Text Style Transfer Dissection
: Deeper understanding on Text Style Transfer

AI 605 | Final Presentation | 20.07.03

Nyoungwoo Lee, KAIST

Jungsoo Lee, KAIST

NakYil Kim, KAIST

Donghyun Jeon, NAVER

Index

Problem Statement

Problem Statement

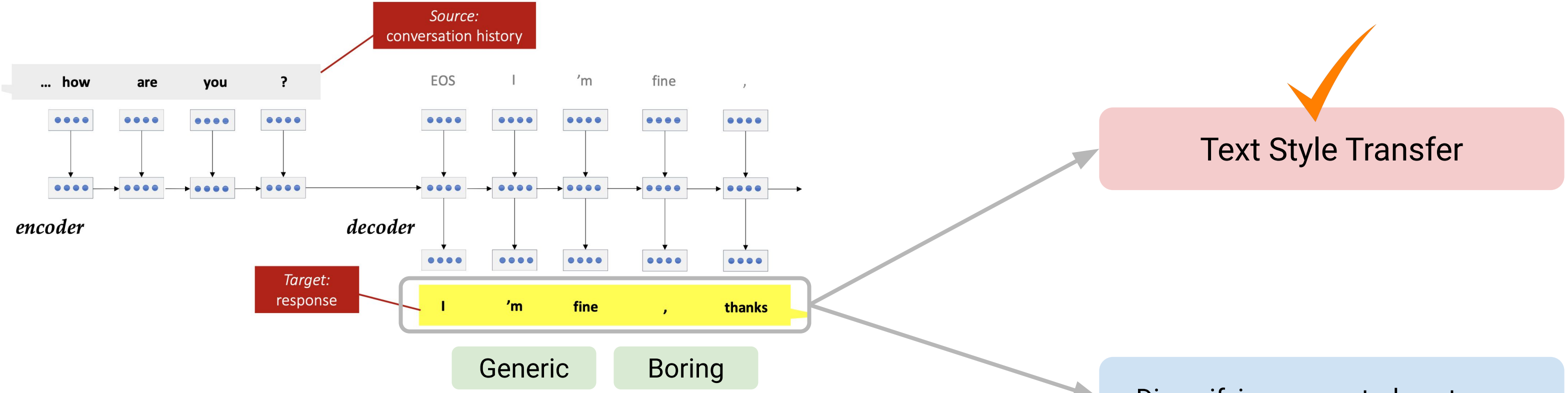
1. What do we want from Conversational AI?

Samantha of <Her>

Jarvis of <Ironman>

Problem Statement

2. Limitation of Language Model



Response Generation
based on Language Model

The general language model still produces **boring** and **generic** responses

Problem Statement

3. Text Style Transfer

Neutral — The music is on.

Negative — The music is terrible here.

Positive — The music is great here.

with **Style of Text**

Pos/Neg | Formality

Text style transfer refers to the study of changing the `style' of an input sentence while preserving the `content', overall information included in the sentence.

Problem Statement

4. What do we think about “style” of text

STYLE
of text

Hip hop

Love song

Problem Statement

5. Research Questions

To generate a more abstract style of sentences

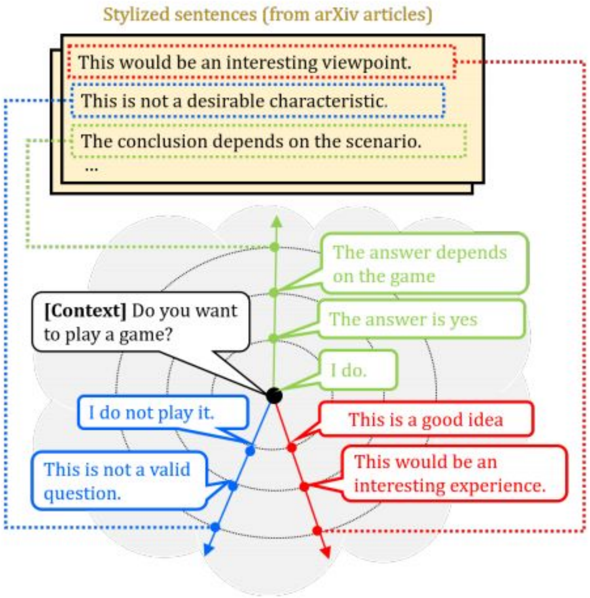
To explore how the model understands style and content.

