Stylized Text Generation

Lili Mou [a]   Olga Vechtomova [w]

[a]University of Alberta
Alberta Machine Intelligence Institute (Amii)
doublepower.mou@gmail.com

[w]University of Waterloo
ovechtomova@uwaterloo.ca

ACL 2020 Tutorial

Lili Mou is admitting

- All-level students
  MSc, PhD, exchanging

- Visiting scholars
  RA, Postdoc

Tutorial Outline

- Introduction
- Style-conditioned text generation
- **Style-transfer text generation**
  - **Parallel supervised**
  - **Non-parallel supervised**
  - **Unsupervised**
- Style-adversarial text generation
- Conclusion

Roadmap of this part

Style-transfer text generation

- Task formulation
- Settings
- Approach overview
- Evaluation

- Detailed discussion on existing work

amii
UNIVERSITY OF ALBERTA
UNIVERSITY OF WATERLOO

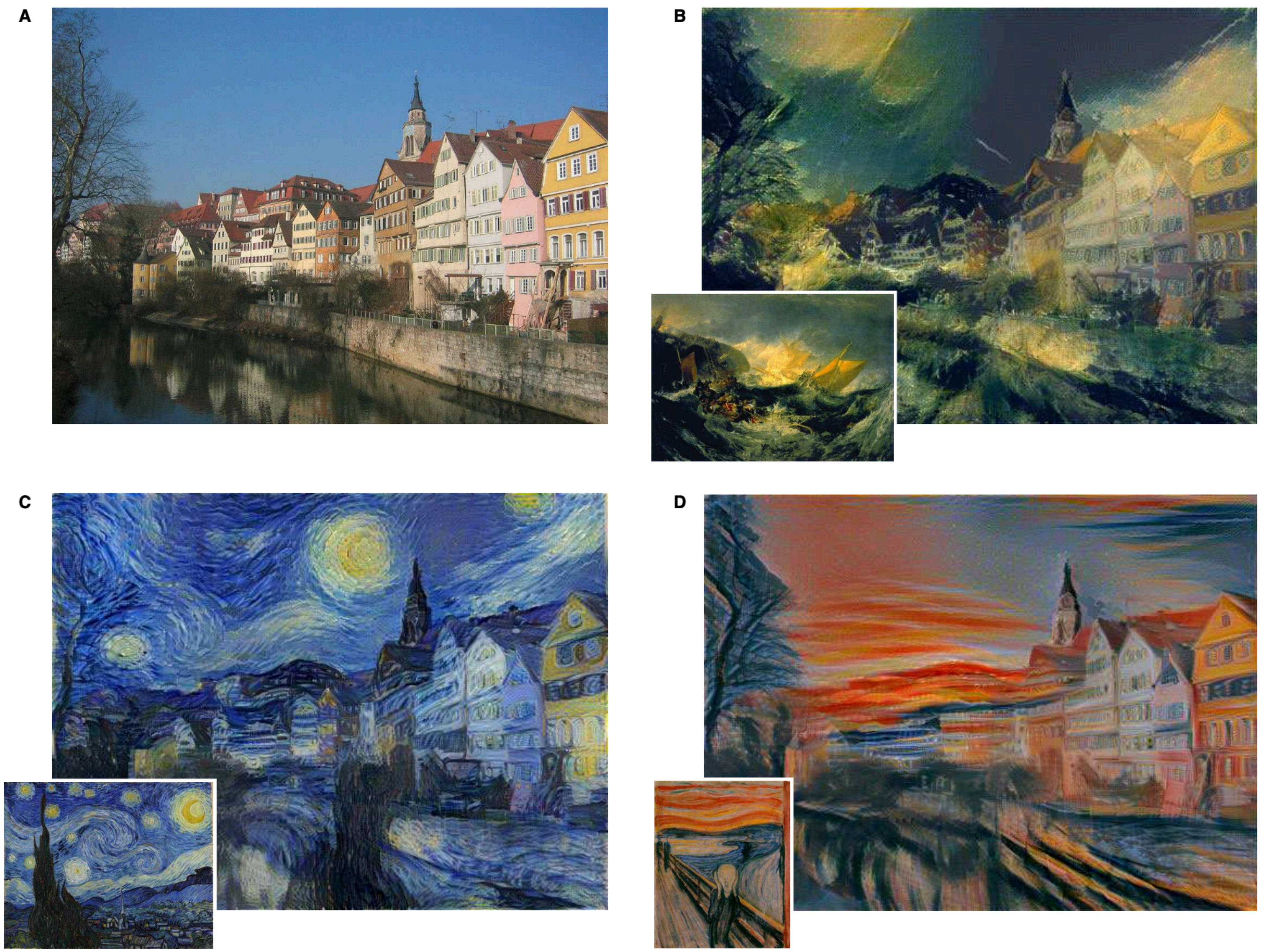
Style-Transfer Generation

Task description

- Input
  - A source sentence $\mathbf{x} = \mathrm{x}_1\mathrm{x}_2\cdots\mathrm{x}_n$
  - The desired style
- Output: A "style-transferred" sentence $\mathbf{y} = \mathrm{y}_1\mathrm{y}_2\cdots\mathrm{y}_m$
- Requirement: $\mathbf{y}$ is in the desired style
  - Usually, $\mathbf{x}$ and $\mathbf{y}$ are **different in "style"**
  - $\mathbf{x}$ and $\mathbf{y}$ share **the same "content"**

Style-Transfer in Computer Vision

Artistic Style Transfer [Gatys+16]



Style-Transfer Tasks in NLP

Sentiment transfer

- Yelp review [Hu+2017]
- Amazon review [Fu+2017]

Input	**Output**
the film is strictly routine !	the film is full of imagination .
after watching this movie , i felt that disappointed .	after seeing this film , i ’m a fan .
the acting is uniformly bad either .	the performances are uniformly good .
this is just awful .	this is pure genius .

Style-Transfer Tasks in NLP

Formality style transfer

- Grammarly's Yahoo Answers Formality Corpus (GYAFC) [Rao&Tetreault, 2018]

Input	**Output**
Wow , I am very dumb in my observation skills	I do not have good observation skills .
i hardly everrr see him in school either usually i see hima t my brothers basketball games .	I hardly ever see him in school . I usually see him with my brothers playing basketball .

Style-Transfer Tasks in NLP

Shakespeare Style Transfer [Xu+2012]

Input	**Output**
I can read my own fortune in my misery.	i can read mine own fortune in my woes .
Good bye, Mr. Anderson.	fare you well , good master anderson .

What is “style” or “content”?

Linguistic Perspective

Defining characteristic	Register	Genre	Style
Textual focus	sample of text excerpts	complete texts	sample of text excerpts
Linguistic characteristics	any lexico-grammatical feature	specialized expressions, rhetorical organization, formatting	any lexico-grammatical feature
Distribution of linguistic characteristics	frequent and pervasive in texts from the variety	usually once-occurring in the text, in a particular place in the text	frequent and pervasive in texts from the variety
Interpretation	features serve important communicative functions in the register	features are conventionally associated with the genre: the expected format, but often not functional	features are not directly functional; they are preferred because they are aesthetically valued

Biber, D., Conrad, S., *Register, Genre, and Style*. Cambridge University Press, 2009

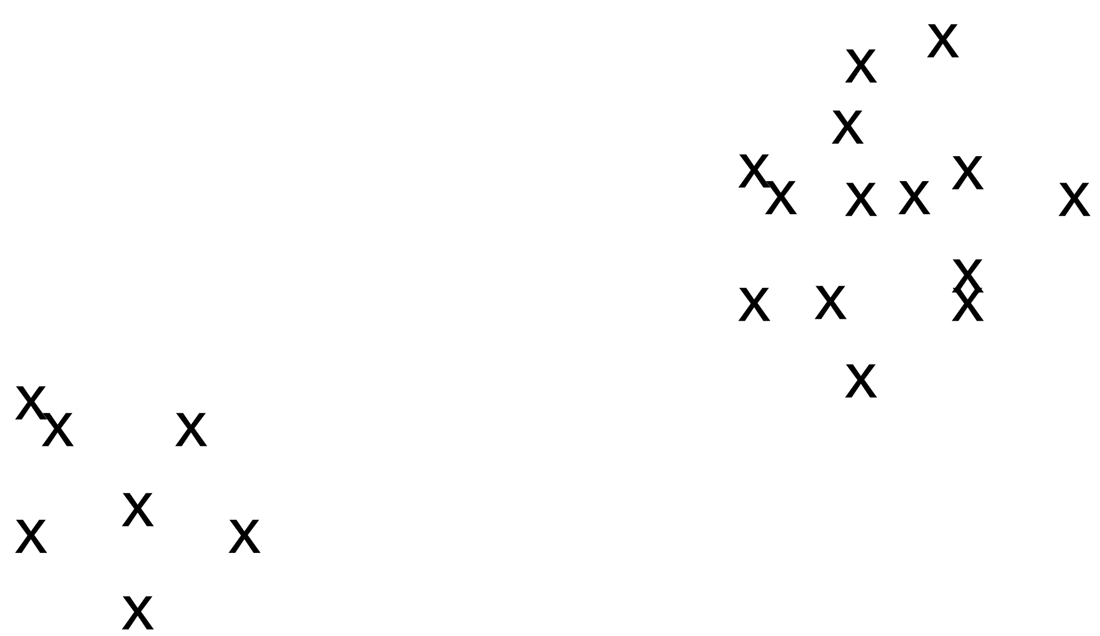
What is “style” or “content”?

More debates

Is “sentiment information” the style or content?

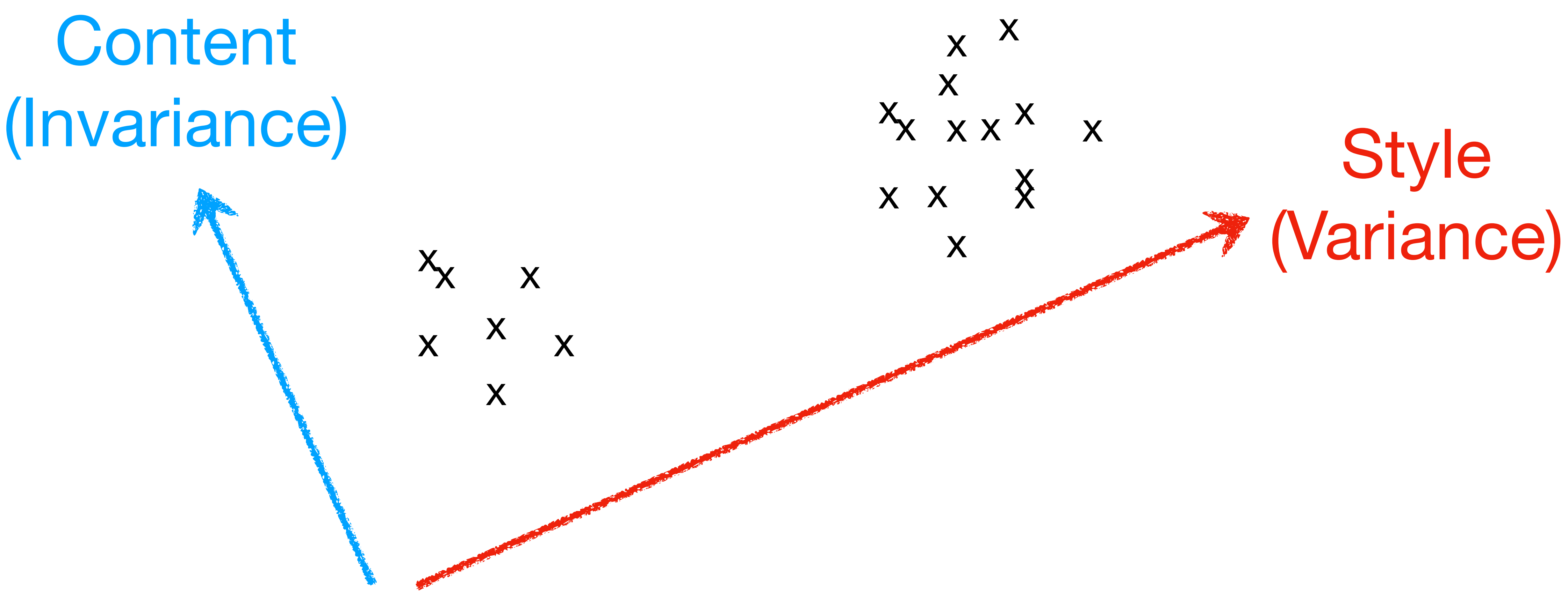
What is “style” or “content”?

An empirical perspective



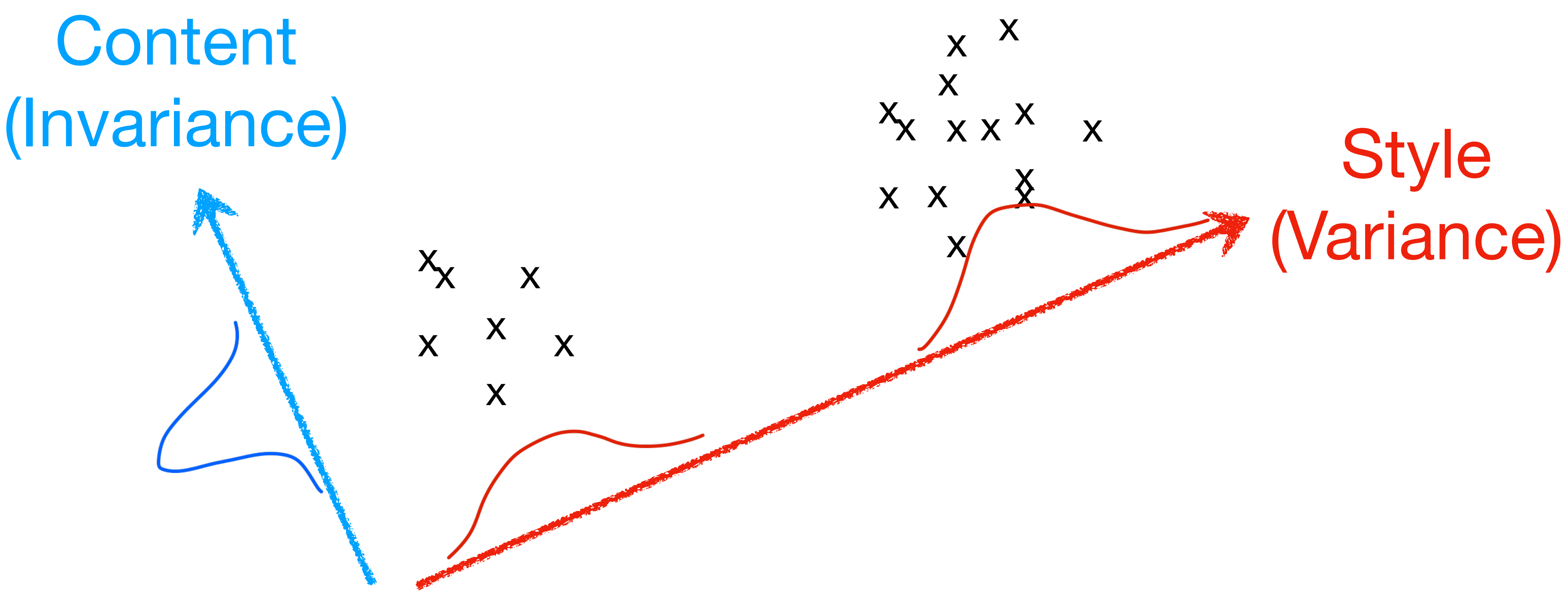
What is “style” or “content”?

An empirical perspective



What is “style” or “content”?

An empirical perspective



Style-Transfer Tasks in NLP

“Content” transfer [Zhao+2018]

- Trained on the Yahoo QA dataset
- Variance = Content, topic
- Invariance = Question words, question structure

Science	what is an event horizon with regards to black holes ?
$\Rightarrow$ Music	what is your favorite sitcom with adam sandler ?
$\Rightarrow$ Politics	what is an event with black people ?
Science	take 1ml of hcl (concentrated) and dilute it to 50ml .
$\Rightarrow$ Music	take em to you and shout it to me
$\Rightarrow$ Politics	take bribes to islam and it will be punished .
Science	just multiply the numerator of one fraction by that of the other .
$\Rightarrow$ Music	just multiply the fraction of the other one that 's just like it .
$\Rightarrow$ Politics	just multiply the same fraction of other countries .

Style-Transfer Tasks in NLP

In summary

- Style-transfer is a **well-defined** task
  - from a data perspective
- Goal is to
  - **Preserve the invariance**
  - **Change the variance**
- In this tutorial, we call
  - Variance = style
  - Invariance = content

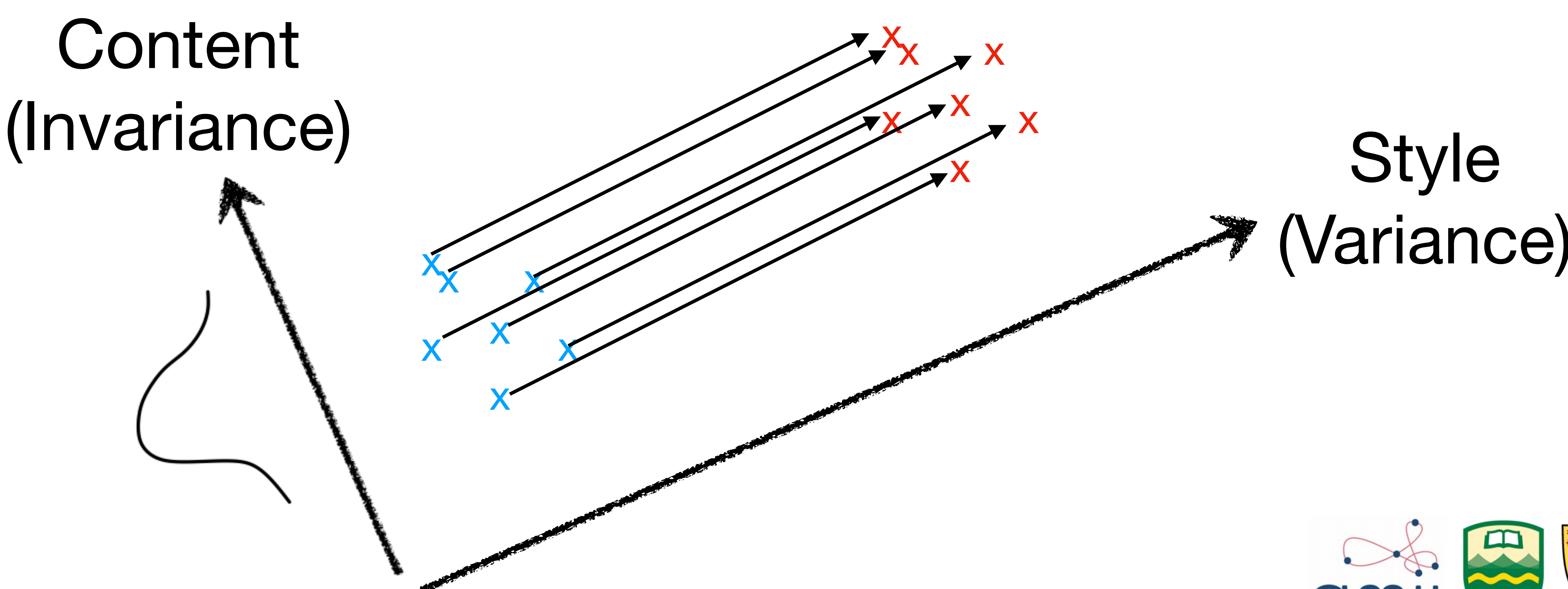
Settings

- Seq2seq supervision
- Non-parallel supervision
- Unsupervised

Settings

- **Parallel supervision**
  - In the training phase, we have parallel corpus of

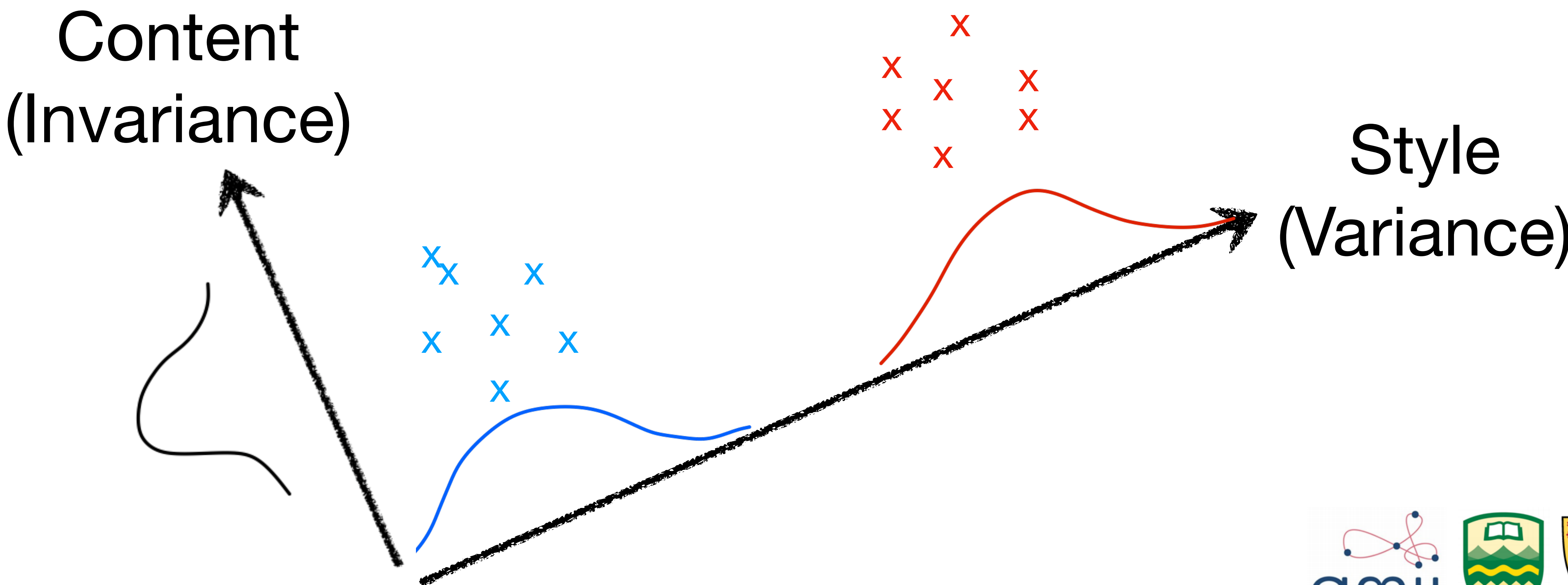
$$\{\mathbf{x}^{(m)}, \mathbf{y}^{(m)}, s^{(m)}\}_{m=1}^{M}$$



Settings

- **Non-parallel supervision**
  - In the training phase, we have non-parallel, style-labeled corpus

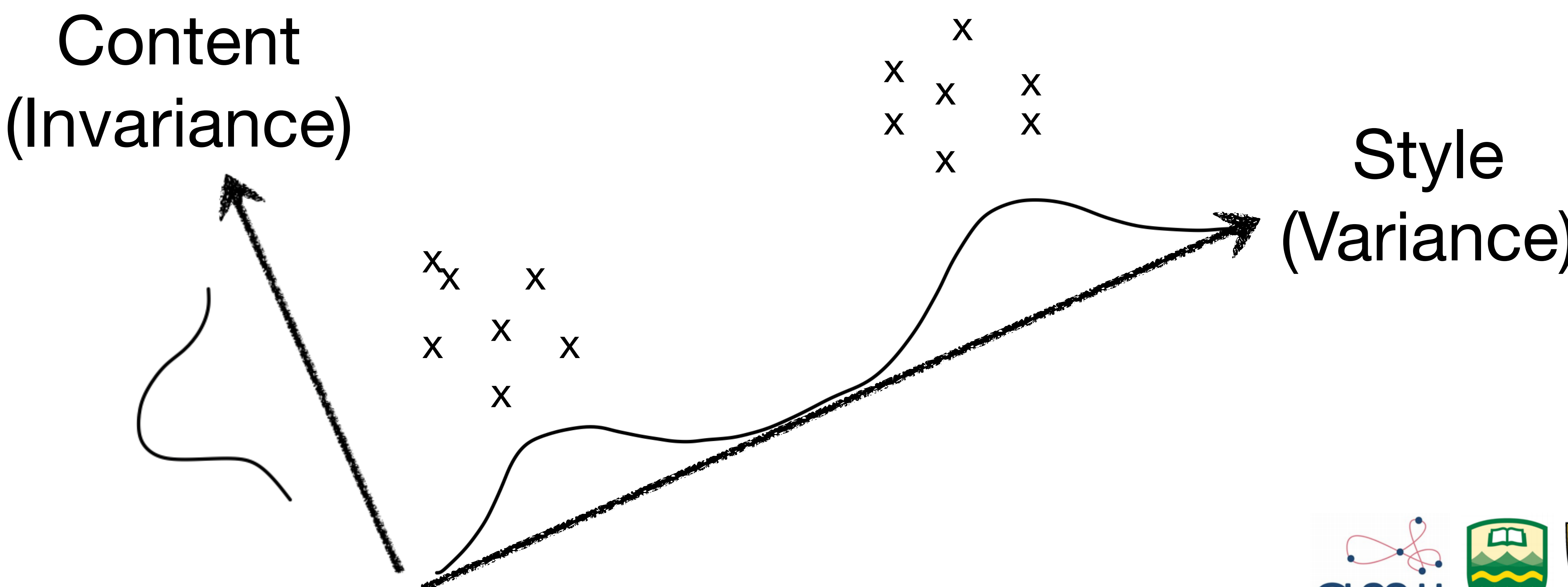
$$\{\mathbf{x}^{(m)}, s^{(m)}\}_{m=1}^{M}$$



Settings

- **Purely unsupervised**
  - In the training phase, we have unlabeled corpus

$$\{\mathbf{x}^{(m)}\}_{m=1}^{M}$$



Settings

- **Multi-attribute style transfer**

	Sentiment		Gender		Category				
SYelp	Positive 266,041	Negative 177,218	Male -	Female -	American -	Asian -	Bar -	Dessert -	Mexican -
FYelp	Positive 2,056,132	Negative 639,272	Male 1,218,068	Female 1,477,336	American 904,026	Asian 518,370	Bar 595,681	Dessert 431,225	Mexican 246,102
Amazon	Positive 64,251,073	Negative 10,944,310	- -	- -	Book 26,208,872	Clothing 14,192,554	Electronics 25,894,877	Movies 4,324,913	Music 4,574,167
Social Media	Relaxed 7,682,688	Annoyed 17,823,468	Male 14,501,958	Female 18,463,789	18-24 12,628,250	65+ 7,629,505			

Subramanian, S., Lample, G., Smith, E.M., Denoyer, L., Ranzato, M.A. and Boureau, Y.L., 2018. Multiple-attribute text style transfer. In *ICLR*, 2018.

