TEXT STYLE TRANSFER IN NLP

S. SHREYA
2018CS10377
COV884

Slides borrowed from Lili Mou a
Olga Vechtomova - ACL 2020
Tutorial on Stylized Text Generation

OVERVIEW

- Task Definition
- Evaluation
- Methods on Parallel Data
- Methods on Non-parallel Data
- Reviews

Task description

- **Input:**
  - A source sentence $\mathbf{x} = x_1x_2x_3..x_n$
  - The desired style
- **Output**: A "style-transferred" sentence $\mathbf{y} = y_1y_2y_3..y_m$
- **Requirement:** y is in the desired style
- Usually x and y are different in **style**
- x and 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 .

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

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 .

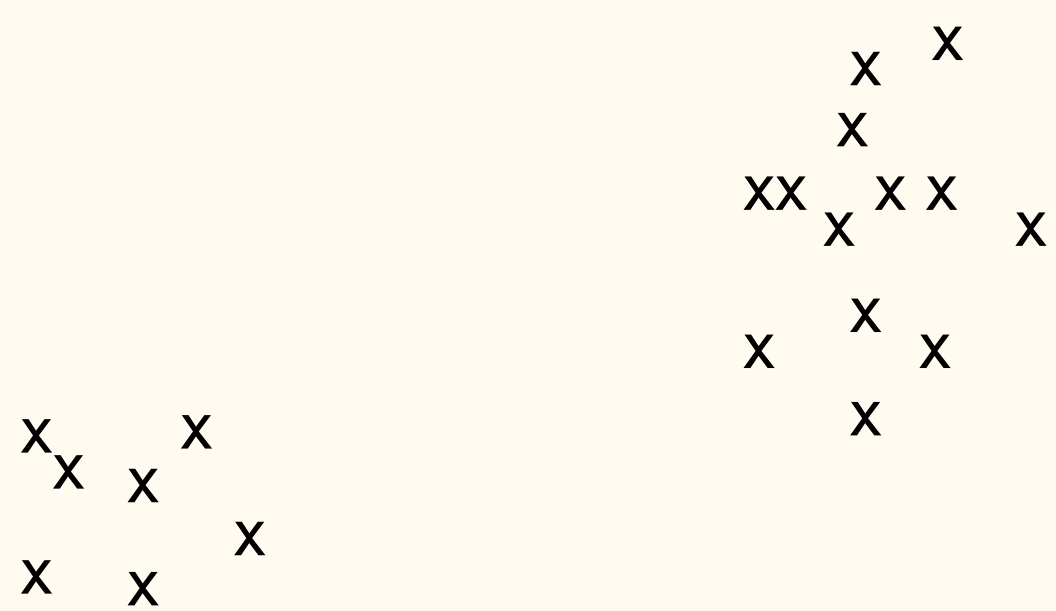
What is “style” or “content”?

More debates

Is “sentiment information” the style or content?

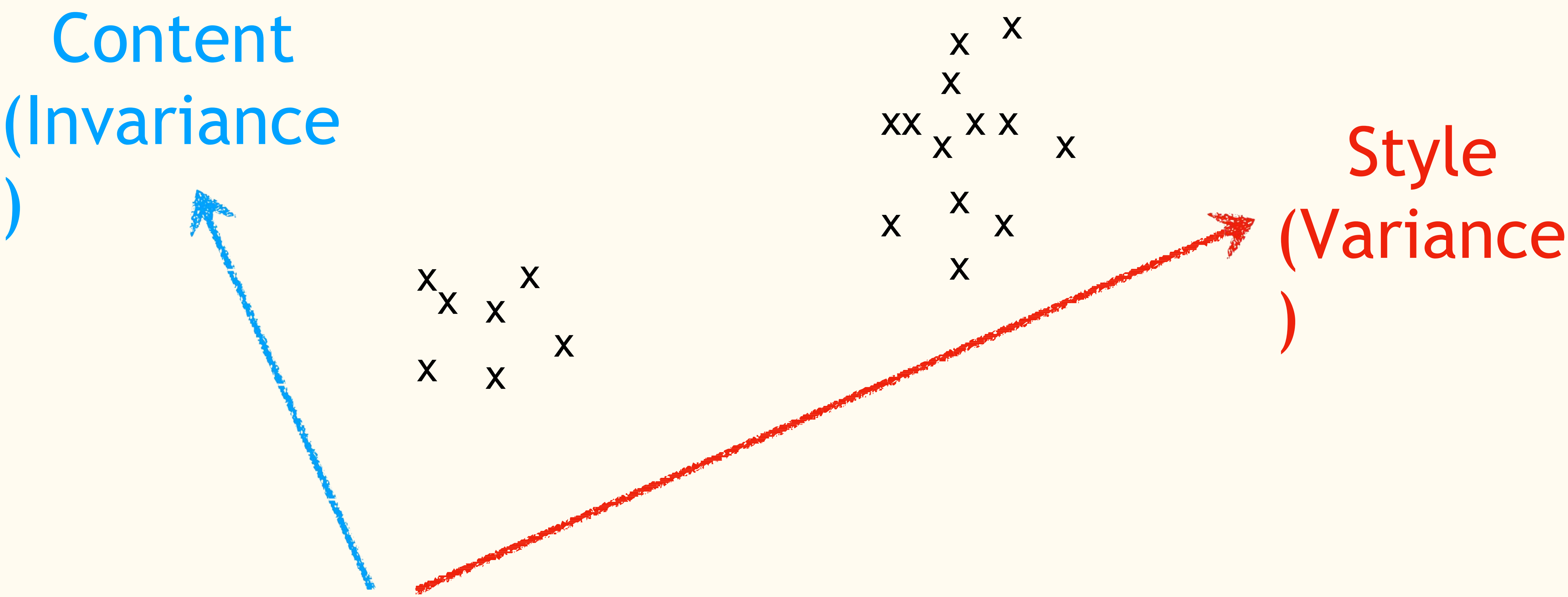
What is “style” or “content”?

An empirical perspective



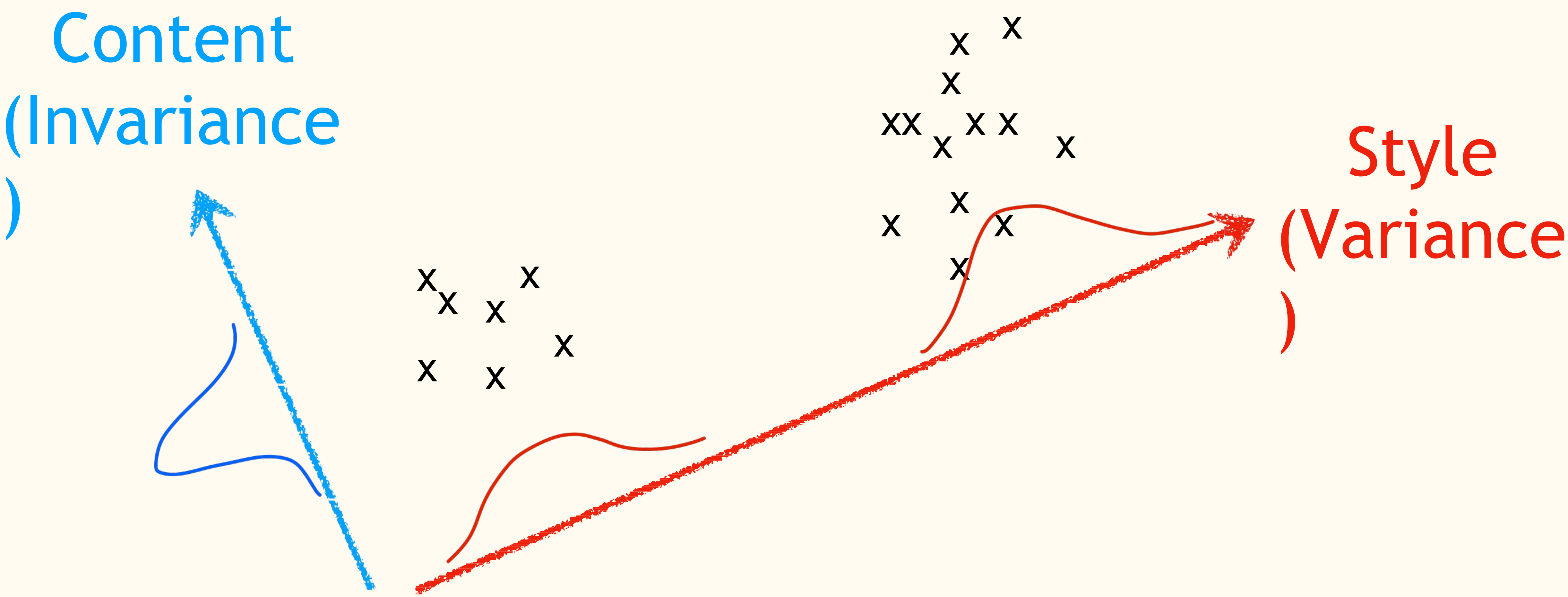
What is “style” or “content”?

