Position Papers of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 81–86
DOI: 10.15439/2025F4661
ISSN 2300-5963 ACSIS, Vol. 44

Adapting CycleGAN architecture for Unpaired Diachronic Text Style Transfer

Adrian Niedziółka-Domański
0009-0003-4797-7484
Maria Curie-Sklodowska University
Plac Marii Curie-Skłodowskiej 5
20-031 Lublin, Poland
Email: niedziolkadadrian@gmail.com

Jarosław Bylina
0000-0002-0319-2525
Maria Curie-Sklodowska University
Plac Marii Curie-Skłodowskiej 5
20-031 Lublin, Poland
Email: jaroslaw.bylina@mail.umcs.pl

***Abstract*—Diachronic text style transfer aims to transform text from one historical period into the style of another while preserving its meaning. However, the scarcity of parallel corpora across time periods makes supervised approaches impractical. In this work, we propose to adapt the CycleGAN architecture, originally developed for unpaired image-to-image translation, to model linguistic change over time. Our method employs a generator and discriminator, both conditioned on temporal information, and trained using a combination of adversarial and cycle-consistency losses. We propose a time-conditioned generative framework that supports both discrete and continuous temporal representations, enabling the model to interpolate between historical language styles. The model is trained on unaligned historical texts and can transform language from any period to another. This approach offers a data-efficient solution for diachronic language modeling and opens new research directions in historical linguistics, digital humanities, and unsupervised style transfer.**

I. Introduction

ONE of the main challenges in working with diachronic textual data lies in the limited availability of directly aligned texts from different historical periods. Unlike modern translation datasets, where sentence-level or even word-level correspondences are often available, historical corpora typically lack such parallel structures. For instance, there is rarely a source with one-to-one correspondence between a text written in Middle English and its equivalent in Modern English. This absence of parallel data complicates efforts to apply conventional supervised methods to historical language normalization, translation, or style transfer tasks. As a result, there is a growing need for methods capable of learning mappings between historical and modern language forms without relying on direct supervision or aligned corpora.

In this paper, we propose adapting the CycleGAN architecture, originally developed for unpaired image-to-image translation [1], to the domain of unpaired diachronic text style transfer. Our goal is to demonstrate that the CycleGAN framework, with appropriate modifications for textual data, can serve as a viable approach to style transformation across time periods.

II. Background and Related Work

CycleGAN is a type of Generative Adversarial Network (GAN) [2] designed for unpaired data translation, originally proposed for image-to-image translation tasks (Zhu et al., 2017) [1]. The model consists of two generators and two discriminators. Each generator learns to map data from one domain to another (e.g., from domain A to B, and from B to A), while each discriminator evaluates whether the generated output appears realistic within its respective domain (Fig. 1).

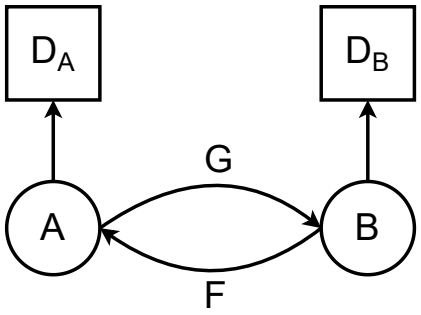

Fig. 1. CycleGan architecture, based on [1]. G and F are generators, A and B are two domains and D_A and D_B are discriminators working in these domains.

A core innovation of CycleGAN is the cycle consistency loss, which ensures that if an input sample is translated to the target domain and then back to the original domain, the result should closely resemble the initial input. This regularization term helps the model retain the core content of the source while adjusting its style to match the target domain.

In formal terms, given two domains X and Y, and two generators $G : X \to Y$ and $F : Y \to X$, the cycle consistency loss is defined as:

$$\mathcal{L}_{cyc}(G, F) = \mathbb{E}_{x \sim p_{data}(x)} \left[\|F(G(x)) - x\|_1 \right] + \mathbb{E}_{y \sim p_{data}(y)} \left[\|G(F(y)) - y\|_1 \right]$$

This encourages $F(G(x)) \approx x$ and $G(F(y)) \approx y$, thereby enforcing that the content of the input is retained after a round-trip translation.

