UCL Natural Language Processing 2018/19 Project: Text Style Transfer from Unsupervised Machine Translation

Dor Hacohen
University College London
dor.hacohen.18@ucl.ac.uk

Anthony Bourached
Trinity College Dublin
bouracha@tcd.ie

Sean Gupta
Imperial College London
ucabas8@ucl.ac.uk

Abstract

We address the task of text style transfer, in which a source text is transformed in such a way as to change its style while preserving its content. Style transfer has been successfully applied to images but is a more challenging endeavour for text due to its discrete nature. We leverage recent techniques in unsupervised machine translation from monolingual, non-parallel corpora to change the artistic genre of songs in the MetroLyrics database from rock to hiphop and vice-versa. This is achieved through projecting lyrics from two distinct styles into a shared embedding space that preserves the content of the lyrics.

1 Introduction

We follow (2). Text style transfer seeks to change style of a source text while at the same time maintaining its semantics. This has been successfully applied to changing informal sentences to formal ones, and to transform the language of traditional Shakespearean plays into modern English. Text style transfer is useful in the following application areas: (4)

- It can aid authors of legal or technical documents in adhering to required stylistic guidelines so the end text conforms to the needs of the expert audience in terms of register (vocabulary used), grammaticality (no slang or uncalled for abbreviations) as well as voice (e.g. passive voice in technical or scientific documents).

- It can enable non-experts to understand technical documents by translating them into simple English. An example of this is the Simple English version of Wikipedia. While in and of itself, the "standard" Wikipedia is considered to be non-technical, it might still be hard to understand and less accessible for laymen. The Simple English version of Wikipedia explains concepts in a more accessible way with a simpler register and less technical jargon, but with everyday language and deploying many more examples. Similarly, many government websites now offer a simplified version of their citizen information.

- It can power educational applications, such as generating modern language version of traditional plays and novels. A well-known examples are the highschool-suitable translations of Shakespearean plays into modern English by Sparknotes.com as students are no longer well-versed in Early Modern English. Similarly, many students of Latin struggle with traditional writings of Cicero and other Roman authors and find simplified comic book versions easier to understand. While one can debate the educational merits of such an approach, it tends to engage students more in the lessons so that they are no longer perceived as boring.

We can thus see that the applications of text style transfer are plentiful. We note however that this is a very new area in natural language processing (NLP) and as such there do not exist any commercially viable products yet. A parallel might be drawn to programs such as Xrumer that aid black-hat engineers in search engine optimisation by rewriting website content in such a way that search engines are fooled to view this as additional "unique" content, hence leading to a better ranking in web searches. However, these kinds of software rely on simple phrase exchange and synonym lookups.

To show the reader what we mean by this, here is the preceding paragraph rewritten by "Spinbot":

We would thus be able to see that the utilizations of content style exchange are ample. We note anyway this is another zone in normal language preparing (NLP) and all things considered there don't exist any economically practical items yet. A parallel may be attracted to projects, for example, Xrumer that guide dark cap builds in site improvement by revamping site content so that web crawlers are tricked to see this as extra "remarkable" content, subsequently prompting a superior positioning in web looks. Nonetheless, these sorts of programming depend on straightforward expression trade and equivalent word queries.

As we can clearly see, this kind of text rewriting might be enough to fool a web crawler but not the human eye. Similar challenges are encountered in text style transfer as we shall see.

Style transfer can be viewed as carrying out a translation of text from a source domain into a target domain. In the field of Artificial Intelligence (AI), machine translation has long and storied history. (3) It is seen as a challenging problem because a good translation necessitates knowledge of the semantic structure of a sentence as well as an understanding of the natural world through commonsense knowledge (problem of ontological engineering, i.e. knowledge representation – no expert systems in use anymore!). However, natural language is usually ambiguous and does not adhere to normative syntactical rules but is in a constant flux.

Historically, rule-based system were deployed to translate from a source to a target language. Rule-based systems work by parsing input sentences according to normal grammatical rules. These might take forms such as "sentence = subject + verb + object" or "subject = article + noun". However, these systems are no longer in practical use and have been supplanted by statistical machine translation, which builds language and translation models from parallel corpora. At a high level, this works as follows. Given the task of translating a source sentence e (for example, in English) into a target sentence f (for example, in French), we need a target languge model (LM) $P(f)$ as well as a translation model $P(f \mid e)$. We seek to find

$$f^\star = \text{argmax}_f P(f \mid e) = \text{argmax}_f P(e \mid f)P(f).$$

This is done by

1. Breaking the English sentence into phrases $e_1, \ldots, e_n$.

2. Choosing a translation f_i for each e_i in accordance with the phrasal probabilities $P(f_i \mid e_i)$.

3. Choosing a permutation of the $f_1, \ldots, f_n$, using the distortions d_i that describe the number of words that f_i has moved with respect to f_{i-1}.

