Unsupervised Text Style Transfer with Content Embeddings

Keith Carlson
Department of Computer Science
Dartmouth College
Hanover, NH 03755
keith.e.carlson.gr@dartmouth.edu

Allen Riddell
Department of Information and Library Science
Indiana University Bloomington
Bloomington, IN 47405

Daniel Rockmore
Departments of Computer Science and Mathematics
Dartmouth College
Hanover, NH 03755
& The Santa Fe Institute
Santa Fe, NM 87501

Abstract

The style transfer task (here *style* is used in a broad "authorial" sense with many aspects including register, sentence structure, and vocabulary choice) takes text input and rewrites it in a specified target style preserving the meaning, but altering the style of the source text to match that of the target. Much of the existing research on this task depends on the use of parallel datasets. In this work we employ recent results in unsupervised cross-lingual language modeling (XLM) and machine translation to effect style transfer while treating the input data as unaligned. First, we show that adding "content embeddings" to the XLM which capture human-specified groupings of subject matter can improve performance over the baseline model. Evaluation of style transfer has often relied on metrics designed for machine translation which have received criticism of their suitability for this task. As a second contribution, we propose the use of a suite of classical stylometrics as a useful complement for evaluation. We select a few such measures and include these in the analysis of our results.

1 Introduction

In this paper we consider the problem of unsupervised *holistic textual style transfer* – both the accomplishment of the task, as well as its evaluation. The "style" of text is roughly the way in which a text communicates its content. It might be thought of as the "voice" characteristic of a given author, an emergent quality that encompasses a wide range of (more or less measurable) characteristics such as register, sentence structure, and vocabulary choice. Holistic style transfer takes a given text – written a priori in one "style" – and then rewrites it (preserving its content) in another style. Holistic style transfer is distinct from more narrow style modification techniques which manipulate specific characteristics of prose such as formality, simplicity, or sentiment.

To illustrate our idea of holistic style consider the following pair of translations of the opening lines of *The Aeneid of Virgil*.

> Arms and the man I sing, the first who came, Compelled by fate, an exile out of Troy...(Humphries et al., 1987)

> I sing of arms and the man who of old from the coasts of Troy came, an exile of fate, to Italy and the shore of Lavinium...(Mackail, 1885)

As another example compare verses from two different "versions" of a fixed verse from the Book of Genesis, the first from the *King James Version*

> And a river went out of Eden to water the garden; and from thence it was parted, and became into four heads.

and the second, the same verse, but in the *New International Version*:

> A river watering the garden flowed from Eden; from there it was separated into four headwaters.

In both example pairs we can see that the content in the passages is the same, but the (holistic) style differs noticeably.

The examples above are effectively examples of a human-executed style transfer. The potential applications of a machine holistic style transfer are numerous. For example, various periodicals often try to have a single "voice" and an unsupervised

Proceedings of Recent Advances in Natural Language Processing, pages 226–233
Sep 1–3, 2021.
https://doi.org/10.26615/978-954-452-072-4_027

style transfer of the kind studied here would enable a staff writer to produce the content required of an article which was then "stylized" per the requirements of the venue. Thus, a style transfer platform could be a high-powered editorial assistant. Such a platform could also assist aspiring writers. All that said, one should not be blind to the more nefarious potential of successful style transfer machinery which could be useful for spoofing an audience to productive, or unproductive writerly ends (Nature, 2020).

One machine learning approach to holistic style transfer is to adopt and adapt the frameworks of translation models, treating each style as a language. Along these lines, much of the existing research on this task depends on the use of parallel datasets, a schema that follows early work in machine translation, but parallel datasets in this domain are in fact rare. This motivates our approach wherein we continue to be inspired by machine translation work and employ recent results in unsupervised cross-lingual language modeling (XLM) to effect holistic style transfer while treating the input data as unaligned, an important next step in advancing this area in light of the scarcity of parallel datasets. Additionally, we show that modifications to this framework which take advantage of the differences between the style transfer and machine translation tasks can improve model performance. Specifically, we add "content embeddings" to the XLM which capture human-specified groupings of subject matter and observe improvement over the baseline model for a range of metrics.