Related Work

Related Work

1. Structuring Latent Spaces for Stylized Response Generation

(Gao et al., 2019)



They designed `**StyleFusion**’ which maps the sentences of the general dataset(e.g. Reddit) and the stylized dataset(e.g. script of drama character) into the same latent space

Stylizes an input sentence with the closest style sentence.

Related Work

2. Domain Adaptive Text Style Transfer

([Li et al., 2019](#))

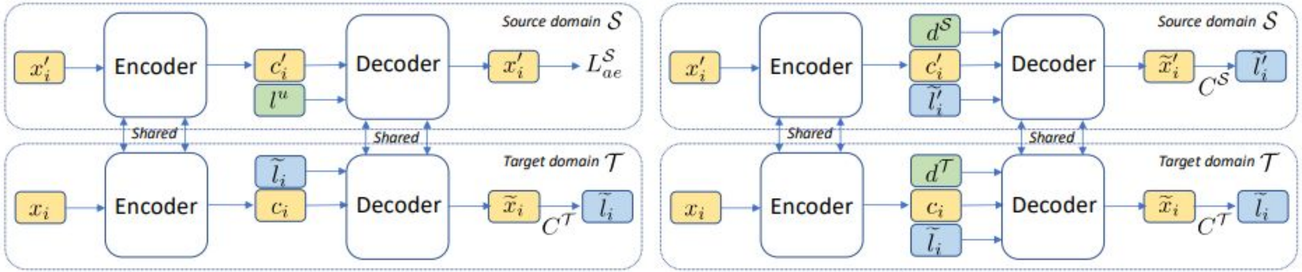


Figure 1: Illustration of the proposed **DAST-C** (left) and **DAST** (right) model. DAST-C learns the generic content information through $L_{ae}^{\mathcal{S}}$ on massive source domain data with unknown style l^u. For DAST, $d^{\mathcal{T}}, d^{\mathcal{S}}$ and $C^{\mathcal{T}}, C^{\mathcal{S}}$ denote domain vectors and domain-specific style classifiers, respectively. Better looked in color.

Autoencoder to maintain content resiliency of the source domain,

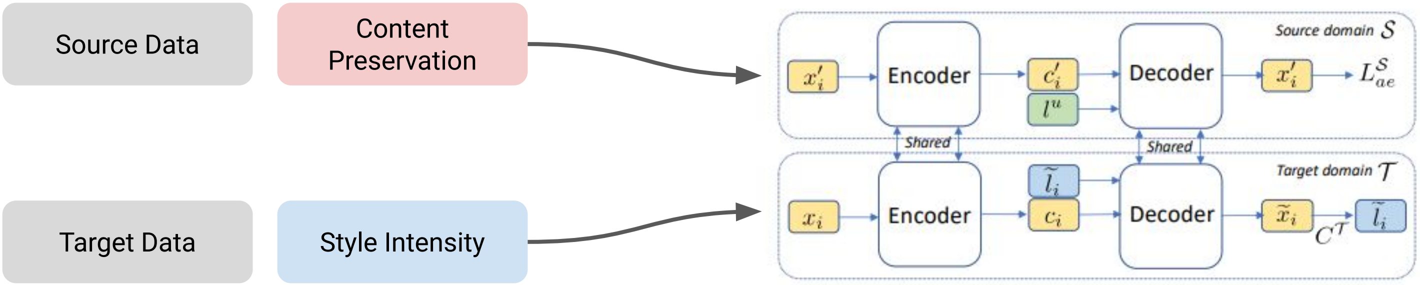
and apply **style classifier** to force meaningful style information into the model

