Conversion of Persian Colloquial Texts into Official Texts using Unsupervised Learning Methods

Progress Report

Dr. Hossein Sameti
Karim Akhavan

Example

من این وسیله رو از مغازه بابام برداشتم.

↓

من این وسیله را از مغازه پدرم برداشتم.

Implementation

- **Task:** Style Transfer[1]
- **Method:** Unsupervised Learning
- **Model:** Transformers
- **Datasets:** Digikala Sentiment Review & Wikipedia Articles

[1] "The task of changing the stylistic properties (e.g., sentiment) of the text while retaining the style-independent content within the context." (Dai, N., Liang, J., Qiu, X., & Huang, X. (2019)

Style Transfer

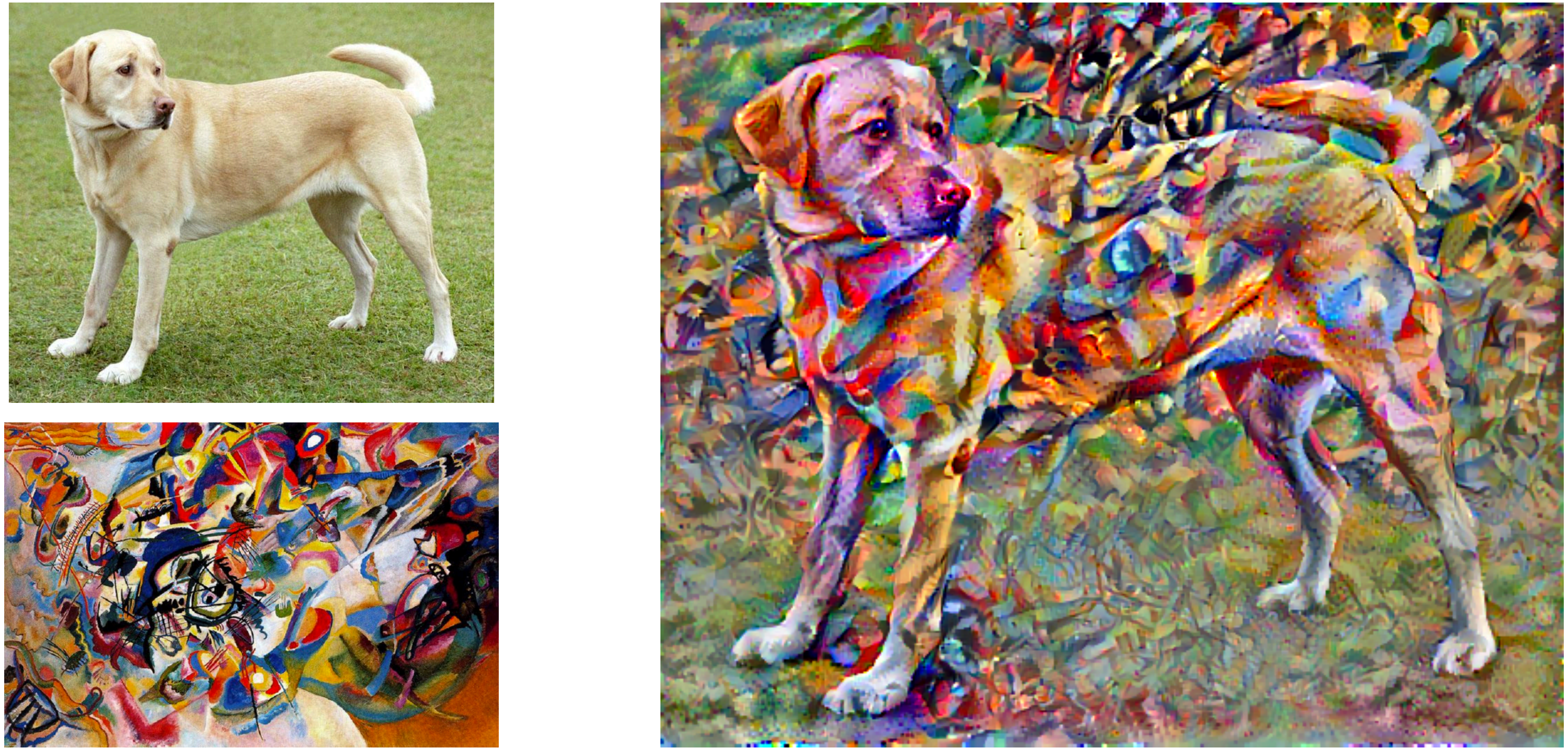
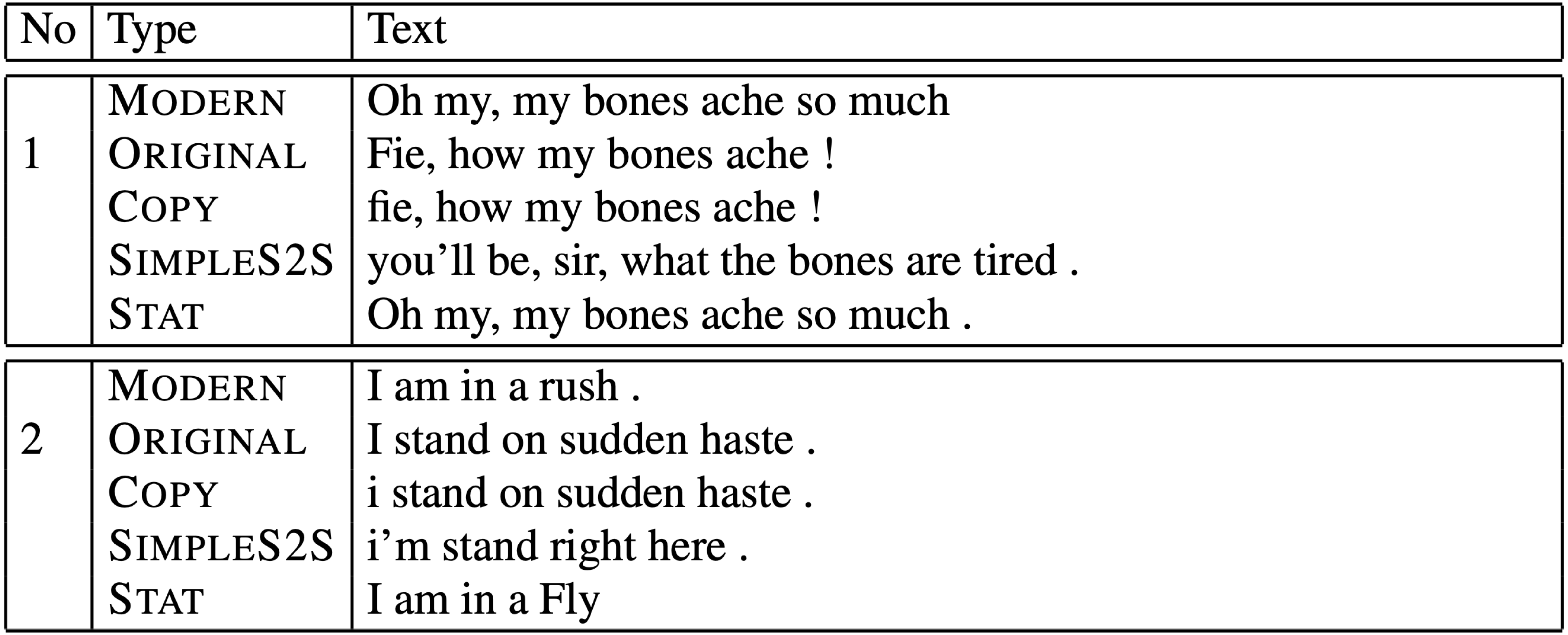