That brings us to the paired challenges resident in evaluting a style transfer technique. This task is complicated by the emergent nature of style. The analogy of style transfer to translation and concomitant efforts to use techniques from machine translation for the style transfer task have inspired the importation of evaluation metrics from machine translation to the style transfer setting, (Xu et al., 2012; Jhamtani et al., 2017; Carlson et al., 2018), although not without criticism (Tikhonov et al., 2019; Xu et al., 2016). As the evaluation of a style transfer task should a priori measure the similarities of source and target texts to their "native environments", it seems natural to bring to bear some of the techniques from the field of *stylometry*, a discipline focused on the quantitative analysis of textual style. Stylometry (or stylometrics) was born of a nineteenth century effort to settle – quantitatively – scholarly dispute around the temporal ordering of Plato's Dialogues (Lutoslawski, 1897)). For this task, over 500 individual and measurable textual characteristics were identified. Since that time stylometrics have been used (most famously) to address questions of disputed authorship (see e.g., Mosteller and Wallace (1964); Boyd and Pennebaker (2015)). If we imagine a system which perfectly performs style transfer as we have defined it, then the output of the system – in terms of its individual characteristics – should be stylistically indistinguishable from text written by the author whose style is targeted. It thus seems natural to use a range of stylometric measures used in the past to distinguish between authors' styles as an evaluation for the performance of such a system. This line of reasoning motivates a second contribution of this work wherein we introduce the idea of using stylometric measures for evaluation. We evaluate our systems using several stylometric measures in addition to the more commonly previously used metrics and show that the stylometric approach is a useful domain specific complement to translation-based metrics for the evaluation of the complex, subtle, and important task of style transfer.

2 Related Work

Style transfer has some long roots. It is possible to frame the early work on text simplification (e.g., Specia (2010)) or paraphrasing Xu et al. (2012) as a form of style transfer. Style transfer research makes use of a range of datasets for training and evaluation. Examples include the corpus of Shakespearean plays and their "translations" into contemporary English (Xu et al., 2012) for paraphrasing and a corpus of Wikipedia pages and their simplified versions (Zhu et al., 2010) which is used for the general task of text simplification.

The more stylistic features that are incorporated into building a model, the closer it gets to the kind of holistic effort we have described above. To that end, we highlight Ficler and Goldberg (2017) wherein a supervised style transfer model is developed which focuses on the modification of prose with respect to six aspects of style, including register, sentiment, focalisation, and prolixity. A broader approach for supervised holistic style transfer is addressed in Carlson et al. (2018); Xu et al. (2012). They make use of a model that depends on a corpus of versions of the Bible, a priori aligned through the canonical and shared structuring of

Book, Chapter, and verse, to learn the differences between examples written in different styles.

Unsupervised methods pose new challenges for style transfer. Previous related work uses unsupervised training for generating text in a particular style. This includes the generation of stylized text (Hu et al., 2017) and modification of the sentiment or formality of prose (Shen et al., 2017; Li et al., 2018; Gong et al., 2019; Li et al., 2019). There have also been advances in the use of unsupervised approaches for machine translation. Many of these rely on the idea of back-translation (Artetxe et al., 2017; Lample et al., 2018) to automatically generate a synthetic parallel from unaligned data. Lample and Conneau (2019) uses this concept along with a novel cross-lingual language model objective for pre-training to achieve impressive performance on the unsupervised translation task.

3 Experiments

3.1 Data

Our work makes use of eight cleaned and aligned public domain versions of the Bible introduced in Carlson et al. (2018) and made available on Github. (That paper mentions the availability of thirty-four versions, but twenty-six of them have copyright restrictions that restricts their distribution.) These represent eight different English writing styles. The texts are divided hierarchically (and canonically), into version, book, chapter and verse, so that the verses from different versions are parallel. For our unsupervised work we do not take advantage of the alignment during training, but the alignment does enable an objective evaluation of our output.

