Text Style Transfer & metrics

ВШЭ ФКН, NLP

Шабалин Александр

Style Transfer

Формальный ↔ Неформальный
Проза ↔ Стихотворение
Позитивный ↔ Негативный
Стиль определенного автора

Стиль – произвольный атрибут текста.
Текст = контекст + стиль

Style Transfer

Формальный ↔ Неформальный
Проза ↔ Стихотворение
Позитивный ↔ Негативный
Стиль определенного автора

Стиль – произвольный атрибут текста.
Текст = контекст + стиль

Суммаризация: Длинный текст ↔ Короткий
Машинный перевод: En ↔ Ge

Style Transfer

Формальный ↔ Неформальный
Проза ↔ Стихотворение
Позитивный ↔ Негативный
Стиль определенного автора

Нет параллельного корпуса

Стиль – произвольный атрибут текста.
Текст = контекст + стиль

Суммаризация: Длинный текст ↔ Короткий
Машинный перевод: En ↔ Ge

Есть параллельный корпус

Neural Machine Translation (NMT)

$$\sum_{i=1}^{T} \log p_\theta(y_i | y_{<i}, X) \to \max_\theta$$

X – текст на исходном языке

Y – текст на целевом языке

Neural Machine Translation (NMT)

$$\sum_{i=1}^{T} \log p_\theta(y_i | y_{<i}, X) \to \max_\theta$$

X – текст на исходном языке

Y – текст на целевом языке

Можно ли использовать неразмеченные данные?

Neural Machine Translation (NMT)

Back-Translation:

1. Обучить модель на том, что есть.
2. Использовать ее для перевода непараллельных данных.
3. Доучить модель обратного перевода на том, что получилось.

Neural Machine Translation (NMT)

Back-Translation:

1. Обучить модель на том, что есть.
2. Использовать ее для перевода непараллельных данных.
3. Доучить модель обратного перевода на том, что получилось.

Iterative Back-Translation

На каждом шаге переводим текст в одну сторону, а затем обратно, используя исходный текст в качестве таргета.

Neural Machine Translation (NMT)

Back-Translation:

1. Обучить модель на том, что есть.
2. Использовать ее для перевода непараллельных данных.
3. Доучить модель обратного перевода на том, что получилось.

Iterative Back-Translation

На каждом шаге переводим текст в одну сторону, а затем обратно, используя исходный текст в качестве таргета.

Denoising AutoEncoder

Аугментируем текст и просим модель восстановить его.

Как измерить качество модели?

Ошибка не подходит, потому что:

- Оперирует на уровне токенов, а не текстов
- Может зависит от размера батча и от длины текста
- Недостаточно хорошо коррелирует с человеческими оценками

Neural Machine Translation (NMT)

- Исходный текст
- Сгенерированный перевод
- Эталонный перевод

Neural Machine Translation (NMT)

- Исходный текст
- Сгенерированный перевод
- Эталонный перевод - 1
- Эталонный перевод - 2
- …

Neural Machine Translation (NMT)

- Исходный текст
- Сгенерированный перевод
- Эталонный перевод - 1
- Эталонный перевод - 2
- …

BLEU, 2002. Все еще лучшая метрика.

BLEU (BiLingual Evaluation Understudy)

Метрика для измерения похожести текста на набор других.

S – машинный текст

$\hat{S}$ – эталонный текст

BLEU (BiLingual Evaluation Understudy)

Метрика для измерения похожести текста на набор других.

S – машинный текст

$\hat{S}$ – эталонный текст

Example

Candidate: the the the the the the the.

Reference 1: The cat is on the mat.

Reference 2: There is a cat on the mat.

Modified Unigram Precision $= 2/7$

$p_n(\hat{S}; S)$ – **precision**; доля n-грамм слов в S, которые присутствуют в $\hat{S}$.

Если эталонных текстов несколько, то проверяется присутствие в каждом тексте.

BLEU (BiLingual Evaluation Understudy)

$$BLEU_w(\hat{S}; S) := BP(\hat{S}; S) \cdot \exp\left(\sum_{n=1}^{\infty} w_n \ln p_n(\hat{S}; S)\right)$$

w – веса для разных n-грамм. Обычно считают **1,2,3,4-граммы**.