Fig 1. Neural style transfer. Tensorflow. https://www.tensorflow.org

Style Transfer



No	Type	Text
1	Modern	Oh my, my bones ache so much
	Original	Fie, how my bones ache !
	Copy	fie, how my bones ache !
	SimpleS2S	you'll be, sir, what the bones are tired .
	Stat	Oh my, my bones ache so much .
2	Modern	I am in a rush .
	Original	I stand on sudden haste .
	Copy	i stand on sudden haste .
	SimpleS2S	i'm stand right here .
	Stat	I am in a Fly

Source: Shakespearizing Modern Language Using Copy-Enriched Sequence-to-Sequence Models (Jhamtani et al, 2017)

Style Transfer

- A machine translation task or even (controlled) text generation.

- Usually through **disentangling** sentence representations in a shared latent space (by using an adversarial approach to learn latent representations, ignoring stylistic informations). A **decoder** is then fed with the latent representation along with **attribute labels** to generate a variation of the input sentence with different attributes.

- Most studies have focused on changing only one attribute.

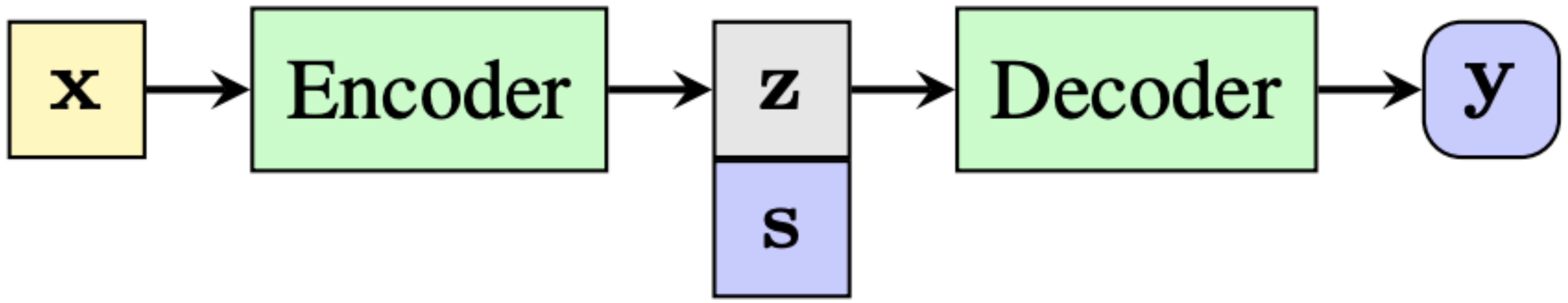
Previous Works

- Converting Persian Colloquial Text to Official Text at the Level of Grammar and Vocabulary. (RajabPour,. Bahrani, 2018)
- Converting Persian Colloquial Texts with the Help of N_grams. (Armin,. Shamsfard, 2011)
- **“Encoder-decoder” frameworks with RNN as both encoder & decoder** (Most common): The encoder maps the text into a style-independent latent representation (vector representation), and the decoder generates a new text with the same content but a different style from the disentangled latent representation plus a style variable.
    - Inferring a latent representation for the input sentence & manipulating the style of the generated sentence based on this learned latent representation → Issue: The model which has assumed a fixed size latent representation cannot utilize the information from the source sentence anymore.
    - **Cross-aligned Auto-encoder** with adversarial training for learning shared latent content and separated latent style distributions. Used variational auto-encoder (VAE) as base model and leveraged an adversarial training scheme where a binary CNN-based discriminator is used to evaluate whether a transferred sentence is real or fake, ensuring that transferred sentences match real sentences in terms of target style. (Shen et al., 2017)
    - **Generative model** which combines variational auto-encoders and holistic attribute discriminators for effective imposition of semantic structures. Used a style classifier to directly enforce the desired style in the generated text. (Hu et al., 2017). Many works in 2018 followed their approach (based on encoder-decoders (Fu et al., 2018; John et al., 2018; Zhang et al., 2018a,b))
    - **Variational Auto-encoder** (VAE), using non-parallel data. (Mueller et al., 2017)
- **Style Transfer through Back-Translation:** 1. Use back-translation to rephrase the sentence and reduce the effect of the original style 2. Generate from the latent representation, using separate style-specific generators controlling for style
- **Copy-Enriched Sequence-to-Sequence Models :** Shakespearizing Modern Language Using Copy-Enriched Sequence-to-Sequence Models (with parallel data) (Jhamtani et al., 2017). Also used a dictionary providing mapping between Shakespearean words and modern English words to enhance pre-trained word embeddings.

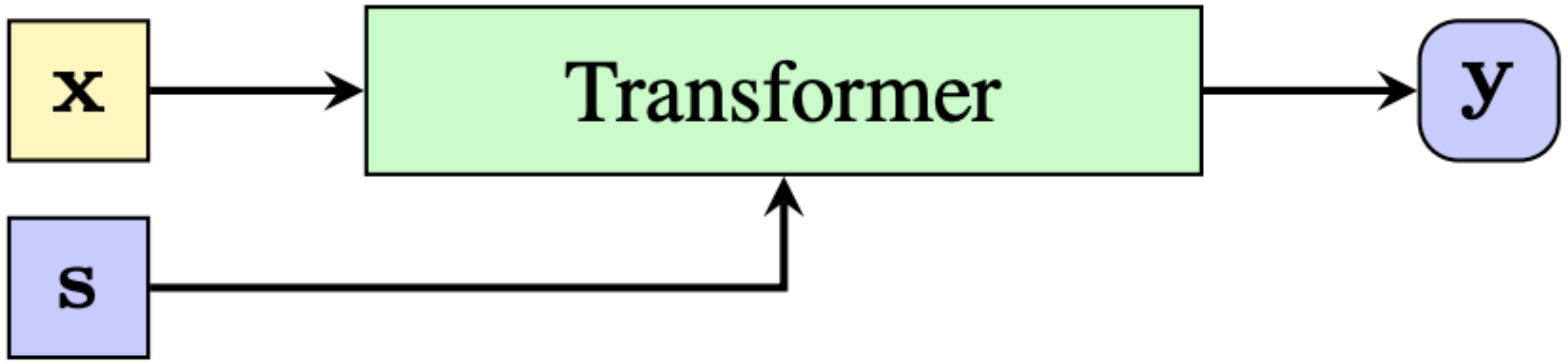
Previous Works

- Dai et al. (2019): Suggest the use of **Transformers** (which they call “Style Transformer”). In contrast with previous works, it doesn’t assume a disentangled latent representation for manipulating the sentence style. Instead, they've used a transformer which learns a mapping function f(x, s) ⇒ x^, where x is a sentence and s is a style control variable.
  - Discriminator transformer assists style transformer in generating the input sentence (By distinguishing between the styles)
  - For evaluation, there is a reference same size as the test dataset which is used to score (BLEU) the output of model

Previous Works



(a) Disentangled Style Transfer



(b) Style Transformer

Thank You for Your Attention!