Our major methodological advance is the introduction of another coarse level of hierarchy which we call *content*, which we use to modify the language model. We see this kind of coarse labeling as an approach which is broadly generalizable to situations in which fine-scaled parallel alignment does not exist. In the case of the Bible, we use nine "divisions" of the Bible which are classical groupings of thematically similar texts.[1] See Table 1 for the divisions used. We do not use the exact data splits detailed in Carlson et al. (2018), but instead split the data as required by the formulation of our models. We use some books of the YLT (Young's Literal Translation) and BBE (Bible in Basic English) versions for validation and testing as style transfer between these versions was identified as the "hardest" task in Carlson et al. (2018). The validation set contains the BBE and YLT versions of 1 Kings, Zephaniah, Mark, and Colossians. The testing set contains the BBE and YLT versions of Judges, 1 Samuel, Philippians, and Hebrews. The remaining books from BBE and YLT and all books from the other six (publicly available) Bible versions make up the training data.

The parallel texts allow for automatic and objective evaluation of translations. While the models we describe can be generalized to other non-parallel datasets, in those cases objective evaluation would be more difficult.

Division	Books
Pentateuch	Genesis, Exodus, Leviticus, Numbers, Deuteronomy
History	Joshua, Judges, Ruth, 1 Samuel, 2 Samuel, 1 Kings, 2 Kings, 1 Chronicles, 2 Chronicles, Ezra, Nehemiah, Esther
Poetry	Job, Psalms, Proverbs, Ecclesiastes, Song of Solomon
Major Prophets	Isaiah, Jeremiah, Lamentations, Ezekiel, Daniel
Minor Prophets	Hosea, Joel, Amos, Obadiah, Jonah, Micah, Nahum, Habakkuk, Zephaniah, Haggai, Zechariah, Malachi
Gospels & Acts	Matthew, Mark, Luke, John, Acts
(Pauline) Epistles	Romans, 1 Corinthians, 2 Corinthians, Galatians, Ephesians, Philippians, Colossians, 1 Thessalonians, 2 Thessalonians, 1 Timothy, 2 Timothy, Titus, Philemon, Hebrews
General Epistles	James, 1 Peter, 2 Peter, 1 John, 2 John, 3 John, Jude
Revelation	Revelation

Table 1: Our partition of Bible books into divisions.

3.2 Baseline System

Lample and Conneau (2019) introduced a method for cross-lingual language model pretraining from non-parallel data[2]. Their model, XLM, feeds token, position, and language embeddings to a Transformer model (Vaswani et al., 2017) which tries to predict masked words. This task, Masked Language Modeling (MLM), was introduced by Devlin et al. (2018) and unsupervised translation was

[1] There is no authoritative partition into divisions, but there are many similar varieties. Our choice among these options is somewhat arbitrary, but has historical and disciplinary support. An example of Old Testament divisions which match ours can found at http://www.scriptureman.com/ot.gif and our New Testament at http://jpatton.bellevue.edu/inspired-table2.jpg

[2] Code found at: https://github.com/facebookresearch/XLM

demonstrated as an application of these pretrained language models. We use the XLM as our baseline.

In our experiment, we treat each version of the Bible in the data as a language. So the embeddings fed to the Transformer for MLM training are position and token embeddings as before, and version embeddings replacing the language embeddings of the original system. Our transformer architecture has embeddings of length 512, 6 layers, 8 attention heads and a 0.1 dropout rate.

We train the language model from scratch until the perplexity of the validation data for the BBE→YLT version has stopped improving. We then use this pretrained language model to initialize Transformers for both the encoder and decoder of our machine translation(style transfer) model and train on the task of unsupervised translation until the BLEU score of the validation data for the BBE→YLT task has stopped improving. This design is based on that used by Lample and Conneau (2019) in the original paper. We call these models "XLM".