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 ?
⇒ Music	what is your favorite sitcom with adam sandler ?
⇒ Politics	what is an event with black people ?
Science	take 1ml of hcl (concentrated) and dilute it to 50ml .
⇒ Music	take em to you and shout it to me
⇒ Politics	take bribes to islam and it will be punished .
Science	just multiply the numerator of one fraction by that of the other .
⇒ Music	just multiply the fraction of the other one that 's just like it .
⇒ 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 presentation, we call
  - Variance = style
  - Invariance = content

Settings

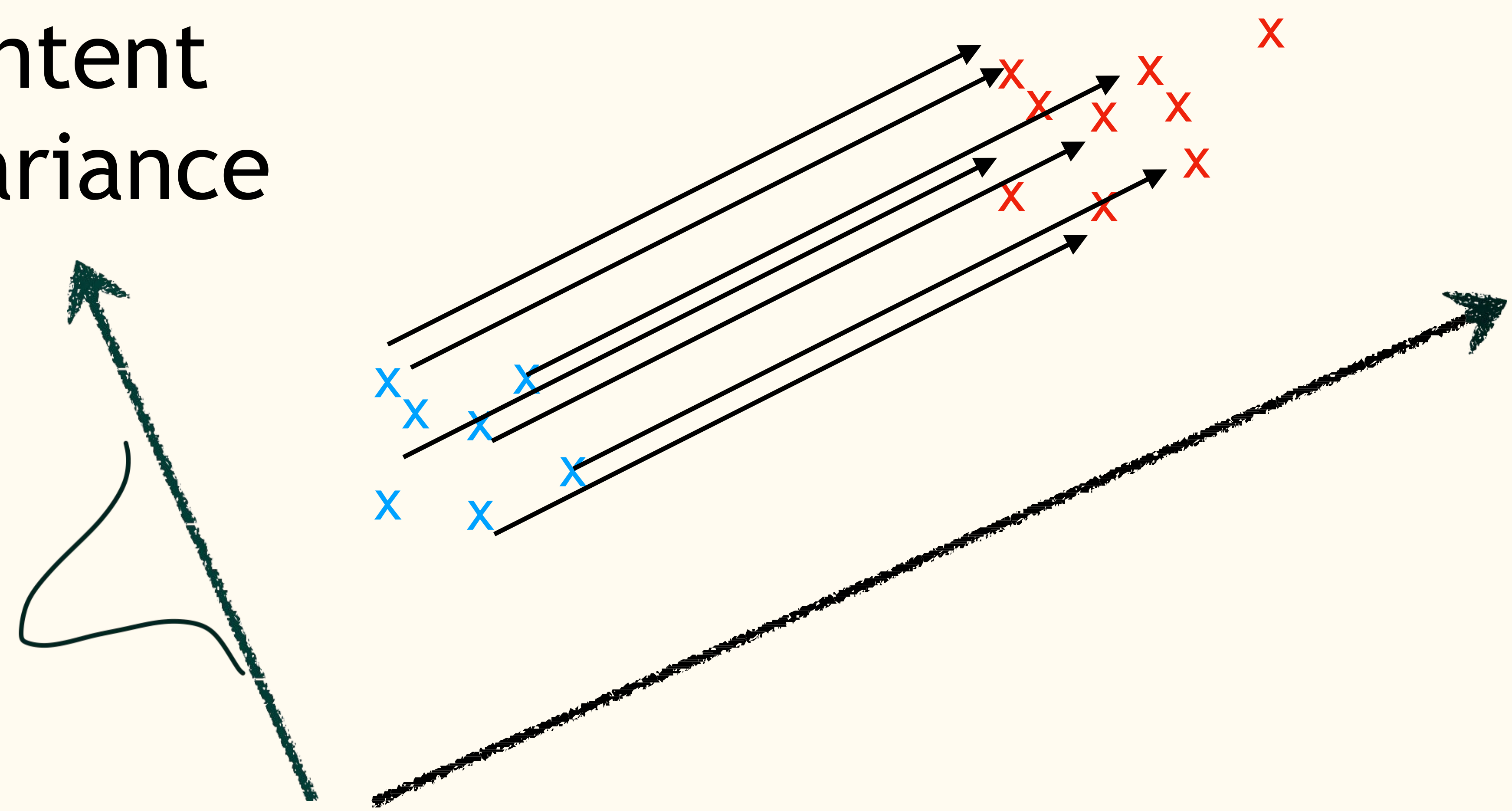
- Seq2seq supervision
- Non-parallel supervision
- Unsupervised

Settings

- **Parallel supervision**

- In the training phase, we have parallel corpus of

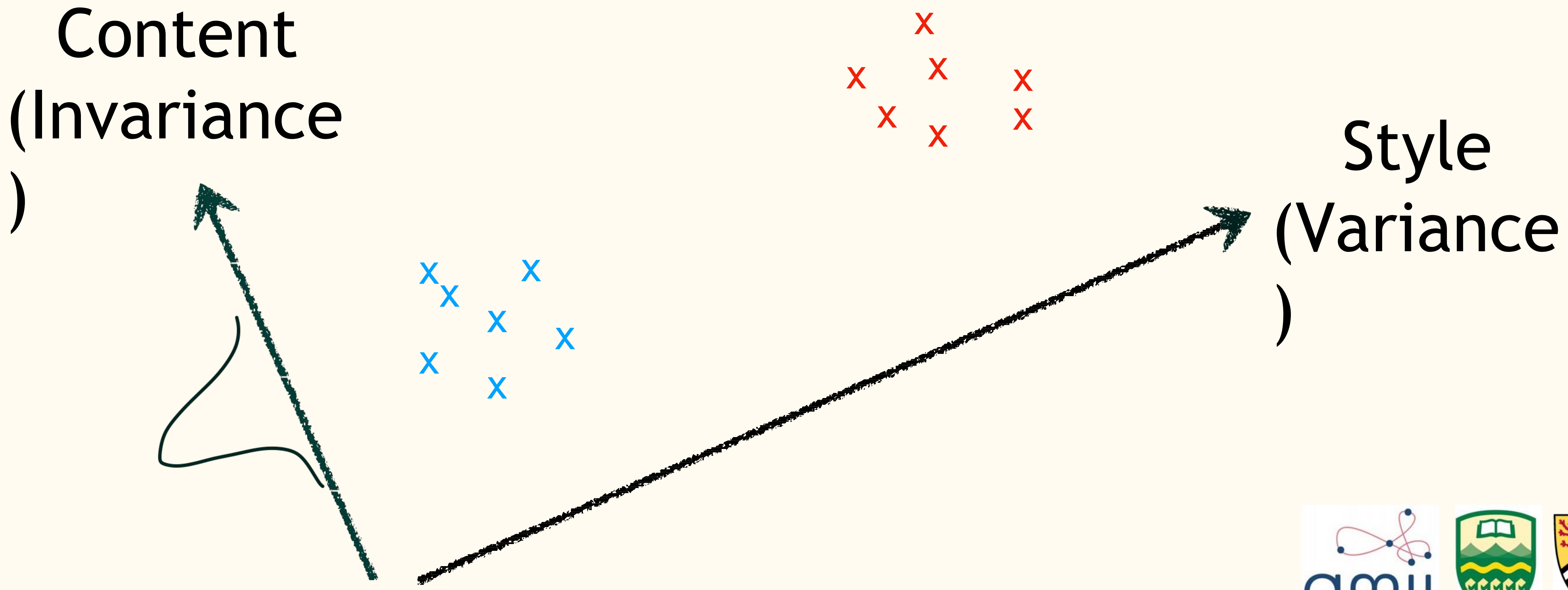
$$\{\mathbf{x}^{(m)}, \mathbf{y}^{(m)}, s^{(m)}\}_{m=1}^{M}$$



Settings

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

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

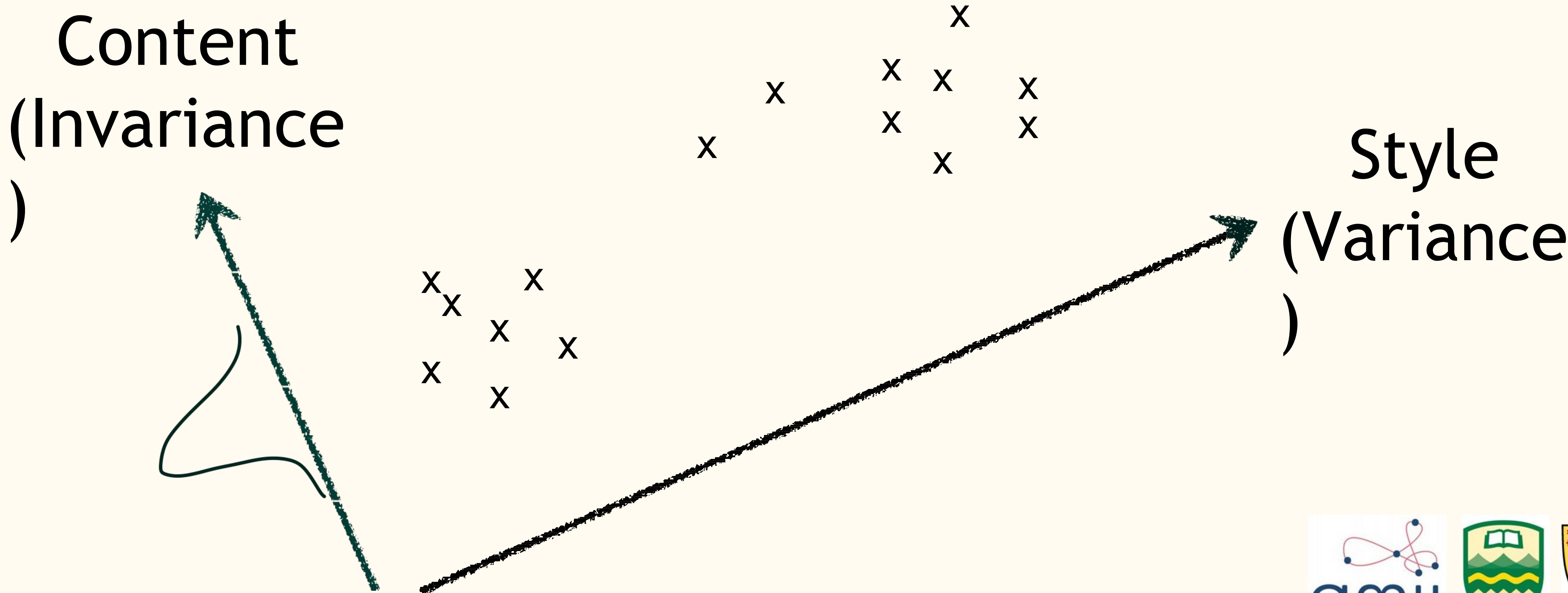


Settings

- **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

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 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.