Штраф за краткость (Brevity penalty)

$$BP(\hat{S}; S) := e^{-(r/c-1)^+}$$

r – длина эталонного текста, c – сгенерированного

Если модель генерирует короткие тексты, то BLEU уменьшается.

Недостатки BLEU

- BLEU не важен порядок слов и их смысл
- Плохо работает для морфологически богатых языков
- Требует много эталонных текстов
- Много различных имплементаций

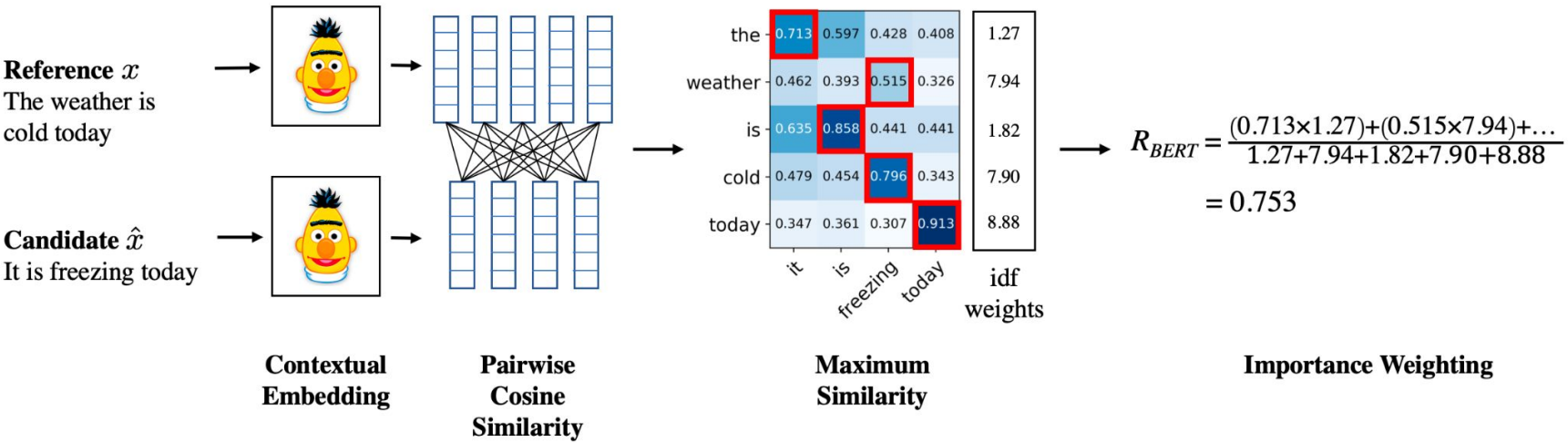
Вариации BLEU

- **ROUGE** – F1-мера вместо precision (суммаризация)
- **METEOR** – precision + recall, учитывается порядок слов и синонимы (машинный перевод)

BERT-score

Метрика на основе контекстных эмбеддингов языковой модели.

- Использует семантическую информацию
- Для некоторых задач лучше коррелирует с оценками человека



https://arxiv.org/pdf/1904.09675.pdf

Style Transfer

Style Transfer

Формальный ↔ Неформальный
Проза ↔ Стихотворение
Стиль определенного автора
Позитивный ↔ Негативный

Текст = контекст + стиль

Стилем может быть любая характеристика корпуса текстов.

Style Transfer

Формальный ↔ Неформальный
Проза ↔ Стихотворение
Стиль определенного автора
Позитивный ↔ Негативный

Текст = контекст + стиль

Стилем может быть любая характеристика корпуса текстов.

В отличие от NMT тут (почти) не бывает параллельных корпусов.

Style Transfer Datasets

Непараллельные:

- Yelp, Amason – тональность
- GYAFC – вежливость

Параллельные:

- ParaDetox
- Bibles
- Shakespeare

- GYAFC validation
- Yelp validation

Как мерить качество переноса стиля?