Approach

Approach

1. Research Questions

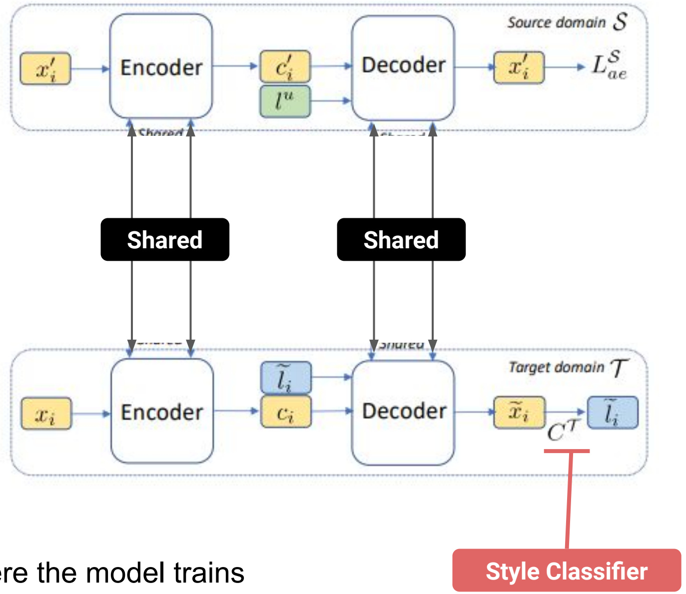
1. The range of style in a text
2. Which part should the model consider for preserving the content



Approach

2. Replication paper

Source Data | Content Preservation



In the **source corpus**, the model proceeds with training to reconstruct the input sentence to the original sentence,

This process can enhance the content preservation of the sentence.

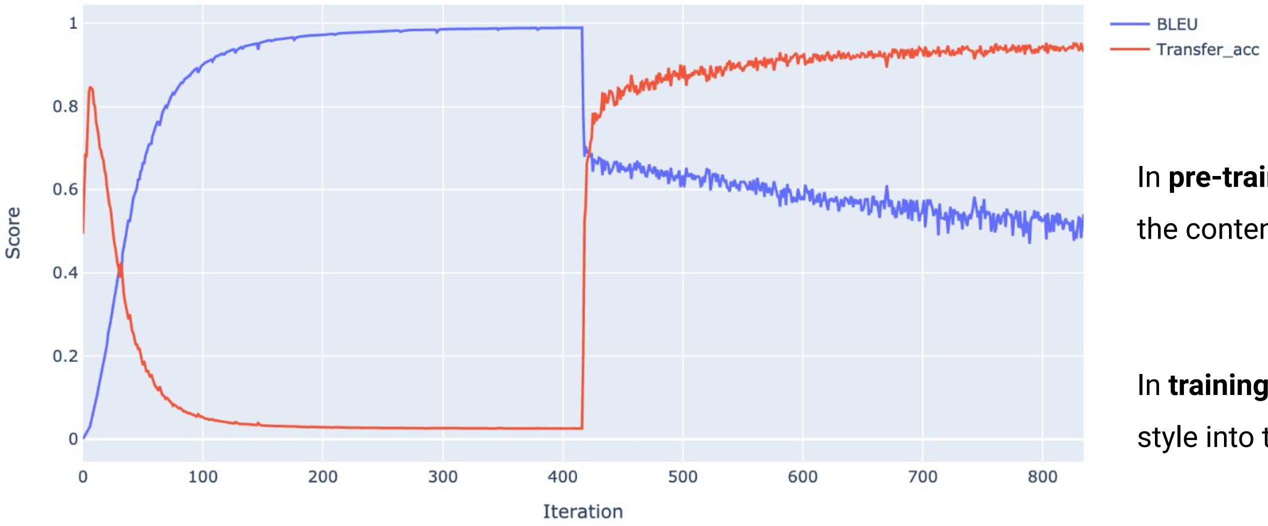
Target Data | Style Intensity

In the **target corpus** is a labeled dataset with a specific set of styles, where the model trains to change one style sentence to another style.

The style of the changed sentence can be verified using the pre-trained style classifier.