To maximise the joint phrasal and distortion probabilities

$$P(f, d \mid e) = \prod_i P(f_i \mid e_i)P(d_i),$$

we carry out a BEAM search, which maintains a fixed number of hypothesis that are iteratively built up while moving forward in the parse tree. To facilitate this procedure, we use parallel corpora in the source and target language. These are distinguished from non-parallel corpora in that a source sentence is aligned (matched) with a corresponding target sentence in a way a translation is. The texts are split up into sentences, which are subsequently aligned across both corpora. Next, phrases are aligned to generate the phrasal translation probabilities, and then the distortion probabilities are extracted. In the final step, the expectation-maximisation (EM) algorithm is used to iteratively improve the estimates of phrasal translation and distortion probabilities.

In principle, this kind of architecture/methodology can also be used to do text-style transfer. However, off-the-shelf statistical machine translation software such as Google Translation is not well-suited to this due to the following challenges:

1. The lack of parallel corpora. Many classical writings and government texts (such as the HANSARD EU parliamentary proceedings) are readily available in a multitutde of languages, there are only very few collections of text that solely differ in their style.

2. The separation of content from style. A style transfer system is needed that translates text in such a way as to preserve its content.

3. Suitable evaluation metrics. How one is to judge what makes a good style-translated text sentence is hard to define lest we make use of human subjective judgment. BLEU and PINC scores (defined below) only take us so far.

We will now survey previous work on text style transfer in order to explore how these challenges have been addressed by other authors.

2 Related Work

In style transfer more generally, most work has been done on images, most famously using CycleGANs. (8) However, these architectures are not 1-1 transferable to the text domain due to the discrete nature of text and different, mostly one-dimensional spatial structure. We hence refer the reader to (8) for work on image style transfer.

2.1 Parallel Text Style Transfer

Rao and Tetreault (7) generate parallel pairs of formal and informal sentences from Yahoo answers using Amazon Mechanical Turk. They compare a rule-based system (that for example capitalises words at the start of sentence and that expands contractions) with a phrase-based model and a neural model, which is a bidirectional LSTM encoder-decoder model with attentation and uses GloVe word embeddings. They use human scoring and off-the-shelf formality, fluency and meaning classifiers for evaluation, as well as BLEU and PINC. An example can be seen in figure 1.

Xu et al. (4) translate Shakespeare's plays into modern English, leveraging existing style transfers on Sparknotes.com. Standard statistical machine translation is used in combination with dictionary-based paraphrasing, as well as an alternative model based on out-of-domain monolingual data. In addition to their base

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.*

Figure 1: Examples of Style Transfer from the Yahoo Corpus

metric BLEU, they also propose additional metrics that are not reliant on parallel corpora, namely cosine similarity, a LM style metric, as well as a logistic regression style metric.

2.2 Non-Parallel Data

Fu et al. (9) were the first to address style transfer without parallel data. They propose two metrics based on transfer strength and content preservation, which moderately correlate with human scoring. Their models are a multi-decoder Seq2Seq model (where the encoder captures the content of the input, and the multi-decoder generates outputs in different styles), and a model with the same encoding strategy but that jointly trains style embeddings that augment the encoder content representations, and only learns one single decoder. The authors test their models on a paper-news title transfer, and an Amazon review dataset. See figure 3 for some examples.

2.3 Word Translation

In (10), the authors demonstrate how to build a bilingual dictionary between two languages without using any parallel corpora through aligning monolingual embedding spaces in an unsupervised fashion. In imaage 4 A there are two distributions of word embeddings we want to align. In B, we learn a rotation matrix W through adversarial learning that aligns the two distributions. This mapping W is further refined in C by minimising an energy function corresponding to a spring system between frequent words that serve as anchor points. Finally in D, we translate by using the mapping W and a distance metric that expands the space with a high point density so that "hubs" (e.g. the word "cat") become less close to other word vectors.

