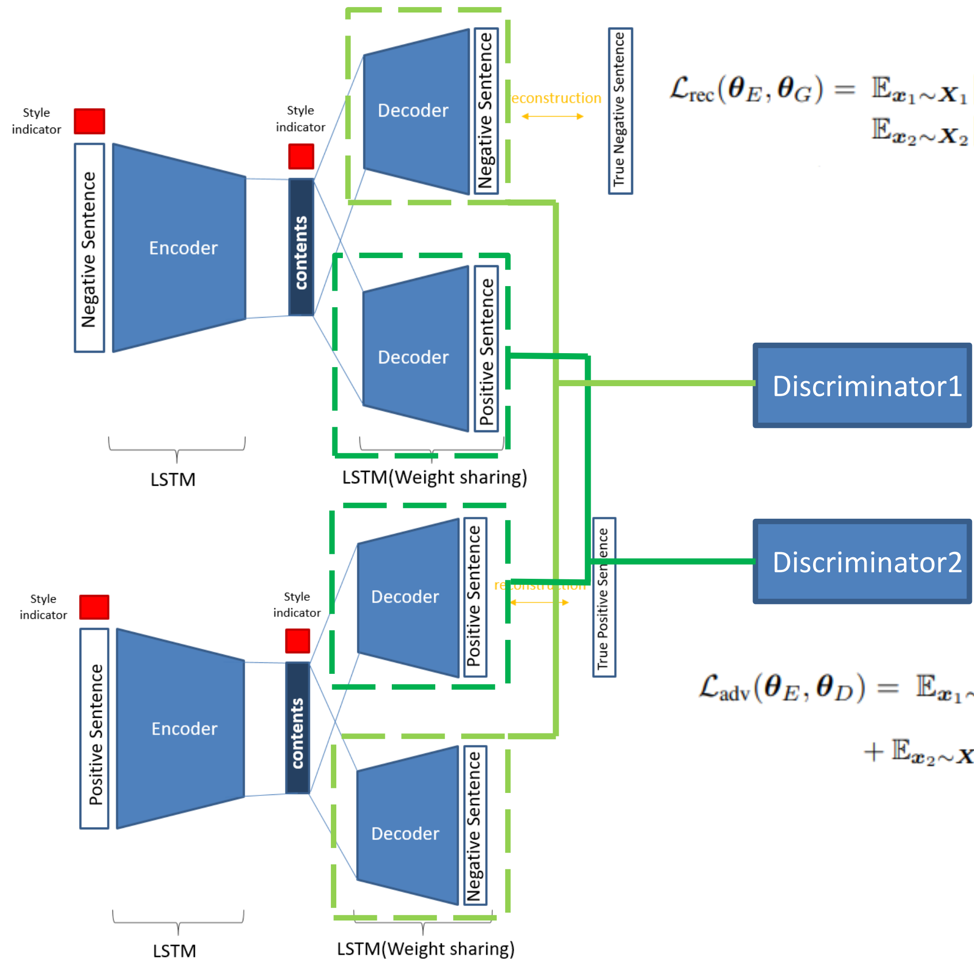


1 Recap

Cross Alignment (Shen et al.)

Style 1 -> Style 2

Style 2 -> Style 1



$$\mathcal{L}_{\text{rec}}(\boldsymbol{\theta}_E, \boldsymbol{\theta}_G) = \mathbb{E}_{\boldsymbol{x}_1 \sim \boldsymbol{X}_1}[-\log p_G(\boldsymbol{x}_1 | \boldsymbol{y}_1, E(\boldsymbol{x}_1, \boldsymbol{y}_1))] + \mathbb{E}_{\boldsymbol{x}_2 \sim \boldsymbol{X}_2}[-\log p_G(\boldsymbol{x}_2 | \boldsymbol{y}_2, E(\boldsymbol{x}_2, \boldsymbol{y}_2))]$$

$$\mathcal{L}_{\text{adv}}(\boldsymbol{\theta}_E, \boldsymbol{\theta}_D) = \mathbb{E}_{\boldsymbol{x}_1 \sim \boldsymbol{X}_1}[-\log D(E(\boldsymbol{x}_1, \boldsymbol{y}_1))] + \mathbb{E}_{\boldsymbol{x}_2 \sim \boldsymbol{X}_2}[-\log(1 - D(E(\boldsymbol{x}_2, \boldsymbol{y}_2)))]$$

2 Motivation

- Existing methods do not generalize to generation with multiple attribute controls
- Previous works have mostly focused on assessing the attribute compatibility of generated sentences