Approach

3. Preliminary Study



Pre-Training

Training

In **pre-training** phase, the model learns to preserve the content of the massive source sentence

In **training** phase, the model is trained to change the style into the opposite style

Approach

3. Preliminary Study

Model	Input (negative)	Transferred sentence (negative-to-positive)
DASTC	service was slow, disinterested	service was delicious, satisfying
	skip this place and stay on the strip	delicious this place and stay on the strip
	I would give this review negative stars	I definitely give perfect review terrific stars
	service was lacking, waitress didn't know the menu	service was delicious, waitress did blast know the menu
Metaphor	swimming in the grease from the beef topping	refreshing in the inside from the beef topping

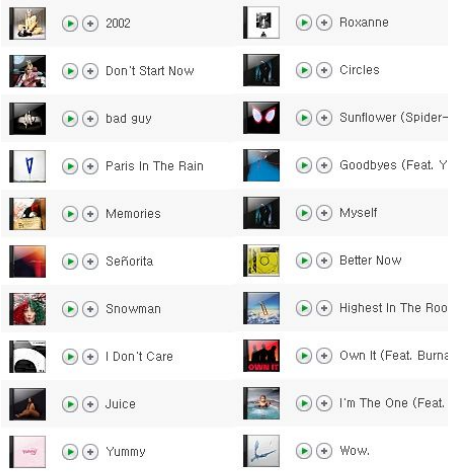
Model	Input (positive)	Transferred sentence (positive-to-negative)
DASTC	just tried their food, it was amazing!	just tried their food, it was terrible!
	prices are great and the food was excellent	prices are awful and the food was disappointing
	very clean and attentive and also very quick	very clean and outdated and also very disappointing
	very good lunch, service, and atmosphere	very low lunch, service, and atmosphere
	we really enjoy everything we've had here	we really left everything we've had here

[Table 1] Transferred sentences on IMDB movie review to Yelp dataset of the preliminary result, where **red** denotes negative term and **blue** denotes positive term

Approach

4. Approach

1. Effects of early stopping - Love & Hip hop lyrics dataset
2. Effects of pre-trained epoch - Love & Hip hop lyrics dataset
3. Effects of weights in style loss - Love & Hip hop lyrics dataset
4. Further experiment - Hillary & Trump twitter dataset



Love & Hip hop lyrics dataset

Trump & Hillary twitter dataset

Experiment

Experiment

1. Effects of early stopping

	Style	Lyrics
Input	**love**	and i swear by the **moon** and the stars in the sky ill be there
early stopping	hip-hop	and i swear by the **mc** and the stars in the am ill not stan
Last epoch	hip-hop	and i swear by the **kick** and the stars in the **police** ill not stan
Input	**love**	every **inch** of your **skin** is a holy **gray** I have got to find
early stopping	hip-hop	every **hip** of your **back** is a holy **hip** I have got to find
Last epoch	hip-hop	every **mcs** of your back is a **gangsta hip** I have got to **joint**
Input	**love**	i never thought that you would be the one to **hold my heart**
early stopping	hip-hop	i never thought that you would be the one to **ask joint plug**
Last epoch	hip-hop	i never thought that you would be the one to **jock swift swift**
Input	**love**	if she changes her mind this is the first place she will **go**
early stopping	hip-hop	if she changes her mind this is the first place she can **properly**
Last epoch	hip-hop	if she **because** her mind this is the first place she **police uhhuh**