3.3 Model with Content Embeddings

Using Bible divisions as a grouping of content similarity, we modify the XLM embedding structure accordingly and include a *content embedding* in addition to the token, position, and language (style) embeddings. In a different context other considerations or structural organization may suggest a different articulation of content. This additional embedding is treated similarly to the three embeddings in the baseline system. The input of each token passed to the Transformer is the combination of four embeddings instead of three. Just as in the XLM, these embeddings are updated during the training process. Our intuition is that for some datasets, the model may have difficulty distinguishing whether differences in language arise because of differences in the style of writing, or differences in the content. By providing training data where both style and content are designated, we anticipate that the model will be better able to reproduce the differences which are style-specific. Similar intuition has led to other approaches which allow a model to learn style and content separately (Fu et al., 2017; Zhang et al., 2018).

In this new formulation, we provide all four embeddings to the Transformer and then train towards the MLM objective as before. We call this model "XLM + Content" (see Figure 1) . We use the same parameter settings as in the "XLM" model and as before, we stop training of the language model when the perplexity of BBE→YLT evaluation task has stopped improving. Once again this transformer which was pretrained on the MLM task is used to initialize the encoder and decoder of a machine translation/style transfer model. This transfer model continues training until the BLEU score of the evaluation data BBE→YLT has stopped improving. Note that the alignment (parallel nature of the texts) makes possible the BLEU scoring.

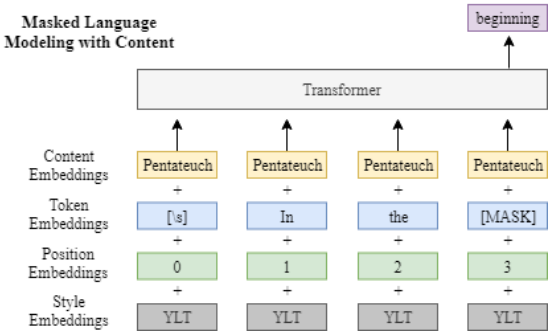

Figure 1: "XLM + Content" model training on the MLM objective. Based on Figure 1 of Lample and Conneau (2019). The choice of types for content embeddings are human-assigned before training as seen in Table 1.

.

4 Results

4.1 Evaluation Metrics

The existence of parallel texts allows us to evaluate our results using the standard translation quality measures BLEU (Papineni et al., 2002) and PINC (Chen and Dolan, 2011), which reward similarity to the target and dissimilarity to the source respectively. PINC was created from a desire to "measure lexical dissimilarity with the source sentence" and its creators say "In essence, it is the inverse of BLEU" (Chen and Dolan, 2011). The results of these evaluations can be seen in Table 2.

We find that our model with content embeddings has a higher (better) PINC score for all four test books, indicating that it has more aggressively made changes than the baseline system. "XLM + Content" also attains a sizeably Higher BLEU score on Philippians and Hebrews. The BLEU score for the other two test books are similar between the two systems.

Test Book	Source	XLM	XLM+Content
Judges	16.18(0)	**26.5(39.93)**	26.1(44.89)
1 Samuel	14.75(0)	24.21(39.72)	**24.36(44.40)**
Philippians	18.29(0)	20.56(25.50)	**22.82(29.83)**
Hebrews	12.27(0)	15.88(29.73)	**17.44(34.70)**

Table 2: The BLEU (PINC) scores of the unmodified source and the output of each model for each test book. All scores are when translating from Bible version YLT to Bible version BBE.

YLT Exclusive Words
unto, flee, fleeth, hath, thine, hast, thus, midst, thy, inheritance, cometh, ye, also, shall, doth, thou, jehovah

Table 3: Words which are frequent in YLT but do not appear in BBE. Used for the *frequent idiosyncratic words* evaluation.