Approach Overview

- **Parallel supervision**
  - Translation-inspired models
    - Phrase-based
    - Neural Seq2Seq
  - Difficulties: small training data
    - Regularization
    - Semi-supervised learning
- Non-parallel supervision
- Unsupervised

Approach Overview

- **Parallel supervision**
- **Non-parallel supervision**
  - Content preserving
    - Adversarial loss, Back-translation
  - Style transferring
    - Style words, style features, style-specific decoder
- Unsupervised

Approach Overview

- **Parallel supervision**
- **Non-parallel supervision**
- **Unsupervised**
  - Disentangling features
  - Pinpointing style-specific features

Automatic Evaluation

- Reference available
  - BLEU, ROUGE, etc.
- Reference unavailable
  - Style-transfer performance
    - Accuracy of a third-party style classifier
  - Content-preservation performance
    - Cosine similarity, word-overlapping rate, self-BLEU
- Auxiliary metric
  - Fluency

Human Evaluation

- Pairwise annotation
  - E.g., Win, Lose, Tie
- Pointwise annotation
  - E.g., 1—5 scale
- Annotation criteria
  - Overall quality
  - Individual aspect
    - Transfer accuracy
    - Content preserving
    - Fluency

Parallel Supervision for Style-Transfer Generation

Shakespeare ⇒ Modern English

		Modern English	Shakespeare
The Matrix	Agent Smith	Good bye, Mr. Anderson.	fare you well , good master anderson .
The Matrix	Morpheus	I'm trying to free your mind, Neo. But I can only show you the door. You're the one that has to walk through it.	i 'll to free your mind , neo. but i can but show you the door. you 're the one that hath to tread it .
Raiders of the Lost Ark	Belloq	Good afternoon, Dr. Jones.	well met , dr. jones .
Raiders of the Lost Ark	Jones	I ought to kill you right now.	i should kill thee straight .

Xu, W., Ritter, A., Dolan, B., Grishman, R. and Cherry, C. Paraphrasing for style. In *COLING*, 2012.

Dataset Collection

	corpus	initial size	aligned size	No-Change BLEU
Modern	http://nfs.sparknotes.com	31,718	21,079	24.67
Early modern	http://enotes.com	13,640	10,365	52.30

Note: BLEU reflects style similarity if content is given

Xu, W., Ritter, A., Dolan, B., Grishman, R. and Cherry, C. Paraphrasing for style. In *COLING*, 2012.

Approaches

- **Phrase-based machine translation (PBMT)**
  - Word alignment: GIZA++ (Och and Ney, 2003)
  - Decoding: Moses (Koehn et al., 2007)
- **PBMT + External Dictionary**
  - 68,709 phrase/word pairs from http://www. shakespeareswords.com
  - Phrase translation probabilities = frequencies of the translation words/phrases in the target language
  - Put it to PBMT
- **PBMT + Ouf-of-domain monolingual corpus**

Xu, W., Ritter, A., Dolan, B., Grishman, R. and Cherry, C. Paraphrasing for style. In *COLING*, 2012.

Formality Style Transfer

Formal $\Longleftrightarrow$ Informal

Informal:	*I'd say it is punk though.*
Formal:	*However, I do believe it to be punk.*
Informal:	*Gotta see both sides of the story.*
Formal:	*You have to consider both sides of the story.*

Dataset construction

- Yahoo answers (Entertainment & Music **and** Family & Relationships)
- Manual rating (Informal vs Formal)
- Manual rewriting (Informal → Formal)

		Informal to Formal		*Formal to Informal*	
	Train	Tune	Test	Tune	Test
E&M	52,595	2,877	1,416	2,356	1,082
F&R	51,967	2,788	1,332	2,247	1,019

Rao, S., Tetreault, J. Dear Sir or Madam, May I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer. In *NAACL-HLT*, 2018.

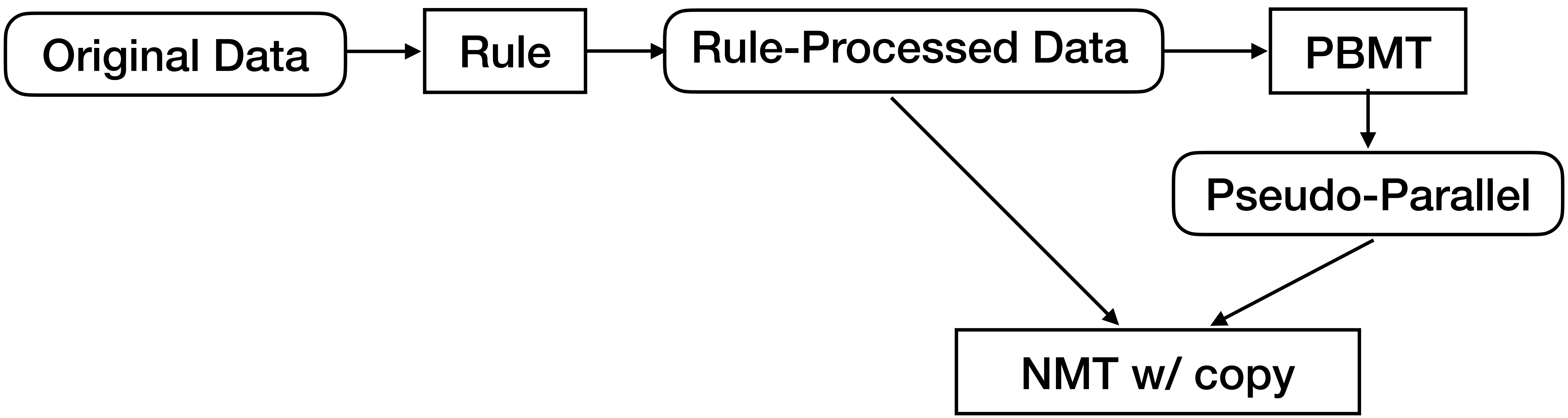
Approaches

- Rule-based
  - E.g., capitalization, punctuations, spelling
- PBMT, NMT (w/ and w/o copy)
- Generating pseudo-parallel corpora
  - Train PBMT, and use it to generate
  - Source $\Rightarrow \widehat{\text{Target}}$
  - Target $\Rightarrow \widehat{\text{Source}}$

Rao, S., Tetreault, J. Dear Sir or Madam, May I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer. In *NAACL-HLT*, 2018.

Results

Model	Formality		Fluency		Meaning		Combined		Overall		
	Human	PT16	Human	H14	Human	HE15	Human	Auto	BLEU	TERp	PINC
Original Informal	*-1.23*	*-1.00*	*3.90*	*2.89*	–	–	–	–	*50.69*	*0.35*	*0.00*
Formal Reference	0.38	0.17	4.45	3.32	4.57	3.64	5.68	4.67	100.0	0.37	69.79
Rule-based	-0.59	-0.34	4.00	3.09	**4.85**	**4.41**	5.24	4.69	61.38	0.27	26.05
PBMT	-0.19*	0.00*	3.96	3.28*	4.64*	4.19*	5.27	4.82*	67.26*	**0.26**	44.94*
NMT Baseline	**0.05***	0.07*	4.05	**3.52***	3.55*	3.89*	4.96*	**4.84***	56.61	0.38*	**56.92***
NMT Copy	0.02*	**0.10***	4.07	3.45*	3.48*	3.87*	4.93*	4.81*	58.01	0.38*	56.39*
NMT Combined	-0.16*	0.00*	**4.09***	3.27*	4.46*	4.20*	**5.32***	4.82*	**67.67***	**0.26**	43.54*



Rao, S., Tetreault, J. Dear Sir or Madam, May I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer. In *NAACL-HLT*, 2018.

Better Using Rules

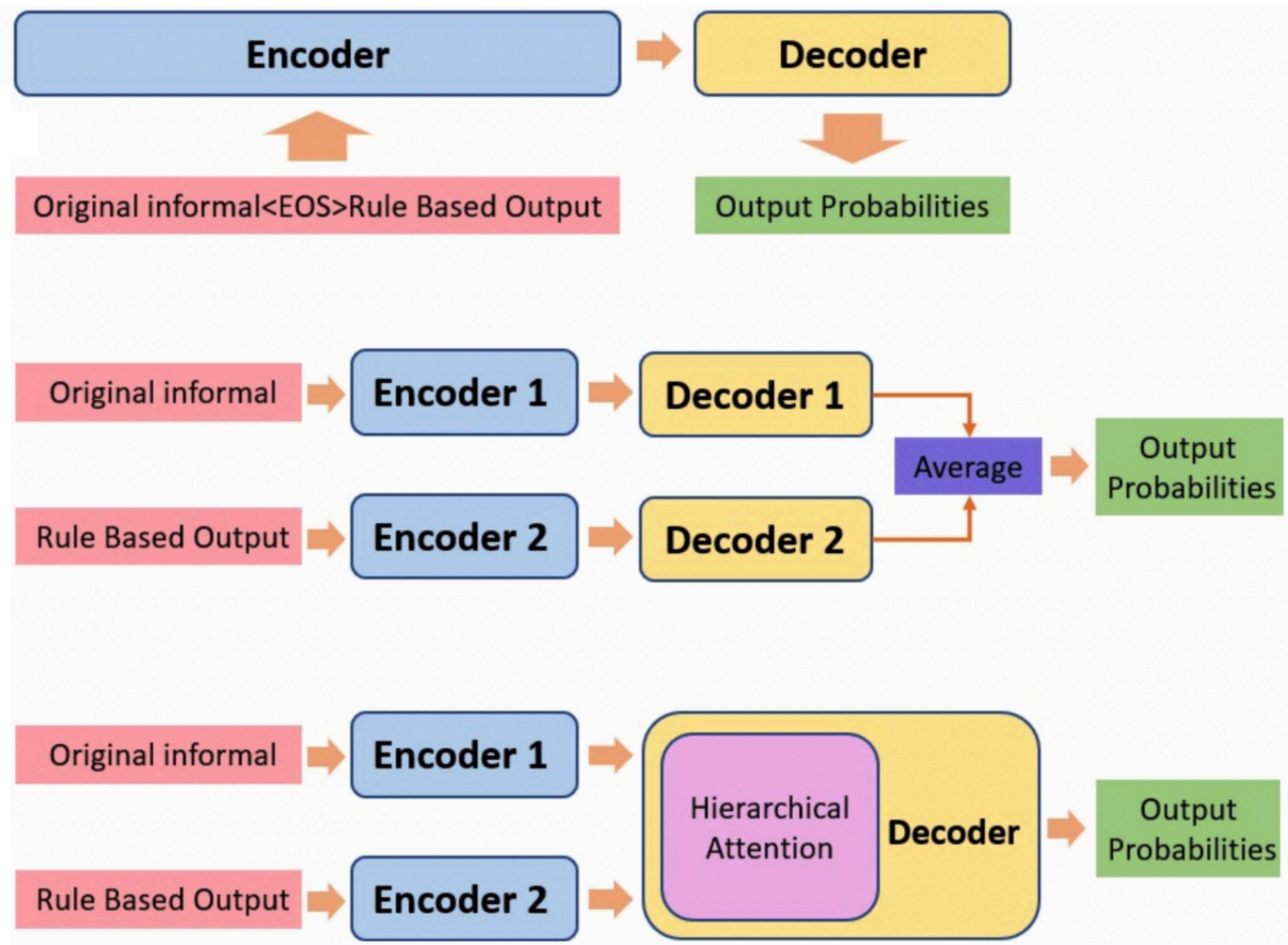
- Observations
  - Rule-processed data are the Markov blanket
  - Some entities (esp. not proper nouns) may be recognized incorrectly

Attempt#1: Input concatenation

(works the best in experiments)

Attempt#2: Decoder ensemble

Attempt#3: Hierarchal attention



Wang, Y., Wu, Y., Mou, L., Li, Z. and Chao, W. Harnessing Pre-Trained Neural Networks with Rules for Formality Style Transfer. In *EMNLP-IJCNLP*. 2019.

Summary for
Parallel-Supervision Style Transfer

- Seq2Seq-style training works
- Difficulties: data sparseness
  - Dictionaries
  - Rules
  - Data augmentation

Non-Parallel Supervision for Style-Transfer Generation

Hu et al. [2017]

- Movie Reviews
  - Positive vs. Negative

the film is strictly routine !
the film is full of imagination .

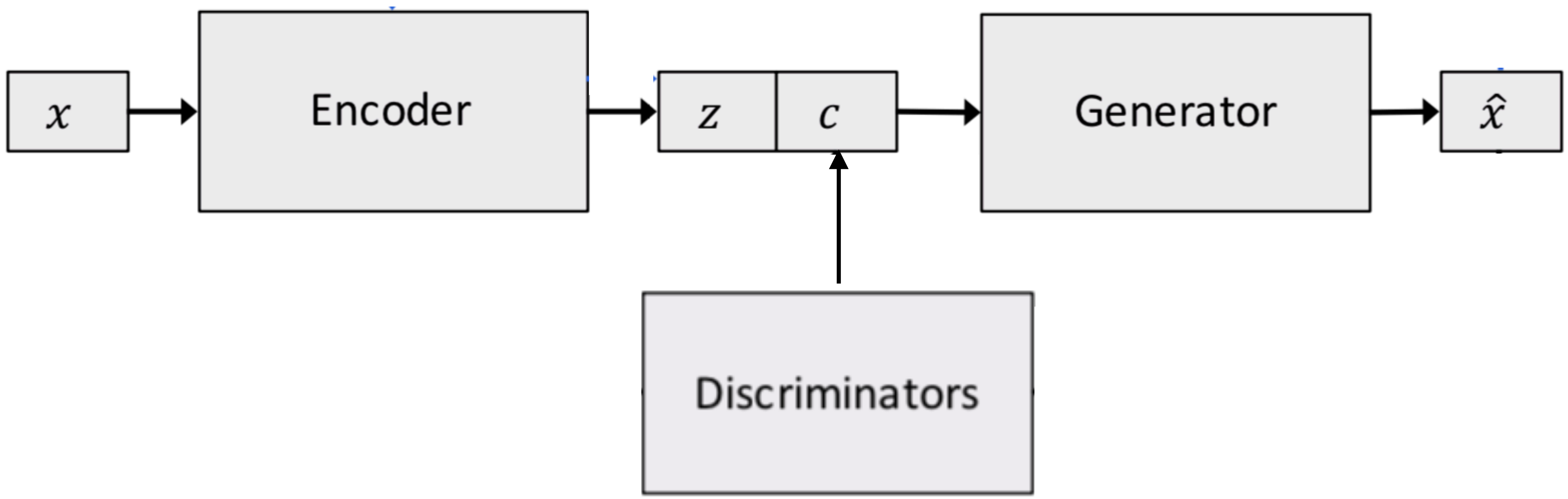
after watching this movie , i felt that disappointed .
after seeing this film , i 'm a fan .

the acting is uniformly bad either .
the performances are uniformly good .

this is just awful .
this is pure genius .

Hu, Z, Yang, Z, Liang, X, Salakhutdinov, R, Xing, EP.
Toward controlled generation of text. In *ICML*, 2017.

Hu et al. [2017]



- Variational auto-encoder with latent space
  - Structured latent space c [style code]
  - Unstructured latent space z [remaining info]
- Discriminator: classifying the style

Hu, Z, Yang, Z, Liang, X, Salakhutdinov, R, Xing, EP.
Toward controlled generation of text. In *ICML*, 2017.

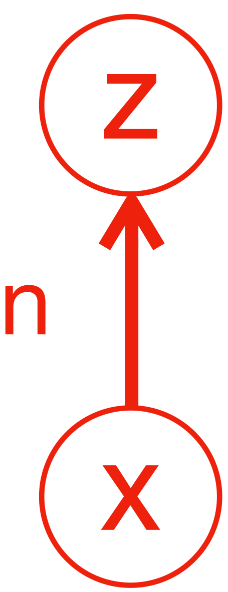
Variational Autoencoder

- Model

Recognition

$$p(x, z) = q(x)q(z \mid x)$$

Recognition



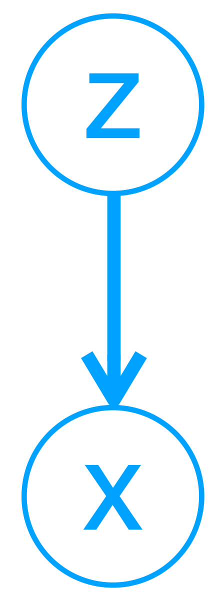
$$\phi$$

Generation/Reconstruction

$$p(x, z) = p(z)p(x \mid z)$$

$p(z)$

$p(x \mid z)$



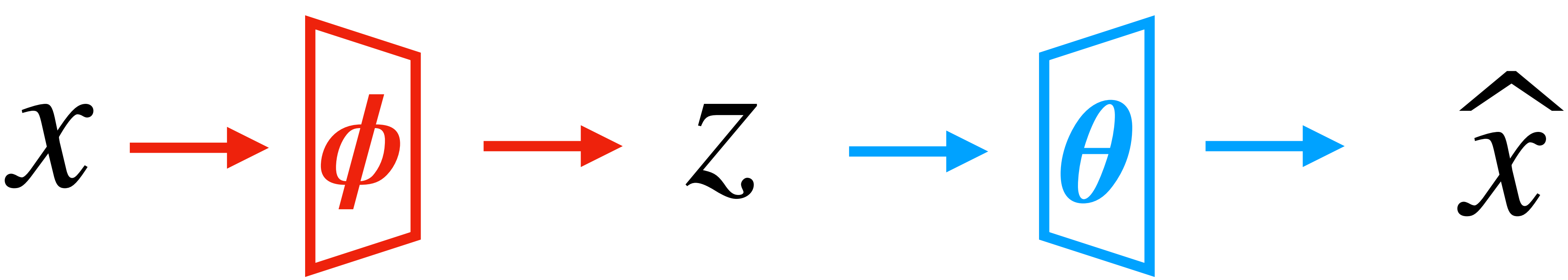
$$\theta$$

Kingma, D.P. and Welling, M., 2013. Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*.

Variational Autoencoder

- Training objective
  - Maximizing the lower bound of log-likelihood
  - Equivalent to expected reconstruction, penalized by a KL term

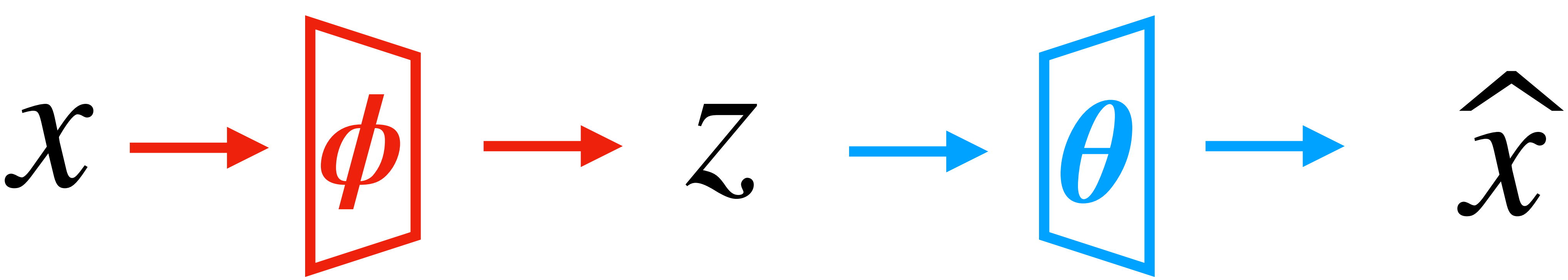
$$J = \mathbb{E}_{z \sim q_{\phi}(z|x)}[-\log p_{\theta}(x \mid z)] + \mathrm{KL}(q_{\phi}(z \mid x) \| p(z))$$



Variational Autoencoder

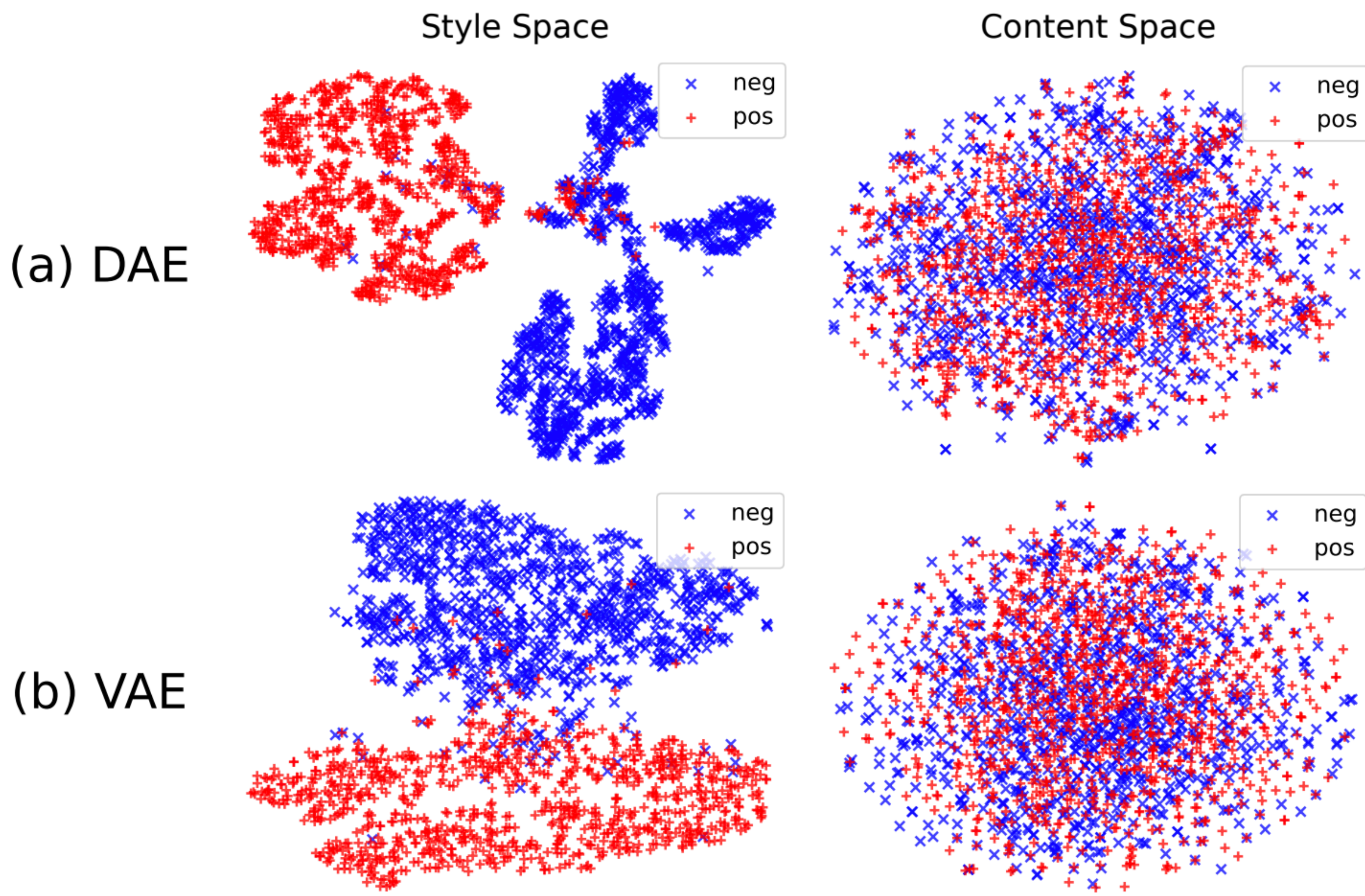
- Define a prior $p(z) = \mathcal{N}(0,1)$
- Define the posterior familiar $q(z\,|\,x) = \mathcal{N}(\mu, \operatorname{diag} \sigma^2)$
  - where μ and σ are predicted by the encoder (recognition)

$$J = \mathop{\mathbb{E}}_{z \sim q_{\boldsymbol{\phi}}(z|x)} \left[-\log p_{\boldsymbol{\theta}}(x\,|\,z)\right] + \mathrm{KL}(q_{\boldsymbol{\phi}}(z\,|\,x) \| p(z))$$



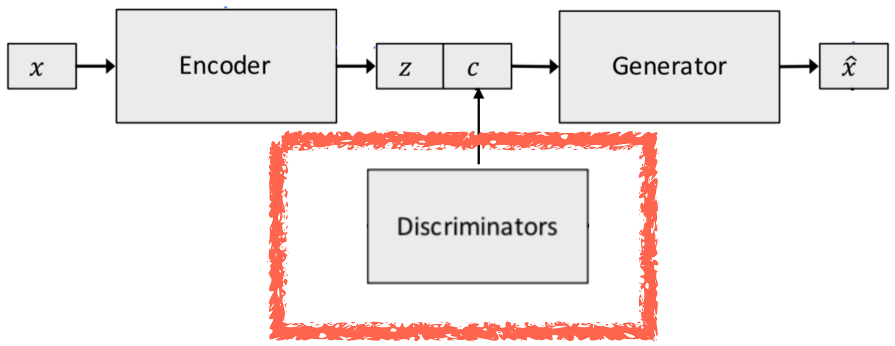
Variational Autoencoder

- VAE is widely used in style-transfer generation
  - Especially good for sampling and manipulation of z



John, V., Mou, L., Bahuleyan, H. and Vechtomova, O. Disentangled representation learning for text style transfer. In *ACL*, 2018.