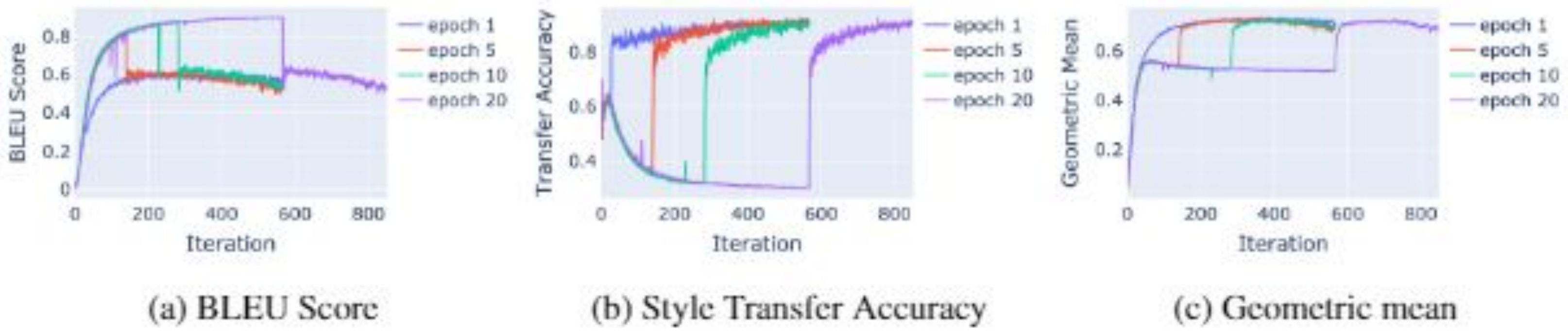
[Table 2] Results of text style transfer with song lyrics with early stopping or not

Experiment

2. Effects of pre-trained epoch

Pre-trained Epochs	Domain Acc.	Style Transfer Acc.	BLEU Score	Geometric Mean
1	0.9952	0.8676	0.5976	0.7201
5	0.9933	0.9043	0.5826	0.7258
10	0.9934	0.8613	0.6114	0.7257
20	0.9899	0.8567	0.6000	0.7169

[Table 3] Quantitative analysis results of generated sentences with variation of pre-trained epochs



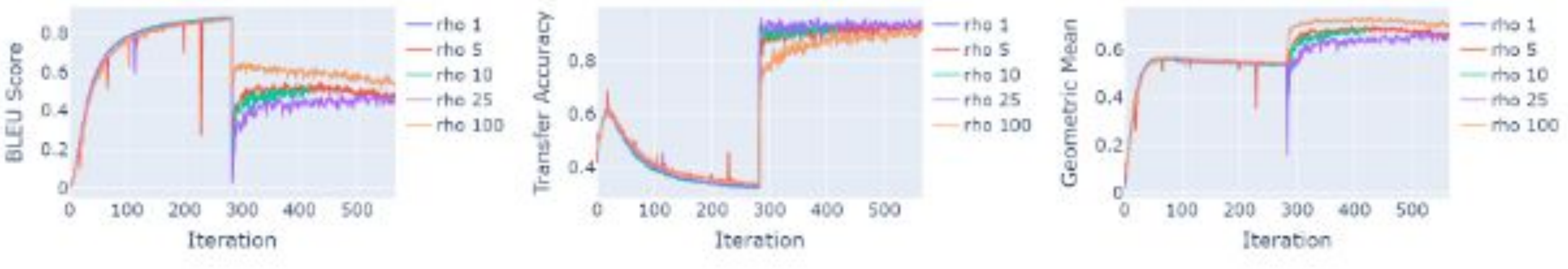
[Figure 1] BLEU score, style transfer accuracy and geometric mean graphs of generated sentences with variation of pre-trained epochs

Experiment

3. Effects of weights ρ in style loss

ρ	Domain Acc.	Style Transfer Acc.	BLEU Score	Geometric Mean
1	0.9923	0.8491	0.6119	0.7207
5	0.9950	0.9031	0.5101	0.6785
25	0.9925	0.9327	0.4274	0.6309
100	0.9585	0.9585	0.3871	0.6090

[Table 4] Quantitative analysis results of generated sentences with variation of variation of value ρ



(a) BLEU Score (b) Style Transfer Accuracy (c) Geometric mean

[Figure 2] BLEU score, style transfer accuracy and geometric mean graphs of generated sentences with variation of value ρ