2.4 Lampel et al. 2018

Lampel et al. (5) propose a model that takes sentences from monolingual corpora in two distinct languages and maps them into the same latent space. The model then learns to reconstruct both languages from this shared feature space, which has the advantage of not requiring parallel sentence pairs during training. They compare model performance on different baseline models using BLEU scores on the WMT'14 and Multi30k-Task1 datasets for English-French and English-German language pairs. While parallel datasets were only used for evaluation, the model selection was done using a surrogate metric. This was the average between the BLEU scores obtained by letting the model translate from either language to the other and then back to the original language, and the same but starting from the other language. The authors demonstrate that this metric correlates very strongly with BLEU scores from parallel corpora for early stopping and reasonably well for hyperparameter selection. The following model selection criterion was used:

$$\begin{aligned} &MS\left(e, d, \mathcal{D}_{src}, \mathcal{D}_{tgt}\right) \\ =&\frac{1}{2} E_{x \sim \mathcal{D}_{src}}\left[\mathrm{BLEU}\left(x, M_{src \to tgt} \circ M_{tgt \to src}(x)\right)\right] + \\ &\frac{1}{2} E_{x \sim \mathcal{D}_{tgt}}\left[\mathrm{BLEU}\left(x, M_{tgt \to src} \circ M_{src \to tgt}(x)\right)\right]. \end{aligned}$$

Here, $\mathcal{D}_{src}$ and $\mathcal{D}_{tgt}$ are the source and target domains respectively, x is the input sequence and M is the MT model.

An adversarial regularisation term enforces that the source and target languages latent domains have the same distribution. The model tries to fool a discriminator, which is trained to classify the original language from the latent space representation. The authors demonstrate that their method achieves remarkable performance rivalling that from similar translation systems trained on parallel sentence pairs.

Both the encoder and the decoder are bidirectional LSTMs with attention and minimise an objective function $\mathcal{L}_{auto}$ for each language, which measures their ability to reconstruct a noisy version of the input sentence (from noisy source $\to$ latent $\to$ source and noisy target $\to$ latent $\to$ target) with a term that penalises the ability to classify the origin language in the latent space ($\mathcal{L}_{adv}$).

Subsequently, the encoder (source $\to$ latent) is combined with the decoder (latent $\to$ target) to yield a machine translation model from the source to the target. In the same way, the encoder (target $\to$ latent) is combined with the decoder (latent $\to$ source) to give a target to source model. Calling the model $M^{(t)}$ at time step t, $M^{(t)}$ is used to make a cross-domain loss function ($\mathcal{L}_{cd}$) by translating source $\to$ target $\to$ source and comparing to the original input. This loss is then used to update the discriminating parameters. This can then be used to produce a better $M^{(t+1)}$ at the next iteration t.

The final objective function is given by

$$\begin{aligned} &\mathcal{L}(\theta_{\mathrm{enc}}, \theta_{\mathrm{dec}}, \mathcal{Z}) = \\ &\quad \lambda_{auto}\left[\mathcal{L}_{auto}\left(\theta_{\mathrm{enc}}, \theta_{\mathrm{dec}}, \mathcal{Z}, src\right) + \right. \\ &\quad \left. \mathcal{L}_{auto}\left(\theta_{\mathrm{enc}}, \theta_{\mathrm{dec}}, \mathcal{Z}, tgt\right)\right] + \\ &\quad \lambda_{cd}\left[\mathcal{L}_{cd}\left(\theta_{\mathrm{enc}}, \theta_{\mathrm{dec}}, \mathcal{Z}, src, tgt\right) + \right. \\ &\quad \left. \mathcal{L}_{cd}\left(\theta_{\mathrm{enc}}, \theta_{\mathrm{dec}}, \mathcal{Z}, tgt, src\right)\right] + \\ &\quad \lambda_{adv} \mathcal{L}_{adv}\left(\theta_{\mathrm{enc}}, \mathcal{Z} | \theta_D\right). \end{aligned}$$

Here λ_{auto}, λ_{cd}, and λ_{adv} are hyper-parameters weighting the importance of the auto-encoding, cross-

Source	Speaker	Input	Output
Romeo & Juliet	Benvolio	He killed your relative, brave Mercutio, and then young Romeo killed him.	he slew thy kinsman , brave mercutio , and then young romeo kill him .
Romeo & Juliet	Romeo	I can read my own fortune in my misery.	i can read mine own fortune in my woes .