A good example of this process involves translating images of horses to zebras and back again. Even without paired examples (i.e., no exact horse-zebra image pairs), the network

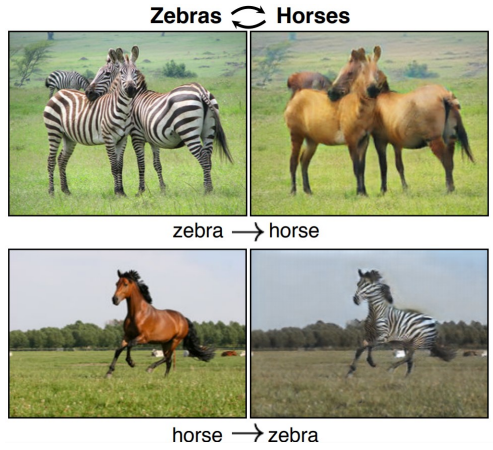

Fig. 2. Example of image to image translation using CycleGan presented in original paper [1].

learns meaningful transformations through adversarial learning combined with the cycle consistency constraint (Fig. 2).

While CycleGAN enables unpaired translation between two domains, it does not scale efficiently to scenarios involving multiple domains. Each pair of domains would require separate generator and discriminator pairs, which would make the model increasingly complex and computationally expensive. In contrast, StarGAN [3] extends the CycleGAN framework to support multi-domain translation within a single unified architecture. StarGAN achieves this by conditioning the generator and discriminator on domain labels, enabling style transfer across many categories using a shared set of parameters. The generator $G(x, c')$ takes an input sample x and a target domain label c', and produces an output in the desired style. The discriminator not only distinguishes real from fake samples but also predicts their domain label (Fig. 3).

In this work, although our primary architecture is inspired by CycleGAN, we also leverage the principles of StarGAN to investigate possible multi-era style transformation tasks. This enables flexible style transfer across multiple historical stages, effectively allowing the model to map between linguistic variants from different centuries using a unified, conditional architecture - without the need to train separate models for each specific pair of eras.

A. Adaptations of CycleGAN for Unsupervised Text Style Transfer

Although CycleGAN was originally proposed for unpaired image-to-image translation, its underlying principles have inspired a number of adaptations in the field of natural language processing and also in tasks involving unsupervised text style transfer. The goal of these adaptations is to utilize CycleGAN's ability to learn mappings between two domains without the need for aligned or parallel training data, a feature especially relevant when working with diachronic corpora or stylistically divergent text.

One of the contributions in this direction is the work by Huang et al. [4], where they proposed a Cycle-Consistent Adversarial Autoencoder model designed specifically for unsupervised text style transfer. Their method combines an autoencoder with cycle consistency loss and adversarial training, allowing the model to keep the semantic content while changing the writing style.

Lorandi et al. [5] proposed a more direct application of the CycleGAN architecture to text style transfer, where they focused on sentiment transformation between positive and negative expressions. Their model, called TextCycleGAN, works without paired data and uses cycle-consistent adversarial training to learn bidirectional mappings between different texts. Although they use a fairly basic LSTM design for both generators and discriminators, their results on the Yelp dataset achieve strong sentiment accuracy and fluency, proving that CycleGAN can work effectively with text data.

Similarly, Wang et al. [6] used CycleGAN for a more structured task: converting abstracts into conclusions in scientific papers. By considering abstracts and conclusions as different stylistic domains, they demonstrated that CycleGAN can learn style transformation patterns within specialized types of text, using only unpaired data.

These studies show the growing potential of CycleGAN-inspired architectures for text style transfer tasks. By demonstrating that effective stylistic transformations can be achieved in an unsupervised manner, without the need for aligned or parallel corpora, they lay the groundwork for extending such approaches to more complex linguistic domains. They provide a strong basis for exploring how CycleGAN-based models might perform in the context of diachronic language data, where the scarcity of parallel examples across historical periods makes supervised approaches very hard to implement.

This motivates our own research idea, in which we adapt the CycleGAN framework to perform style transfer between different variants of the language over the centuries.

III. Proposed Method

A. Task Formulation

The goal of this work is to perform unpaired diachronic text style transfer. That is, given a piece of text written in the linguistic style of a certain historical period, e.g., the 15th century, we aim to generate a version of that text that retains its original meaning but is expressed in the linguistic style of a different period, such as the 21st century.