3 Method

Formulation

Attribute: Difference between given two corpora (i.e. tense, sentiment)

Content: Information in the corpora that is not captured by the attribute

$$D = \{(x^n, l^n)\}_{n=1}^{N}$$

Sentence with labeled attributes

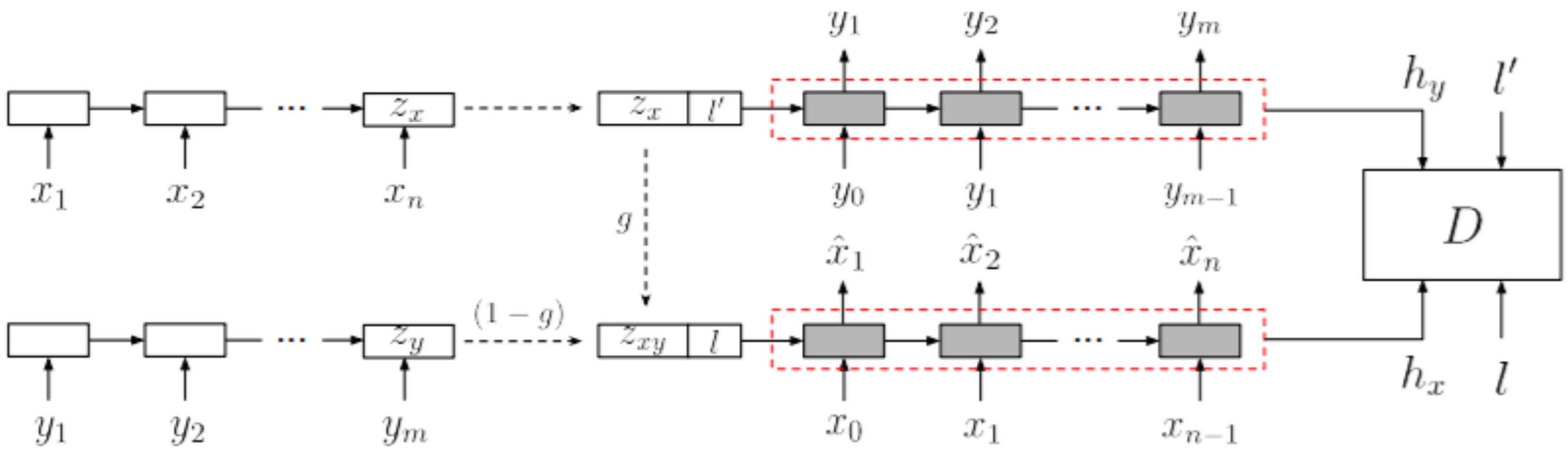
$$l' = (l_1, ..., l_K)$$

Define K attributes of interest

Tense Sentiment

3 Method

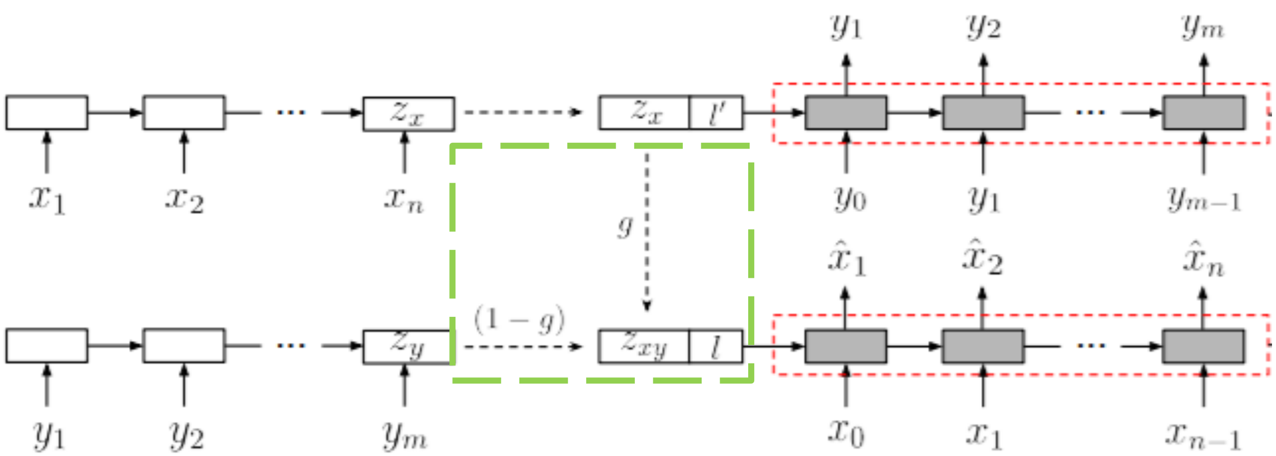
General Pipeline



- Generator consists of Encoder and Decoder which are RNN
- Reconstruction Loss for content compatibility and adversarial loss for attribute compatibility