Figure 2: Examples of Style Transfer from the Shakespeare Corpus

source	positive: all came well sharpened and ready to go .
auto-encoder:	→negative: all came well sharpened and ready to go .
multi-decoder:	→negative: all came around , they did not work .
style-embedding:	→negative: my ⟨NUM⟩ and still never cut down it .
source	negative: my husband said it was obvious so i had to return it .
auto-encoder:	→positive: my husband said it was obvious so i had to return it .
multi-decoder:	→positive: my husband was no problems with this because i had to use .
style-embedding:	→positive: my husband said it was not damaged from i would pass right .

Figure 3: Style Transfer on paper-news titles and Amazon reviews

domain and adversarial loss. $\mathcal{Z}$ is the latent space domain.

Overall, the authors successfully show that this unsupervised technique is superior to supervised ones for datasets with fewer than 100,000 parallel sentences. However, it is not discussed if an unsupervised method could rival supervised models that have access to larger corpora of parallel sentences. This work demonstrates that it is possible to build universal cross-lingual encoders that can encode any sentence into a shared embedding space. This development aims to mitigate the English-centric bias.

2.5 Lampel et al. EMNLP

In (14), the authors extend their previous work on fully unsupervised machine translation by proposing a much simpler and more effective initialisation scheme for related languages, and by identifying three key principles of unsupervised machine translation and applying them to a phrase-based machine translation model (PBSMT), which outperforms their neural MT model (NMT). They achieve the best performance by ensembling their PBSMT and NMT. Their results are especially promising for low-resource language pairs such as ENglish-Urdu and English-Romanian, where byte-pair encodings are sparser. (Byte-pair encoding is a simple form of data compression (introduced in 1994 by Gage (15)) where the most common pair of consecutive bytes of data is replaced with another byte not occurring within the same data. To rebuild the original data, a replacement table is created.)

Their three key principles are depicted in figure 4. Figure A shows two monolingual datasets where markers correspond to sentences.

- Suitable initialisation of the translation models: the two distributions are roughly aligned through word-by-word (WBW) translation with an inferred bilingual dictionary as explained above (see figure B)

- Language modelling: To infer the structure of the data, a LM is learned independently in each domain (see figure C)

- Back-translation: An observed source sentence is translated using the current source-to-target model (see figure D). From this translation, the target-to-source model is used reconstruct the sentence in the original language. The difference between the original sentence and its reconstruction is used as the error signal to train the target-to-source parameters. The same method is also applied in the source-to-target direction.

The algorithm is given in figure 5. Note that now we use a slightly simplified loss compared to the previous paper as we no longer have the adversarial loss term L_{adv}.

3 Methods

We follow the architecture of Lampel et al. (12) most recent (2019) unsupervised machine translation model: Cross-lingual Language Model Pretraining. In which they demonstrate the impact of cross lingual language model (XLM) pre-training on supervised and unsupervised machine translation. In (12) the authors provide a cross-lingual implementation of BERT (Bidirectional Encoder Representations from Transformers). BERT is a contextual embedding (unlike Word2Vec or GloVe that are context-free). This cross-lingual implementation mitigates the English-centric bias in available corpora and encodes any sentence in any language into a shared embedding space. All languages are processed

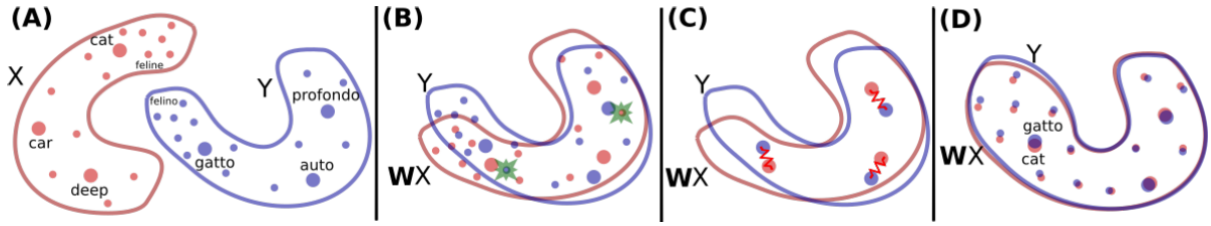

Figure 4: Embedding Strategy in Unsupervised MT

Algorithm 1: Unsupervised MT

1 **Language models:** Learn language models P_s and P_t over source and target languages;
2 **Initial translation models:** Leveraging P_s and P_t, learn two initial translation models, one in each direction: $P^{(0)}_{s\to t}$ and $P^{(0)}_{t\to s}$;
3 **for** k=1 to N **do**
4 **Back-translation:** Generate source and target sentences using the current translation models, $P^{(k-1)}_{t\to s}$ and $P^{(k-1)}_{s\to t}$, factoring in language models, P_s and P_t;
5 Train new translation models $P^{(k)}_{s\to t}$ and $P^{(k)}_{t\to s}$ using the generated sentences and leveraging P_s and P_t;
6 **end**