4.1.1 Stylometry-Inspired Evaluation

This combination of BLEU and PINC scores for evaluating style transfer in text has been used in other work (Xu et al., 2012; Jhamtani et al., 2017; Carlson et al., 2018), but not without criticism (Tikhonov et al., 2019; Xu et al., 2016). Arguably, style transfer – especially for the situation in which there is no parallel (aligned) text – cries out for new kinds of measures. As mentioned in the Introduction, we believe that classical stylometric measures provide a natural source of appropriate options. Some approaches to stylometry are structural, while others focus on word usage frequency. For example, function word-based approaches[3] have proved to be a useful (partial) fingerprint for authorial style in some cases (see e.g., (Mosteller and Wallace, 1964; Binongo, 2003)).

Thus inspired we augment the use of BLEU and PINC through several stylometrically inspired metrics. The first is the identification of *frequent idiosyncratic words*, words that seem simultaneously characteristic of one style but absent or rare in another. This form of bespoke evaluation checks to see if 17 frequent words with known translations have been correctly translated in the YLT→BBE test task. All the words occur frequently and exclusively in YLT. Examples include *unto*, *hath*, *flee*, *doth* and the full list can be seen in Table 3. These words occur 2,522 times in YLT source lines in the test set. In this test, a YLT→BBE translation is counted as correct if the BBE version does *not* include the idiosyncratic word from the YLT line. Accuracy scores in this evaluation increase with the complexity of the model: 99.3% ("XLM") and 99.8% ("XLM + Content").

In addition to this test of frequent idiosyncratic words, we analyze the entire test set of source, reference, and model outputs with a few other simple stylometrics: number of multi-syllable words, average number of syllables per word, average number of letters per word, and number of complex words (Dale and Chall, 1948). The results can be seen in Table 4. On all 4 of these evaluation metrics we find that the model modified to include separate content embeddings (XLM+Content) is closer to the target BBE than is the unmodified XLM model. This analysis provides further evidence that the content embeddings are enabling the model to produce better results.

[3]"Function words" are "common" words and possess little or no information about content. Examples include prepositions, articles, etc.

4.2 Example Outputs

Table 5 shows two test data example inputs and their targets alongside the corresponding outputs of our systems. In the first example, note that both outputs correctly remove the use of quotation marks as is consistent with the BBE target and modernize the archaic *Thou* and *dost*. The "XLM + Content" also correctly changes the word *testify* to *witness*. In the second example, the "XLM + Content" model correctly changes *age-during* to *eternal*.

5 Conclusion and Future Work

The task of holistic textual style transfer requires a system to take text in a native (source) style as input and then rewrite the text, retaining the meaning while changing the style consistent with a specified target. In many potential applications this task will need to be performed in contexts where there is no parallel data which captures the styles of interest available for training. Examples range from the journalistic (writing articles in a given editorial style) to the literary (writing the style or voice of a given author). Contexts such as these have large corpora of source and target examples, but – presumably – no source/target pairings.

In this work we demonstrated that a modern unsupervised machine translation technique could be applied to unsupervised holistic textual style transfer in the context of different styles (well known and publicly available versions) of the Bible. We

	YLT(Source)	BBE(Ref)	XLM	XLM + Content
Multi-Syllable Words	3596	2231	2597	**2369**
Syllables Per Word	1.31	1.23	1.25	**1.24**
Letters Per Word	3.96	3.75	3.8	**3.75**
Complex Words (Dale and Chall, 1948)	13189	7967	9480	**8946**

Table 4: Comparison of simple stylometric measures on whole test set.