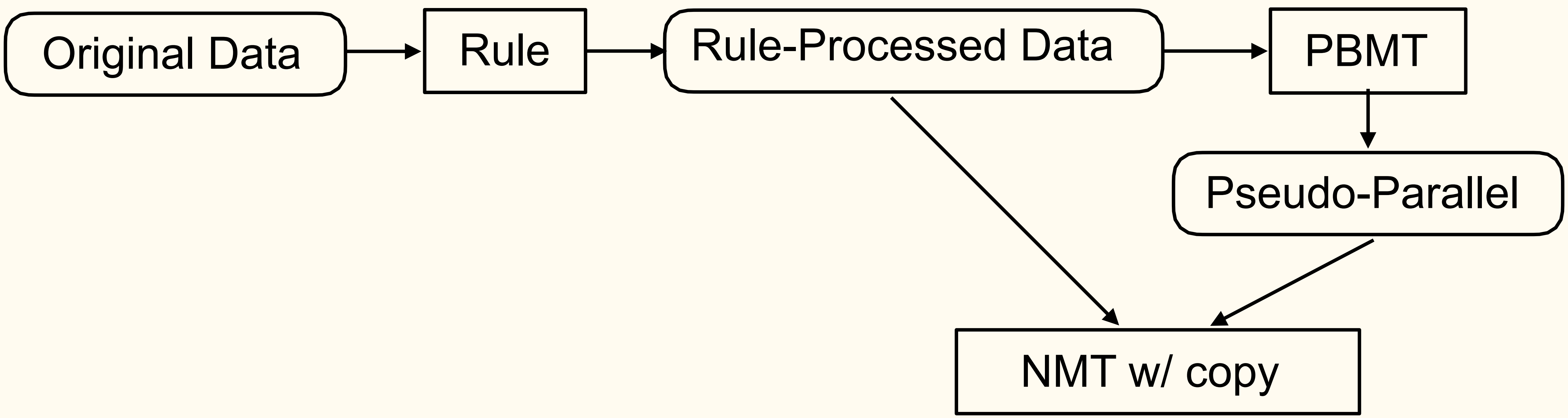
Approaches

- 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 => Târget
  - Target => 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

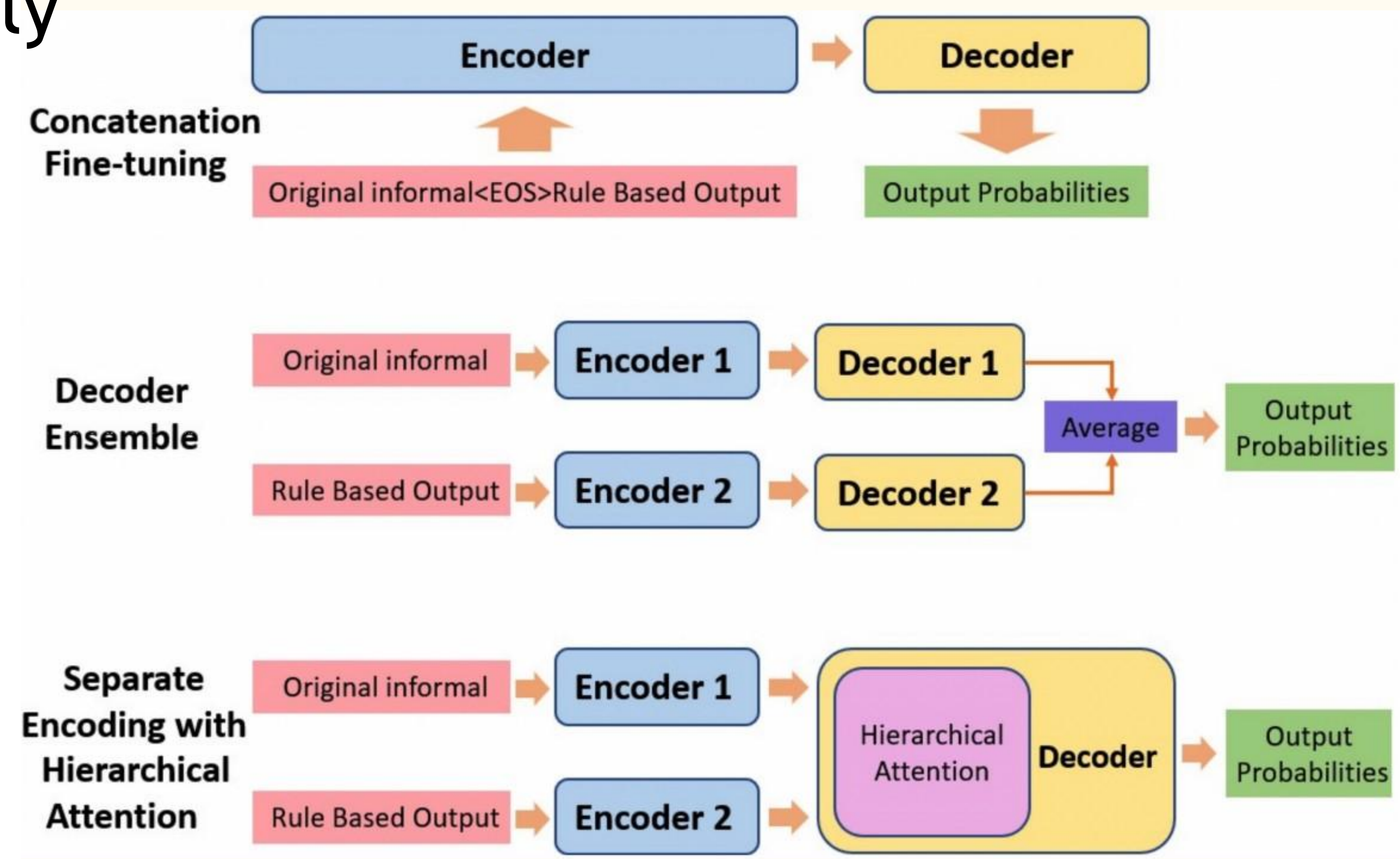
- Observations
  - Rule-processed data are the Markov blanket
  - Some entities (esp. not proper nouns) may be recognized incorrectly

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

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 .

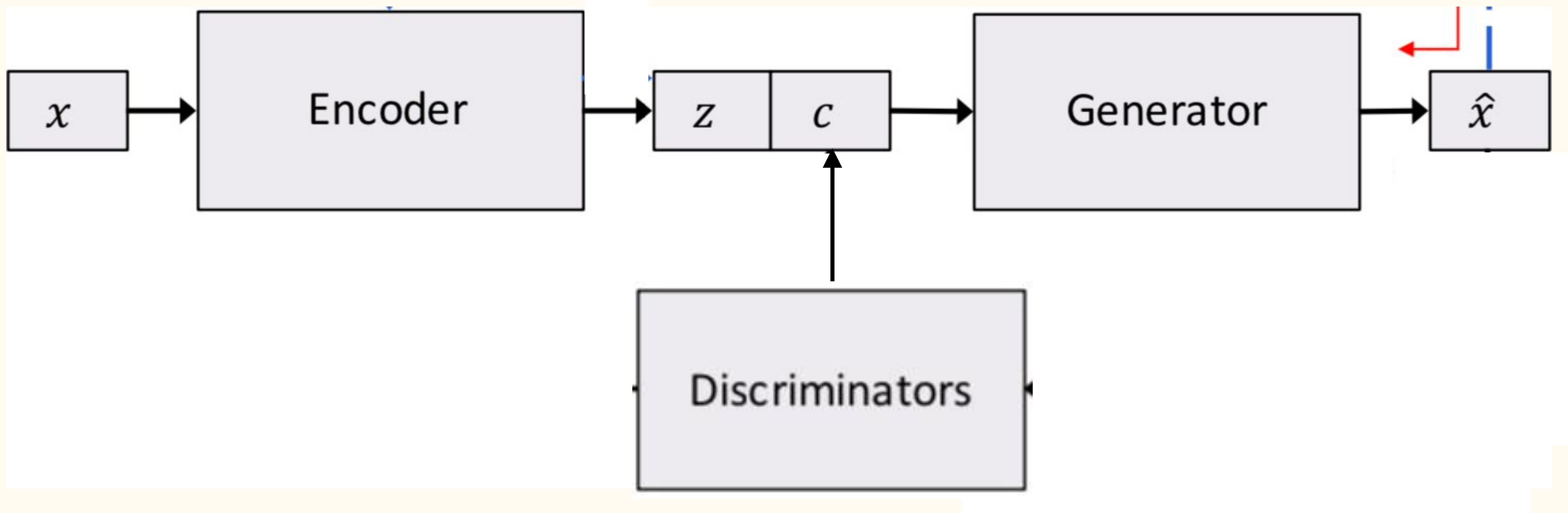
after watching this movie , i felt that disappointed .
after seeing this film , i 'm a fan .

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

Hu et al. [2017]

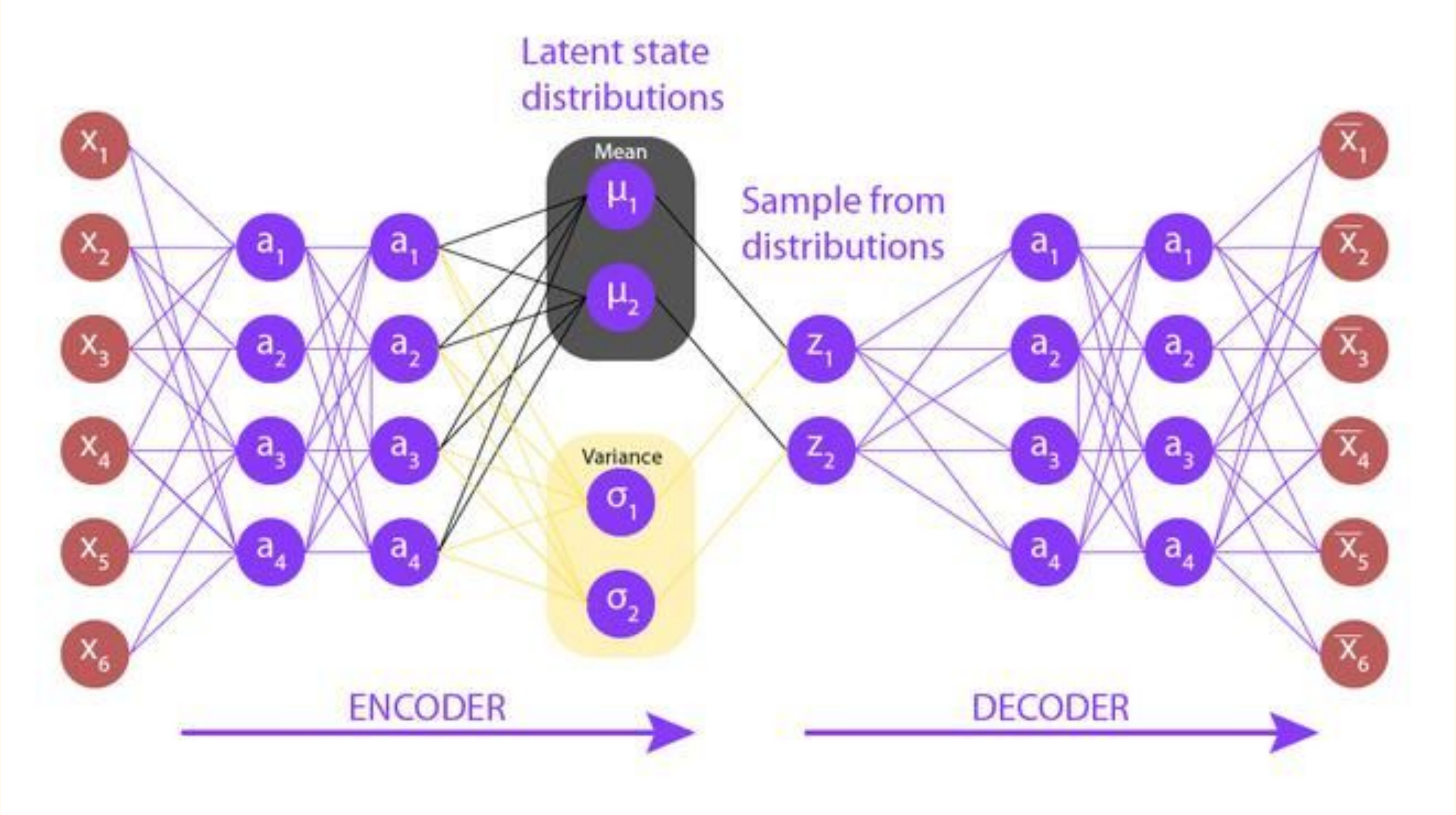


- Variational auto-encoder with latent space
  - Structured latent space c [style code]
  - Unstructured latent space z [remaining info]
- Discriminator: classifying the style