Figure 5: Unsupervised MT Algorithm Pseudocode

with the shared vocabulary from Byte Pair Encodings (BPEs).

They demonstrate how cross-lingual models as pretraining can be used to obtain: a better initialization of sentence encoders for zero-shot cross-lingual classification for supervised and unsupervised objectives. We are particularly interested in the latter as our style transfer objectives can only correspond to unsupervised techniques. Thus the two pretraining techniques that apply to monolingual data: Causal Language Modeling (CLM) and Masked Language Modeling (MLM) were used here. The prior uses a Transformer language model trained to model the probability of a word given the previous words in a sentence $P\left(w_t|w_1,\ldots,w_{t-1},\theta\right)$. The latter masks random elements in the sequence and models the probability of the masked element from previous and future words $P\left(w_m|w_1,\ldots,w_m,\ldots w_t,\theta\right)$.

Lampel et al. show that for unsupervised machine translation, that MLM pretraining is extremely effective. They reach a new state of the art of 34.3 BLEU on WMT'16 German- English, outperforming the previous best approach by more than 9 BLEU. Also making less drastic improvements with supervised machine translation.

We use the following metrics.

1. For a candidate sentence c of length C and a reference sentence r of length R, BLEU is defined as

$$\text{BLEU}(c,r) = \left[\prod_{n=1}^{N} P(n)\right]^{1/N} \times \text{BP}$$

where

$$P(n) = \frac{|\text{ngrams}_c \cap \text{ngrams}_r|}{|\text{ngrams}_c|}$$

measures ngram overlap between the candidate and the reference sentence. ngram_s and ngram_c are the lists of ngrams in the source and target sentence.

$$\text{BP} = \min\left(1, \exp\left(\frac{C-R}{C}\right)\right)$$

is a a brevity penalty. BLEU measures semantic adequacy and fluency. The candidate sentence might be a proposed translation and the reference sentence might be a target translation. Note that to compute BLEU, a parallel corpus is needed. We hence use backtranslation as a replacement for the parallel corpus (style transfer from the source to the target domain and then back to the target domain).

2. For a source sentence s and a candidate sentence c, PINC is defined as

$$\text{PINC}(s,c) = \frac{1}{N}\sum_{n=1}^{N}\left(1 - \frac{|\text{ngrams}_s \cap \text{ngrams}_c|}{|\text{ngrams}_c|}\right)$$

where N is the maximum ngram considered. It computes the percentage of ngrams present in the target but not the source sentence and is thus a measure of the lexical dissimilarity between both sentences. Note that no constraint is placed on sentence length as in BLEU. Also, parallel corpora are not needed.

3. Finally, there is human inspection to obtain a qualitative understanding of the model performance. This is especially crucial for song text style transfer as we note the inadequacy of BLEU and PINC to fully capture the essence of what we think makes a good style transfer. In particular, note