Crucially, we assume there are no parallel corpora linking these periods. That means we do not have direct sentence-level alignments between time periods. This makes the task a fully unsupervised sequence transformation problem.

B. Formal Setting

Formally, let $\mathbb{T}$ denote the temporal domain associated with linguistic style. We consider two possible representations of time:

1) Discrete time domain:

$$\mathbb{T}_{\text{disc}} = \{t_1, t_2, \ldots, t_n\}$$

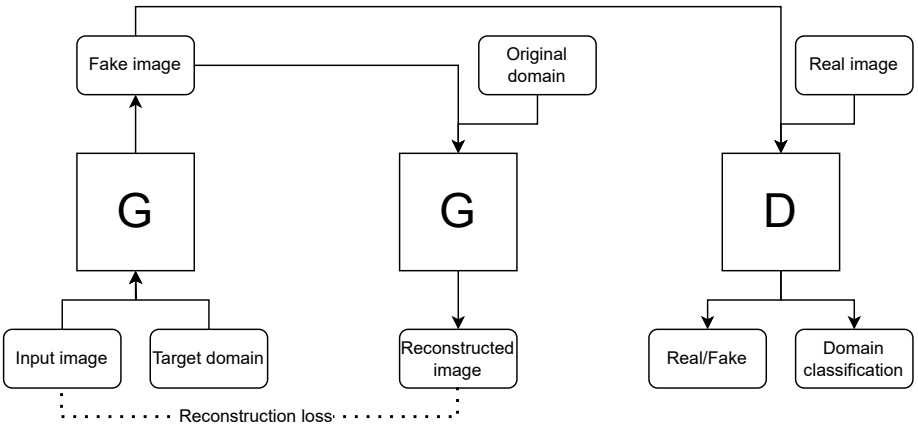

Fig. 3. Overview of StarGAN, consisting of two modules, a discriminator D and a generator G, based on [3].

where each t_i corresponds to a fixed historical period (e.g., 15th century, 16th century, etc.), or to broader linguistic eras (e.g., Old English, Middle English, Modern English).

2) Continuous time domain:

$$\mathbb{T}_{\text{cont}} \subset \mathbb{R}$$

where time is modeled as a real-valued scalar, such as the year or century of origin. This formulation allows the model to reason about intermediate or underrepresented styles and enables smooth interpolation across time.

In this work, we emphasize the continuous representation due to its potential for fine-grained modeling of historical language change. However, the proposed framework remains compatible with discrete labels, which may be more practical in cases where time annotations are coarse or categorical.

Let $x \in X_t$ denote a text sample originating from time period $t \in \mathbb{T}$. Our objective is to learn a generative function:

$$G(x, t') \to \hat{x}_{t'}$$

where t' is the target time period and $\hat{x}_{t'}$ is a text that preserves the meaning of x but adopts the linguistic characteristics of period t'.

To ensure that the model preserves the semantic content of the input, we adopt a cycle-consistency mechanism inspired by StarGAN, where a single shared generator G is used for both forward and reverse transformations. Specifically, given a source text x from time period t, we first translate it to the target style t', and then we use the same generator to map $\hat{x}_{t'}$ back to the original style t:

$$G(G(x, t'), t) \approx x$$

This should allow us to enforce that the transformation is approximately invertible, encouraging the generator to preserve content while altering only the stylistic features associated with time.

C. Model Losses

Our model is suppose to be trained using a combination of adversarial and cycle-consistency objectives, adapted for the temporal style transfer task.

1) Adversarial Loss: Rather than using separate discriminators for each time domain (as in CycleGAN), we employ a single shared discriminator D that is conditioned on the target time period t'. Its objective is to perform **real/fake classification**–that is, to determine whether a given sentence is a genuine example from time t' or a synthetic sample generated by the model. By conditioning on t', the discriminator learns to judge the temporal authenticity of the input relative to the specified style period.