Experiment

3. Effects of weights ρ in style loss

ρ		Lyrics
	Input	and i swear by the **moon** and the stars in the sky ill be there
1	Early stop	and i swear by the moon and the **short** in the style ill be there
5	Early stop	and i swear by the **rhythm** and the stars in the **suckers** ill be there
25	Early stop	and i swear **as** the **role** and the stars in the **cell not phase** there
100	Early stop	and i swear by the moon and the stars **in the jungle be street heard**
	Input	every **inch** of your **skin** is a **holy gray** I have got to find
1	Early stop	every **mcs** of your **brothers** is a **dandy uhh** I have got to find
5	Early stop	every **slick** of your **vocal** is a **few uncle** I have got to find
25	Early stop	every **stain** of your **skin** is a **sayin off** I am got to find
100	Early stop	**visiting livin** of your **skin** is a **old soul** I have got to find
	Input	oh her eyes her eyes make the **stars** look like they are not shinin
1	Early stop	oh her eyes her eyes make the **stars** look like they are not
5	Early stop	oh she eyes her eyes make the **o** look like people are not
25	Early stop	o her can to bear make the **stars** look like they are not
100	Early stop	oh her eyes her eyes make the **school** look like they are not

[Table 5] Results of text style transfer with song lyrics with different ρ

Experiment

4. Further experiment - Hillary & Trump twitter dataset

	Style	Sentence	
Input	**Hillary**	it took years but will be the year of the first **woman** POTUS great day to be an American **woman**	proponent
	Trump	it took years but will be the winner of the very **presidential person** my speech for very many the	
Input	**Hillary**	I want you to know that I see you and I am with you Hillary to the Latino community at	
	Trump	I want you to know that I am me today to me with I hope to the amazing amp	
Input	**Hillary**	if fighting for **affordable child care** and paid family leave is playing the woman card, then deal me	policy
	Trump	if fighting for **jobs care** than jobs is setting the opposite card then, will join me	
Input	**Trump**	the ratings for the **republican national convention** were very good but for the final night my speech great thank you	party
	Hillary	the ratings for the **democratic congressional convention** is also happening on my heart for this speech my democratic friend	
Input	**Trump**	this is why I would think the **unions** would support	proponent
	Hillary	this is why I would think the **women** would support	
Input	**Trump**	we are going to bring steel and **manufacturing** back to Indiana	policy
	Hillary	we are going to bring the **community** and build back to Indiana	

[Table 6] Results of text style transfer with Twitter data of Hillary and Trump

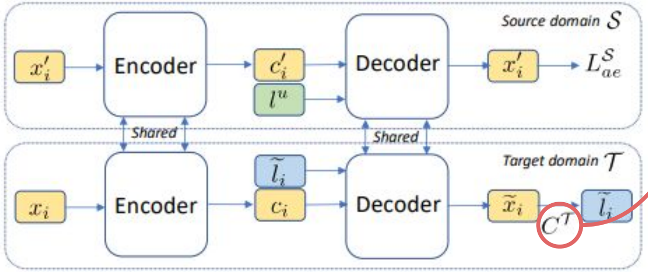
Discussion

Discussion

1. Improvement of the style classifier

Target Dataset	Style set	Pre-trained style classifier accuracy
Yelp	positive, negative	0.9317
Lyrics	love song, hip hop	0.8701
Tweet	Trump, Hillary	0.8453