Hu et al. [2017]

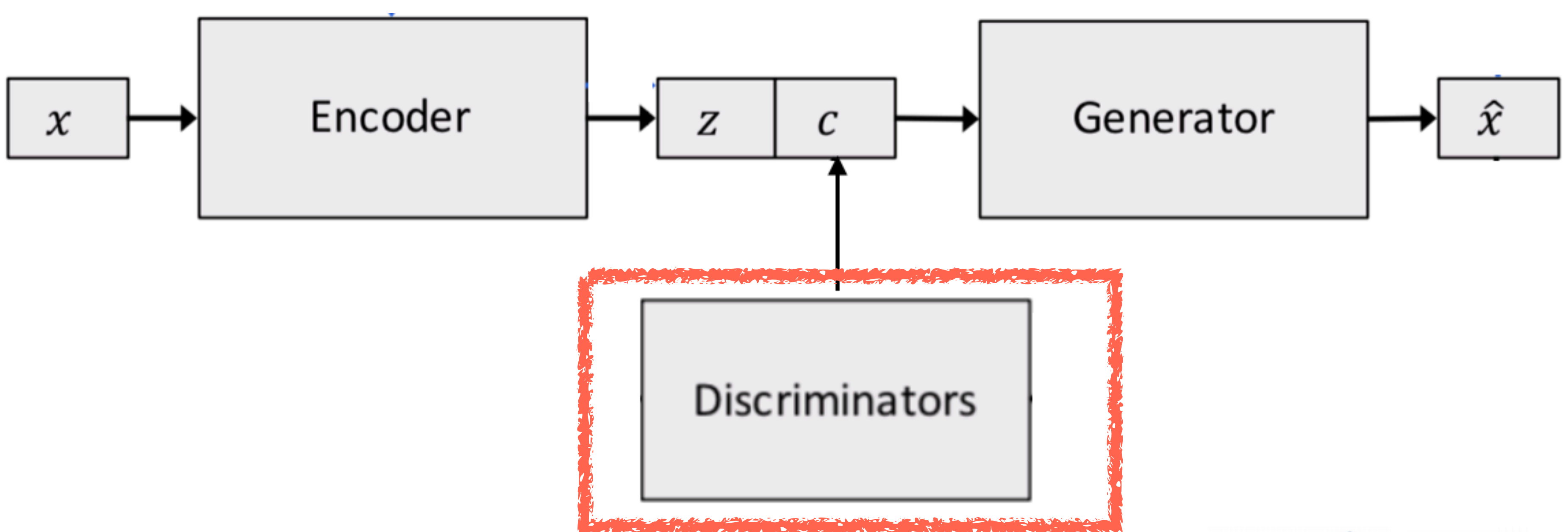


Training the discriminator $\min_{\boldsymbol{\theta}_D} \mathcal{L}_D = \boxed{\mathcal{L}_s} + \lambda_u \mathcal{L}_u$

w/ real labeled data

$$\mathcal{L}_s(\boldsymbol{\theta}_D) = \mathbb{E}_{\mathcal{X}_L}\left[\log q_D(\boldsymbol{c}_L|\boldsymbol{x}_L)\right]$$

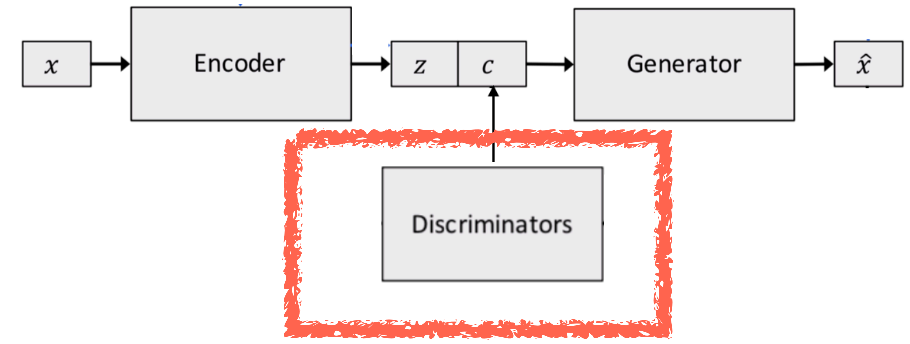
[How well does the encoder classifier the style(s) as $\boldsymbol{c}$?]



Hu, Z, Yang, Z, Liang, X, Salakhutdinov, R, Xing, EP. Toward controlled generation of text. In *ICML*, 2017.

amii

Hu et al. [2017]



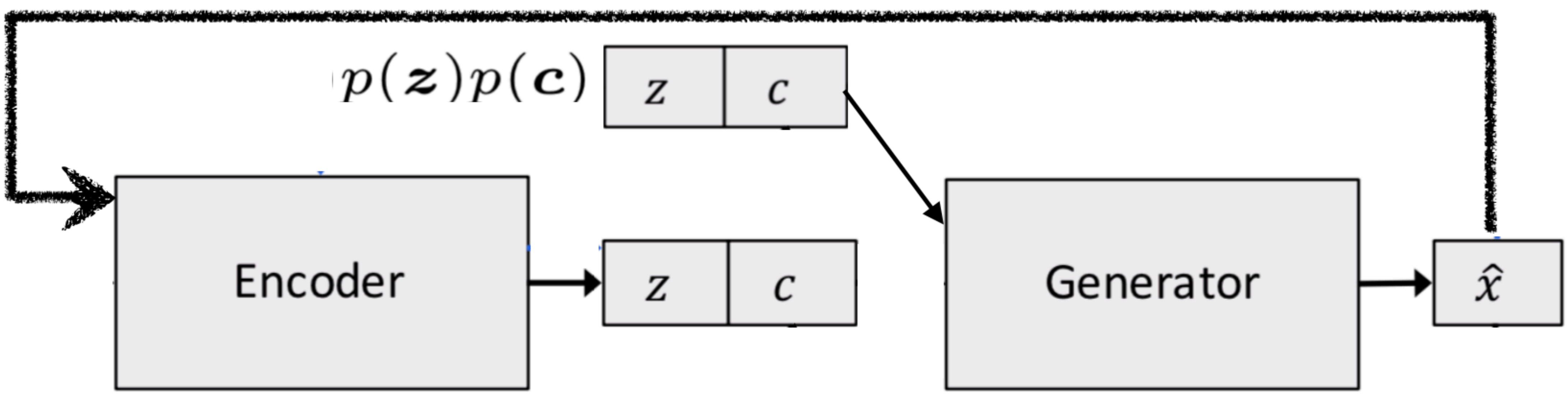
Training the discriminator $\min_{\boldsymbol{\theta}_D} \mathcal{L}_D = \mathcal{L}_s + \lambda_u \boxed{\mathcal{L}_u}$

w/ generated data from VAE

$$\mathcal{L}_u(\boldsymbol{\theta}_D) = \mathbb{E}_{p_G(\hat{\boldsymbol{x}}|\boldsymbol{z},\boldsymbol{c})p(\boldsymbol{z})p(\boldsymbol{c})}\left[\log q_D(\boldsymbol{c}|\hat{\boldsymbol{x}}) + \beta \mathcal{H}(q_D(\boldsymbol{c}'|\hat{\boldsymbol{x}}))\right]$$

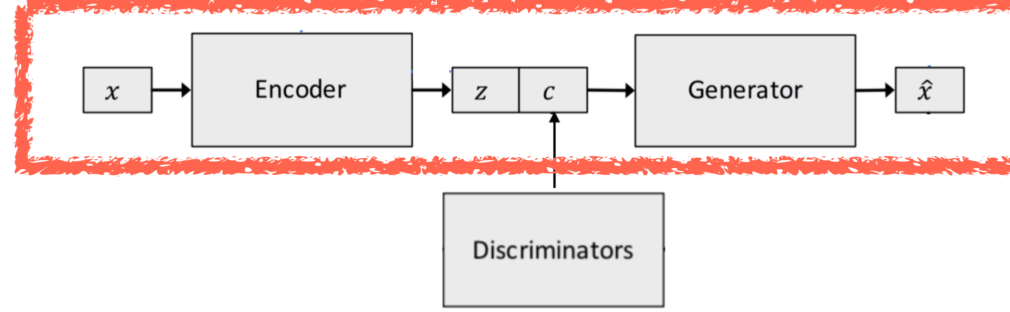
[How well does the model preserve style info after a cycle of $[z, \boldsymbol{c}] \rightarrow x \rightarrow \boldsymbol{c}$?

softmax deterministic approx.



Hu, Z, Yang, Z, Liang, X, Salakhutdinov, R, Xing, EP.
Toward controlled generation of text. In *ICML*, 2017.

Hu et al. [2017]

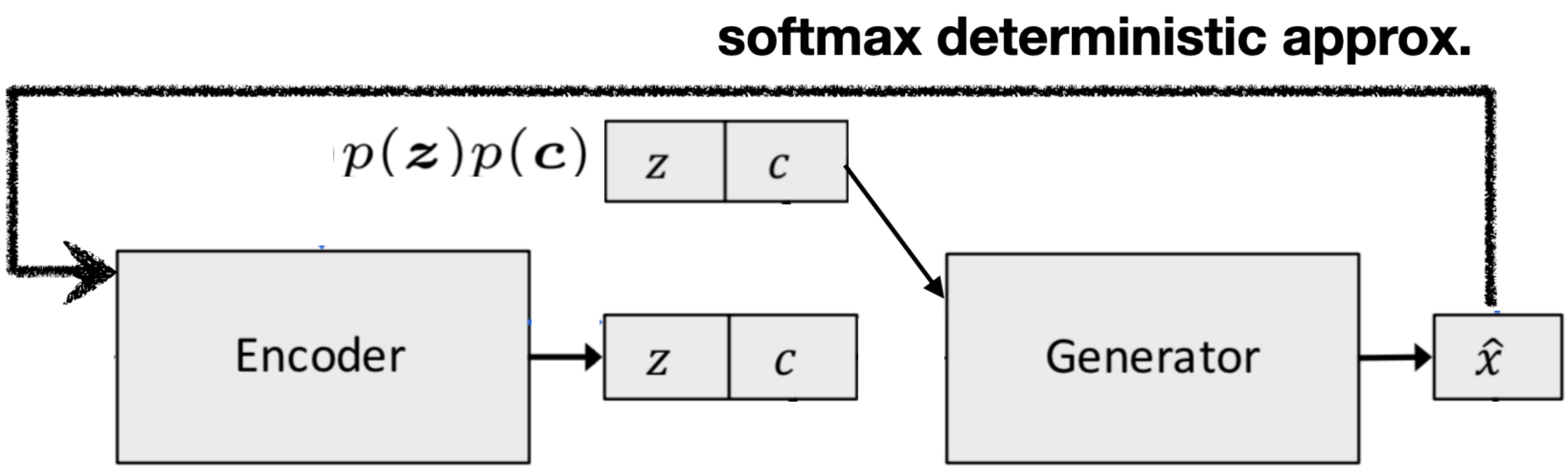


Training the generator

$$\min_{\boldsymbol{\theta}_G} \mathcal{L}_G = \mathcal{L}_{\text{VAE}} + \lambda_c \mathcal{L}_{\text{Attr},c} + \lambda_z \mathcal{L}_{\text{Attr},z}$$

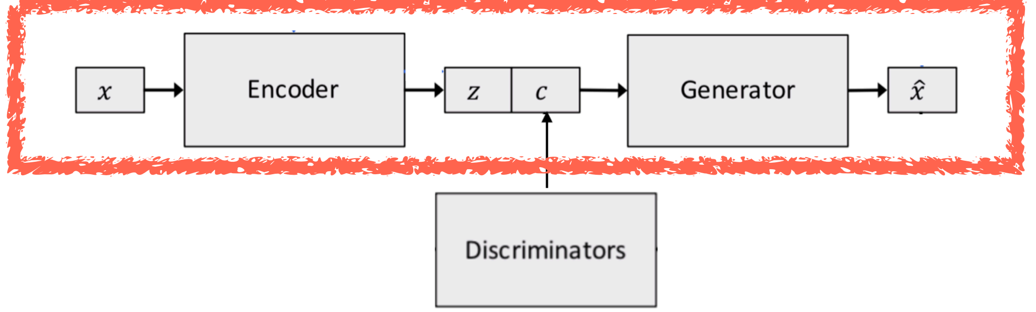
$$\mathcal{L}_{\text{Attr},c}(\boldsymbol{\theta}_G) = \mathbb{E}_{p(\boldsymbol{z})p(\boldsymbol{c})} \left[\log q_D(\boldsymbol{c}|\widetilde{G}_\tau(\boldsymbol{z}, \boldsymbol{c})) \right]$$

$$\mathcal{L}_{\text{Attr},z}(\boldsymbol{\theta}_G) = \mathbb{E}_{p(\boldsymbol{z})p(\boldsymbol{c})} \left[\log q_E(\boldsymbol{z}|\widetilde{G}_\tau(\boldsymbol{z}, \boldsymbol{c})) \right]$$



Hu, Z, Yang, Z, Liang, X, Salakhutdinov, R, Xing, EP.
Toward controlled generation of text. In *ICML*, 2017.

Essence of this work

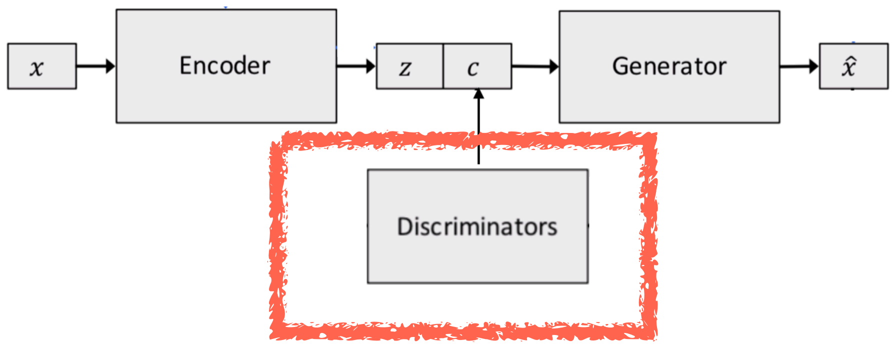


- VAE loss
  - “sentence—latent—sentence” **reconstruction**
- Alleged structured/unstructured attribute loss
  - “latent — soft sentence — latent” **reconstruction**

[mainly serving as regularization]

Hu, Z, Yang, Z, Liang, X, Salakhutdinov, R, Xing, EP. Toward controlled generation of text. In *ICML*, 2017.

Essence of this work



- VAE loss
  - “sentence—latent—sentence” reconstruction
- Alleged structured/unstructured attribute loss
  - “latent — soft sentence — latent” reconstruction

    [mainly serving as regularization; no ablation test was given]
- The semantic “grounding” of c and/or z
  - Given by style classifier/discriminator c

Hu, Z, Yang, Z, Liang, X, Salakhutdinov, R, Xing, EP.
Toward controlled generation of text. In *ICML*, 2017.

(Cross)-Alignment

- Setup and notations
  - Discrete style variable $y \in \{y_1, y_2\}$
    - Might be embedded, externally specified, not encoded
  - VAE-encoded content variable $\boldsymbol{z}$
  - Sentence $\boldsymbol{x}$

$$\begin{tikzcd} \boldsymbol{x} \arrow[r] & \boldsymbol{z} \arrow[r] & \hat{\boldsymbol{x}} \\ & y \arrow[ur] & \end{tikzcd}$$

Shen, T., Lei, T., Barzilay, R. and Jaakkola, T. Style transfer from non-parallel text by cross-alignment. In *NIPS*, 2017.

(Cross)-Alignment

- Setup and notations
  - Discrete style variable $y \in \{y_1, y_2\}$
    - Might be embedded, externally specified, not encoded
  - VAE-encoded content variable z
  - Sentence $\boldsymbol{x}$

$$\begin{tikzcd} \boldsymbol{x} \arrow[r] & z \arrow[r] & \hat{\boldsymbol{x}} \\ & y \arrow[ur] & \end{tikzcd}$$

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

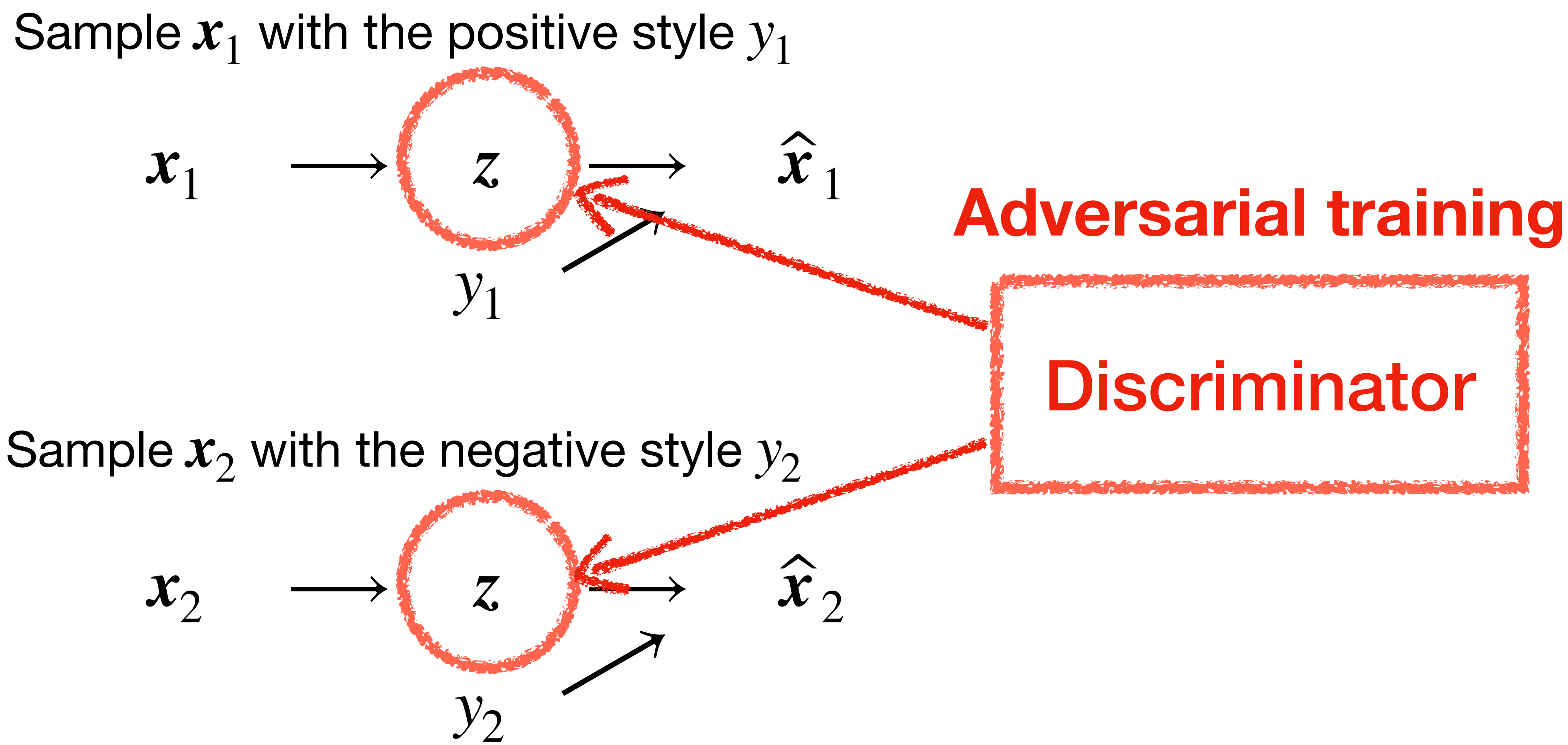
$$+ \quad \mathcal{L}_{\text{KL}}(\boldsymbol{\theta}_E) = \mathbb{E}_{\boldsymbol{x}_1 \sim \boldsymbol{X}_1}[D_{\text{KL}}(p_E(\boldsymbol{z} | \boldsymbol{x}_1, \boldsymbol{y}_1) \| p(\boldsymbol{z}))] + \mathbb{E}_{\boldsymbol{x}_2 \sim \boldsymbol{X}_2}[D_{\text{KL}}(p_E(\boldsymbol{z} | \boldsymbol{x}_2, \boldsymbol{y}_2) \| p(\boldsymbol{z}))]$$

VAE loss

Shen, T., Lei, T., Barzilay, R. and Jaakkola, T. Style transfer from non-parallel text by cross-alignment. In *NIPS*, 2017.

(Cross)-Alignment

- Variant #1: Aligned VAE



Shen, T., Lei, T., Barzilay, R. and Jaakkola, T. Style transfer from non-parallel text by cross-alignment. In *NIPS*, 2017.

(Cross)-Alignment

- Variant #1: Aligned VAE

Sa

Sa

ng

Adversarial learning on some space z

Input: samples $z_1^{(n)}$ from generative distribution G_{θ_1}

samples $z_2^{(n)}$ from generative distribution G_{θ_2}

Loop until convergence

Train a discriminator D_{θ_D} on $z_1^{(n)}$ and $z_2^{(n)}$ by

$$J_D(\theta_{\text{dis}}) = \mathbb{E}_{z_1 \sim G_{\theta_1}}[-\log D(z_1)] + E_{z_2 \sim G_{\theta_2}}[-\log(1 - D(z_2))]$$

Train generative models θ_1 and θ_2 by

$$J_{adv}(\theta_1, \theta_2) = -J_D$$

Shen, T., Lei, T., Barzilay, R. and Jaakkola, T. Style transfer from non-parallel text by cross-alignment. In *NIPS*, 2017.

(Cross)-Alignment

- Variant #1: Aligned VAE

Sa

ng

Sa

Adversarial learning on some space z

Input: samples $z_1^{(n)}$ from generative distribution G_{θ_1}

samples $z_2^{(n)}$ from generative distribution G_{θ_2}

- Adversarial training is a min-max game on z
- Overall goal is

$$\min_G \max_D (-J_D)$$

Ideally, after adv training,

z should be indistinguishable from G_{θ_1} and G_{θ_2}

In short, adversarial training pushes two distributions together with their samples.

Shen, T., Lei, T., Barzilay, R. and Jaakkola, T. Style transfer from non-parallel text by cross-alignment. In *NIPS*, 2017.