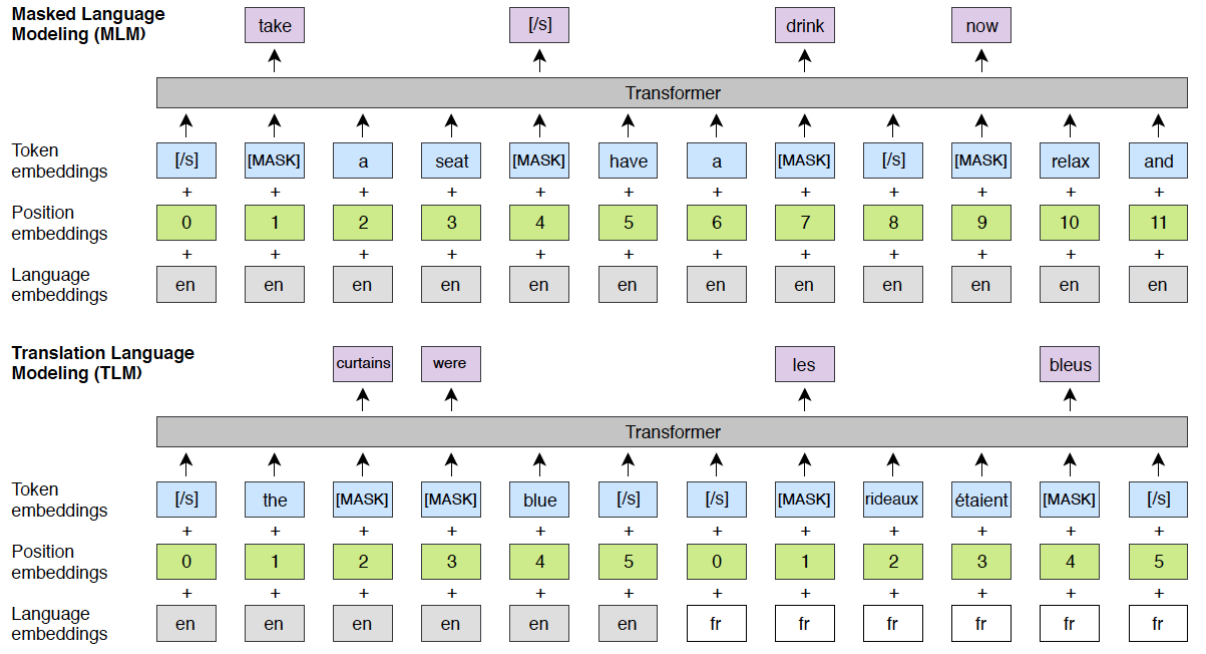

Figure 6: Cross-Lingual Language Model Embedding

that a high BLEU or a high PINC alone is not sufficient. Just replicating the input sentence yields a perfect BLEU and a zero PINC. Generating the reference sentence does the opposite.

4 Experiments

Figures (7 and 8) show the BLEU and PINC scores used to decide early stopping. PINC was used as a satisfaction metric to ensure that the model learned wasn't trivial (ie the input is being changed). The optimisation metric was the BLEU score where the higher the score idicated a better model. A number of epochs of 5 was consider the best point at which to 'early stop' training.

The BLEU and PINC scores shown here are as a percentage.

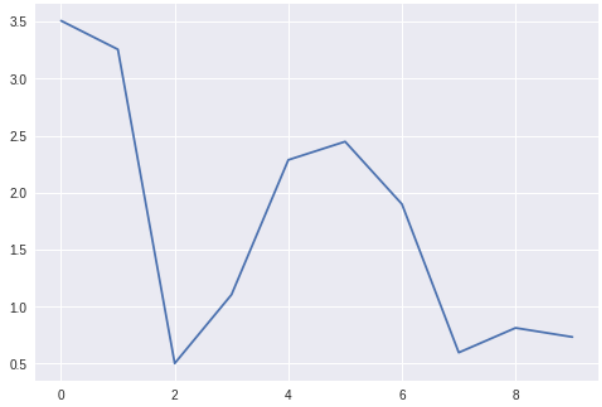

Figure 7: BLEU scores as a function of number of epochs.

5 Results and Discussion

We achieve a BLEU score of 1.34 evaluated over 1000 sentences. Where we have multiplied conventional BLEU score by 100. As is discussed, computational power was a major limiting factor in the project.

5.1 Basic unsupervised song style transfer has been demonstrated

We give the following example style-transferred lyrics.

Yesterday - The Beatles. Original (rock):

```
Yesterday
All my troubles seemed so far away
Now it looks as though they're here to s
Oh, I believe in yesterday
Suddenly
I'm not half the man I used to be
There's a shadow hanging over me
```