The generator $G(x, t')$ attempts to transform a text sample x from its original time period t into the style of target time t'. The discriminator then assesses whether the result is:

1) **Authentic**, and
2) **Temporally consistent** with t'

To train this system adversarially, we define the adversarial loss as follows:

$$\mathcal{L}_{adv} = \underbrace{\mathbb{E}_{x' \sim p_{\text{data}}(x'|t')}[\log D(x', t')]}_{\text{real samples from target time } t'} + \underbrace{\mathbb{E}_{x \sim p_{\text{data}}(x|t),\ t \neq t'}[\log(1 - D(G(x, t'), t'))]}_{\text{generated samples styled for } t'}$$

This formulation encourages the discriminator to correctly distinguish real samples from generated ones. Specifically, it rewards the discriminator for identifying genuine examples from time t', and penalizes it when it fails to detect synthetic ones. The generator, conversely, is optimized to fool the discriminator into classifying its outputs as authentic. So it ensures that G learns to generate text indistinguishable from true samples belonging to the target time period t'.

2) Cycle-Consistency Loss: To ensure that the semantic content of the text is preserved during style transfer across dif-

ferent time periods, we impose a cycle-consistency constraint using the same generator G. Formally, this loss is defined as:

$$\mathcal{L}_{cyc} = \mathbb{E}_{x,t,t'} \left[\|G(G(x, t'), t) - x\|_1 \right]$$

This term encourages the model to reconstruct the original input text x after sequentially transforming it to a different temporal style t' and then back to its original style t. By minimizing this reconstruction error, the model is guided to produce style-transferred outputs that maintain the original meaning and content, rather than simply generating stylistically plausible but semantically unrelated text. In essence, cycle-consistency enforces that the transformations are invertible and that the core semantic information remains stable across diachronic style mappings.

3) Full Objective: The overall training objective combines the adversarial loss and cycle-consistency loss to jointly optimize generator G and discriminator D. Formally, the full loss function is given by:

$$\mathcal{L} = \mathcal{L}_{adv} + \lambda_{cyc}\mathcal{L}_{cyc}$$

where λ_{cyc} is a hyperparameter that balances the importance of cycle-consistency relative to the adversarial loss.

The generator G aims to minimize this combined loss, learning to produce temporally consistent and semantically faithful style transfers, while the discriminator D is trained to maximize the adversarial loss, improving its ability to distinguish real from generated samples conditioned on the target time period.

Thus, the model should achieve effective unpaired diachronic text style transfer by encouraging realistic temporal style generation and content preservation simultaneously.

D. Proposed Model

Our proposed model adopts a transformer-based architecture, drawing inspiration from CycleGAN and StarGAN, specifically designed for diachronic text style transfer. The model consists of two main components:

- Generator $G(x, t')$
- Discriminator $D(x, t')$

The generator will probably be a conditional transformer encoder-decoder model [7] that transforms a given input text x from its original time period t into the style of a target time t'. The temporal condition t' will be injected into the model in the decoder part.

The discriminator will most likely be a decoder-only transformer that evaluates whether the input x is a real sample drawn from the target time period t' or a generated one. It receives the time condition t' as additional input and is trained to perform binary classification (real/fake) with respect to this condition.

Fig. 4 illustrates the overall model architecture, including the generator and discriminator modules, temporal conditioning flow, and the cycle path used during training.

Algorithm 1 Training Procedure

Require: Training corpus $\mathcal{D} = \{(x_i, t_i)\}$ with time labels
for each minibatch $\{(x_i, t_i)\}_{i=1}^{N}$ sampled from $\mathcal{D}$ **do**
  for each x_i in minibatch **do**
    Select a target time $t'_i \in \mathbb{T} \setminus \{t_i\}$ uniformly at random
    Generate transformed sentence: $\hat{x}_{t'_i} \leftarrow G(x_i, t'_i)$
    Reconstruct original: $\hat{x}_{t_i} \leftarrow G(\hat{x}_{t'_i}, t_i)$
  end for
  Compute adversarial loss $\mathcal{L}_{adv}$
  Compute cycle-consistency loss $\mathcal{L}_{cyc}$
  Update discriminator D to maximize $\mathcal{L}_{adv}$
  Update generator G to minimize $\mathcal{L}_{adv} + \lambda_{cyc}\mathcal{L}_{cyc}$
end for

E. Training and Evaluation

The proposed model is trained end-to-end using a combination of adversarial and cycle-consistency losses. As shown in algorithm 1, training proceeds by iterating over mini-batches of text samples drawn from the training corpus $\mathcal{D} = \{(x_i, t_i)\}$, where each sample is annotated with its corresponding time period t_i.