Style Accuracy (точность совпадения стиля)

- Обучить классификатор стиля (часто переобучается)
- Perplexity языковой модели
- Zero-shot LLM (часто работает лучше всего)

Content Preservation (сохранение контента)

- Self-BLEU
- embedding-based metric

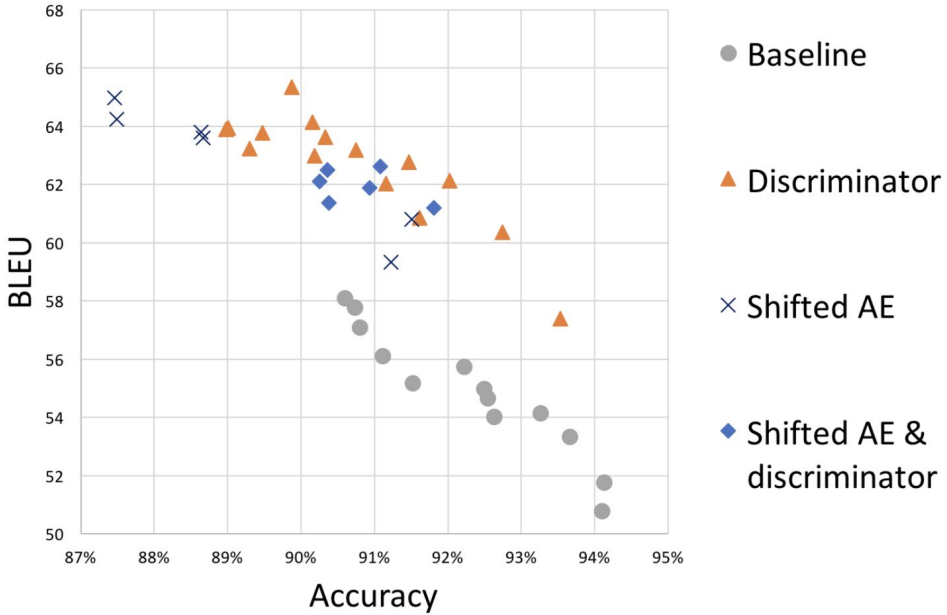
Language Fluency (качество текста в целом)

- Perplexity
- Обучить классификатор (RoBERTa, обученная на CoLA)

https://aclanthology.org/2022.acl-long.469/

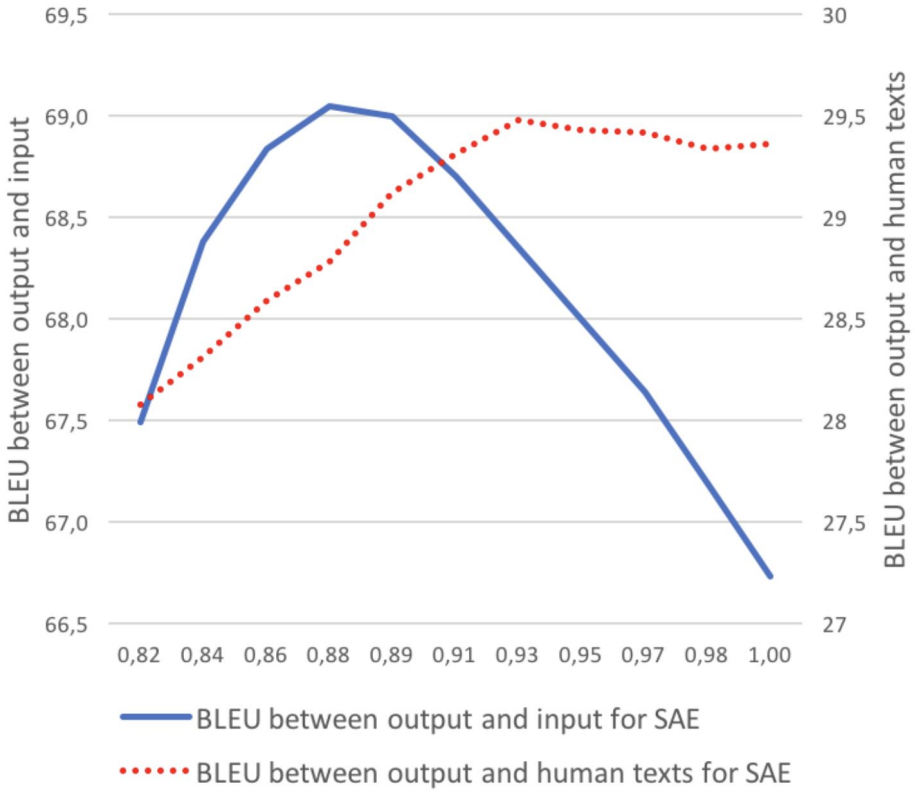
Стиль vs Контекст

Две метрики противоречат друг другу. Часто модель либо сохраняет контекст, либо меняет стиль.