(Cross)-Alignment

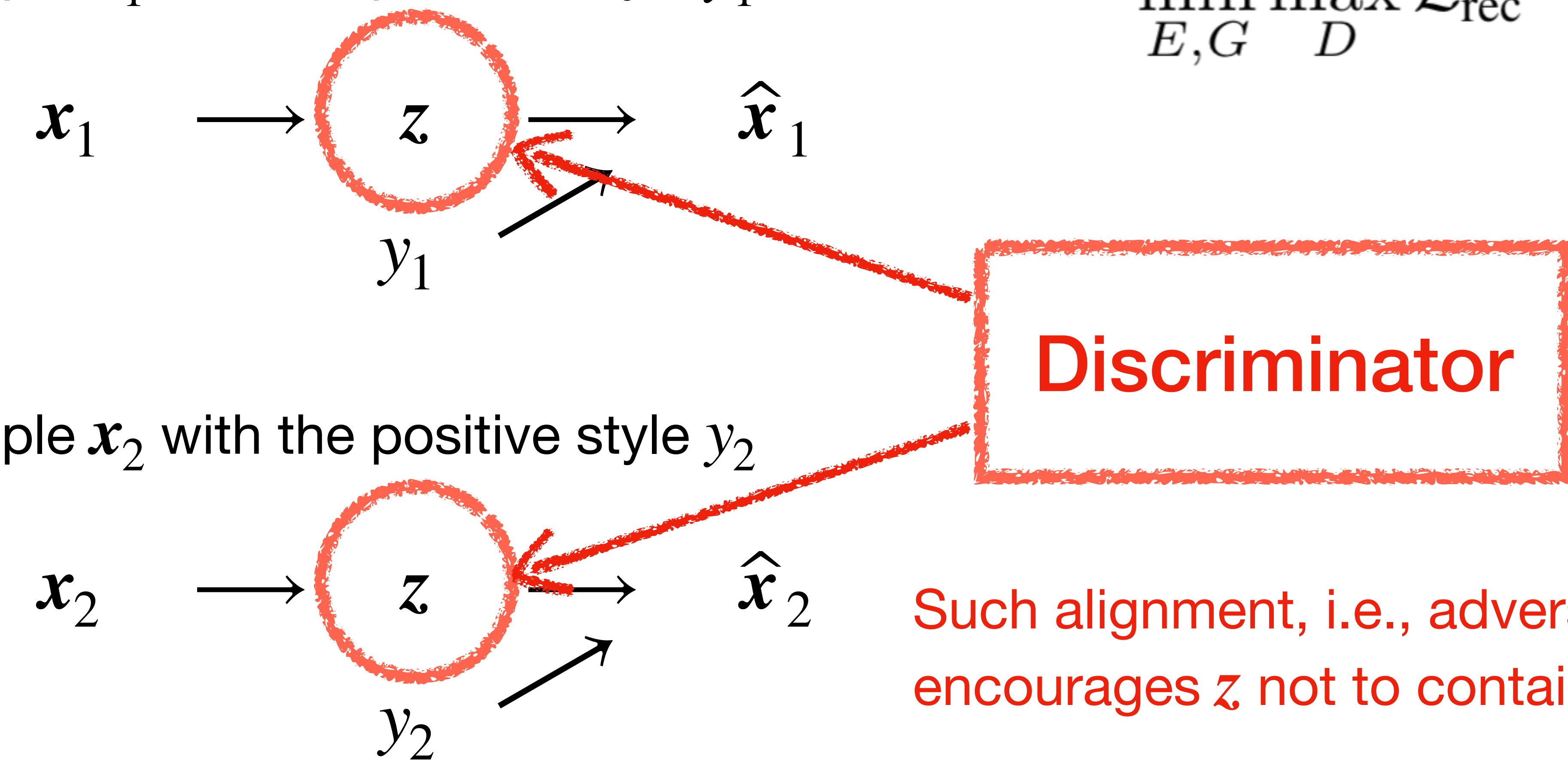
- Variant #1: Aligned VAE

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

$$\min_{E,G} \max_{D} \mathcal{L}_{\text{rec}} - \lambda \mathcal{L}_{\text{adv}}$$

Sample $\boldsymbol{x}_1$ with the positive style y_1

Sample $\boldsymbol{x}_2$ with the positive style y_2

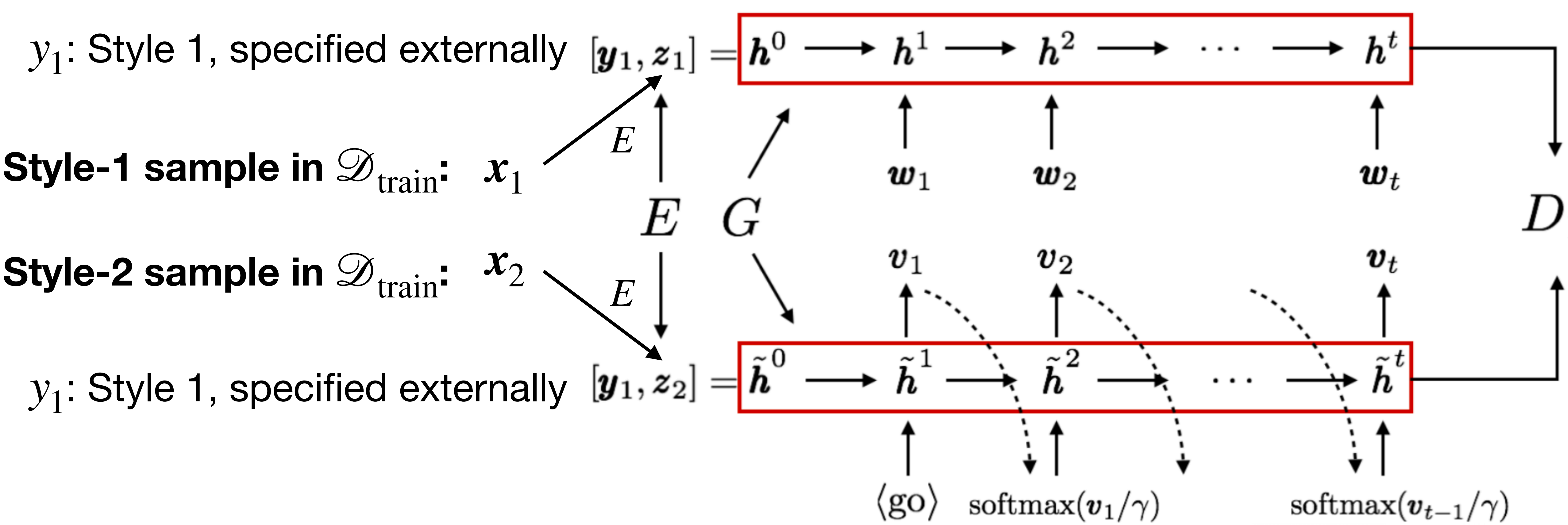


Such alignment, i.e., adversarial training encourages z not to contain style information

Shen, T., Lei, T., Barzilay, R. and Jaakkola, T. Style transfer from non-parallel text by cross-alignment. In *NIPS*, 2017.

(Cross)-Alignment

- Variant #2: **Cross**-aligned VAE
  - Incorporate style-transfer generation into training
  - Perform two adversarial trainings on
    - Style 1 sentence VS. Style 2→1 transferred sentence (example below)
    - Style 2 sentence VS. Style 1→2 transferred sentence (example omitted)



Shen, T., Lei, T., Barzilay, R. and Jaakkola, T. Style transfer from non-parallel text by cross-alignment. In *NIPS*, 2017.

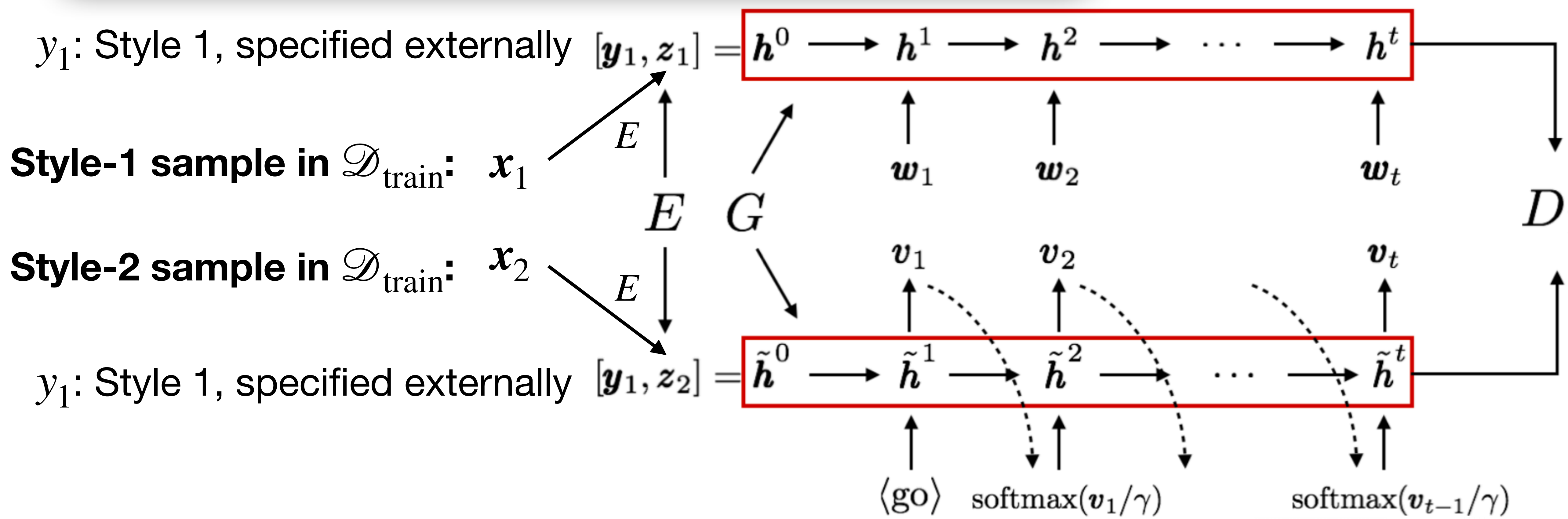
(Cross)-Alignment

- Variant #2: **Cross**-aligned VAE
  - Incorporate style-transfer generation into training
  - Perform two adversarial trainings on

$$\mathcal{L}_{\text{adv}_1} = -\frac{1}{k}\sum_{i=1}^{k}\log D_1(\boldsymbol{h}_1^{(i)}) - \frac{1}{k}\sum_{i=1}^{k}\log(1 - D_1(\tilde{\boldsymbol{h}}_2^{(i)}))$$

ce (example below)

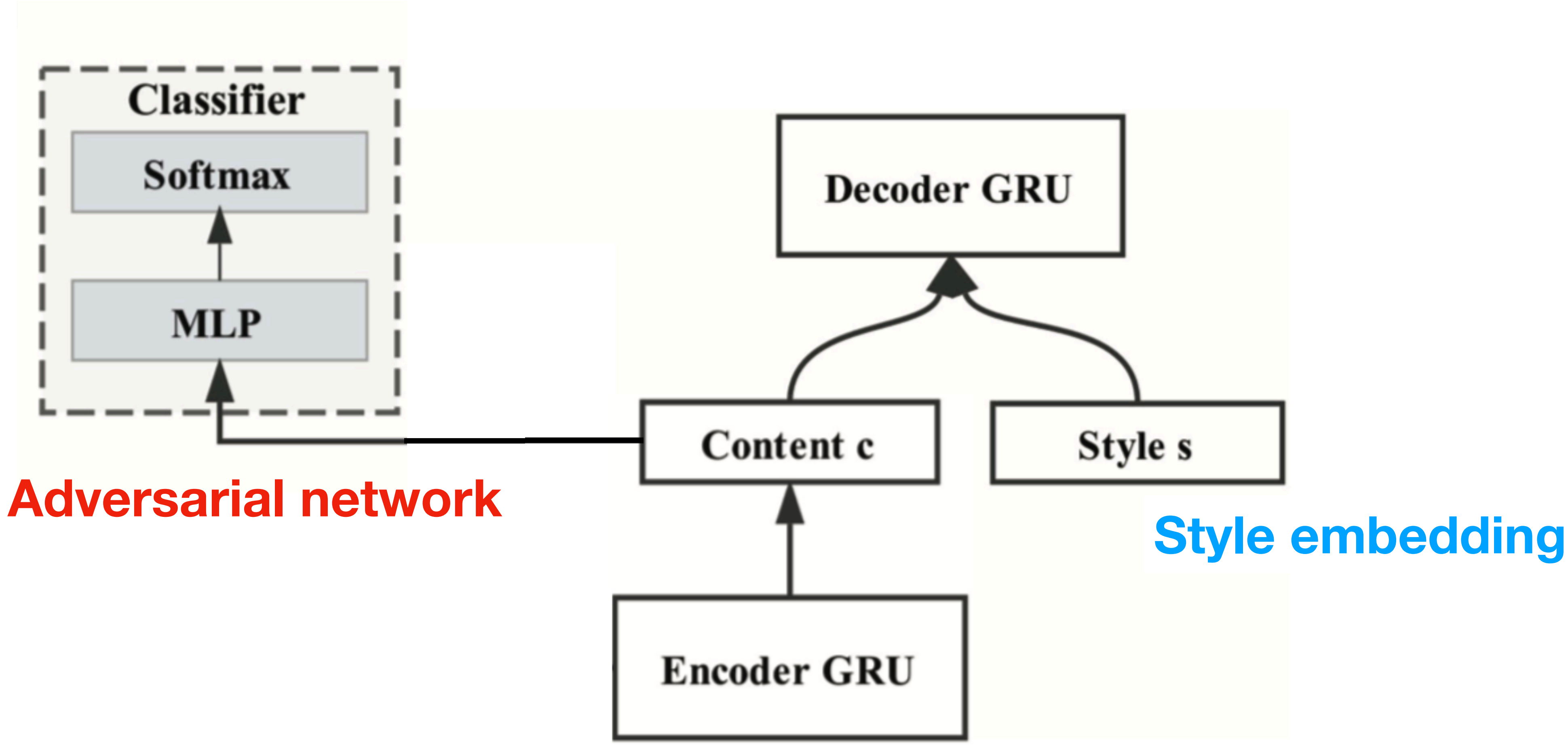
ce (example omitted)



Shen, T., Lei, T., Barzilay, R. and Jaakkola, T. Style transfer from non-parallel text by cross-alignment. In *NIPS*, 2017.

Fu et al. [2018]

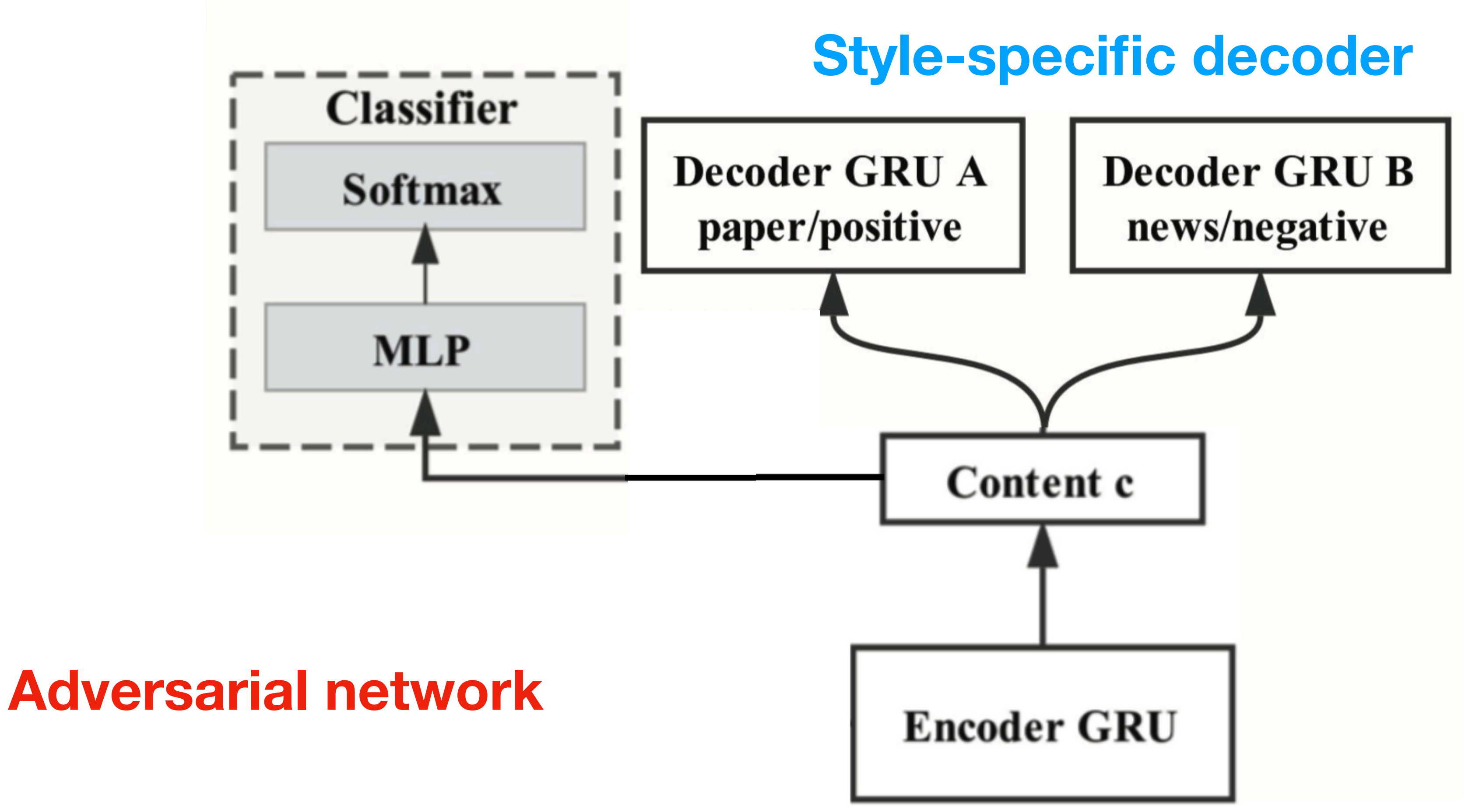
- Model #1: Style embedding



Fu, Z., Tan, X., Peng, N., Zhao, D. and Yan, R. Style transfer in text: Exploration and evaluation. In *AAAI*, 2018.

Fu et al. [2018]

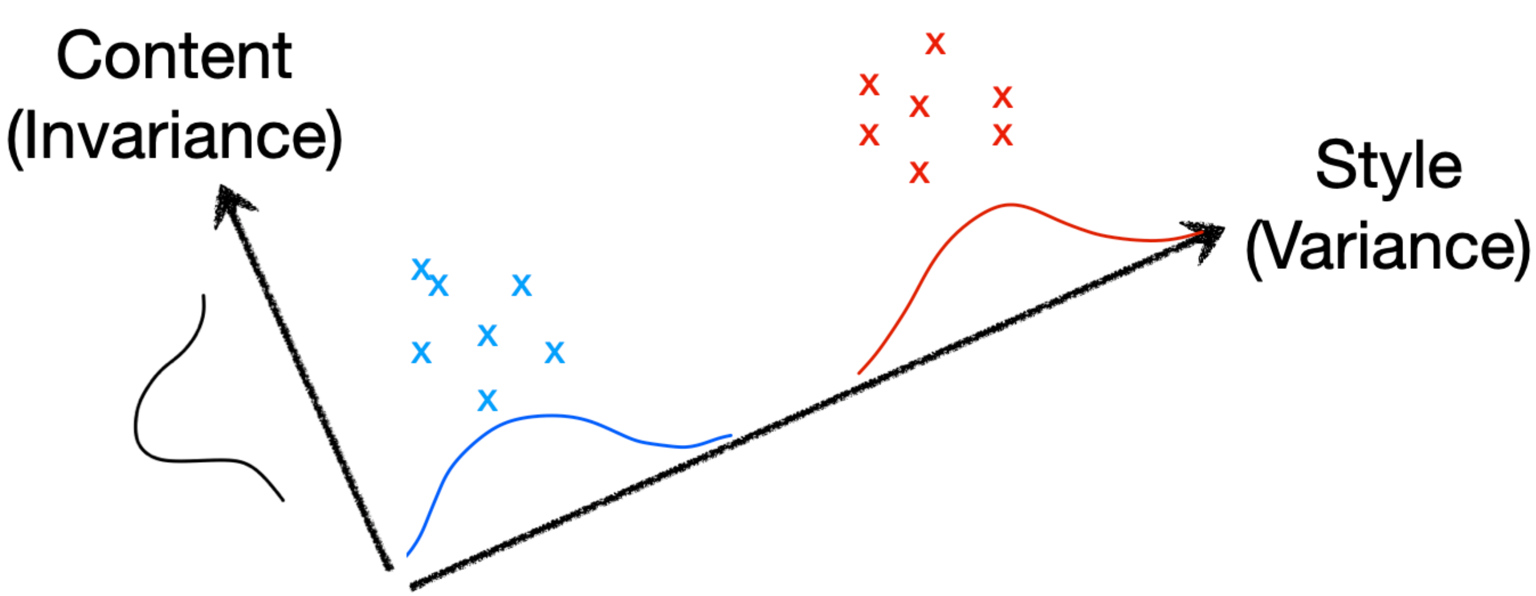
- Model #2: Style-specific decoder



Fu, Z., Tan, X., Peng, N., Zhao, D. and Yan, R. Style transfer in text: Exploration and evaluation. In *AAAI*, 2018.

Summary so-far

Model	Style treatment	Content Treatment
Hu et al. [2017]	Style classification	—
Cross-alignment [Shen et al. 2017]	Style embedding	Adv training based on style-transferred hidden states
Fu et al. [2018]	Style embedding	Adv training
	Style-specific decoder	



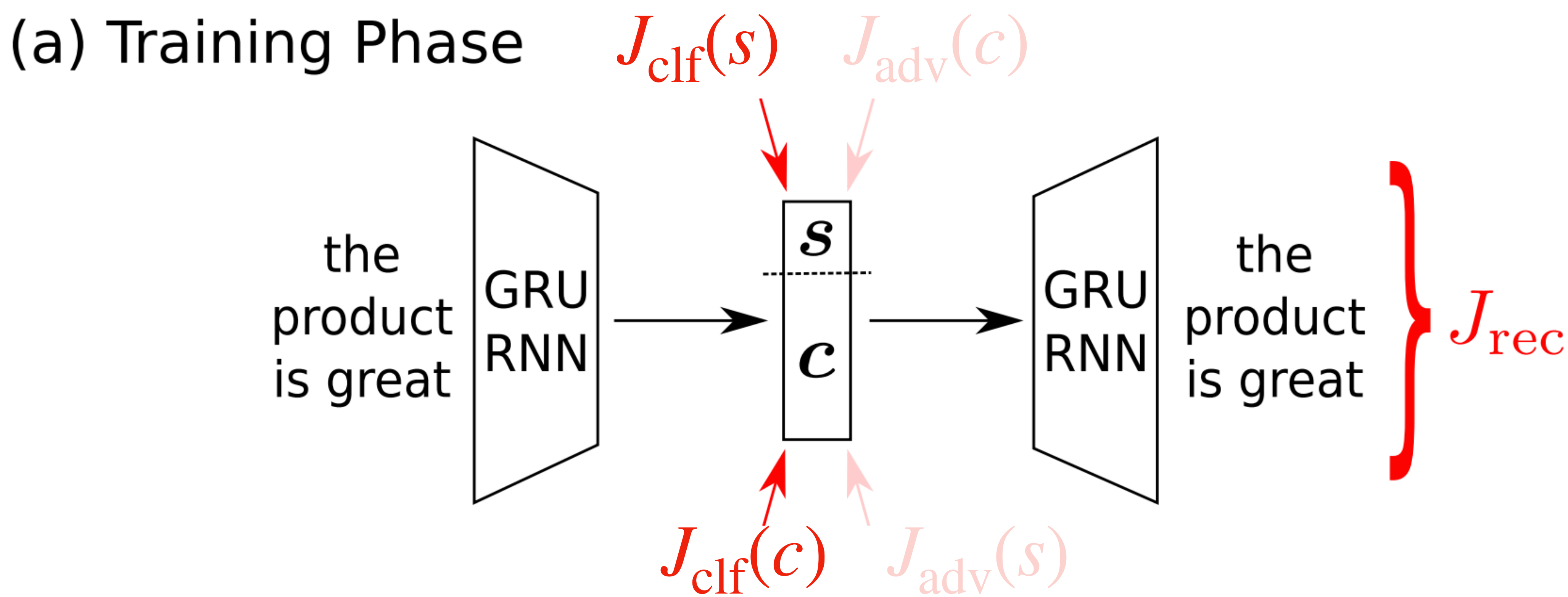
Some Thoughts

- For the **style** treatment
  - Style embedding/decoder
  - Removing style
  - Only works with very discrete styles
- For **content** treatment
  - Inadequate. E.g., adv training
    - Discourages no style information, but
    - Does not enhance content.
- Some of our thought
  - Encode style info (not by embedding)
  - Auxiliary losses can be applied to both content and style

Some Thoughts

- For the **style** treatment
  - Style embedding/decoder
  - Removing style
  - Only works with very discrete styles
- For **content** treatment
  - Inadequate. E.g., adv training
    - Discourages no style information, but
    - Does not enhance content.
- Some of our thoughts
  - Encode style info (not by embedding)
  - Auxiliary losses can be applied to both content and style

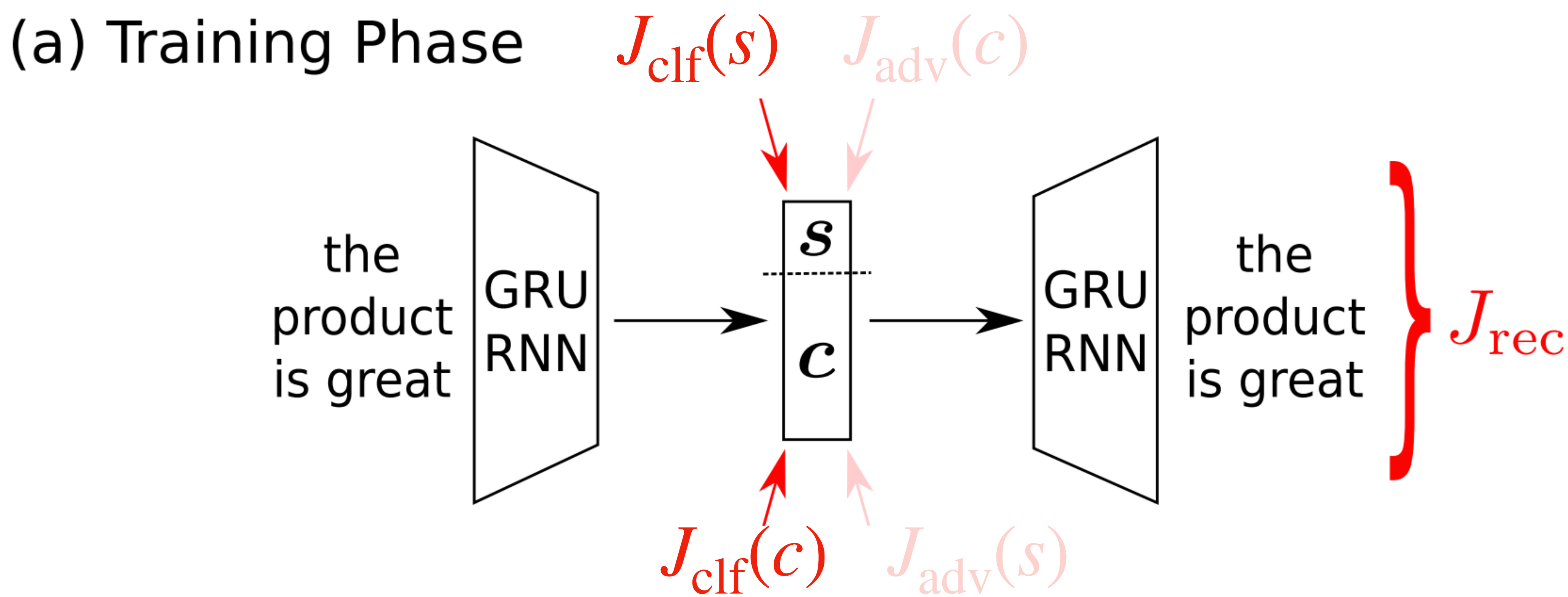
Disentangling Approach



- **Classification loss** ensures a space contains desired info
  - $J_{\text{clf}}(s)$: applied to **style** space, to classifier style
  - $J_{\text{clf}}(c)$: applied to **content** space, to classifier content
- But what is content classification?