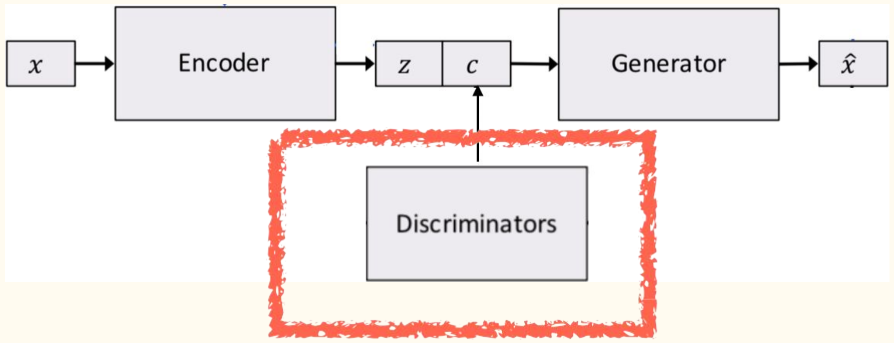
Hu, Z, Yang, Z, Liang, X, Salakhutdinov, R, Xing, EP. Toward controlled generation of text. In *ICML*, 2017

Variational Auto Encoder

- Autoencoders - encode data efficiently through a bottleneck architecture.
- InVAE, the encoder outputs a probability distribution in the bottleneck layer instead of a single output value.
- Use KL Divergence Loss



Hu et al. [2017]

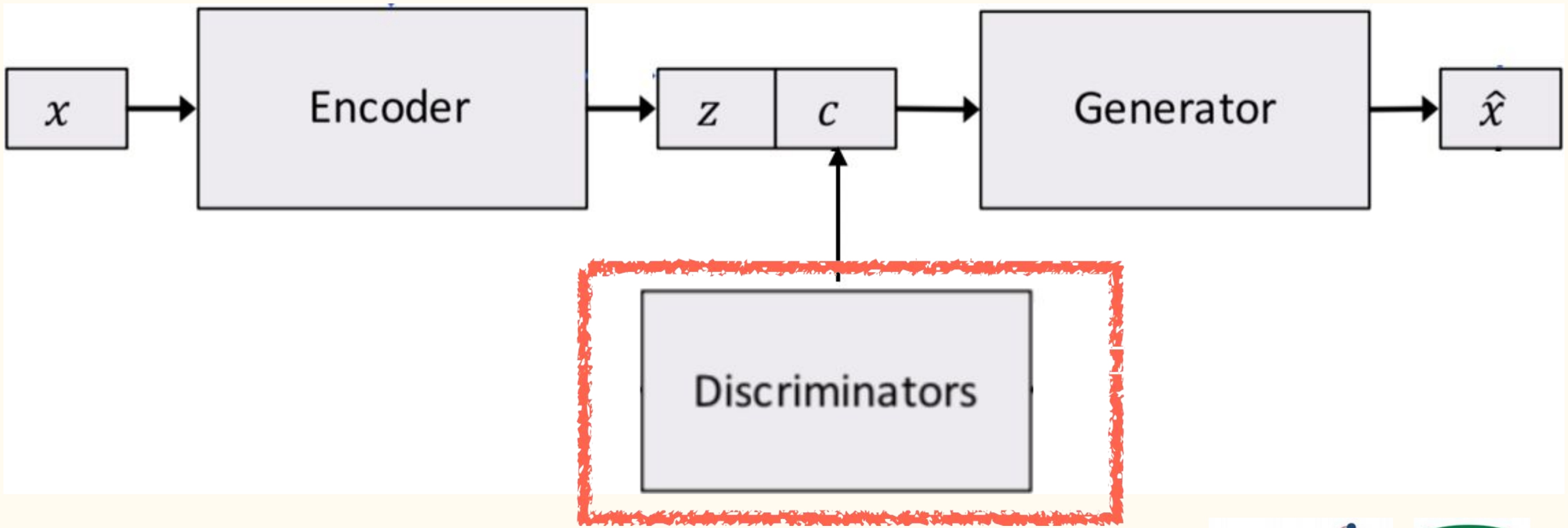


Training the discriminator w/ real labeled data

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

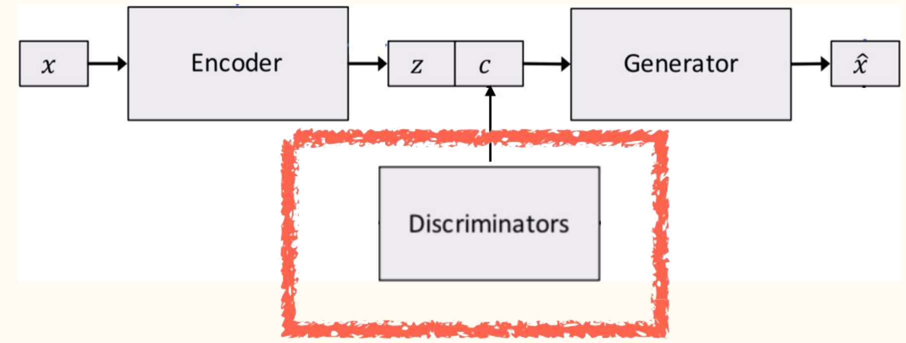
$$\mathcal{L}_s(\boldsymbol{\theta}_D) = \mathbb{E}_{\mathcal{X}_L}\left[\log q_D(\boldsymbol{c}_L|\boldsymbol{x}_L)\right]$$

[How well does the encoder classifier the style(s) as ?]



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

Hu et al. [2017]

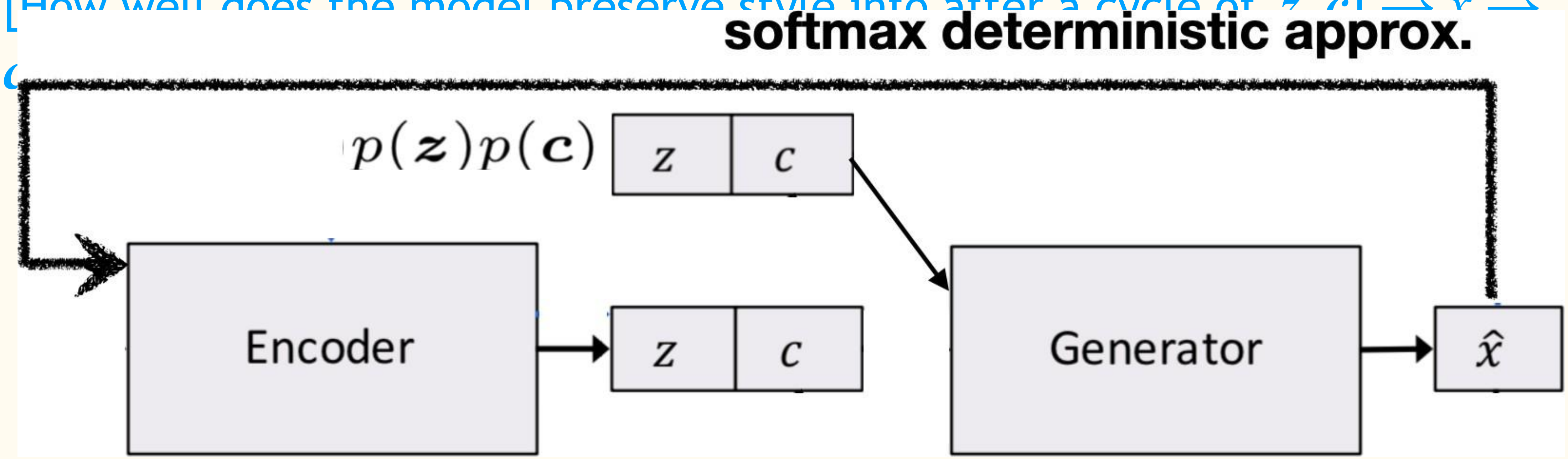


Training the discriminator w/ generated data from VAE

$$\min_{\theta_D} \mathcal{L}_D = \mathcal{L}_s + \lambda_u \boxed{\mathcal{L}_u}$$

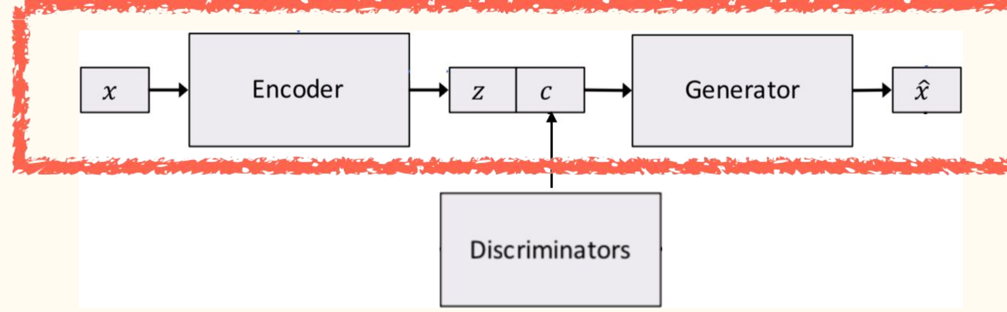
$$\mathcal{L}_u(\boldsymbol{\theta}_D) = \mathbb{E}_{p_G(\hat{\boldsymbol{x}}|\boldsymbol{z},\boldsymbol{c})p(\boldsymbol{z})p(\boldsymbol{c})}\left[\log q_D(\boldsymbol{c}|\hat{\boldsymbol{x}}) + \beta\mathcal{H}(q_D(\boldsymbol{c}'|\hat{\boldsymbol{x}}))\right]$$