https://arxiv.org/pdf/1908.06809v2.pdf

Self-BLUE очень так себе



https://arxiv.org/pdf/1908.06809v2.pdf

Методы для переноса стиля

- Template-based
- Latent space search
- Disentanglement-based
- Unsupervised NMT-like

Template-based

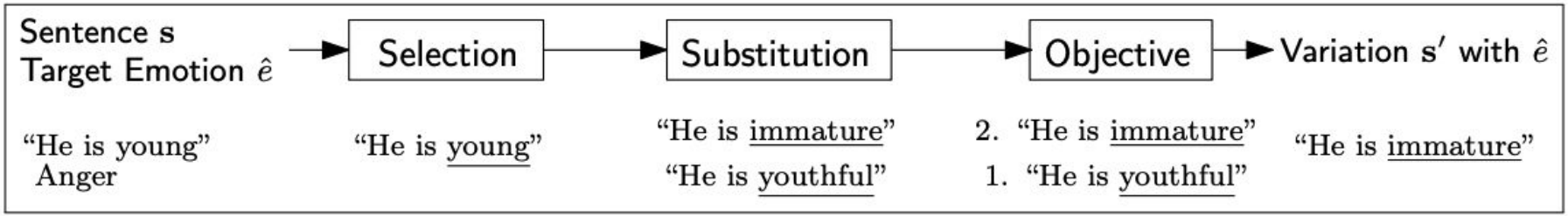


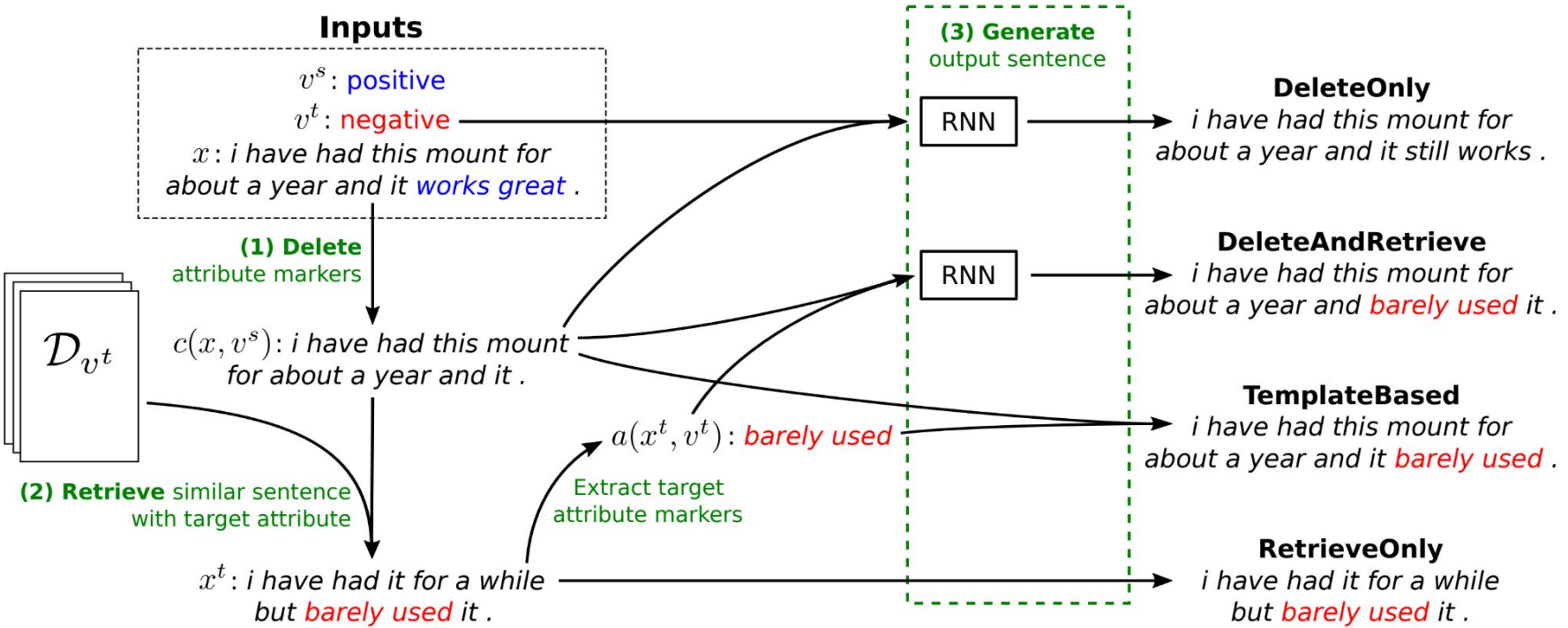
Figure 2: Pipeline model architecture. The selection module marks tokens to substitute, the substitution module retrieves candidates and perform substitution. The objective ranks and scores variations.