John, V., Mou, L., Bahuleyan, H. and Vechtomova, O. Disentangled representation learning for text style transfer. In *ACL*, 2018.

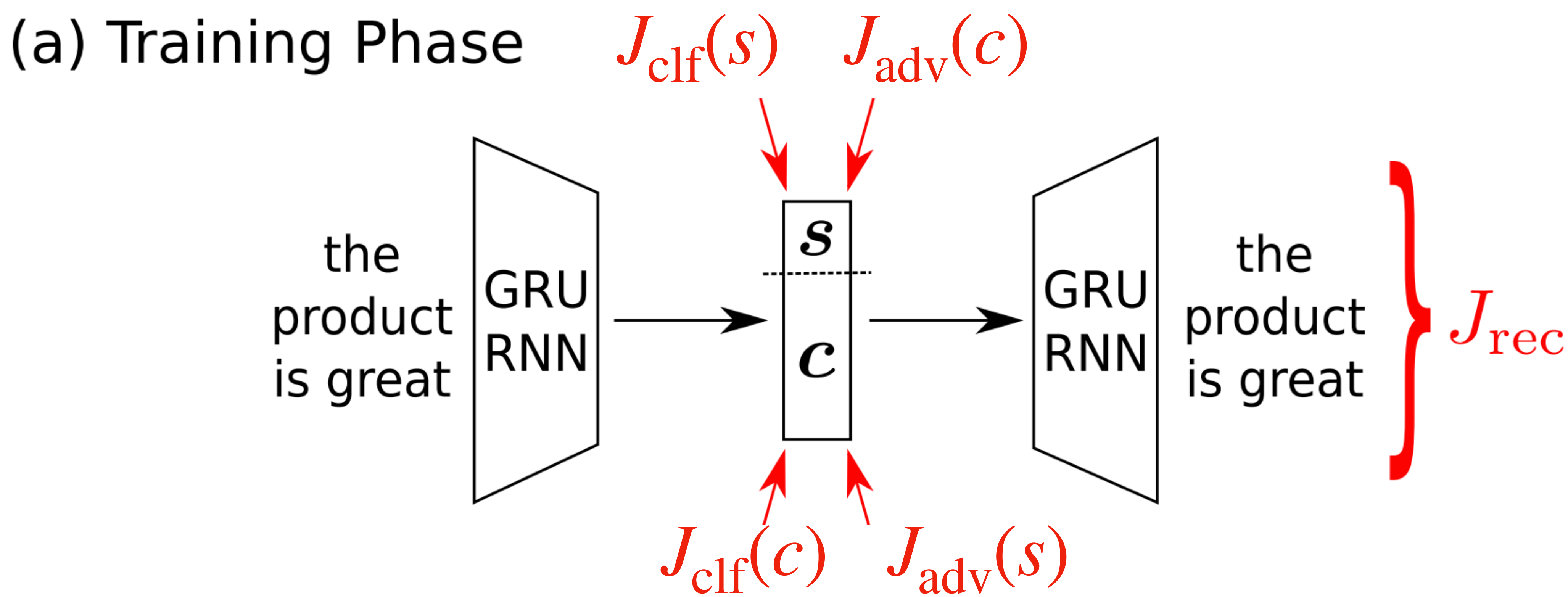
Disentangling Approach



- **Classification loss** ensures a space contains desired info
  - $J_{\mathrm{clf}}(s)$: applied to **style** space, to classifier style
  - $J_{\mathrm{clf}}(c)$: applied to **content** space, to classifier content
- But what is content classification?
  - BoW excl. style words and stop words

John, V., Mou, L., Bahuleyan, H. and Vechtomova, O. Disentangled representation learning for text style transfer. In *ACL*, 2018.

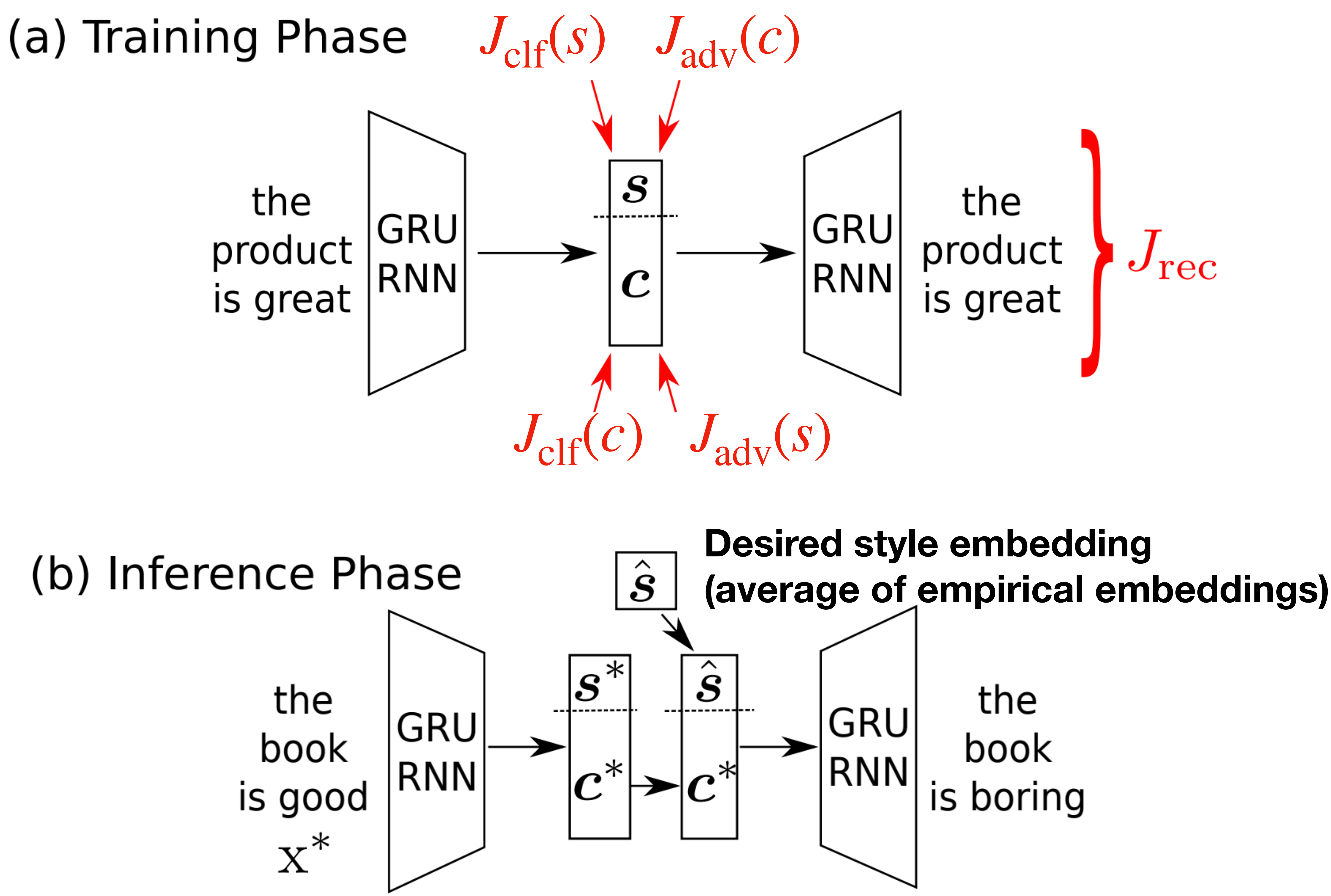
Disentangling Approach



- **Adversarial loss** ensures a space does not contain undesired info
  - $J_{\text{adv}}(s)$: applied to **content** space, in order NOT to classifier style
  - $J_{\text{adv}}(c)$: applied to **style** space, in order NOT to classifier content

John, V., Mou, L., Bahuleyan, H. and Vechtomova, O. Disentangled representation learning for text style transfer. In *ACL*, 2018.

Disentangling Approach

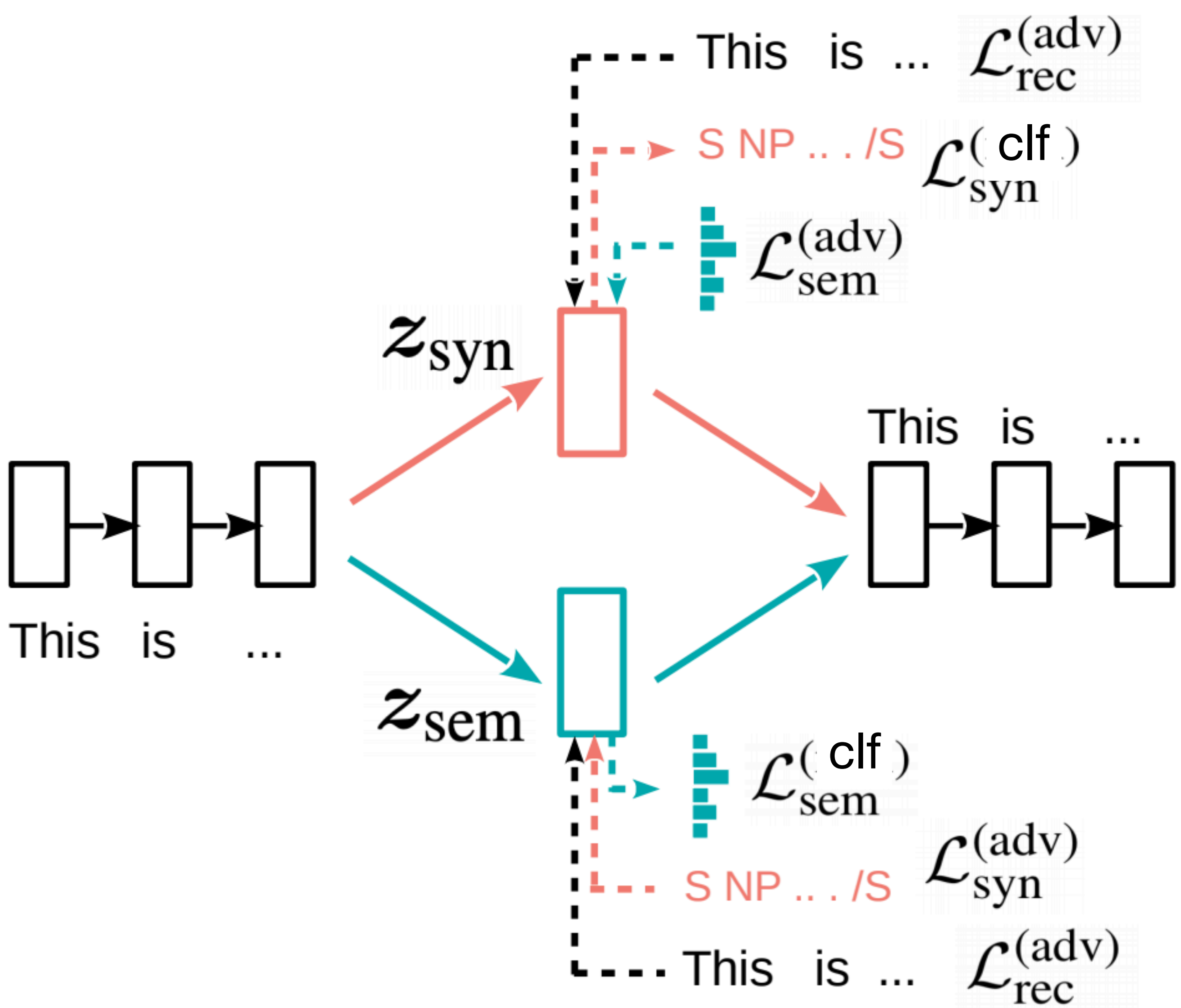


John, V., Mou, L., Bahuleyan, H. and Vechtomova, O. Disentangled representation learning for text style transfer. In *ACL*, 2018.

Non-Categorical Style Transfer

- Such **disentangling** approach works with non-categorical "styles"

- Example: syntax vs. content



- What is **content**?
  - Bag-of-words (BoW)

- What is **syntax**?

$$p(\boldsymbol{x}) = \int p(\boldsymbol{z}_{\text{sem}}, \boldsymbol{z}_{\text{syn}}) p(\boldsymbol{x} | \boldsymbol{z}_{\text{sem}}, \boldsymbol{z}_{\text{syn}}) \, \mathrm{d}\boldsymbol{z}_{\text{sem}} \, \mathrm{d}\boldsymbol{z}_{\text{syn}}$$

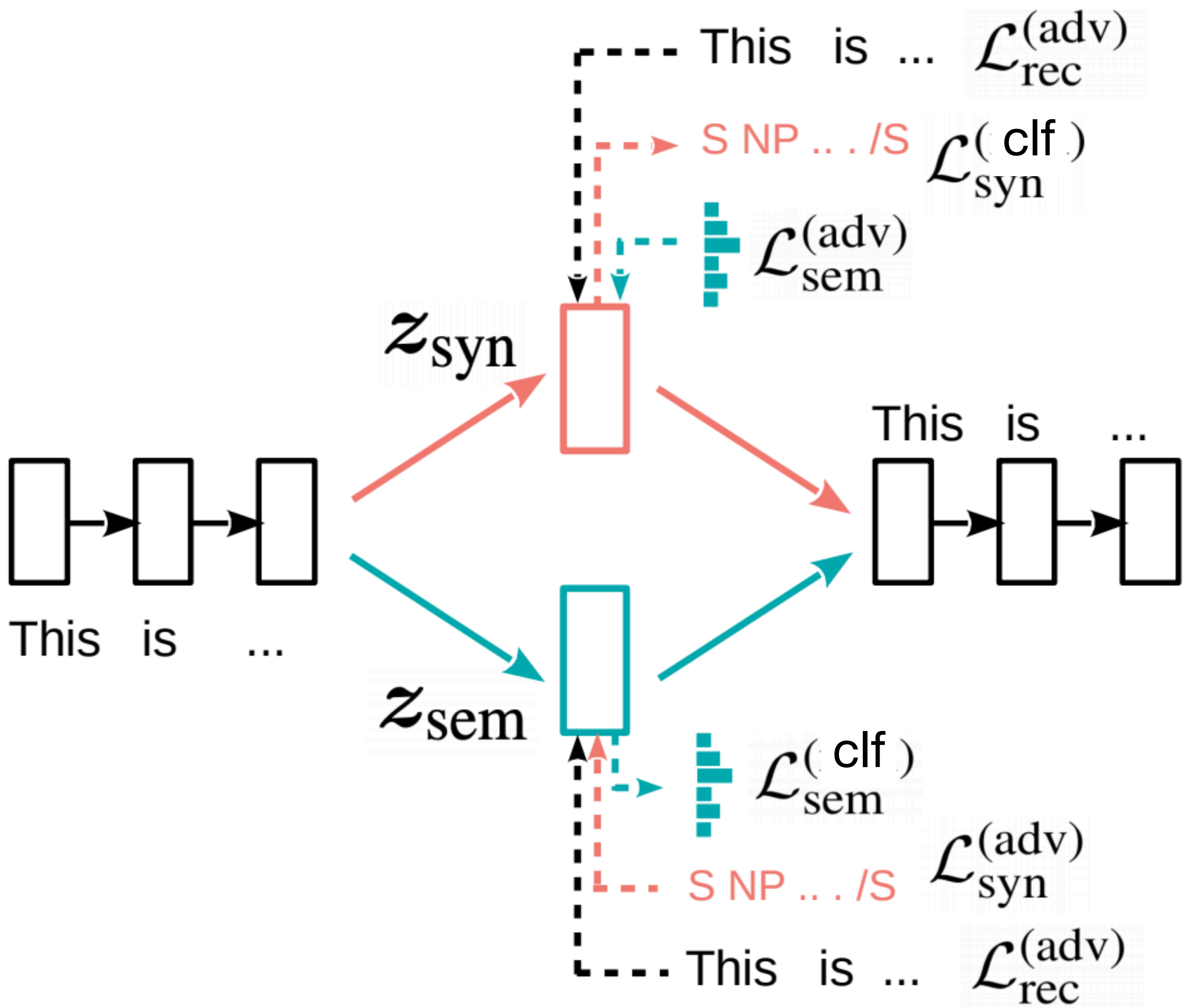
$$= \int p(\boldsymbol{z}_{\text{sem}}) p(\boldsymbol{z}_{\text{syn}}) p(\boldsymbol{x} | \boldsymbol{z}_{\text{sem}}, \boldsymbol{z}_{\text{syn}}) \, \mathrm{d}\boldsymbol{z}_{\text{sem}} \, \mathrm{d}\boldsymbol{z}_{\text{syn}}$$

Bao, Y., Zhou, H., Huang, S., Li, L., Mou, L., Vechtomova, O., Dai, X. and Chen, J. Generating sentences from disentangled syntactic and semantic spaces. In *ACL*, 2019.

Non-Categorical Style Transfer

- Such **disentangling** approach works with non-categorical "styles"

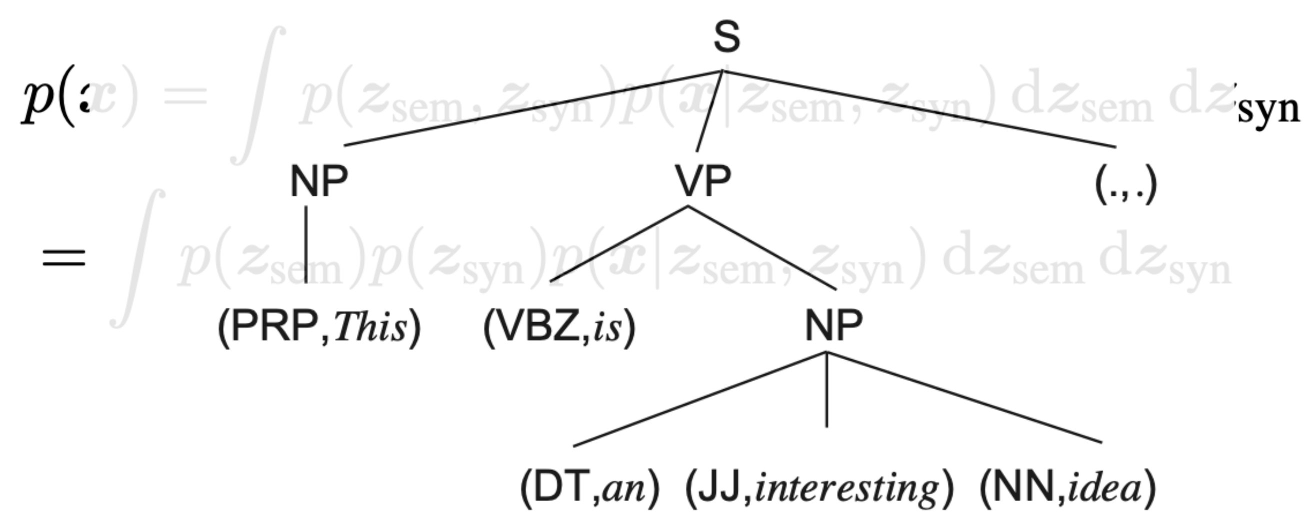
- Example: syntax vs. content



- What is **content**?
  - Bag-of-words (BoW)

- What is **syntax**?

Constituency parse tree



Linearized representation

S NP PRP /NP VP VBZ NP DT JJ NN /NP /VP . /S

Bao, Y., Zhou, H., Huang, S., Li, L., Mou, L., Vechtomova, O., Dai, X. and Chen, J. Generating sentences from disentangled syntactic and semantic spaces. In *ACL*, 2019.

Applications

- **Paraphrase generation** (by posterior sampling)
  - Syntax should vary
  - Semantics should be preserved

$$z_{\text{syn}} \sim \operatorname{argmax} p(z_{\text{syn}} \mid \boldsymbol{x})$$

$$z_{\text{sem}} = \operatorname{argmax} p(z_{\text{sem}} \mid \boldsymbol{x})$$

- Syntax transfer

$$\text{Dec}\left(z_{\text{syn}}=\text{Enc}(\text{Ref}_{\text{syn}}),\ z_{\text{sem}}=\text{Enc}(\text{Ref}_{\text{sem}}) \right)$$

Semantic and Syntactic Providers		Syntax-Transfer Output	
Ref$_{\text{syn}}$:	There is an apple on the table.	**VAE**:	The man is in the kitchen.
Ref$_{\text{sem}}$:	The airplane is in the sky.	**DSS-VAE**:	There is a airplane in the sky.
Ref$_{\text{syn}}$:	The shellfish was cooked in a wok.	**VAE**:	The man was filled with people.
Ref$_{\text{sem}}$:	The stadium was packed with people.	**DSS-VAE**:	The stadium was packed with people.
Ref$_{\text{syn}}$:	The child is playing in the garden.	**VAE**:	There is a person in the garden.
Ref$_{\text{sem}}$:	There is a dog behind the door.	**DSS-VAE**:	A dog is walking behind the door.

Bao, Y., Zhou, H., Huang, S., Li, L., Mou, L., Vechtomova, O., Dai, X. and Chen, J. Generating sentences from disentangled syntactic and semantic spaces. In *ACL*, 2019.

Applications

- **Paraphrase generation** (by posterior sampling)
  - Syntax should vary
  - Semantics should be preserved

$$z_{\text{syn}} \sim \operatorname{argmax} p(z_{\text{syn}} \mid \boldsymbol{x})$$

$$z_{\text{sem}} = \operatorname{argmax} p(z_{\text{sem}} \mid \boldsymbol{x})$$

- Syntax transfer

$$\text{Dec}\big(\, z_{\text{syn}}=\text{Enc}(\text{Ref}_{\text{syn}}),\ z_{\text{sem}}=\text{Enc}(\text{Ref}_{\text{sem}})\,\big)$$

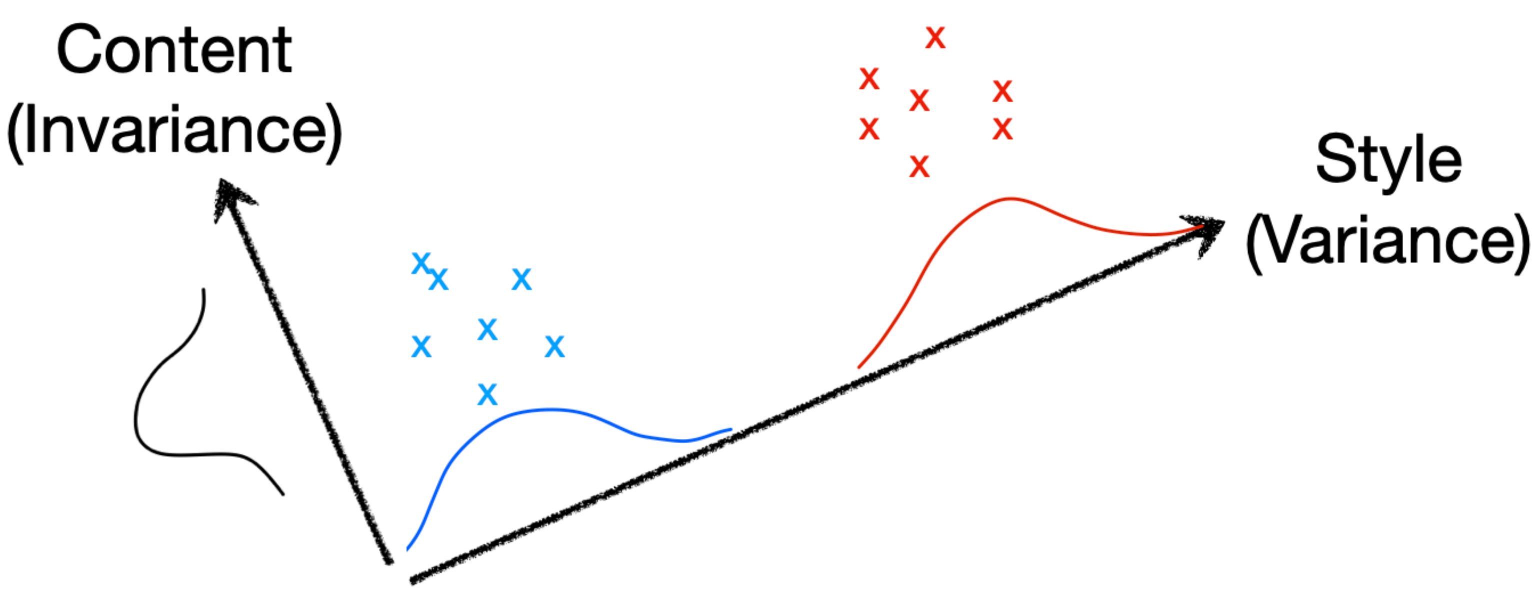
Semantic and Syntactic Providers		Syntax-Transfer Output	
Ref$_{\text{syn}}$:	There is an apple on the table.	**VAE**:	The man is in the kitchen.
Ref$_{\text{sem}}$:	The airplane is in the sky.	DSS-VAE:	There is a airplane in the sky.
Ref$_{\text{syn}}$:	The shellfish was cooked in a wok.	VAE:	The man was filled with people.
Ref$_{\text{sem}}$:	The stadium was packed with people.	DSS-VAE:	The stadium was packed with people.
Ref$_{\text{syn}}$:	The child is playing in the garden.	VAE:	There is a person in the garden.
Ref$_{\text{sem}}$:	There is a dog behind the door.	**DSS-VAE**:	A dog is walking behind the door.