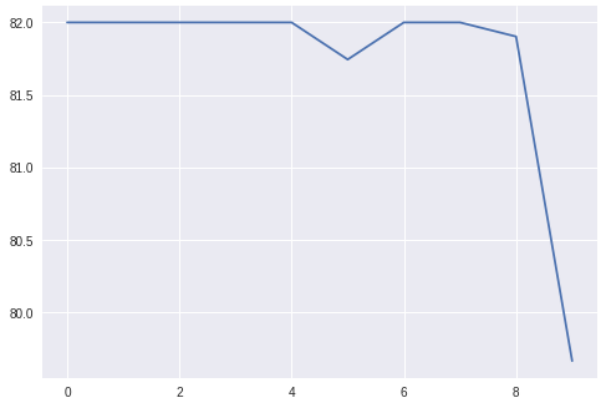

Figure 8: PINC scores as a function of number of epochs.

```
Oh, yesterday came suddenly
Why she had to go, I don't know
She wouldn't say
I said something wrong
Now I long for yesterday
Yesterday
Love was such an easy game to play
Now I need a place to hide away
Oh, I believe in yesterday
```

Translation (hip-hop):

```
Now Go simply
Now I
I not a yesterday
I was
had such believe something seemed
as the hanging as to suddenly
an though an though in so though so
We easy place ain't dim yesterday
I got me I got me
We got to stay here to away
You know the nigga to hide to hide
to hide
You know the nigga to be to to to
to used to be
You know the nigga is not to be
You know the way is not a nigga
You know the nigga is not a nigga,
```

In Da Club - 50 Cent. Original (hip-hop):

```
Go, go, go, go go, go, go, shawty
It's your birthday
We gon' party like it's yo birthday
We gon' sip Bacardi like it's your
birthday
And you know we don't give a fuck
It's not your birthday!
You can find me in the club, bottle
full of bub
Look mami I got the X if you into
taking drugs
I'm into having sex, I ain't into
making love
So come give me a hug if you into
to getting rubbed
You can find me in the club, bottle
full of bub
Look mami I got the X if you into
taking drugs
I'm into having sex, I ain't into
making love
So come give me a hug if you into
to getting rubbed
```

Translation (rock):

```
And We 're Go Go Go Go Go Go And And
We 're And
you can not anytime your bumpin
into your Go
I know having sip your birthday I
we got me Bacardi like we got me
I'm like simply simply simply simply
ain't like
I'll give the X, my you ain't
wasting show you a a a wasting
I'm a down if ain't dim if a dim
into your birthday
I'm not a bottle you're making your
birthday
I'm not a full into love into into
into love
I'm not a love of taking to taking
out of
I'm not a one, you're getting
drugs bub getting drugs
I'm not a thing, you're rubbed
rubbed
I'm not a thing, you're my mind's
not a good
```

We notice the following features of our style transfers:

1. Use of genre-specific vocabulary (hip-hop: "nigga", "ain't", rock: "go go go")
2. Some syntactic structure ("You know the nigga is not a nigga").
3. Some content preservation.
4. Repetitive sentence structures at the end.

5.2 CycleGANs are not well-suited to text style transfer

In addition, we carried out some experiments using existing CycleGAN code for text style transfer. However, we quickly realised that CycleGANs did not yield satisfactory results. This is due to the following two reasons:

1. Discrete nature of text. Images are continuous objects in the sense that pixel intensities in RGB are approximately continuous (e.g. numbers between 0 and 255) with an underlying monotonic structure. Stated differently, given two pixels of different hues of blue, we can calculate what their average hue is and the resulting pixel "makes sense", i.e. is a valid pixel. However, we obviously cannot form a simple average of words like "cat" and "house". If we try to do this in an embedding space (e.g. from Word2Vec or GloVe), the average usually will not correspond to a specific word. We could define the "average word" to be the nearest neighbour word of the numerical average in the embedding space, but this yields introduces a great deal of noise into our "word algebra". It might be interesting though for future work to explore this aspect further, seeing as we do not yet have a thorough understanding of the information geometric structure of embedding spaces.

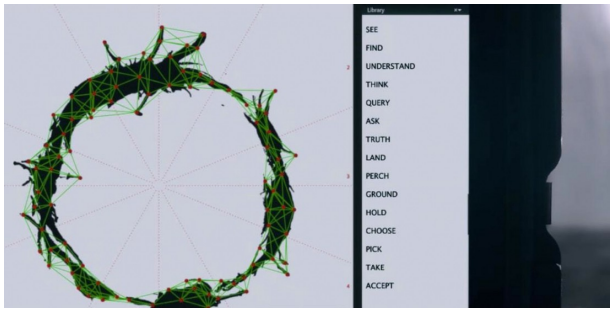

Figure 9: Alien script from the film Arrival

2. One-dimensional spatial structure of text. Note that in two-dimensional images, neighbouring pixels lie not just to the left or right of a pixel but also to the top, bottom and diagonally offset. In text, there are only two neighbouring words to a given word, and these tend to bear the closest semantic and syntactic similarity with the given word. While some more structure is present (e.g. rhyming last words of successive sentences), most of the textual information geometry remains one-dimensional. This is in notable contrast to the Alien script from the movie Arrival, which has more two-dimensional structure. We hence hypothesise that CycleGANs can be more successfully deployed for text style transfer for this type of script.