$$\text{score}(c, \hat{e}) = \cos(\hat{\mathbf{e}}, \mathbf{c}) - \frac{1}{|E| - 1} \sum_{\bar{e} \in E \setminus \hat{e}} \cos(\bar{\mathbf{e}}, \mathbf{c})$$

$$\mathrm{f}(\mathbf{s}, \mathbf{s}', \hat{e}) = \lambda_1 \cdot \text{emo}(\mathbf{s}', \hat{e}) + \lambda_2 \cdot \text{sim}(\mathbf{s}, \mathbf{s}') + \lambda_3 \cdot \text{flu}(\mathbf{s}')$$

https://aclanthology.org/2020.socialnlp-1.6.pdf

Template-based



https://aclanthology.org/N18-1169.pdf

Template-based

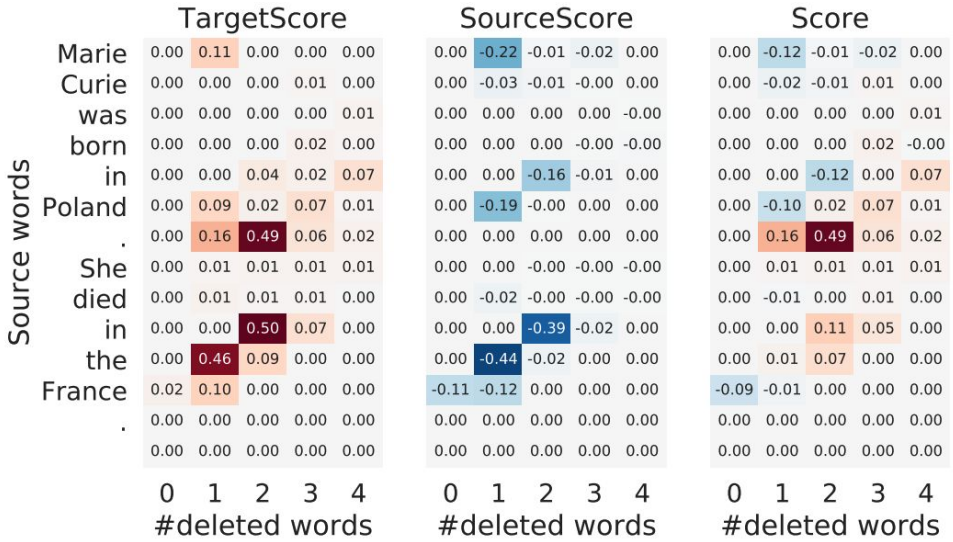


Figure 1: MASKER replaces span ". *She*" by "*and [PAD] [PAD] [PAD]*", resulting in the following fused sentence: *Marie Curie was born in Poland and died in the France .*

Random Sample of Correct MASKER Predictions	
Source	so far i 'm not really impressed .
Prediction	so far i 'm really impressed .
Source	either way i would never recommend buying from camping world .
Prediction	either way i would recommend buying from camping world .
Source	this is a horrible venue .
Prediction	this is a great venue .
Source	this place is a terrible place to live !
Prediction	this place is a great place to live !
Source	i 'm not one of the corn people .
Prediction	i 'm one of the corn people .
Source	this is easily the worst greek food i 've had in my life .
Prediction	this is easily the best greek food i 've had in my life .
Source	the sandwich was not that great .
Prediction	the sandwich was great .
Source	its also not a very clean park .
Prediction	its also a very clean park .