Insider's knowledge: Currently only works with compatible syntax

Bao, Y., Zhou, H., Huang, S., Li, L., Mou, L., Vechtomova, O., Dai, X. and Chen, J. Generating sentences from disentangled syntactic and semantic spaces. In *ACL*, 2019.

Summary so-far

Model	Style treatment	Content Treatment
Hu et al. [2017]	Style classification	—
Cross-alignment [Shen et al. 2017]	Style embedding	Adv training based on style-transferred hidden states
Fu et al. [2018]	Style embedding Style-specific decoder	Adv training
Disentangling [John+'19; Bao+'19]	Style classification Content adversarial	Content adversarial Style classifiation



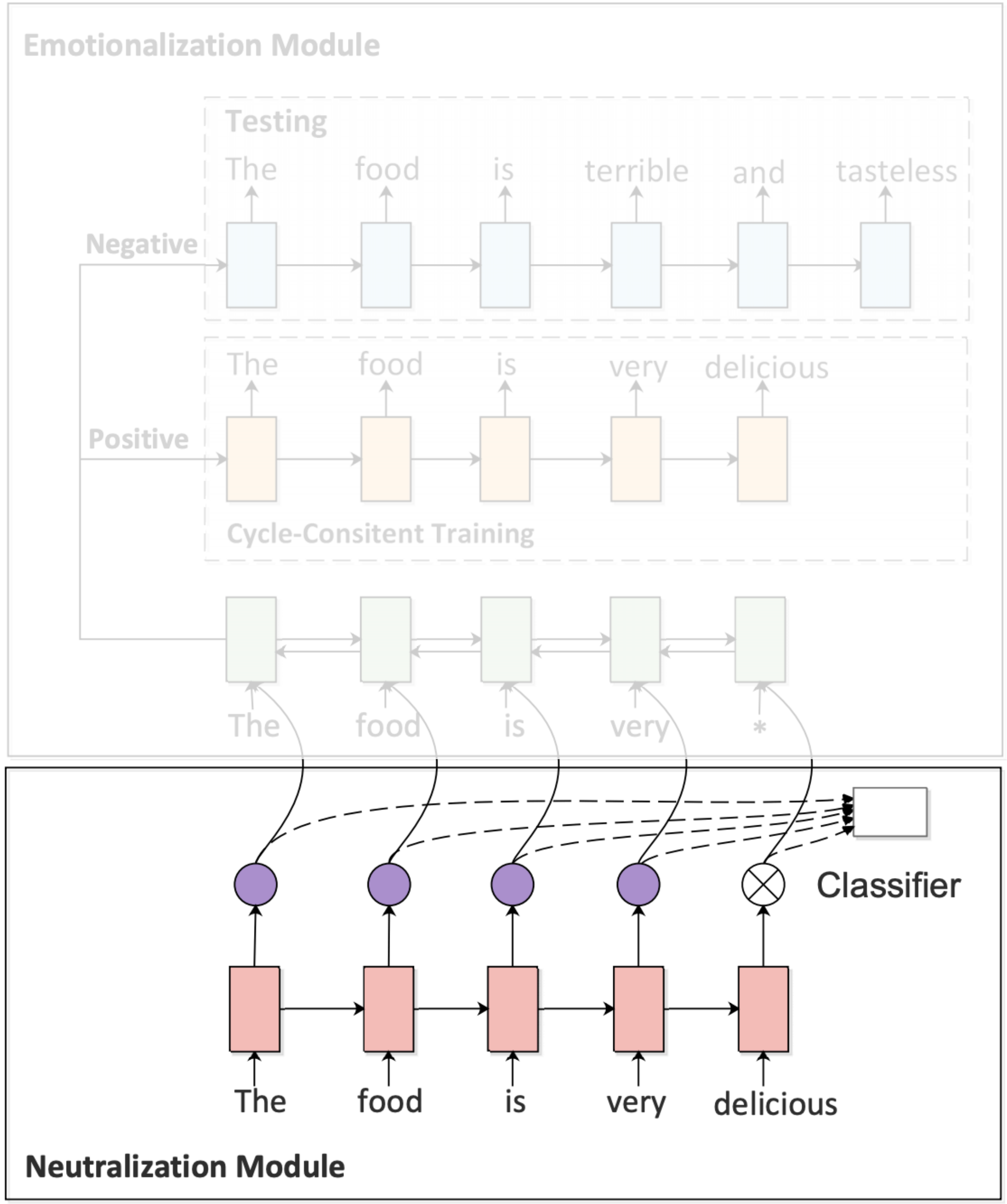
Cycled RL

- Module#1:

Extracting style-**neutral** words

  - Train a sentiment classifier w/ attention
  - Thresholding attention to select **style-neurtral** words

- Module#2: Reconstructing



Xu, J., Sun, X., Zeng, Q., Ren, X., Zhang, X., Wang, H. and Li, W. Unpaired sentiment-to-sentiment translation: A cycled reinforcement learning approach. In *ACL*, 2018.

Cycled RL

- Module#1:

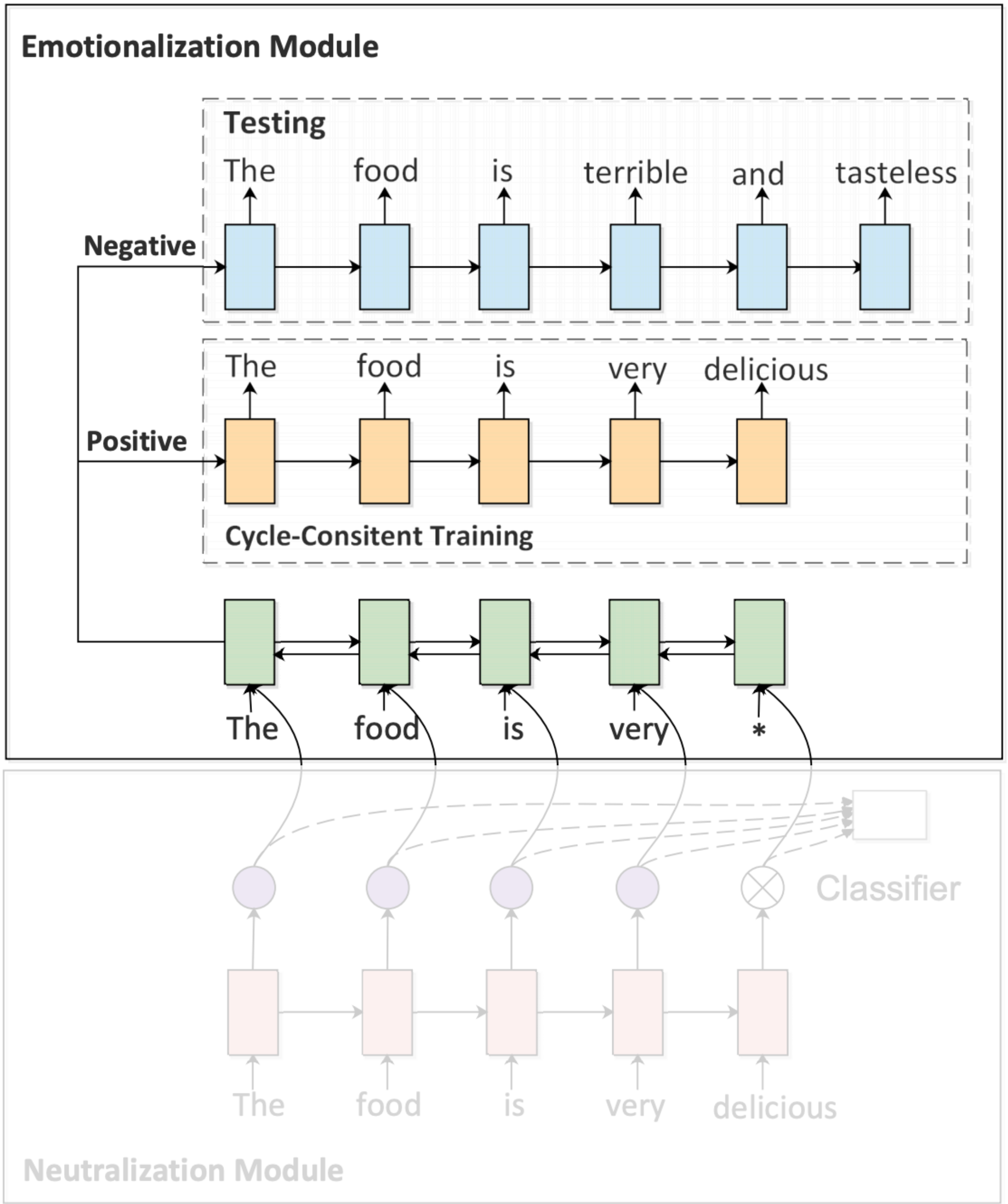
Extracting style-**neutral** words

- Module#2: Reconstructing

  style-**rich** sentences

  from style-**neutral** words

  (with style-specific decoders)



Xu, J., Sun, X., Zeng, Q., Ren, X., Zhang, X., Wang, H. and Li, W. Unpaired sentiment-to-sentiment translation: A cycled reinforcement learning approach. In *ACL*, 2018.

Cycled RL

- Module#1: Extracting style-**neutral** words
- Module#2: Reconstructing style-**rich** sentences
  - Cycle consistency to refine style-word extractor
  - Cross-entropy for training the decoder

Module#1

Module#2 (multi-decoders)

Positive sentence → **Sentiment-neural subsentence** → **Reconstructed positive** / **Reconstructed negative**

The food is very delicious

The food is very *

The food is very delicious

The food is terrible and tasteless

$$\text{Reward} = (1+\beta^2)\frac{\text{BLEU} \cdot \text{StyleClassification}}{\beta^2 \cdot \text{BLEU} + \text{StyleClassification}}$$

(harmonic mean)

Xu, J., Sun, X., Zeng, Q., Ren, X., Zhang, X., Wang, H. and Li, W. Unpaired sentiment-to-sentiment translation: A cycled reinforcement learning approach. In *ACL*, 2018.

amii

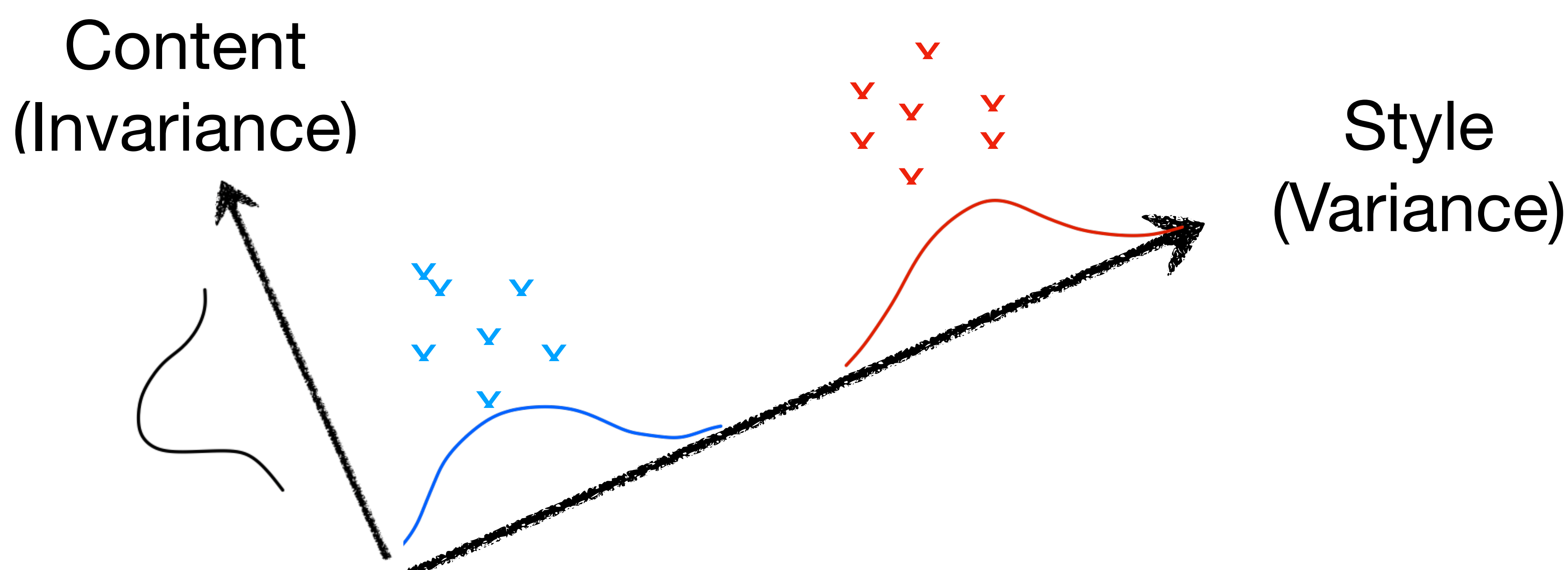
A Quick Detour to REINFORCE

- RL works with discrete actions (e.g., which words to generate)
- REINFORCE is commonly used in NLP
  - Sample your action
  - If the result is good, enhance/reinforce it
  - If the result is not good, enhance it in an opposite way

(supervised learning with reward as weight)

Delete-Retrieve-Generate

- **General idea**
  - Detect and delete style-rich phrases
  - Retrieve similar sentences with the target style
  - Generate a style-transferred sentence
- **Assumption**
  - a roughly aligned sentence can be retrieved in training data



Li, J., Jia, R., He, H. and Liang, P. Delete, retrieve, generate: A simple approach to sentiment and style transfer. In *NAACL-HLT*, 2018.

Delete-Retrieve-Generate

- **Detecting style-rich phrases** (called attribute marker)
  - Counting n-gram frequency

$$s(u,v) = \frac{\mathrm{count}(u, \mathcal{D}_v) + \lambda}{\left(\sum_{v' \in \mathcal{V}, v' \neq v} \mathrm{count}(u, \mathcal{D}_{v'})\right) + \lambda},$$

(for style v and n-gram u)

  - Thresholding
- Example

Li, J., Jia, R., He, H. and Liang, P. Delete, retrieve, generate: A simple approach to sentiment and style transfer. In *NAACL-HLT*, 2018.

Delete-Retrieve-Generate

- **Detecting style-rich phrases** (called attribute marker)
  - Counting n-gram frequency

$$s(u, v) = \frac{\mathrm{count}(u, \mathcal{D}_v) + \lambda}{\left(\sum_{v' \in \mathcal{V}, v' \neq v} \mathrm{count}(u, \mathcal{D}_{v'})\right) + \lambda};$$

(for style v and n-gram u)

  - Thresholding

- **Example**

i have had this mount for about a year and it works great .

$\downarrow$ **Delete**

i have had this mount for about a year and it .

Li, J., Jia, R., He, H. and Liang, P. Delete, retrieve, generate: A simple approach to sentiment and style transfer. In *NAACL-HLT*, 2018.

Delete-Retrieve-Generate

- **Retrieve a similar sentence in the desired style**

$$x^{\mathrm{tgt}} = \operatorname*{argmin}_{x' \in \mathcal{D}_{v^{\mathrm{tgt}}}} d(c(x, v^{\mathrm{src}}), c(x', v^{\mathrm{tgt}}))$$

x' in the training set with the designed style

c(,): content words of a sentence

d: distance metric

- Attempt#1: tf·idf-based overalap
- Attempt#2: Euclidean distance of embeddings (used for different model variants)

- **Example**

i have had this mount for about a year and it works great .

Delete

i have had this mount for about a year and it .

Model#1: RetrieveOnly

Retrieve

i have had it for a while but barely used it .

Li, J., Jia, R., He, H. and Liang, P. Delete, retrieve, generate: A simple approach to sentiment and style transfer. In *NAACL-HLT*, 2018.

Delete-Retrieve-Generate

- **Retrieve a similar sentence in the desired style**

$$x^{\text{tgt}} = \underset{x' \in \mathcal{D}_{v^{\text{tgt}}}}{\operatorname{argmin}}\, d(c(x, v^{\text{src}}), c(x', v^{\text{tgt}}))$$

x' in the training set with the designed style

c(,): content words of a sentence

d: distance metric

- Attempt#1: tf·idf-based overalap
- Attempt#2: Euclidean distance of embeddings (used for different model variants)

- **Example**

i have had this mount for about a year and it works great .

↓ **Delete**

i have had this mount for about a year and it .

↓ **Retrieve**

Model#1: RetrieveOnly

i have had it for a while but barely used it .

Li, J., Jia, R., He, H. and Liang, P. Delete, retrieve, generate: A simple approach to sentiment and style transfer. In *NAACL-HLT*, 2018.

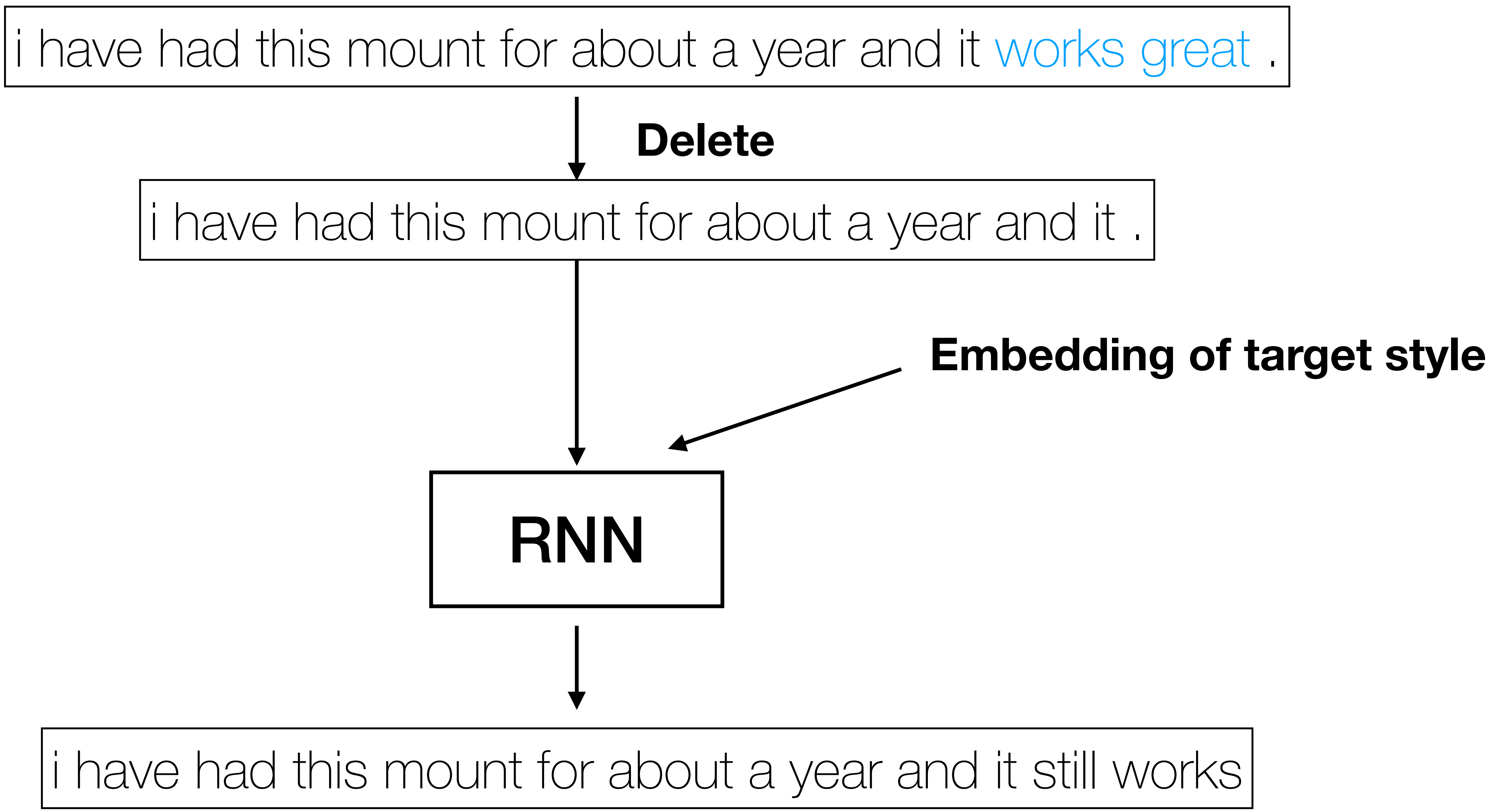
Delete-Retrieve-Generate

- **Model#1: Template**
  - Some naive swapping of attribute markers
  - May yield ungrammatical sentences

Li, J., Jia, R., He, H. and Liang, P. Delete, retrieve, generate: A simple approach to sentiment and style transfer. In *NAACL-HLT*, 2018.

Delete-Retrieve-Generate

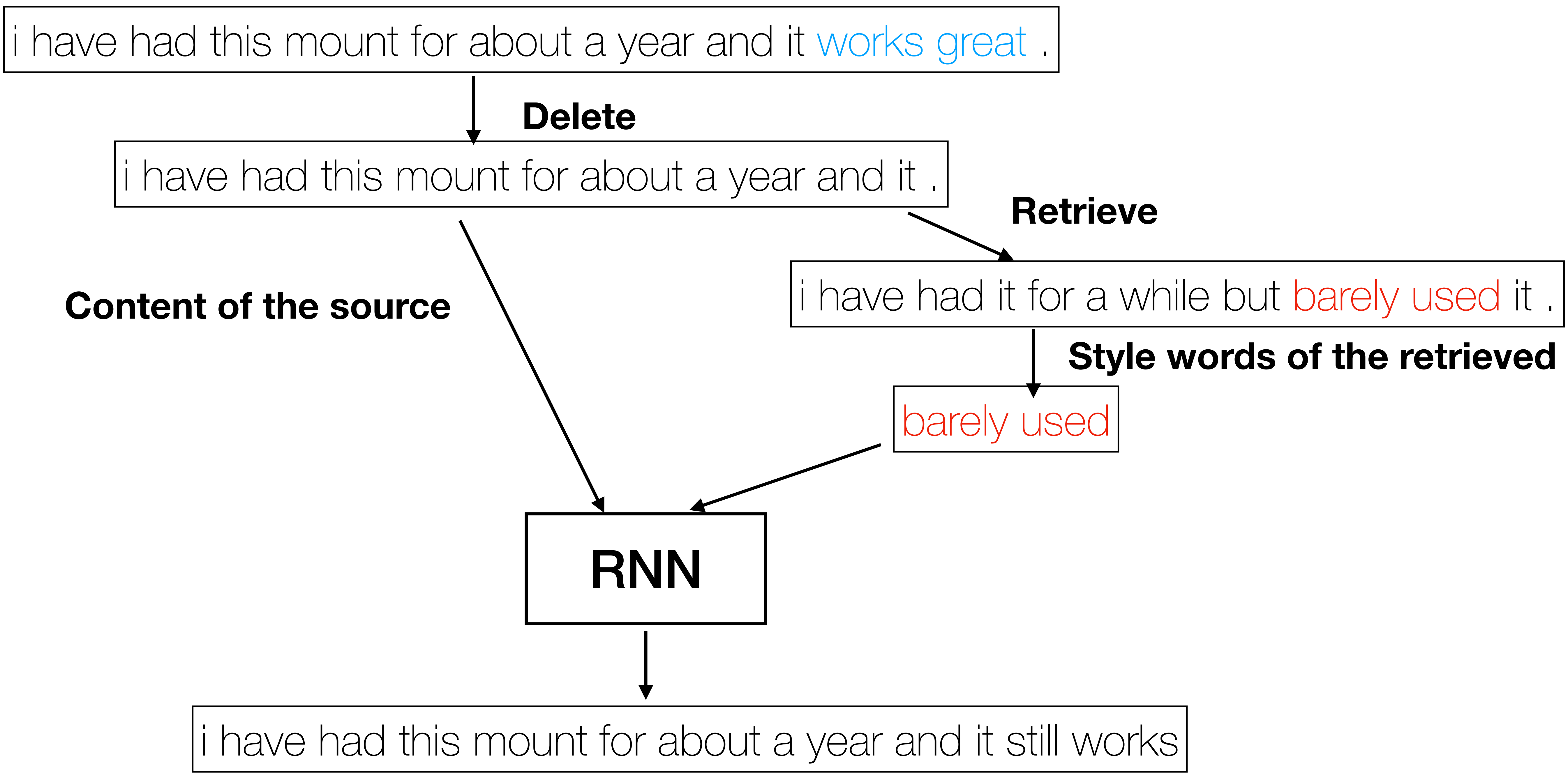
- **Model#2: Delete+Generate**



Li, J., Jia, R., He, H. and Liang, P. Delete, retrieve, generate: A simple approach to sentiment and style transfer. In *NAACL-HLT*, 2018.