5.3 More computational resources are needed

Lyrics style transfer is decidedly harder than Yahoo Answers style transfer or Shakespearean novel style transfer. Due to the availability of one GPU for only a couple of hours, we had to stop our training early. However, with access to more computational power we expect to obtain much high BLEU and PINC scores, as well as appeal of the generated lyrics to the human eye. In comparison, the authors in (12) used 64 Volta GPUs for the language modelling, and 8 Volta GPUs for the machine translation task, trained over several days.

6 Conclusion and Future Work

The contributions of this paper were the following:

1. Adaptation of the Facebook XLM code to our specific computational environment.
2. Novel application to song lyrics style transfer.
3. Adaptation of Cycle GANs to text style transfer.

We conclude the following.

- It seems more difficult to carry out song text style transfer than formal-informal style transfer or Shakespeare-to-modern English style transfer. This is due to the greater stylistic variety between different musical genres in terms of:
  - Register
  - Sentence and song length
  - Grammaticality and use of contractions
  - Use of rhyme and verse
  - Ipso-facto, de-facto and ex-post facto lack of parallel corpora

  We thus hypothesise that ipso facto even larger training corpora are needed for successful song style transfer, as well as more computational resources. The commercial viability of text style transfer systems for arbitrary corpora will thus depend on cloud integration.
- CycleGANs seem best suited to image style transfer but fail to capture the discrete, largely one-dimensional spatial structure of text.
- In-principle song style transfer has been demonstrated.

We recommend that future work on this area focus on replicating our approach using shared, cross-lingual embedding spaces but using more computational resources that were not available to us. In particular, it will be valuable to investigate further whether more closely-aligned genres are easier to carry out style transfer for, such as pop and rock.

References

[1] Kishore Papineni, Salim Roukos, Todd Ward, Wei-Jing Zhu, BLEU: a Method for Automatic Evaluation of Machine Translation. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, July 2002, pp. 311-318

[2] Dor Hacohen, Anthony Bourached, Sean Gupta: UCL NLP course 2018/19 Project 2: Text Style Transfer from Unsupervised Machine Translation.

[3] Russell, Stuart J., and Peter Norvig. Artificial intelligence: a modern approach. Malaysia; Pearson Education Limited,, 2016.

[4] Xu, Wei, et al. "Paraphrasing for style." Proceedings of COLING 2012 (2012): 2899-2914.

[5] Lample, Guillaume, et al. "Unsupervised machine translation using monolingual corpora only." arXiv preprint arXiv:1711.00043 (2017).

[6] https://www.kaggle.com/gyani95/380000-lyrics-from-metrolyrics/home, last accessed 24th March 2019

[7] Rao, Sudha, and Joel Tetreault. "Dear sir or madam, may I introduce the YAFC corpus: corpus, benchmarks and metrics for formality style transfer." arXiv preprint arXiv:1803.06535 (2018).

[8] Zhu, Jun-Yan, et al. "Unpaired image-to-image translation using cycle-consistent adversarial networks." Proceedings of the IEEE International Conference on Computer Vision. 2017.

[9] Fu, Zhenxin, et al. "Style transfer in text: Exploration and evaluation." Thirty-Second AAAI Conference on Artificial Intelligence. 2018.

[10] Conneau, Alexis, et al. "Word translation without parallel data." arXiv preprint arXiv:1710.04087 (2017).

[11] Shen, Tianxiao, et al. "Style transfer from non-parallel text by cross-alignment." Advances in neural information processing systems. 2017.

[12] Guillaume Lample and Alexis Conneau, Cross-lingual Language Model Pretraining. CoRR abs/1901.07291 2019.

[13] https://spinbot.com, last accessed 24th March 2019

[14] Lample, Guillaume, et al. "Phrase-based neural unsupervised machine translation." arXiv preprint arXiv:1804.07755 (2018).

[15] Gage, Philip. "A new algorithm for data compression." The C Users Journal 12.2 (1994): 23-38.