3 Method

Content Compatibility



- Reconstruction Loss is an interpolation of auto-encoding loss and back-translation loss
- Auto-encoding loss has pitfall of incurring simple copying given an input sentence
- Back-translation loss can misguide the early stage learning since the contents of y and x would not match

3 Method

Content Compatibility

$y \sim p_G(\cdot|z_x, l')$ # y is l' attributed sentence, Encoder yields z

$\mathcal{L}^{ae}(x,l) = -\log p_G(x|z_x,l)$ # reconstructing x back to it's original

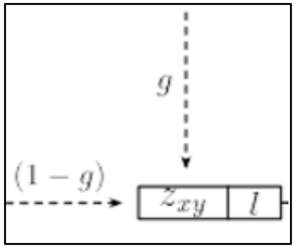
$\mathcal{L}^{bt}(x,l) = -\log p_G(x|z_y,l)$ # reconstructing y(different attribute) to x

$\Downarrow$

$\mathcal{L}^{int} = \mathbb{E}_{(x,l)\sim p_{\text{data}}, y\sim p_G(\cdot|z_x,l')}[-\log p_G(x|z_{xy},l)]$ # proposed loss

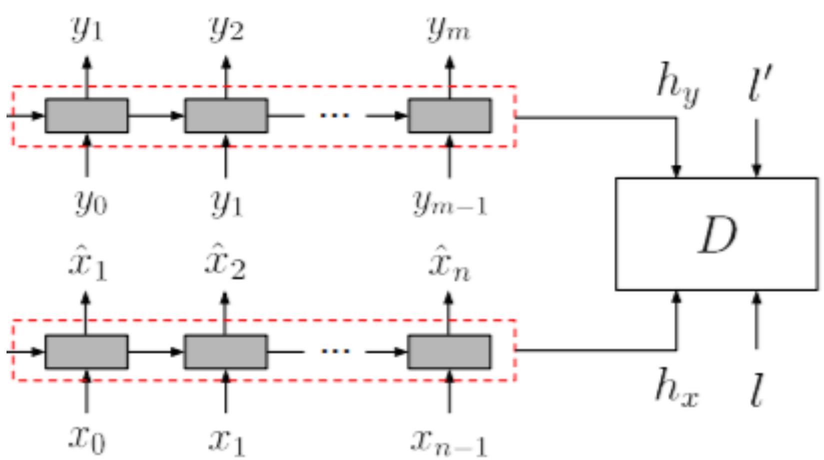
$z_{xy} = g \odot z_x + (1-g) \odot z_y$

Implicitly enforce the z to be attribute-independent



3 Method

Attribute Compatibility



- Adversarial loss encourages generating realistic and attribute compatible sentences

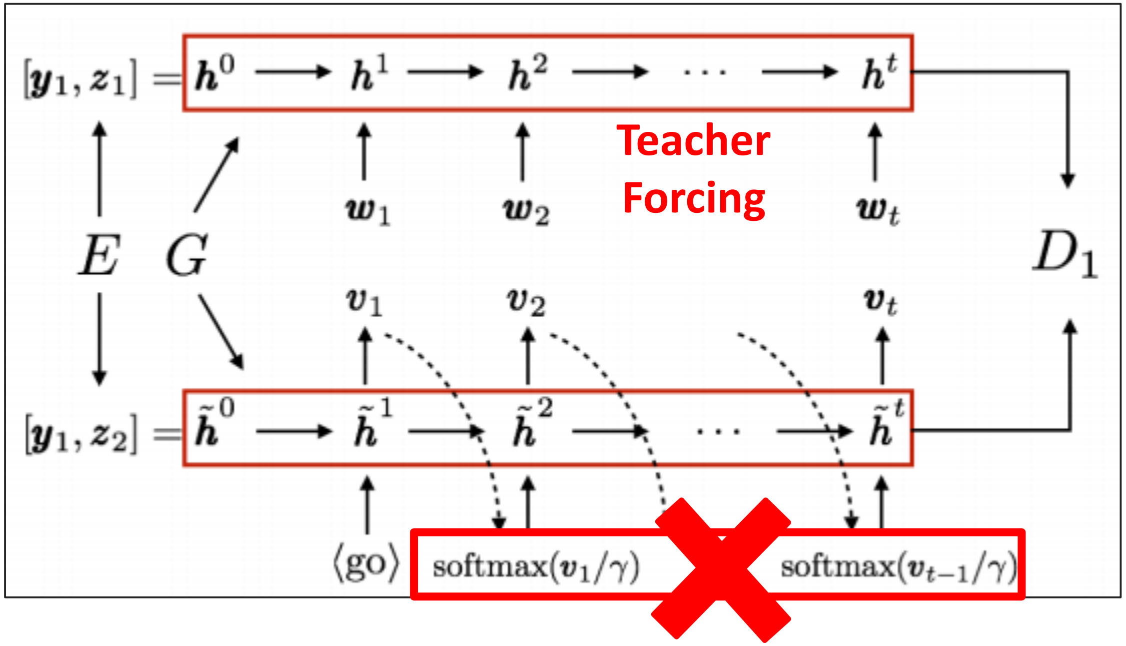
$$\mathcal{L}^{\text{adv}} = \min_G \max_D \mathbb{E}_{(x,l)\sim p_{\text{data}}, y\sim p_G(\cdot|z_x,l')}[\log D(h_x,l) + \log(1 - D(h_y,l'))]$$