[How well does the model preserve style info after a cycle of $z, c] \to x \to$



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

Hu et al. [2017]

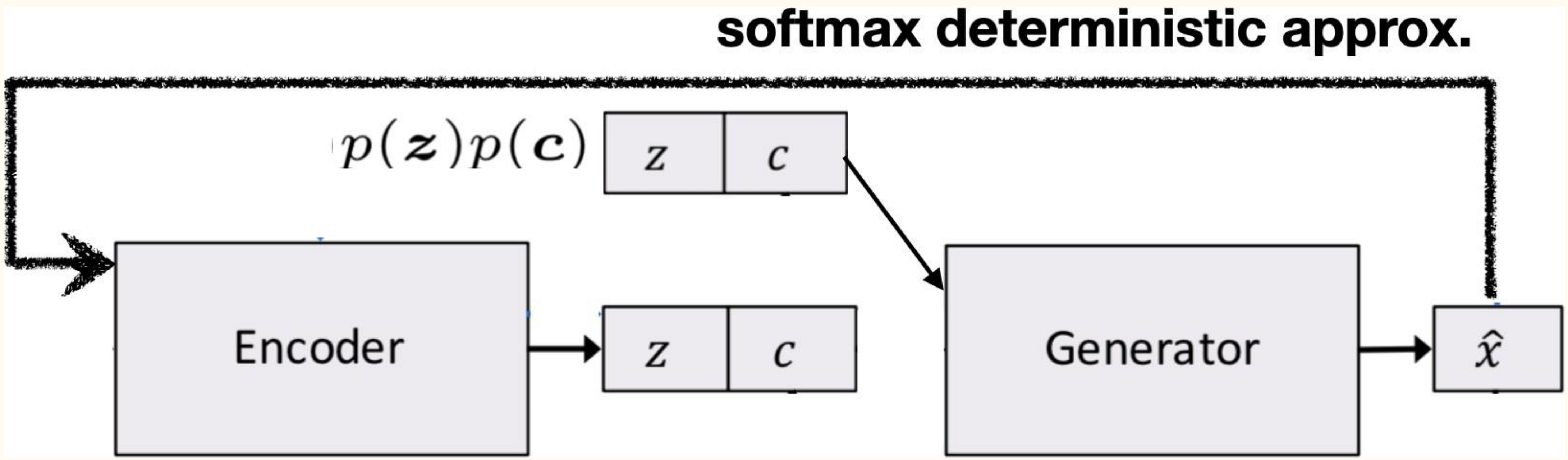


Training the generator

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

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

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



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

(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
  - Sentence

$$\boldsymbol{x} \rightarrow z \rightarrow \boldsymbol{x}$$

$$y$$

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
  - Sentence

$$\boldsymbol{x} \longrightarrow \boldsymbol{z} \longrightarrow \boldsymbol{x}$$

$$y \nearrow$$

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

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

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

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

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

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

$\boldsymbol{x}_1 \longrightarrow \boldsymbol{z} \longrightarrow \hat{\boldsymbol{x}}_1$

y_1

Discriminator

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

$\boldsymbol{x}_2 \longrightarrow \boldsymbol{z} \longrightarrow \hat{\boldsymbol{x}}_2$

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.

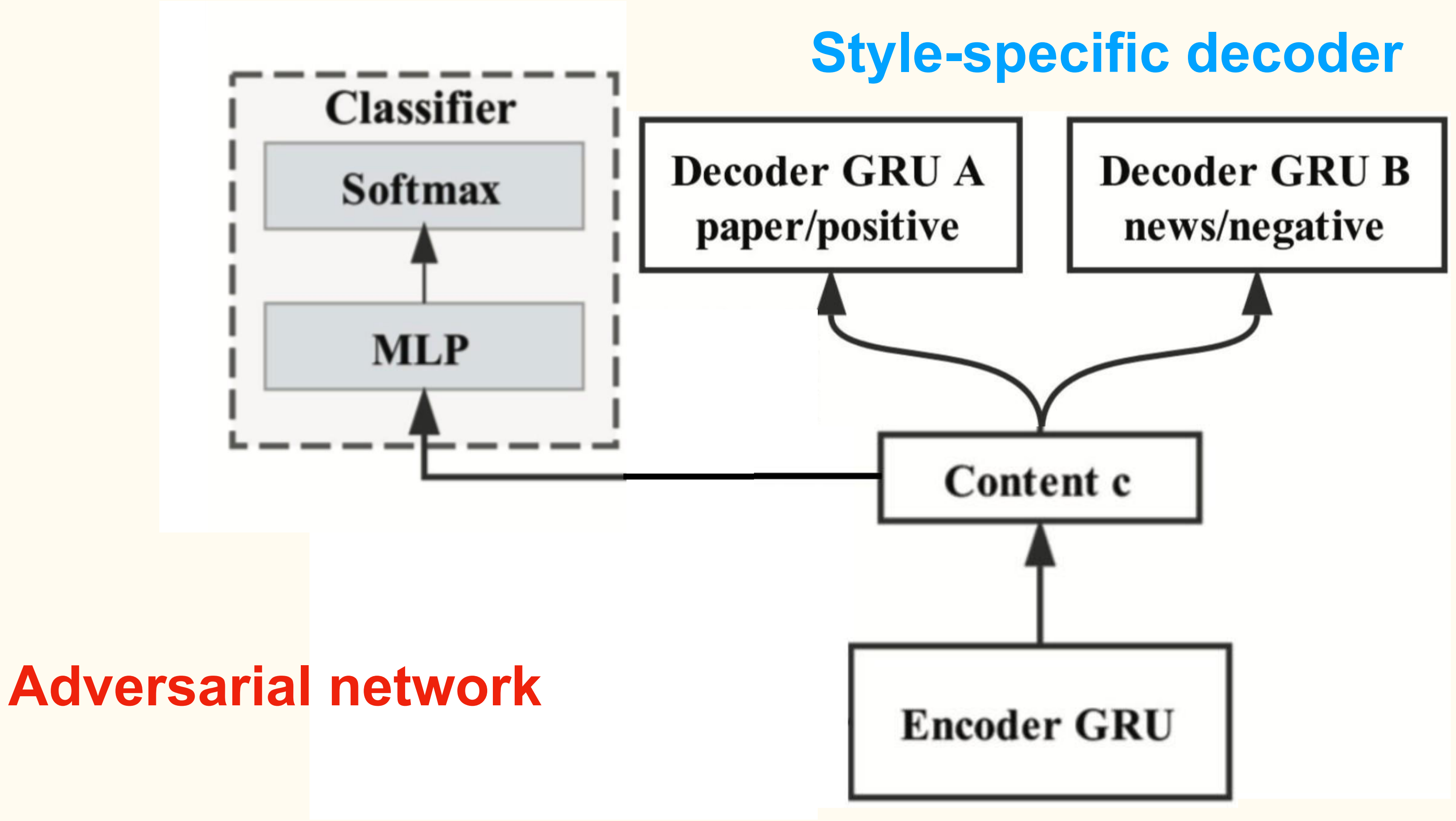
(Cross)-Alignment

- 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
    - Style 2 sentence VS. Style 1 → 2 transferred sentence

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 variant: Style-specific decoder

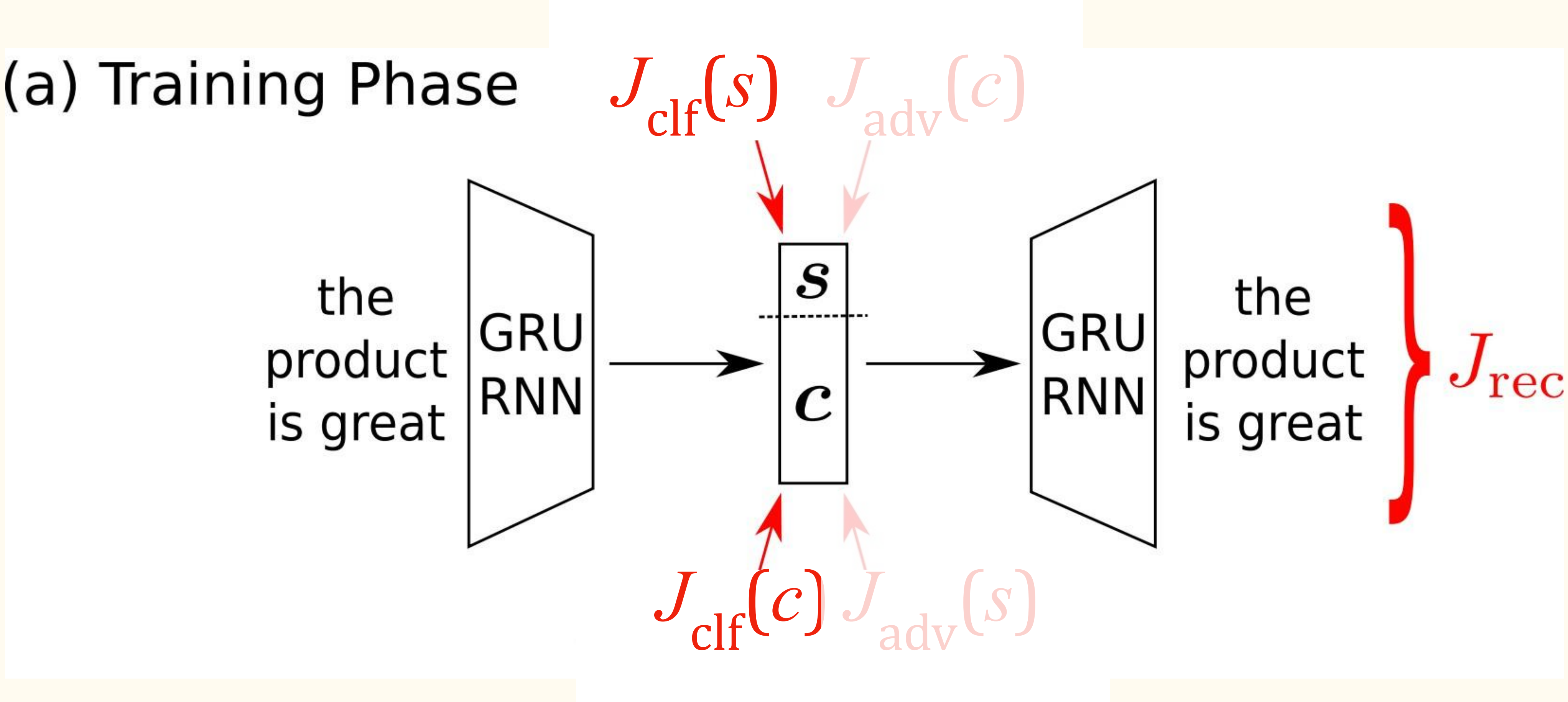


Fu, Z., Tan, X., Peng, N., Zhao, D. and Yan, R. Style transfer in text: Exploration and evaluation. In *AAAI*, 2018.