For each input sentence x_i in a mini-batch, a target time t'_i is randomly sampled from the set of available time labels, excluding the original time t_i. The generator G then transforms the sentence into the style of the target time, producing $\hat{x}_{t'} = G(x, t')$. To enforce semantic preservation, this generated sample is passed again through the generator, that is now conditioned on the original time period to reconstruct the source sentence: $\hat{x}_t = G(\hat{x}_{t'}, t)$, where $\hat{x}_t \approx x$.

After all forward and backward transformations are completed for the mini-batch, two loss functions are computed:

- The adversarial loss $\mathcal{L}_{adv}$ encourages the discriminator D to distinguish real samples from generated ones, while guiding the generator to produce temporally consistent and realistic outputs.
- The cycle-consistency loss $\mathcal{L}_{cyc}$ enforces that the content of the original sentence is preserved across the round-trip transformation between time styles.

The discriminator is updated to maximize the adversarial loss, while the generator is updated to minimize a weighted combination of both losses: $\mathcal{L}_{adv} + \lambda_{cyc}\mathcal{L}_{cyc}$.

Due to the lack of parallel diachronic corpora, automatic evaluation is challenging. We propose the following evaluation strategies:

- **Temporal Classification Accuracy**: A pretrained time classifier can be used to assess whether generated samples are stylistically consistent with the target time period.
- **Cycle Reconstruction Error**: Content preservation can be approximated, for example, by measuring the L_1 distance between the input sentence and its reconstruction after a cycle pass.
- **Human Evaluation**: Expert evaluations by historians or linguists can offer valuable insights into the fluency,

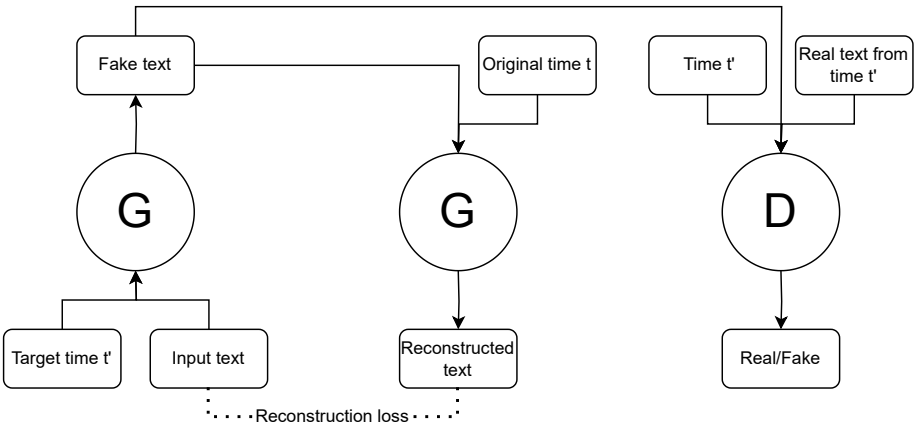

Fig. 4. Proposed model architecture

semantic accuracy, and historical authenticity of the generated text.

This training strategy should help the model learn a smooth, time-aware text style transformation function that can generalize across different historical periods, even without parallel supervision. By conditioning both the generator and discriminator on a continuous temporal index, the model may learn to recognize subtle patterns in the evolution of linguistic features over time. Instead of memorizing fixed mappings between specific time periods, the generator will learn to interpolate and extrapolate stylistic attributes across the temporal space. This should allow flexible text generation at arbitrary points in the historical timeline, including periods for which little or no direct training data exists.

IV. Expected contribution

This paper introduces a new framework for diachronic text style transfer, allowing sentences written in the style of one historical era to be transformed into stylistically consistent versions from another period. Unlike conventional style transfer approaches that often depend on aligned or parallel corpora, our method operates in a fully unsupervised manner, enabling training on naturally occurring, unaligned historical texts. This marks a substantial advancement in tackling the issues of the scarcity of alligned data in diachronic text corpora.

A core innovation of our model lies in its use of continuous time representations to condition both the generator and the discriminator. Rather than assigning fixed domain labels (e.g., "15th century" or "modern English"), time is treated as an input variable, allowing the model to learn smooth, temporally-aware transitions between language styles. This enables more granular control over the generated outputs and allows the model to capture linguistic change over time as a continuous process based on continuous data, rather than relying on discrete class divisions. Moreover, by viewing time as a continuous factor, the model could potentially predict extrapolate how language might evolve and even create believable future versions of it.