Delete-Retrieve-Generate

- **Model#3: Delete+Retrieve+Generate**

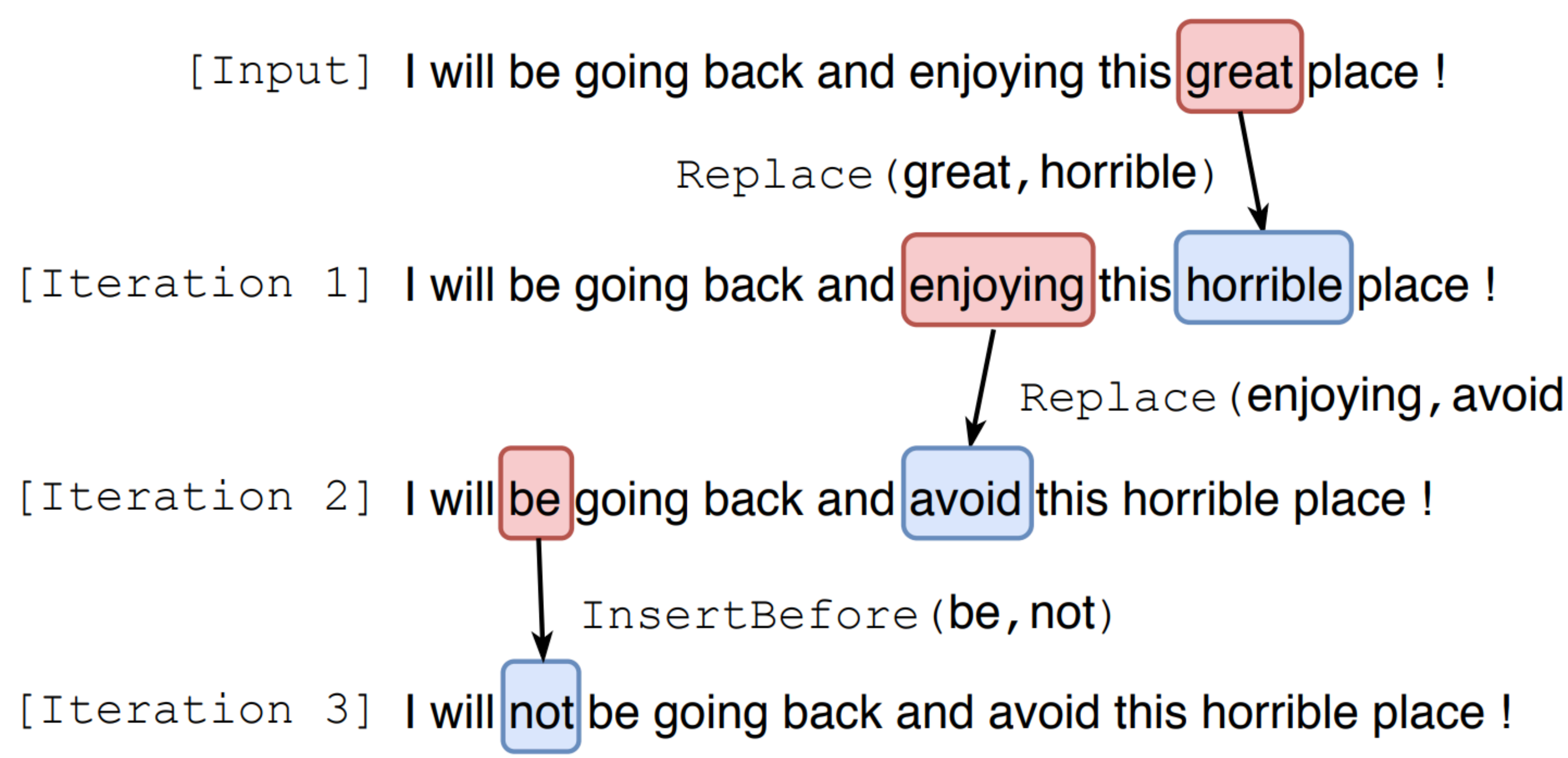


Li, J., Jia, R., He, H. and Liang, P. Delete, retrieve, generate: A simple approach to sentiment and style transfer. In *NAACL-HLT*, 2018.

RL for Edit

- General idea
  - Define a reward function
  - Search towards it
  - REINFORCE learns appropriate operations

Module	Operation
IF_{ϕ_1}	**I**nsert a word $\hat{w}$ in **F**ront of the position
IB_{ϕ_2}	**I**nsert a word $\hat{w}$ **B**ehind the position
Rep_{ϕ_3}	**Rep**lace it with another word $\hat{w}$
DC	**D**elete the **C**urrent word
DF	**D**elete the word in **F**ront of the position
DB	**D**elete the word **B**ehind the position
Skip	Do not change anything



Wu, C., Ren, X., Luo, F. and Sun, X. A Hierarchical Reinforced Sequence Operation Method for Unsupervised Text Style Transfer. In *ACL*, 2019.

RL for Edit

- General idea
  - Define a reward function
  - REINFORCE learns appropriate operations

[Input] I will be going back and enjoying this great place !

Replace (great, horrible)

[Iteration 1] I will be going back and enjoying this horrible place !

Replace (enjoying, avoid)

[Iteration 2] I will be going back and avoid this horrible place !

InsertBefore (be, not)

[Iteration 3] I will not be going back and avoid this horrible place !

Wu, C., Ren, X., Luo, F. and Sun, X. A Hierarchical Reinforced Sequence Operation Method for Unsupervised Text Style Transfer. In *ACL*, 2019.

RL for Edit

- **Hierarchical Actions**
  - High-level: selecting the word to edit
  - Low-level: an edit operator (and, if needed, a candidate word)

Module	**Operation**
IF_{ϕ_1}	**I**nsert a word $\hat{w}$ in **F**ront of the position
IB_{ϕ_2}	**I**nsert a word $\hat{w}$ **B**ehind the position
Rep_{ϕ_3}	**Rep**lace it with another word $\hat{w}$
DC	**D**elete the **C**urrent word
DF	**D**elete the word in **F**ront of the position
DB	**D**elete the word **B**ehind the position
Skip	Do not change anything

Wu, C., Ren, X., Luo, F. and Sun, X. A Hierarchical Reinforced Sequence Operation Method for Unsupervised Text Style Transfer. In *ACL*, 2019.

RL for Edit

- **Reward** (one-step rollout for training)
  - High-level: selecting the word to edit

$$R_{\text{style}} = \lambda_{\text{style}} \left[p(s_2 | \hat{\boldsymbol{x}}_2) - p(s_2 | \boldsymbol{x}_1) \right]$$

$\hat{\boldsymbol{x}}_2$: transfer candidate $\boldsymbol{x}_1$: Input

[Encouraging a larger change of sentiment]

Pretrained by attention-based style classifier

  - Low-level:
    - Action prediction (policy not learned)
    - Candidate word
      - Insertion: $R_{\text{lm}} + R_{\text{conf}}$
      - Replacement: $R_{\text{lm}} + R_{\text{conf}} + R_{\text{rec}}$

Wu, C., Ren, X., Luo, F. and Sun, X. A Hierarchical Reinforced Sequence Operation Method for Unsupervised Text Style Transfer. In *ACL*, 2019.

RL for Edit

- **Inference:** Search towards the objective

$$\mathrm{LM}_2(\hat{\boldsymbol{x}}_2) \cdot p(s_2|\hat{\boldsymbol{x}}_2)^\eta$$

  - Sample position and, if needed, a candidate word by the learned policy
  - Sample operator uniformly

Loop until the stopping criterion is satisfied

Wu, C., Ren, X., Luo, F. and Sun, X. A Hierarchical Reinforced Sequence Operation Method for Unsupervised Text Style Transfer. In *ACL*, 2019.

DualRL

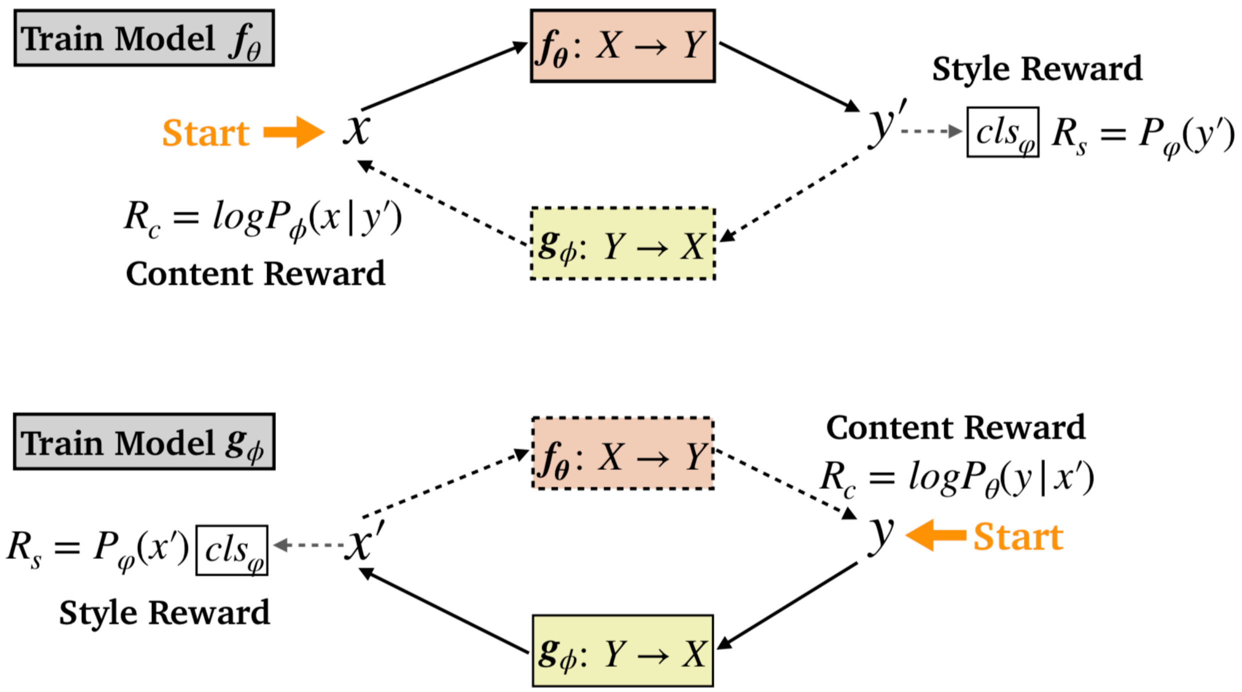
- **Idea:** Deal with output sentence directly

- Style reward $R_s = P(s_y | \boldsymbol{y'}; \varphi)$
- Content reward $R_c = P(\boldsymbol{x} | \boldsymbol{y'}; \boldsymbol{\phi})$

- Overall reward

$$R = (1 + \beta^2) \frac{R_c \cdot R_s}{(\beta^2 \cdot R_c) + R_s}$$

- Then, train a Seq2Seq model



Luo F, Li P, Zhou J, Yang P, Chang B, Sui Z, Sun X. A Dual Reinforcement Learning Framework for Unsupervised Text Style Transfer. *IJCAI*, 2019.

DualRL

- **Idea:** Deal with output sentence directly

- Cold start problem
  - Train a template-based baseline [Li *et al.*, 2018]
  - Experience replay of the last model snapshot

Algorithm 2 The annealing pseudo teacher-forcing algorithm for dual reinforcement learning.

Initialize the iteration interval p
for each iteration $i = 1, 2, ..., M$ **do**
  ▷ *Start to train model* $\boldsymbol{f_\theta}$
  Update parameter $\boldsymbol{\theta}$ via RL based on Eq. 4
  if $i \ \% \ p = 0$ **then** ▷ *Pseudo Teacher-Forcing*
    Generate a pair of data $(\boldsymbol{x}_i', \boldsymbol{y}_i)$, where $\boldsymbol{x}_i' = \boldsymbol{g}(\boldsymbol{y}_i)$
    Update $\boldsymbol{\theta}$ using data $(\boldsymbol{x}_i', \boldsymbol{y}_i)$ via MLE
  end if
  ▷ *Start to train model* $\boldsymbol{g_\phi}$
  Update parameter $\boldsymbol{\phi}$ via RL similar to Eq. 4
  if $i \ \% \ p = 0$ **then** ▷ *Pseudo Teacher-Forcing*
    Generate a pair of data $(\boldsymbol{y}_i', \boldsymbol{x}_i)$, where $\boldsymbol{y}_i' = \boldsymbol{f}(\boldsymbol{x}_i)$
    Update $\boldsymbol{\phi}$ using data $(\boldsymbol{y}_i', \boldsymbol{x}_i)$ via MLE
  end if
  Exponential increase in p based on Eq. 5
end for

Luo F, Li P, Zhou J, Yang P, Chang B, Sui Z, Sun X. A Dual Reinforcement Learning Framework for Unsupervised Text Style Transfer. *IJCAI*, 2019.

Summary so-far

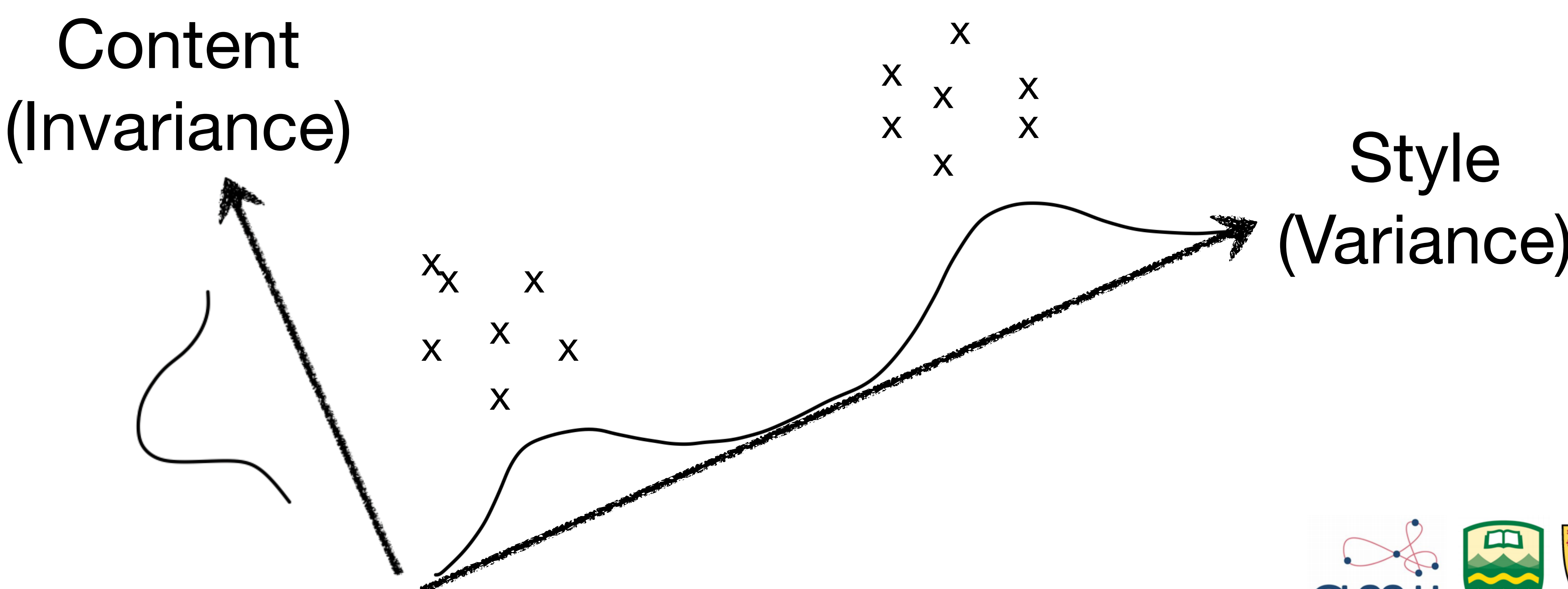
Model	Style treatment	Content Treatment
Hu et al. [2017]	Style classification	—
Cross-alignment [Shen et al. 2017]	Style embedding	Adv training based on style-transferred hidden states
Fu et al. [2018]	Style embedding Style-specific decoder	Adv training
Disentangling [John+'19; Bao+'19]	Style classification + Content adversarial	Content adversarial + Style classification
CycleRL [Xu+2018]	Delete style words + Multi-decoder	Content words for reconstruction Cycle Consistency for extractor
Del-Retr-Gen [Li et al., 2018]	Delete style phrases +Retrieve for target style	Content words for reconstruction
RL-Edit [Wu et al., 2019]	Search obj $\mathrm{LM}_2(\hat{\boldsymbol{x}}_2) \cdot p(s_2 \mid \hat{\boldsymbol{x}}_2)^\eta$	Training reward of reconstruction $R_{\mathrm{lm}} + R_{\mathrm{conf}} + R_{\mathrm{rec}}$
Dual RL [Luo et al., 2019]	Style reward $R_s = P(s_y \mid \boldsymbol{y}'; \varphi)$	Content reward $R_c = P(\boldsymbol{x} \mid \boldsymbol{y}'; \boldsymbol{\phi})$

Unsupervised
Style-Transfer Generation

Settings

- **Unsupervised supervision**
  - In the training phase, we have unlabeled corpus

$$\{\mathbf{x}^{(m)}\}_{m=1}^{M}$$

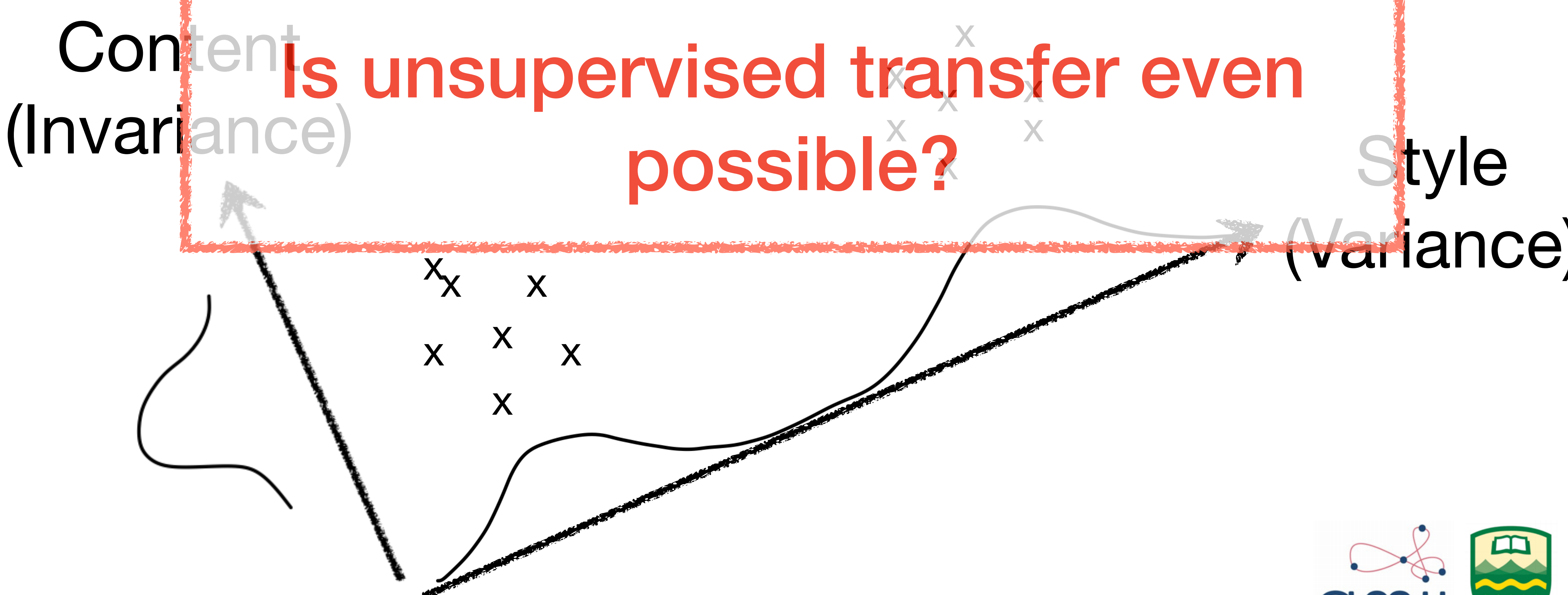


Settings

- **Unsupervised supervision**
  - In the training phase, we have unlabeled corpus

$$\{\mathbf{x}^{(m)}\}_{m=1}^{M}$$

Is unsupervised transfer even possible?



Unsupervised Disentanglement

- ***β*-VAE** $\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - \beta\, D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z}))$

with $\beta > 1$. (If $\beta = 1$, then standard VAE)

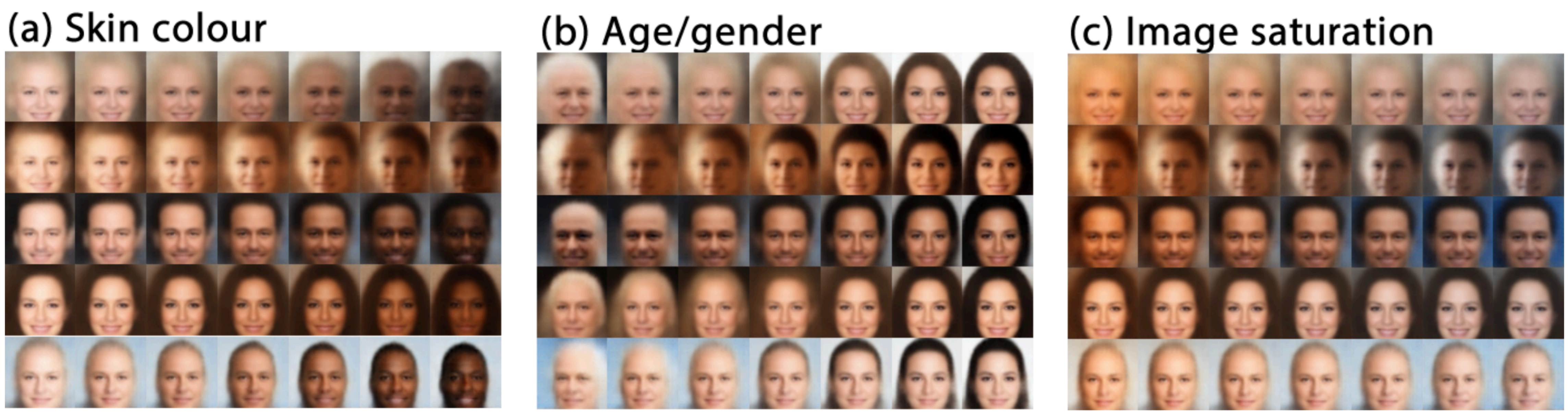
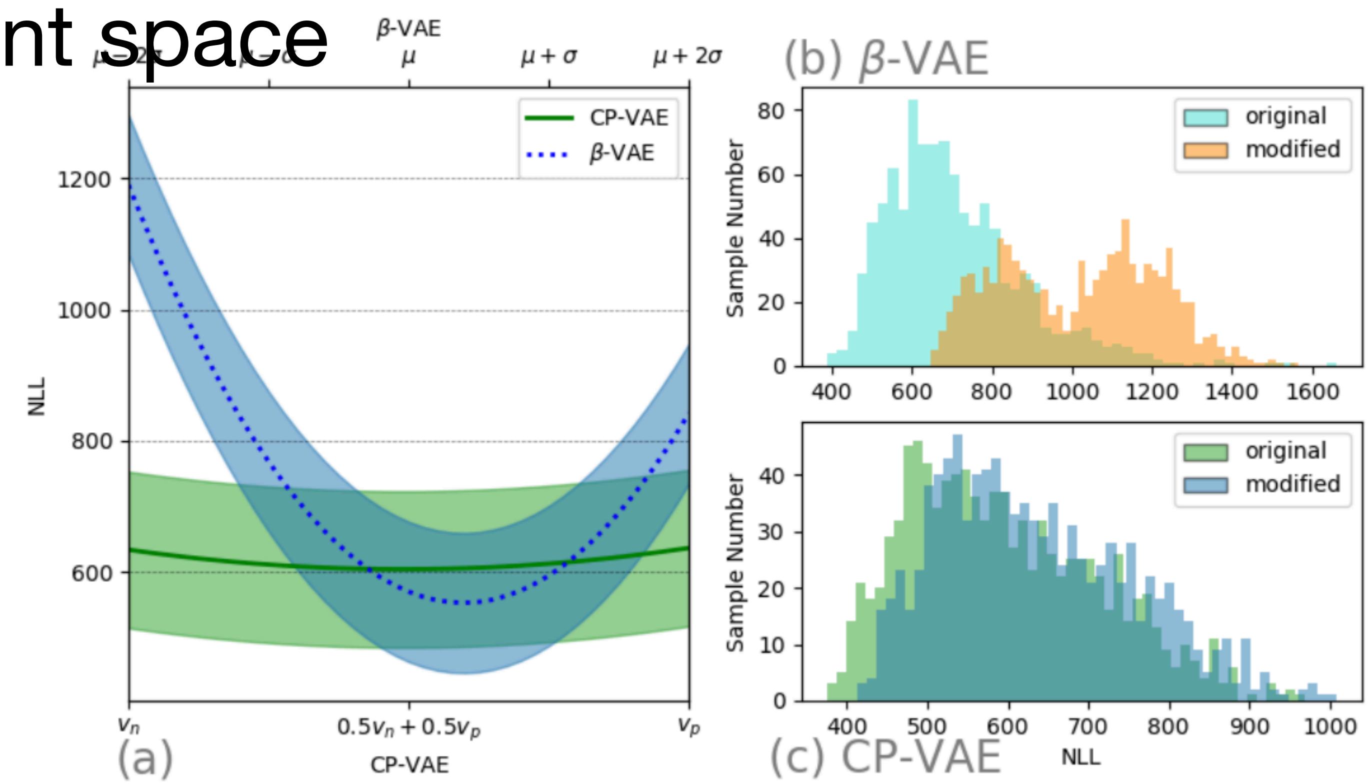


Figure 4: **Latent factors learnt by β-VAE on celebA:** traversal of individual latents demonstrates that β-VAE discovered in an unsupervised manner factors that encode skin colour, transition from an elderly male to younger female, and image saturation.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S. and Lerchner, A. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. *ICLR*, 2017.