https://aclanthology.org/2020.emnlp-main.699.pdf

Latent space search

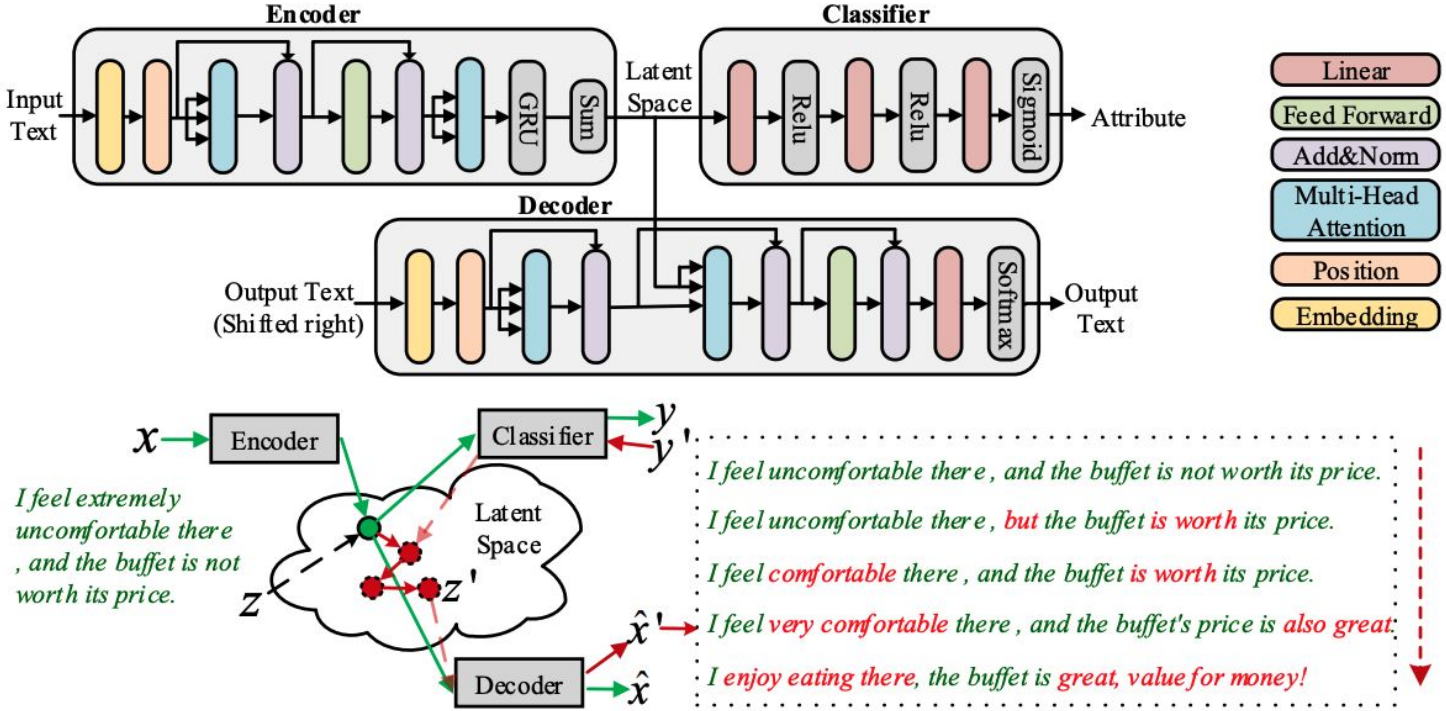


Figure 1: Model architecture.