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 further thoughts
  - Encode style info (not by embedding)
  - Auxiliary losses can be applied to both content and style

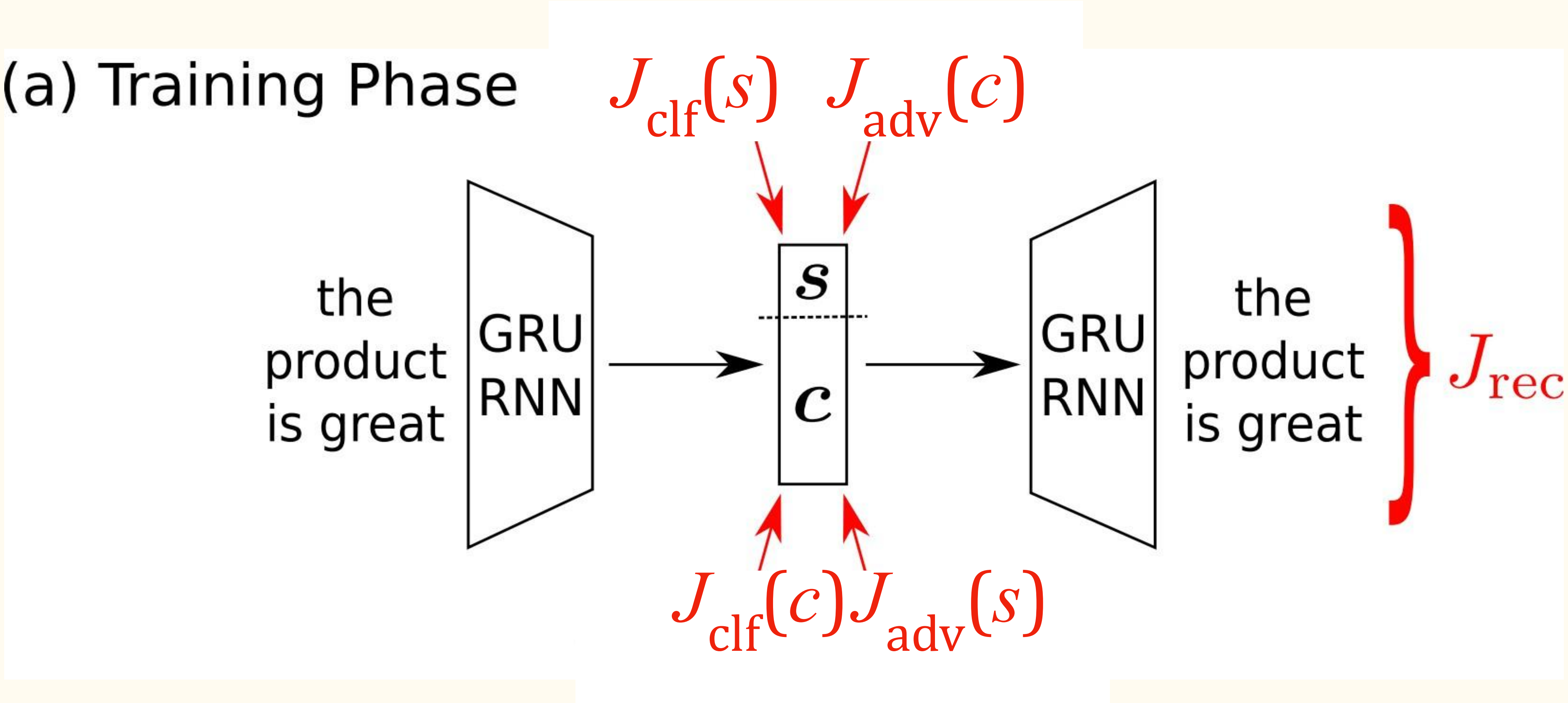
Disentangling Approach



- **Classification loss** ensures a space contains desired info
    - $J_{\mathrm{clf}}(s)$: 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

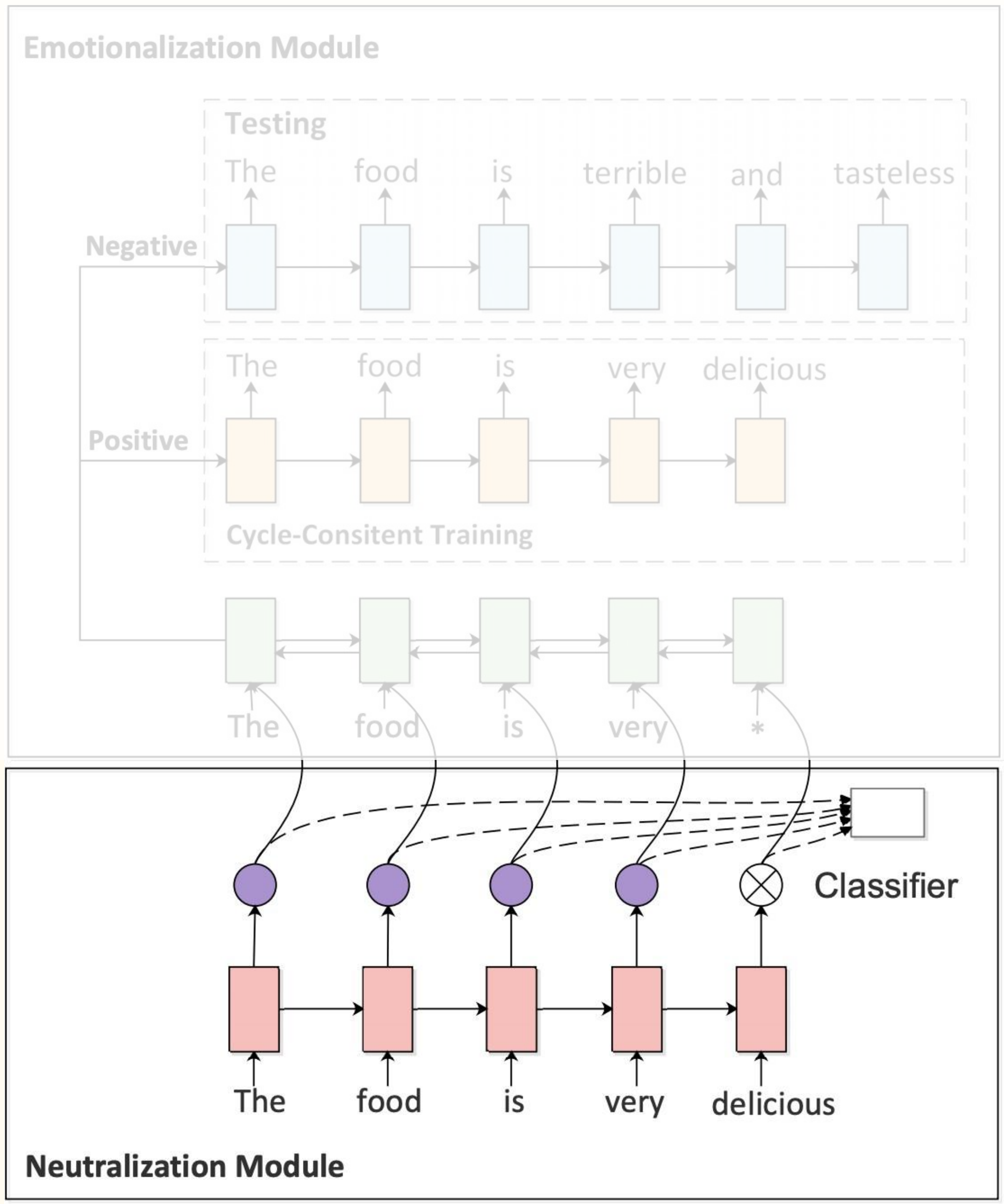
Cycled RL

- Module#1:

Extracting style-**neutral** words

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

Cycled RL

- Module#1:

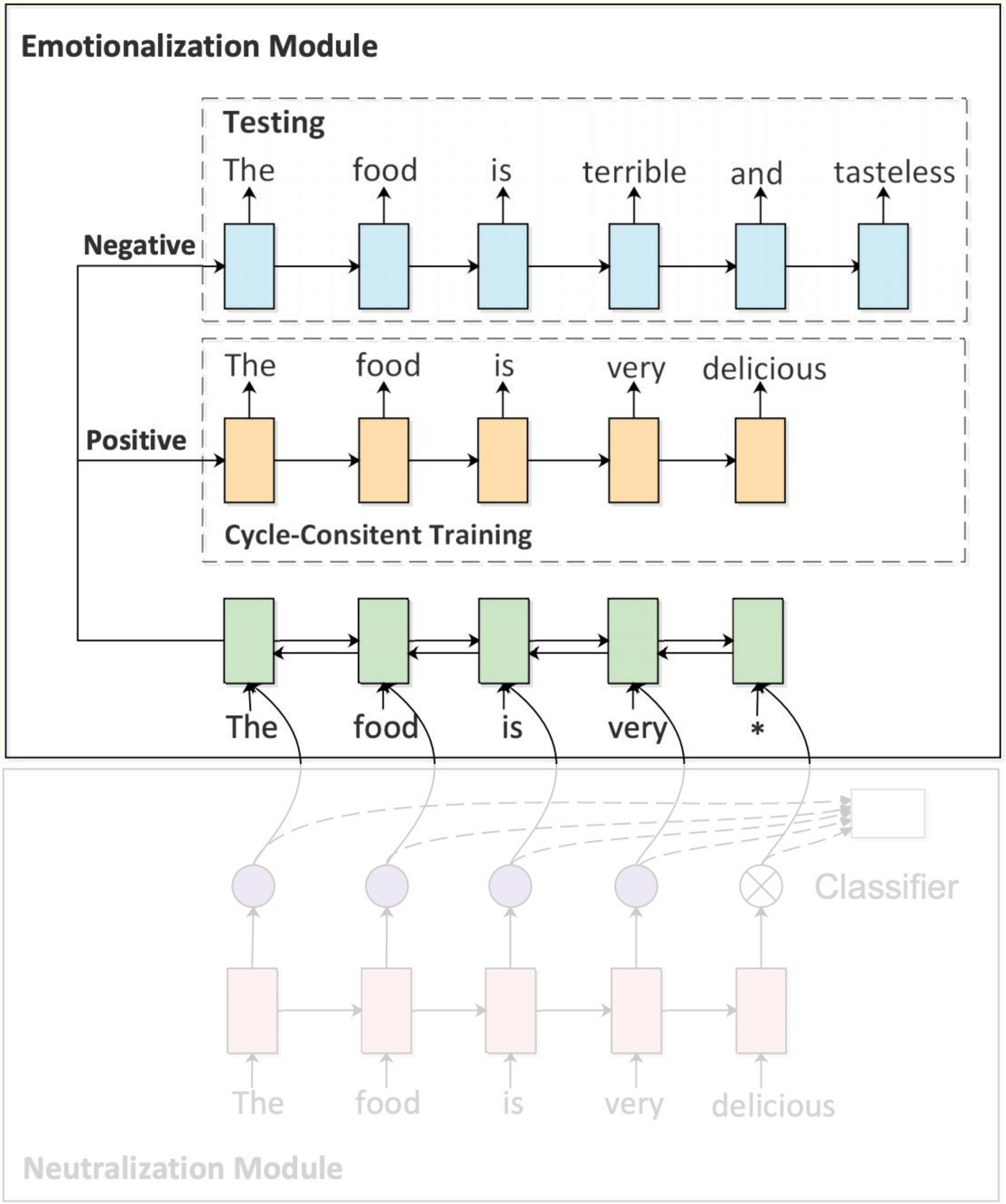
Extracting style-**neutral** words

- 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

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

Sentiment-neural subsentence → 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

A Quick Detour to REINFORCE

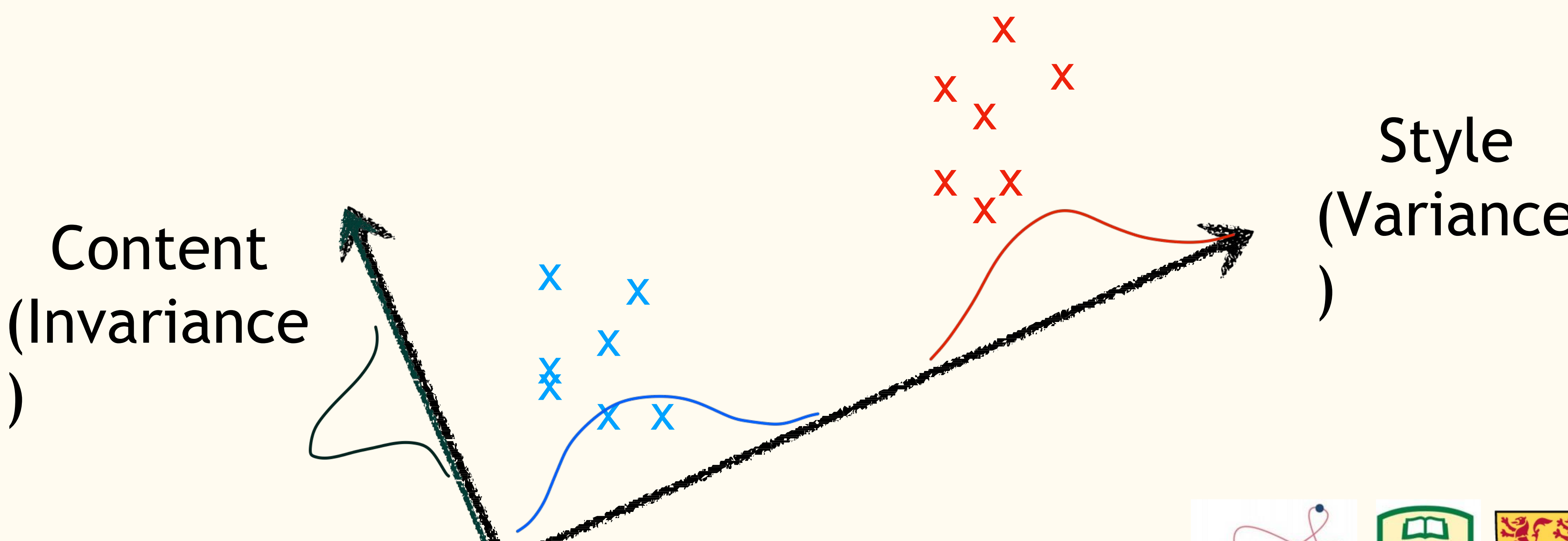
- RL works with discrete actions (e.g., which words to generate)

- 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

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

(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 -gram frequency

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

(for style and n-gram)

  - Thresholding
- **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 .