YLT(Source)	BBE(Ref)	XLM	XLM + Content
6 and one in a certain place did testify fully, saying, 'What is man, that Thou art mindful of him, or a son of man, that Thou dost look after him?	6 But a certain writer has given his witness, saying, What is man, that you keep him in mind? what is the son of man, that you take him into account?	6 And one in a certain place did testify fully, saying, What is man, that you are mindful of him, or a son of man, that you do look after him?	6 And one in a certain place did give witness fully, saying, What is man, that you are mindful of him, or a son of man, that you will go after him?
9 and having been made perfect, he did become to all those obeying him a cause of salvation age-during,	9 And when he had been made complete, he became the giver of eternal salvation to all those who are under his orders;	9 And having been made perfect, he did become to all those who obey him a cause of salvation age-during,	9 And having been made perfect, he gave to all those who keep him a cause of salvation eternal,

Table 5: Examples Outputs of each of the systems with YLT source and BBE target.

show that by adding an additional "content embedding layer" to encode the type of content in text, holistic style transfer is improved. The parallel nature of Bible versions enables us to objectively measure the effect of our innovation of content embedding – improvement is witnessed in terms PINC and BLEU scores that are greater when using content embedding than when not. Specifically, this improves upon the work of Carlson et al. (2018) and makes use of their publicly accessible data. We further introduce new measures of style transfer quality (a simple test of frequent idiosyncratic words as well as source/target comparisons of some basic stylometric measures – number of multi-syllable words, syllables per word, letters per word, number of complex words) as novel evaluations of style transfer, supplementing the traditional – and by some accounts, somewhat flawed – use of the PINC and BLEU metrics in this context. These new measures are a contribution in their own right to the space of evaluation frameworks for style transfer and also support our claim that content embedding improves style transfer.

Future work will need to identify additional datasets that are suitable for research on this task. In particular, having some diversity of parallel corpora for testing style transfer would be of great interest. The structure of the Bible suggests a division of text into specific types of content (which we readily adopt), but other contexts may require a different approach to content labeling and embedding. The broader range of possible stylometric evaluation measures suggests that at least with respect to evaluation, a requirement of perfect evaluation and parallel texts might be relaxed.

While the Bible may seem to be particularly suited to the partition into content classes we employ, we believe this technique can be directly applied to many other textual sources as well. Similar to Bible versions, many translations exist of other classical works such as the epics written by Homer or Dante. In many of these translations alignment does not exist line by line so traditional supervised methods are not applicable. They are however "softly aligned" by book or chapter making content embeddings a natural choice. A model trained on these could then produce Homer's Iliad in the style of a translator who only produced a version of the Odyssey. Similarly, many translations of classic non-English novels exist and this system could be used to create a new translation targeting the style of a particular translator.

Demand for English-to-English style transfer also exists commercially. Examples here include poetry parodies (Zaranka, 1981) , continuations of stories from famous authors (James, 2011), or modernized retellings of stories (Rivers, 2012; McKinley, 2011). In these cases the content of the text either exists publicly or is written by the author. The style however is intentionally changed, either to match the works of another writer, or to remove the idiosyncrasies of the original style. Unsuper-

vised style transfer models could be used to help produce these works.

In addition to these potential applications, our results reinforce the idea that consideration of content and style independently can improve the results of style transfer models. In cases where our technique cannot be directly applied, this provides additional evidence to researchers that finding a way to separate the two may improve results.

In conclusion, this work highlights the utility of the Bible as a dataset for holistic style transfer, demonstrates that unsupervised machine translation methods for holistic style transfer are possible and can be objectively evaluated, provides further evidence – and an actionable methodology – for the idea that learning content independent of style can be beneficial, and proposes the use of classical stylometric measures for evaluation of style transfer systems.

References

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2017. Unsupervised neural machine translation. *arXiv preprint arXiv:1710.11041*.

J. N. G. Binongo. 2003. Who wrote the 15th book of Oz? An application of multivariate analysis to authorship attribution. *Chance*, 16(2):9–17.

Ryan L. Boyd and James W. Pennebaker. 2015. Did Shakespeare write Double Falsehood? Identifying individuals by creating psychological signatures with text analysis. *Psychological Science*, 26(5):570–582.