https://arxiv.org/pdf/1905.12926v2.pdf

Disentanglement-based

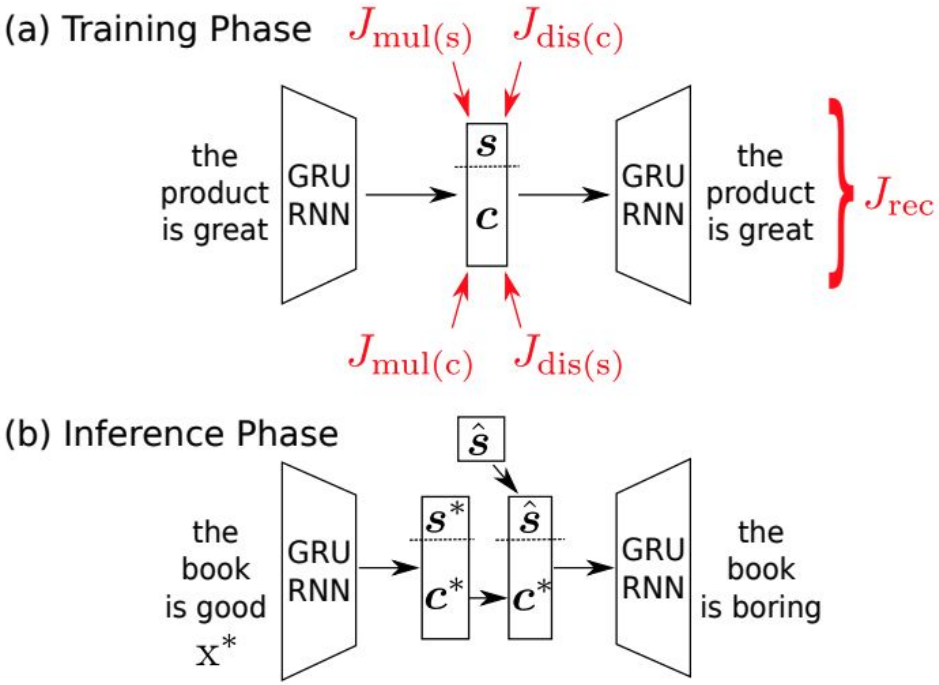
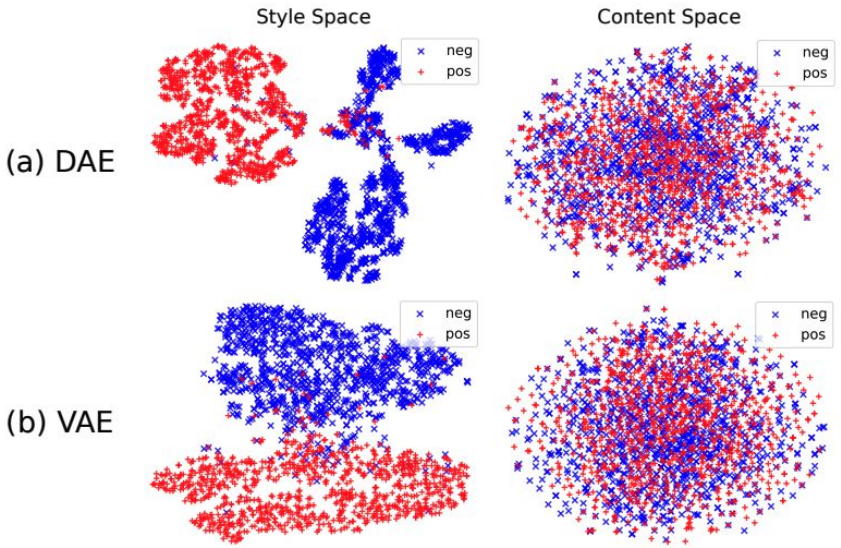
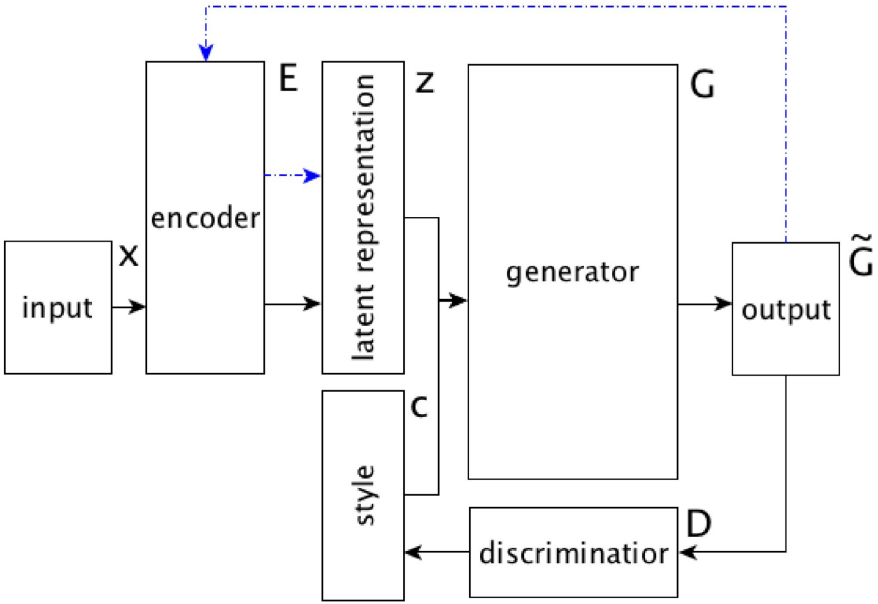