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

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

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

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**

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

Model#1: RetrieveOnly

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

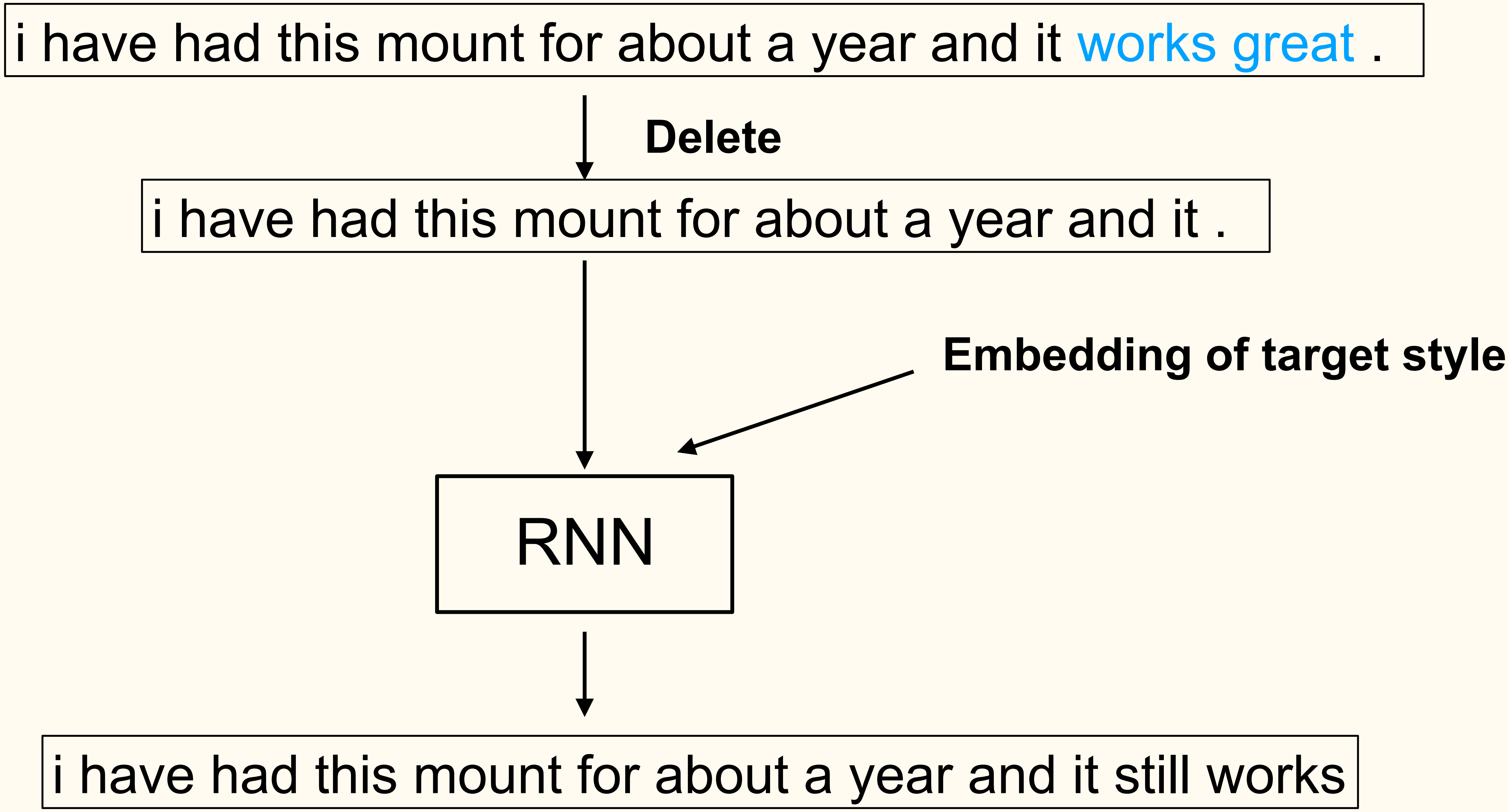
Delete-Retrieve-Generate

- **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

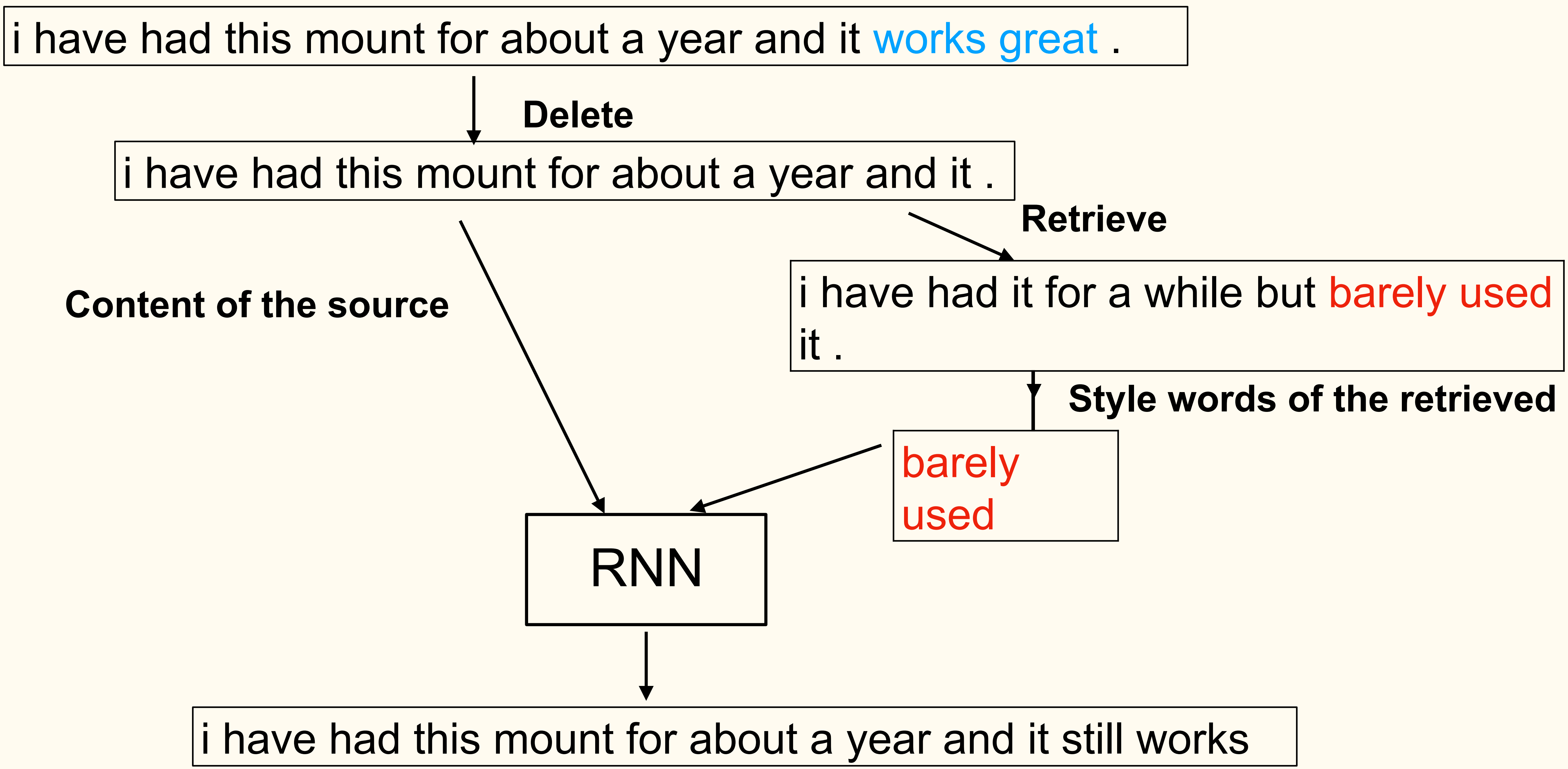
- **Model#2: Delete+Generate**



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.

DualRL

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

- Style reward

$$R_s = P(s_y|\boldsymbol{y'}; \varphi)$$

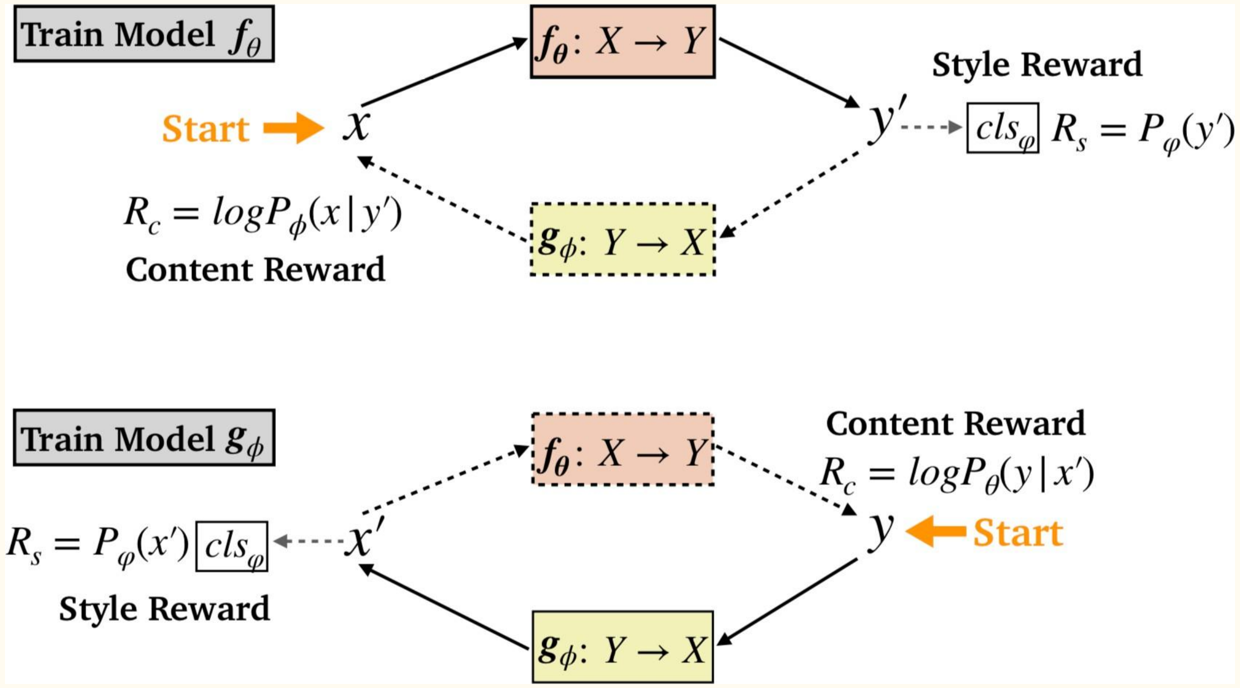
- Content reward

$$R_c = P(\boldsymbol{x}|\boldsymbol{y'}; \boldsymbol{\phi})$$

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

- Overall reward

- 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.

Stable Style Transformer

- **Main Approach**
  - Stage 1 - Deletes tokens that contain style attributes
  - Stage 2 - Encode content tokens and and combine with target style to generate style-transferred sentence
- **Main Innovation**
  - Pick style words using Important Score (
- **Datasets**
  - YELP and Amazon
- **Losses**
  - Reconstruction Loss
  - Style Loss

$$p_{\boldsymbol{x}} = p_{\theta_C}(s|\boldsymbol{x})$$

$$p_{\boldsymbol{x},t_i} = p_{\theta_C}(s|\boldsymbol{x}, t_i)$$

$$IS^k_{t_i} = p_{\boldsymbol{x}^k} - p_{\boldsymbol{x}^k,t_i}$$

$$\mathcal{L}_{rec} = -log\, p_{\theta_E,\theta_G}(\boldsymbol{x}|\boldsymbol{x}^c, s)$$

$$\mathcal{L}_{style} = -log\, p_{\theta_C}(\hat{\boldsymbol{x}} = \hat{s}|\boldsymbol{x}^c, \hat{s})$$

Joosung Lee , Kakao Enterprise Corp., South Korea. . Stable Style Transformer: Delete and Generate Approach with Encoder-Decoder for Text Style Transfer, 2020

Examples

	Yelp (negative to positive)	Yelp (positive to negative)
Input (source)	the food was so-so and very **over priced** for what you get .	these two women are **professionals** .
SST	the service is so-so and very **reasonably priced** for what you get .	these two women are **rude** .
CrossAligned	the food was **fantastic** and very very **nice** for what you .	these two dogs are **hard** down .
StyleEmbedding	the food was so-so and very **over priced** for what you get .	these two pot everywhere was .
DeleteOnly	the food was so-so and very **over priced** for what you get .	i **would n't like** these two women are professionals .
DeleteAndRetrieve	the service is **fantastic** and the food was so-so and the food is **very priced** for what you get .	these two **scam** women are **professionals** .
Back-translation	the food is **delicious** and the staff are very **good** for me .	this place is just **not good** .
UnpairedRL	the food was so-so and very **over priced** for what **great** qualities .	these two women are **great** .
DualRL	the food was **surprising** and very **reasonably priced** for what you get .	these two women are **unprofessional** .
B-GST	the food was **amazing** - so **fresh** and very **good** for what you get .	these two women are **terrible** liars .
G-GST	the food was **priced right** - so **nice** and very **good** for what you get .	these two women are **condescending** .
Human_DRG	the food was **great** and **perfectly priced**	these two women are **not professionals** .
Human_DualRL	the food was **good** and the **price is low** .	these two women are **not professionals** at all
	Amazon (negative to positive)	Amazon (positive to negative)
Input (source)	i have to **lower** the rating another notch .	it seems to be of very **good** quality in its build .
SST	**love** the rating another one ,	it seems to be of very **poor** quality in its build .
CrossAligned	i would **recommend** this for the price .	it s **not be for a good** game for my phone .
StyleEmbedding	i have to **get** by a one market .	it seems to be the num_extend is **good nice** high cases .
DeleteOnly	i have to **lower** the rating and it fits into another notch .	i have previously charged num_num different bt headsets that last num_num hours longer .
DeleteAndRetrieve	i have to **lower** the rating another notch and i **love** it .	initially it was very **good** quality in its build .
B-GST	i have **lower** levels for the other notch .	it seems to be of very **good** quality in taste .
G-GST	i have **lower** the steel another notch .	it seems to be of very **good** value in return .
Human_DRG	i have to **raise** the rating another notch .	it seems to be of very **poor** quality in its build

Reviews - Pros

- Paper proposes a **novel and intuitive** Delete method to remove the style attributes from the input sentence. (Shivangi)
- **Evaluations** performed are strong and **comprehensive**. The approach is tested for content, fluency and style accuracy and meaning. The evaluations are straightforward (eg. using GPT for fluency, BERTScore for meaning) (Harman)
- alpha and beta parameter provide parameters to **control content and style**, allowing user to set how much style or content preservation is required (Rohit)
- Results are **interpretable**. We can clearly see which tokens model considers to be content or style. (Vishal Saley)

Reviews - Cons

- The paper has not tried and tested different types of **pre-trained models**. (Sheshank)
- Style and Content are **not completely independent** of each other, and the deleted words can contain content as well. (Jai, Seshank, Aditya, Daman)
- The paper claims to have "**stable**" performance across all evaluation metrics. While this may be somewhat useful, I also see some other methods are performing way better than this method, (eg in Table 2), while not being unstable. (Harman)

Reviews - Extensions

- Composing various style attributes can be an interesting idea. Eg: Make my sentence formal and angry. (Daman)
- The paper uses a 3 layer model whereas BART base is of 6 layers. The size of the model can be increased to see if there is improvement in performance.(Jai)
- Soft deletion: Instead of hard deletion of tokens, can we delete tokens in a soft manner, so that we can learn the model, end to end? For eg, using attention scores to give importance weighting to the tokens, and then inputting them to the encoder-decoder architecture. (Harman)

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.

Thank you for listening!