$$\mathcal{L}^{\text{adv}} = \min_G \max_D \mathbb{E}_{\substack{(x,l)\sim p_{\text{data}} \\ y\sim p_G(\cdot|z_x,l')}}[2\log D(h_x,l) + \log(1 - D(h_y,l')) + \log(1 - D(h_x,l'))]]$$

$$D(s,l) = \sigma(l_v^T W\phi(s) + v^T\phi(s))$$

3 Method

Soft-Sampling vs Hard-Sampling



- Soft sampling(i.e. gumbel-softmax) is used to back-propagate gradients through sampling process
- Inference performs hard sampling, thus when soft sampling is used at training, gap exists between the dynamics of sequences at training time and sequences hard-sampled at test time

3 Experiment

Evaluation

	Yelp Reviews				IMDB Reviews			
Model	Attribute ↑ Accuracy	Content ↑ B-1	B-4	Fluency ↓ Perp.	Attribute ↑ Accuracy	Content ↑ B-1	B-4	Fluency ↓ Perp.
Ctrl-gen [18]	76.36%	11.5	0.0	156	76.99%	15.4	0.1	94
Cross-align [22]	90.09%	41.9	3.9	180	88.68%	31.1	1.1	63
Ours	**90.50%**	**53.0**	**7.5**	**133**	**94.46%**	**40.3**	**2.2**	**52**

$$f_{\text{content}}(M, M') = 0.5[\mathbb{E}_{x\sim D_{\text{src}}}\text{BLEU}(x, M' \circ M(x)) + \mathbb{E}_{x\sim D_{\text{tgt}}}\text{BLEU}(x, M \circ M'(x))]$$

Restaurant reviews	
negative → positive	
Query	*the people behind the counter were not friendly whatsoever .*
Ctrl gen [18]	the food did n't taste as fresh as it could have been either .
Cross-align [22]	the owners are the staff is so friendly .
Ours	the people at the counter were very friendly and helpful .
positive → negative	
Query	*they do an exceptional job here , the entire staff is professional and accommodating !*
Ctrl gen [18]	very little water just boring ruined !
Cross-align [22]	they do not be back here , the service is so rude and do n't care !
Ours	they do not care about customer service , the staff is rude and unprofessional !

4 Experiment

Evaluation

Mood	Tense	Voice	Neg.	john was born in the camp
Indicative	Past	Passive	No	john was born in the camp .
Indicative	Past	Passive	Yes	john wasn't born in the camp .
Indicative	Past	Active	No	john had lived in the camp .
Indicative	Past	Active	Yes	john didn't live in the camp .
Indicative	Present	Passive	No	john is born in the camp .
Indicative	Present	Passive	Yes	john isn't born in the camp .
Indicative	Present	Active	No	john has lived in the camp .
Indicative	Present	Active	Yes	john doesn't live in the camp .
Indicative	Future	Passive	No	john will be born in the camp .
Indicative	Future	Passive	Yes	john will not be born in the camp .
Indicative	Future	Active	No	john will live in the camp .
Indicative	Future	Active	Yes	john will not survive in the camp .
Subjunctive	Cond	Passive	No	john could be born in the camp .
Subjunctive	Cond	Passive	Yes	john couldn't live in the camp .
Subjunctive	Cond	Active	No	john could live in the camp .
Subjunctive	Cond	Active	Yes	john couldn't live in the camp .

Table 5: Simultaneous control of multiple attributes. Generated sentences for all valid combinations of the input attribute values.