Unsupervised Disentanglement

- **β-VAE for NLP [Xu et al., 2019]**
  - Successfully detecting a latent dimension responsible for sentiment with 90+% accuracy
  - Naïve flipping this dimension does not work
  - Hypothesis: vacancy in latent space



Xu, P., Cao, Y. and Cheung, J.C.K., 2019. Unsupervised Controllable Text Generation with Global Variation Discovery and Disentanglement. *arXiv preprint arXiv:1905.11975*, 2019.

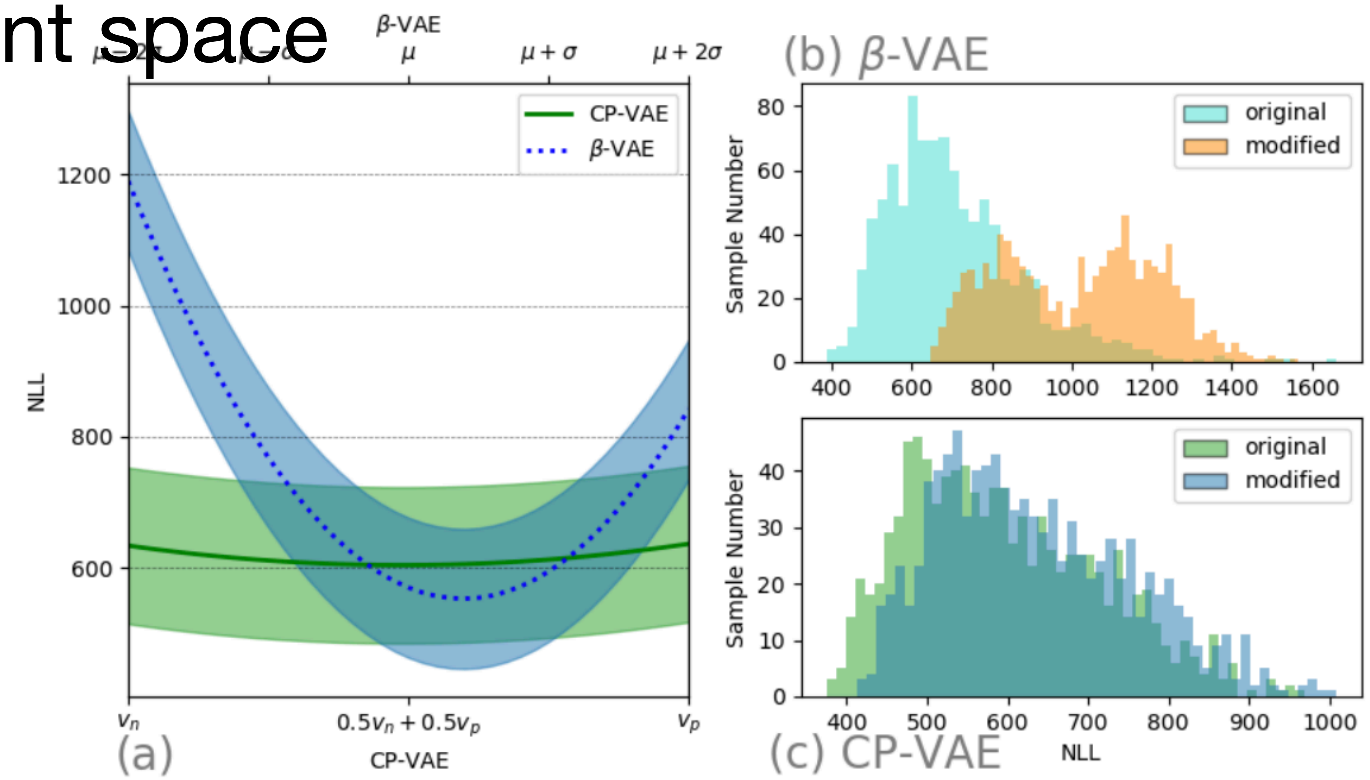
Unsupervised Disentanglement

- **β-VAE for NLP [Xu et al., 2019]**
  - Successfully detecting a latent dimension responsible for sentiment with 90+% accuracy
  - Naïve flipping this dimension does not work

Xu, P., Cao, Y. and Cheung, J.C.K., 2019. Unsupervised Controllable Text Generation with Global Variation Discovery and Disentanglement. *arXiv preprint arXiv:1905.11975*, 2019.

Unsupervised Disentanglement

- **β-VAE for NLP [Xu et al., 2019]**
  - Successfully detecting a latent dimension responsible for sentiment with 90+% accuracy
  - Naïve flipping this dimension does not work
  - Hypothesis: vacancy in latent space



Xu, P., Cao, Y. and Cheung, J.C.K., 2019. Unsupervised Controllable Text Generation with Global Variation Discovery and Disentanglement. *arXiv preprint arXiv:1905.11975*, 2019.

Unsupervised Disentanglement

- **Filling the latent vacancy**
  - Encoding the latent vector in a k-dimensional subspace (e.g., $k = 3$)

$$\boldsymbol{\mu} = \sum_{i=1}^{K} p_i \boldsymbol{e}_i, \quad \sum_{i=1}^{K} p_i = 1, \quad \langle \boldsymbol{e}_i, \boldsymbol{e}_j \rangle = 0, i \neq j, \quad K \leq N$$

  - with soft-penalized orthonormal basis

$$\mathcal{L}_{\mathrm{REG}}(\boldsymbol{x}; \boldsymbol{\phi}_1) = \|\boldsymbol{E}^\top \boldsymbol{E} - \alpha \boldsymbol{I}\|$$

Xu, P., Cao, Y. and Cheung, J.C.K., 2019. Unsupervised Controllable Text Generation with Global Variation Discovery and Disentanglement. *arXiv preprint arXiv:1905.11975*, 2019.

Unsupervised Disentanglement

- **Filling the latent vacancy**
  - Confining latent vectors

N-dimensional space $\Longrightarrow$ k-dimensional simplex

$$\boldsymbol{\mu} = \sum_{i=1}^{K} p_i \boldsymbol{e}_i, \quad \sum_{i=1}^{K} p_i = 1, \quad \langle \boldsymbol{e}_i, \boldsymbol{e}_j \rangle = 0, i \neq j, \quad K \leq N$$

  - Stretching over the simplex

$$\mathcal{L}_{\text{S-REC}}(\boldsymbol{x}; \boldsymbol{\phi}_1) = \mathbb{E}_{\boldsymbol{z}^{(1)} \sim q_{\boldsymbol{\phi}_1}(\boldsymbol{z}^{(1)}|\boldsymbol{x})} \left[\frac{1}{m} \sum_{i=1}^{m} \max(0, 1 - \boldsymbol{h} \cdot \boldsymbol{\mu} + \boldsymbol{h} \cdot \boldsymbol{\mu}_i^{(-)}) \right]$$

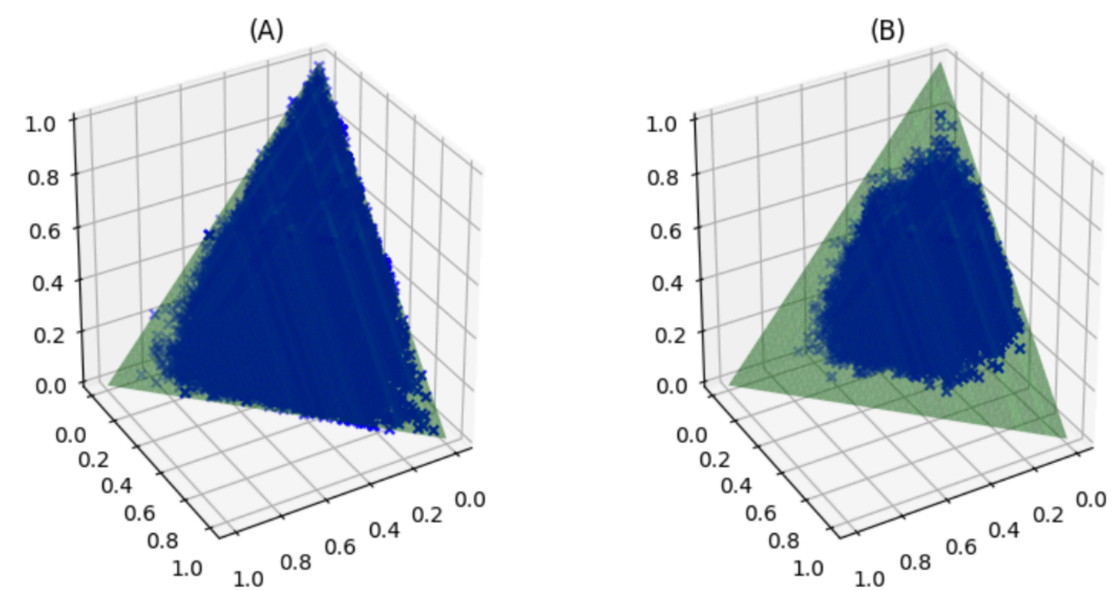
Xu, P., Cao, Y. and Cheung, J.C.K., 2019. Unsupervised Controllable Text Generation with Global Variation Discovery and Disentanglement. *ICML,* 2020.

Unsupervised Disentanglement

- **Filling the latent vacancy**
  - Loss $$\mathcal{L}(\boldsymbol{x};\boldsymbol{\theta},\boldsymbol{\phi}) = \mathcal{L}_{\text{VAE}} + \mathcal{L}_{\text{REG}} + \mathcal{L}_{\text{S-REC}}$$

$$\mathcal{L}_{\text{REG}}(\boldsymbol{x};\boldsymbol{\phi}_1) = \|\boldsymbol{E}^\top\boldsymbol{E} - \alpha\boldsymbol{I}\|$$

$$\mathcal{L}_{\text{S-REC}}(\boldsymbol{x};\boldsymbol{\phi}_1) = \mathbb{E}_{\boldsymbol{z}^{(1)}\sim q_{\boldsymbol{\phi}_1}(\boldsymbol{z}^{(1)}|\boldsymbol{x})}\left[\frac{1}{m}\sum_{i=1}^{m}\max(0, 1 - \boldsymbol{h}\cdot\boldsymbol{\mu} + \boldsymbol{h}\cdot\boldsymbol{\mu}_i^{(-)})\right]$$



Xu, P., Cao, Y. and Cheung, J.C.K., 2019. Unsupervised Controllable Text Generation with Global Variation Discovery and Disentanglement. *ICML,* 2020.

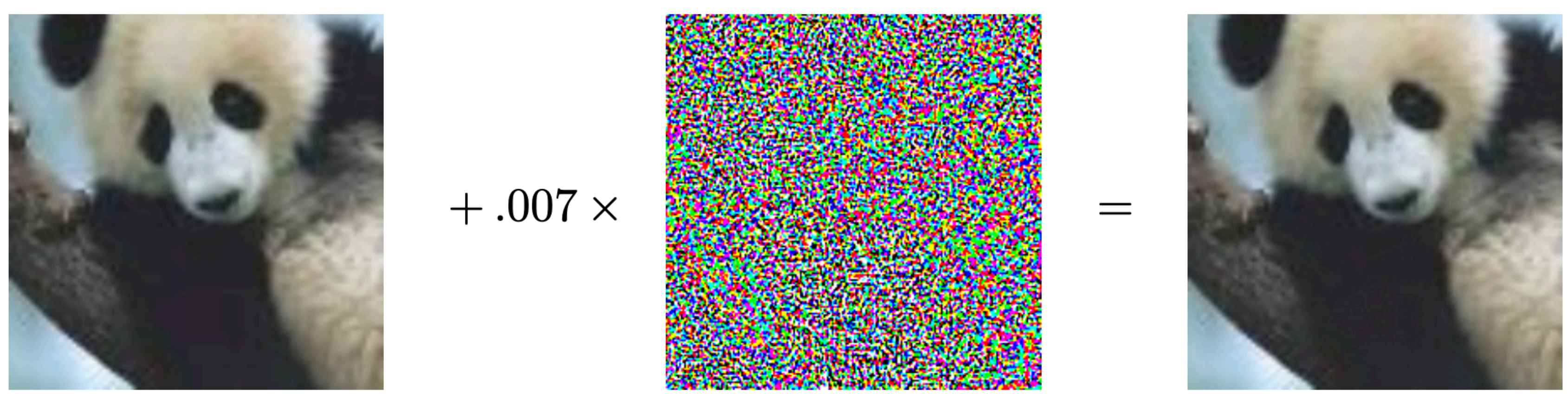
Tutorial Outline

- Introduction
- Style-conditioned text generation
- Style-transfer text generation
- **Style-adversarial text generation**
  - **Character-level attack**
  - **Sentence-level attack**
  - **Word-level attack**
- Conclusion

Adversarial Attack

Task

- “Slightly” change the data, but
- Drastically change a machine learning model’s predictor



$\boldsymbol{x}$
“panda”
57.7% confidence

$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
“nematode”
8.2% confidence

$\boldsymbol{x} +$
$\epsilon\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
“gibbon”
99.3 % confidence

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. and Fergus, R., 2013. Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199.

Adversarial Attacks in Text

- Still to fool a classifier (e.g., some style)
- Need to relax the constraint of being imperceivable
  - Pend additional sentences/phrases
  - Allow typos
  - Allow word changes

Xu, P., Cao, Y. and Cheung, J.C.K., 2019. Unsupervised Controllable Text Generation with Global Variation Discovery and Disentanglement. *ICML,* 2020.

Comparison: style transfer and adversarial attacks

Task	Model Prediction	Human Perception
Text Style Transfer	Changed	**Changed**
Adversarial Attack	Changed	**Not Changed**

Categorization of Adversarial Attacks in NLP

- Sentence-level
- Word-level
- Character-level

Sentence-Level Attacks

- ADDSENT [Jia+2017]
  - Fool a machine reading model by adding one additional sentence to the original texts.
  - requires human engineering
- Experiments on machine comprehension

(strictly speaking: not **style** adversarial)

Article: Super Bowl 50
Paragraph: "*Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV.*"
Question: "*What is the name of the quarterback who was 38 in Super Bowl XXXIII?*"
Original Prediction: John Elway
Prediction under adversary: Jeff Dean

Jia, R. and Liang, P., Adversarial examples for evaluating reading comprehension systems. In EMNLP, 2017.

Character-level attack

- Add, delete, or swap characters
  - Hotflip [Ebrahimi+2018]
  - TEXTBUGGER [Li+2019]

HotFlip

- Gradient-based
- J(x, y) is the loss of model on input x with true output y
- Represent character sequence as
  - $x = [(x_{11}, .., x_{1n}); .. (x_{m1}, ..., x_{mn})], x_{ij} \in \{0,1\}^{|V|}$

South Africa's historic Soweto township marks its 100th birthday on Tuesday in a mood of optimism. 57% **World**
South Africa's historic Soweto township marks its 100th birthday on Tuesday in a moo**P** of optimism. 95% **Sci/Tech**
Chancellor Gordon Brown has sought to quell speculation over who should run the Labour Party and turned the attack on the opposition Conservatives. 75% **World**
Chancellor Gordon Brown has sought to quell speculation over who should run the Labour Party and turned the attack on the o**B**position Conservatives. 94% **Business**

Adversarial examples with a single character change [ebrahimi+2018]

HotFlip

- A **flip** of the j-th character of the i-th word (a→b):
  - $\overrightarrow{v}_{ijb} = (\overrightarrow{0}, \ldots; (\overrightarrow{0}, \ldots (0,.. -1,0,..,1,0)_j, \ldots \overrightarrow{0})_i; \overrightarrow{0}, \ldots)$
  - $x_{ij}^{(a)} = 1$
- Choose the vector with biggest increase in loss
  - $\max \nabla_x J(\mathbf{x}, \mathbf{y})^T \cdot \overrightarrow{v}_{ijb} = \max_{ijb}\left(\frac{\partial J}{\partial x_{ij}}^{(b)} - \frac{\partial J}{\partial x_{ij}}^{(a)}\right)$
- First-order approximation of change in loss
  - $\nabla_{\overrightarrow{v}_{ijb}} J(\mathbf{x}, \mathbf{y}) = \nabla_x J(\mathbf{x}, \mathbf{y})^T \cdot \overrightarrow{v}_{ijb}$
- Character **insertion/deletion** can be treated as a sequence of flips, as characters are shifted to the right/left until the end of the word

Word-Level Attacks

- Add, delete, or swap words in original texts
  - Metropolis-Hastings attack [Zhang+2018] (insert+delete+swap)
  - Universal triggers [Wallace+2019] (insert)

Metropolis-Hastings Attack

- Metropolis-Hastings Algorithm
  - Given the stationary distribution $\pi(x)$ and transition proposal, M-H is able to generate desirable examples from $\pi(x)$
  - A proposal to jump from x to x' is made on the proposal distribution $g(x \mid x')$
  - Proposal acceptance rate:
    - $\alpha(x' \mid x) = \min\{\frac{\pi(x')g(x \mid x')}{\pi(x)g(x' \mid x)}\}$

Metropolis-Hastings Attack

- stationary distribution
  - $\pi(x \mid \tilde{y}) \propto LM(x) \cdot C(\tilde{y} \mid x)$
- Transition proposal
  - Replacement: $$T_r^B(x'|x) = \mathcal{I}\{w^c \in \mathcal{Q}\} \cdot \frac{\pi(w_1, \cdots, w_{m-1}, w^c, w_{m+1}, \cdots, w_n|\tilde{y})}{\sum_{w \in \mathcal{Q}} \pi(w_1, \cdots, w_{m-1}, w, w_{m+1}, \cdots, w_n|\tilde{y})} \quad (3)$$
  - Insertion: insert a random word into the position and then performing replacement
  - deletion: $T_d^B(x' \mid x) = 1$ if $x' = x_{-m}$, where is the sentence after deleting the m-th word, otherwise $T_d^B(x' \mid x) = 0$

Universal Adversarial Triggers

Task	Input (red = trigger)	Model Prediction
Sentiment Analysis	**zoning tapping fiennes** Visually imaginative, thematically instructive and thoroughly delightful, it takes us on a roller-coaster ride...	Positive $\rightarrow$ Negative
	zoning tapping fiennes As surreal as a dream and as detailed as a photograph, as visually dexterous as it is at times imaginatively overwhelming.	Positive $\rightarrow$ Negative

- Universal Attack [Wallace+2019]: the same trigger sequence prepended to every input in the dataset
  - No need to access the model at test time
  - Lower the barrier for an adversary: trigger can be distributed to anyone
- Often transfer across models [Moosavi-Dezfooli+2017]

Universal Adversarial Triggers

- Given a model f, a text input of tokens t, and a target label $\tilde{y}$, the attack aims to concatenate trigger tokens t_{adv} to the front or end of t, such that $f(\mathrm{t}_{adv}; \mathrm{t}) = \tilde{y}$
- Minimize loss for target class $\tilde{y}$ for ***all inputs***
  - $$\arg\min_{\mathbf{t}_{adv}} \mathbb{E}_{\mathbf{t}\sim\mathcal{T}}\left[\mathcal{L}(\tilde{y}, f(\mathbf{t}_{adv}; \mathbf{t}))\right]$$

Conclusion

Topics:

- Style-conditional generation
  - Generate a sentence in a given style
- Style-transfer generation
  - Change a style but keep the content
- Style-adversarial generation
  - Keep the style, but fool the style classifier

Conclusion

Techniques related to stylized text generation

- Variational auto encoder
  - Learning a smooth latent space (good for sampling, manipulation)
- Adversarial training
  - Matching two distributions by empirical samples
- Reinforcement learning
  - Learning with discrete actions

Future Work

Related tasks

- Syntactically controlling
- Text summarization
- Text simplification
- etc.

Future Work

Fundamental machine learning problems

- Disentangling latent space
- Effect search/learning in the word space

Thank you for listening!

Q&A

Acknowledgments

Lili Mou is supported by AltaML, Amii Fellow Program, and Canadian CIFAR AI Chair Program. He is also supported by NSERC DG RGPIN-2020-04465.

References

- Bao, Y., Zhou, H., Huang, S., Li, L., Mou, L., Vechtomova, O., Dai, X. and Chen, J. Generating sentences from disentangled syntactic and semantic spaces. In *ACL*, 2019.
- Biber, D., Conrad, S., *Register, Genre, and Style*. Cambridge University Press, 2009.
- Fu, Z., Tan, X., Peng, N., Zhao, D. and Yan, R. Style transfer in text: Exploration and evaluation. In *AAAI*, 2018.
- Gatys, Leon A., Alexander S. Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *CVPR*, 2016.
- Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S. and Lerchner, A. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. *ICLR*, 2017.
- Hu, Z, Yang, Z, Liang, X, Salakhutdinov, R, Xing, EP. Toward controlled generation of text. In *ICML*, 2017.
- Lample, G., Subramanian, S., Smith, E., Denoyer, L., Ranzato, M.A. and Boureau, Y.L. Multiple-attribute text rewriting. In *ICLR*, 2019.
- John, V., Mou, L., Bahuleyan, H. and Vechtomova, O. Disentangled representation learning for text style transfer. In *ACL*, 2018.
- Kingma, D.P. and Welling, M., 2013. Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*.
- Rao, S., Tetreault, J. Dear Sir or Madam, May I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer. In *NAACL-HLT*, 2018.

References

- Shen, T., Lei, T., Barzilay, R. and Jaakkola, T. Style transfer from non-parallel text by cross-alignment. In *NIPS*, 2017.
- Subramanian, S., Lample, G., Smith, E.M., Denoyer, L., Ranzato, M.A. and Boureau, Y.L., 2018. Multiple-attribute text style transfer. In *ICLR*, 2018.
- Xu, P., Cao, Y. and Cheung, J.C.K., 2019. Unsupervised Controllable Text Generation with Global Variation Discovery and Disentanglement. *ICML,* 2020.
- Xu, J., Sun, X., Zeng, Q., Ren, X., Zhang, X., Wang, H. and Li, W. Unpaired sentiment-to-sentiment translation: A cycled reinforcement learning approach. In *ACL*, 2018.
- Xu, W., Ritter, A., Dolan, B., Grishman, R. and Cherry, C. Paraphrasing for style. In *COLING*, 2012.
- Wang, Y., Wu, Y., Mou, L., Li, Z. and Chao, W. Harnessing Pre-Trained Neural Networks with Rules for Formality Style Transfer. In *EMNLP-IJCNLP*. 2019.
- Li, J., Jia, R., He, H. and Liang, P. Delete, retrieve, generate: A simple approach to sentiment and style transfer. In *NAACL-HLT*, 2018.
- Luo F, Li P, Zhou J, Yang P, Chang B, Sui Z, Sun X. A dual reinforcement learning framework for unsupervised text style transfer. IJCAI, 2019.
- Wu, C., Ren, X., Luo, F. and Sun, X. A Hierarchical Reinforced Sequence Operation Method for Unsupervised Text Style Transfer. In *ACL*, 2019.
- Zhao, J., Kim, Y., Zhang, K., Rush, A.M. and LeCun, Y. Adversarially regularized autoencoders. In *ICML*, 2018.

References

- Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. and Fergus, R., 2013. Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199.
- Moosavi-Dezfooli, S.M., Fawzi, A. and Frossard, P.. Deepfool: a simple and accurate method to fool deep neural networks. In CVPR, 2016.
- Li, J., Chen, X., Hovy, E. and Jurafsky, D. Visualizing and understanding neural models in nlp. In ACL, 2016.
- Li, J., Monroe, W. and Jurafsky, D. Understanding neural networks through representation erasure. arXiv preprint arXiv:1612.08220.
- Ren, S., Deng, Y., He, K. and Che, W. Generating natural language adversarial examples through probability weighted word saliency. In ACL, 2019.
- Moosavi-Dezfooli, S.M., Fawzi, A., Fawzi, O. and Frossard, P. Universal adversarial perturbations. In CVPR, 2017.
- Jia, R. and Liang, P., Adversarial examples for evaluating reading comprehension systems. In EMNLP, 2017.
- Ebrahimi, J., Rao, A., Lowd, D. and Dou, D., 2017. Hotflip: White-box adversarial examples for text classification. In ACL, 2018.
- Zhang, H., Zhou, H., Miao, N. and Li, L., 2019, July. Generating Fluent Adversarial Examples for Natural Languages. In ACL, 2019.
- Wallace, E., Feng, S., Kandpal, N., Gardner, M. and Singh, S. Universal Adversarial Triggers for Attacking and Analyzing NLP. In EMNLP, 2019.
- Li, J., Ji, S., Du, T., Li, B. and Wang, T., Textbugger: Generating adversarial text against real-world applications. In NDSS, 2019.