We adapt adversarial and cycle-consistency learning techniques, originally developed for images (CycleGAN, StarGAN), to the domain of natural language. Our proposed architecture uses a single shared generator trained with a combination of adversarial and cycle-consistency objectives, ensuring that generated sentences not only match the target time's style but also preserve the original semantic content. This allows the model to strike a balance between stylistic transformation and content fidelity, which is critical for meaningful diachronic translation.

This work advances the field of diachronic NLP by introducing a general, data-efficient method for modeling language over time. It creates new opportunities for research in historical language translation and computational philology. By combining ideas from image style transfer and natural language processing, this study provides a foundation for future models that better understand, generate, and adapt text across various historical periods.

Potential applications of our approach include the modernization of historical documents, stylistic harmonization of corpora for linguistic research and speculative modeling of future language evolutions.

V. Possible limitations and future work

Although the proposed model is looking very promising, it can also have several possible limitations. One of the big issues is the lack of large, high-quality diachronic corpora that cover long historical periods. This can limit the variety and reliability of the transformations the proposed model can learn. In addition, the current architecture may not be able to completely capture the complexity of language changes, such as shifts in grammar, meaning, or vocabulary. As a result, the model may focus on surface-level characteristics while overlooking some deeper linguistic structures.

Another limitation occurs when the training data only covers distant time periods, for example, the 15th and 21st centuries. In such cases, the model may produce intermediate linguistic forms that did not exist in the past. The challenge of learning smooth transformations across large temporal gaps becomes evident in this interpolation task. To overcome this issue, temporal conditioning must be carefully designed, potentially incorporating constraints informed by historical linguistic knowledge.

Future research could focus on utilizing outside linguistic knowledge, such as syntactic parsers or etymological databases, to improve semantic preservation and style accuracy. The model could also be adjusted for finer temporal resolutions, like working with decades instead of entire centuries, or expanded to manage multilingual historical corpora.

REFERENCES

[1] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks," in *2017 IEEE International Conference on Computer Vision (ICCV)*. Venice: IEEE, Oct. 2017. doi: 10.1109/ICCV.2017.244. ISBN 978-1-5386-1032-9 pp. 2242–2251. [Online]. Available: http://ieeexplore.ieee.org/document/8237506/

[2] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Networks," Jun. 2014, arXiv:1406.2661 [stat]. [Online]. Available: http://arxiv.org/abs/1406.2661

[3] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "StarGAN: Unified Generative Adversarial Networks for Multi-domain Image-to-Image Translation," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Salt Lake City, UT: IEEE, Jun. 2018. doi: 10.1109/CVPR.2018.00916. ISBN 978-1-5386-6420-9 pp. 8789–8797. [Online]. Available: https://ieeexplore.ieee.org/document/8579014/

[4] Y. Huang, W. Zhu, D. Xiong, Y. Zhang, C. Hu, and F. Xu, "Cycle-Consistent Adversarial Autoencoders for Unsupervised Text Style Transfer," in *Proceedings of the 28th International Conference on Computational Linguistics*, D. Scott, N. Bel, and C. Zong, Eds. Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020. doi: 10.18653/v1/2020.coling-main.201 pp. 2213–2223. [Online]. Available: https://aclanthology.org/2020.coling-main.201/

[5] M. Lorandi, A. Mohamed, and K. McGuinness, "Adapting the CycleGAN architecture for text style transfer." Galway, Ireland: Zenodo, Aug. 2023. doi: 10.5281/zenodo.8268838. [Online]. Available: https://doi.org/10.5281/zenodo.8268838

[6] H. Wang, Y. Lepage, and C. L. Goh, "Unpaired Abstract-to-Conclusion Text Style Transfer using CycleGANs," in *2020 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*. Depok, Indonesia: IEEE, Oct. 2020. doi: 10.1109/ICACSIS51025.2020.9263246. ISBN 978-1-7281-9279-6 pp. 435–440. [Online]. Available: https://ieeexplore.ieee.org/document/9263246/

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is All you Need," in *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html