Keith Carlson, Allen Riddell, and Daniel Rockmore. 2018. Evaluating prose style transfer with the bible. *Royal Sociey Open Science*, 5(10).

David L Chen and William B Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 190–200. Association for Computational Linguistics.

Edgar Dale and Jeanne S Chall. 1948. A formula for predicting readability: Instructions. *Educational research bulletin*, pages 37–54.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jessica Ficler and Yoav Goldberg. 2017. Controlling linguistic style aspects in neural language generation. In *Proceedings of the Workshop on Stylistic Variation*, pages 94–104, Copenhagen, Denmark. Association for Computational Linguistics.

Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2017. Style transfer in text: Exploration and evaluation. *arXiv preprint arXiv:1711.06861*.

Hongyu Gong, Suma Bhat, Lingfei Wu, Jinjun Xiong, and Wen Mei Hwu. 2019. Reinforcement learning based text style transfer without parallel training corpus. In *2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2019*, pages 3168–3180. Association for Computational Linguistics (ACL).

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. 2017. Toward controlled generation of text. In *International Conference on Machine Learning*, pages 1587–1596.

Rolfe Humphries et al. 1987. *The Aeneid of Virgil: A Verse Translation*. Longman Publishing Group.

Phyllis Dorothy James. 2011. *Death Comes to Pemberley: Enhanced Edition*. Faber & Faber.

Harsh Jhamtani, Varun Gangal, Eduard Hovy, and Eric Nyberg. 2017. Shakespearizing modern language using copy-enriched sequence-to-sequence models. *arXiv preprint arXiv:1707.01161*.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *Advances in Neural Information Processing Systems (NeurIPS)*.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5039–5049, Brussels, Belgium. Association for Computational Linguistics.

Dianqi Li, Yizhe Zhang, Zhe Gan, Yu Cheng, Chris Brockett, Bill Dolan, and Ming-Ting Sun. 2019. Domain adaptive text style transfer. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3295–3304.

Juncen Li, Robin Jia, He He, and Percy Liang. 2018. Delete, retrieve, generate: A simple approach to sentiment and style transfer. *arXiv preprint arXiv:1804.06437*.

Wincenty Lutoslawski. 1897. *The Origin and Growth of Plato's Logic*. Longmans, Green and Co, London, New York, Bombay.

John William Mackail. 1885. *The Aeneid of Virgil*, volume 36. Macmillan.

Robin McKinley. 2011. *Beauty*. Random House.

Frederick Mosteller and David L Wallace. 1964. *Inference and Disputed Authorship: The Federalist*. Addison-Wesley, Reading, MA.

Editors Nature. 2020. Next chapter in artificial writing. *Nat. Mach. Intell.*, page 419.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318. Association for Computational Linguistics.

Francine Rivers. 2012. *A Lineage of Grace*. Tyndale House Publishers, Inc.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. In *Advances in Neural Information Processing Systems*, pages 6833–6844.

Lucia Specia. 2010. Translating from complex to simplified sentences. *Computational Processing of the Portuguese Language*, pages 30–39.

Alexey Tikhonov, Viacheslav Shibaev, Aleksander Nagaev, Aigul Nugmanova, and Ivan P. Yamshchikov. 2019. Style transfer for texts: Retrain, report errors, compare with rewrites. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3936–3945, Hong Kong, China. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Wei Xu, Alan Ritter, William B Dolan, Ralph Grishman, and Colin Cherry. 2012. Paraphrasing for style. In *24th International Conference on Computational Linguistics, COLING 2012*.

William Zaranka. 1981. *The Brand-X Anthology of Poetry*. Apple-Wood Books.

Yexun Zhang, Ya Zhang, and Wenbin Cai. 2018. Separating style and content for generalized style transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8447–8455.

Zhemin Zhu, Delphine Bernhard, and Iryna Gurevych. 2010. A monolingual tree-based translation model for sentence simplification. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 1353–1361. Association for Computational Linguistics.