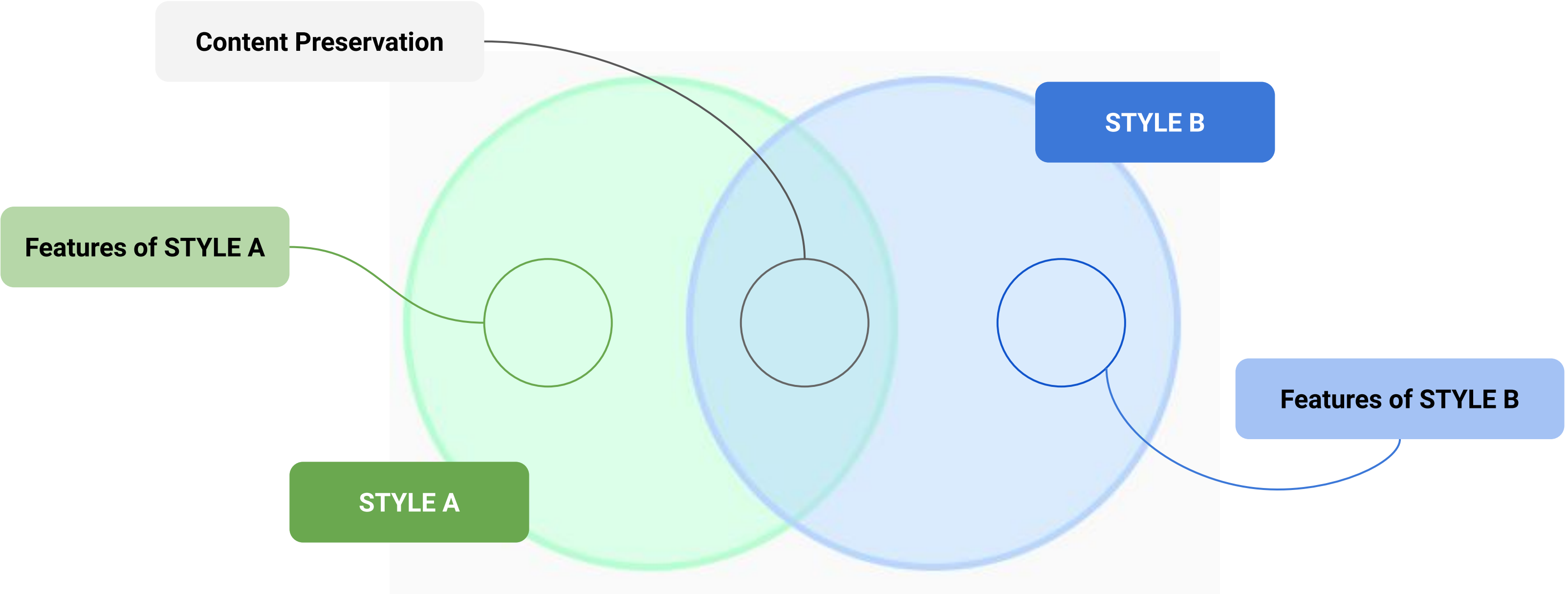
[Table 7] Pre-trained style classifier accuracy of models with different dataset



Style Classifier improvement with pre-trained **BERT**

Discussion

2. Differences in model styles and content recognition by data configuration

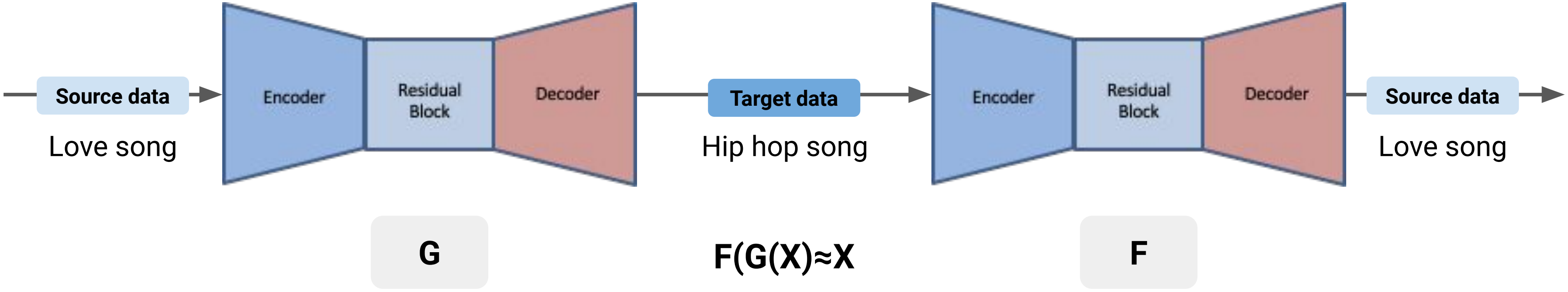


Model recognizes the **CONTENT to preserve** as the words which are <u>not different between datasets</u>

while it regards to **STYLE** as the words which seem <u>different apparently</u>.

Discussion

3. Cycle consistency loss



Adding **cycle consistency loss** on our work

would promote the quality of the stylized sentences

THANK YOU