Figure 1: Overview of our approach.



https://arxiv.org/pdf/1808.04339.pdf

Disentanglement-based



https://arxiv.org/pdf/1908.06809v2.pdf

Disentanglement-based

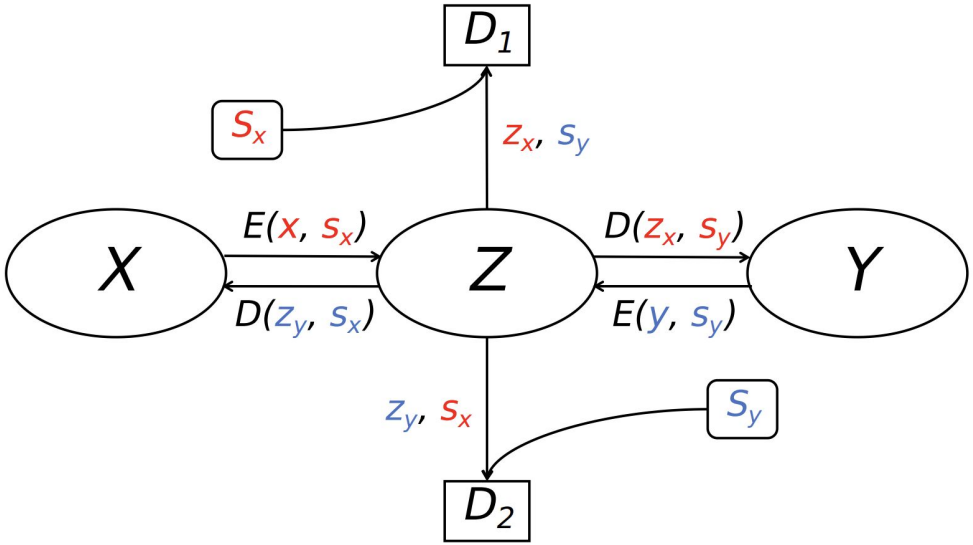


Figure 1: CrossAlign architecture

https://arxiv.org/pdf/2004.11742.pdf

Disentanglement-based

Disentanglement is not happening

λ_{adv}	Discriminator Acc (Train)	Post-fit Classifier Acc (Test)
0	89.45%	93.8%
0.001	85.04%	92.6%
0.01	75.47%	91.3%
0.03	61.16%	93.5%
0.1	57.63%	94.5%
1.0	52.75%	86.1%
10	51.89%	85.2%
fastText	-	97.7%

https://arxiv.org/pdf/1811.00552.pdf

Disentanglement-based

Disentanglement is not happening

λ_{adv}	Discriminator Acc (Train)	Post-fit Classifier Acc (Test)
0	89.45%	93.8%
0.001	85.04%	92.6%
0.01	75.47%	91.3%
0.03	61.16%	93.5%
0.1	57.63%	94.5%
1.0	52.75%	86.1%
10	51.89%	85.2%
fastText	-	97.7%

$$\mathcal{L} = \lambda_{AE} \sum_{(x,y)\sim\mathcal{D}} -\log p_d\Big(x|e(x_c), y\Big) + \lambda_{BT} \sum_{(x,y)\sim\mathcal{D},\tilde{y}\sim\mathcal{Y}} -\log p_d\Big(x|e\Big(d\big(e(x),\tilde{y}\big)\Big), y\Big)$$

https://arxiv.org/pdf/1811.00552.pdf