Measuring Content Preservation in Textual Style Transfer

Stuart Fitzpatrick[1(✉)], Laurence Park[2], and Oliver Obst[2]

[1] School of Computer, Data and Mathematical Sciences, Western Sydney University, Penrith, Australia
S.Fitzpatrick2@westernsydney.edu.au

[2] Centre for Research in Mathematics and Data Science, Western Sydney University, Penrith, Australia
lapark@cdms.westernsydney.edu.au, O.Obst@westernsydney.edu.au

Abstract. Style transfer in text, changing text that is written in a particular style such as the works of Shakespeare to be written in another style, currently relies on taking the cosine similarity of the sentence embeddings of the original and transferred sentence to determine if the content of the sentence, its meaning, hasn't changed. This assumes however that such sentence embeddings are style invariant, which can result in inaccurate measurements of content preservation. To investigate this we compared the average similarity of multiple styles of text from the Corpus of Diverse Styles using a variety of sentence embedding methods and find that those embeddings which are created from aggregated word embeddings are style invariant, but those created by sentence embeddings are not.

Keywords: Style transfer · Content preservation · Cosine similarity

1 Introduction

Making a piece of text that is written in one particular way, its style, be it that of Shakespeare, a tweet or everyday Conversational English and making it appear to be written in another way, such as the works of James Joyce or Mary Shelley, in such as way as to preserve the semantics or meaning of the sentence [6] is the task of textual style transfer.

Style transfer in the domain of natural language processing (NLP) often uses a two-valued metric to measure the success of a model's ability to transfer style: The first of these is Transfer Strength or Transfer Accuracy [5–7], in a nutshell if our task is to transfer from style $\mathbb{A}$ to style $\mathbb{B}$ then the Strength or Accuracy of the style transfer is simply the accuracy at which we successfully transfer the style of the input sentence to our target style. The second of these is Semantic or Content Preservation [5–7], which is put simply, the evaluation of whether the input sentence and the output sentence have the same meaning. For example if Shakespeare is being transferred to Conversational English, then "wherefore art

L. A. F. Park et al. (Eds.): AusDM 2022, CCIS 1741, pp. 3–14, 2022.
https://doi.org/10.1007/978-981-19-8746-5_1

thou Romeo?" should be become "where are you Romeo?" and not just a random sentence from Conversational English (which would be correct style transfer).

Of great difficulty however, is the means by which Semantic Preservation is measured. Previously this has been measured using the Bilingual EvaLuation Understudy (BLEU), however this ultimately proved unfit for the task due to issues such as; the ability to easily manipulate the score of a system [13], discouraging output diversity [15], not upweighting words that are semantically important [14,15] and unreliable correlations between n-gram overlaps and human judgements [2].

To overcome this, an approach of taking the cosine distance (called cosine similarity, in this context) between the sentence embeddings of the input and output sentences [15] was developed. The specific methods that sentences are embedded with vary, but are typically aggregated word embeddings, where the individual words are embedded using word2vec or GloVe [3,4,15].

In previous work, when evaluation of content preservation is desired [3,4,6, 15], each pair of input and output sentences has been mapped to vectors using word embedding methods and then the similarity of the sentences is compared; if the vectors are similar, then it is assumed that the content/meaning of the two sentences is the same.

However, if a style transfer system incorrectly outputs a sentence of the same style as the input, then the sentences will be incorrectly evaluated as being more similar than they are intended to be, and conversely, the nature of a style transfer system requires that the input and output sentences are of different styles, does that not mean then that correct style transfer is penalised when it comes to content preservation? So there is clearly an assumption that sentence embeddings are style invariant.

In this paper, we investigate the assumption that sentence embeddings are inherently style invariant. We make as a research contribution; showing that aggregated word embeddings, but not pre-trained sentence, embeddings are indeed style invariant.

The remainder of this paper is organised as follows; in Sect. 2 we discuss related previous work on content preservation for style transfer, in Sect. 3 we detail our investigation into the style invariance of sentence embeddings, in Sect. 4 we present and discuss our results, and we conclude the paper in Sect. 5.

2 Background and Motivation

There has been some work into finding a method by which content preservation within style transfer (and other related areas such as neural machine translation and semantic similarity) may be measured, namely the use of the cosine similarity on the input and output sentences [15]. In this section we discuss cosine similarity and its use, the disentanglement of style and content and the style invariance assumption of sentence embeddings.

2.1 Cosine Similarity

Let us consider two embedded sentences with respect to a style transfer system, an input $\mathbf{A}$ and an output sentence $\mathbf{B}$. The cosine similarity (distance) between the two sentences is given by:

$$\text{Cossim}(\mathbf{A}, \mathbf{B}) = \cos(\theta_{\mathbf{A},\mathbf{B}}) = \frac{\langle \mathbf{A}, \mathbf{B} \rangle}{\|\mathbf{A}\| \cdot \|\mathbf{B}\|} \quad \mathbf{A}, \mathbf{B} \in \mathbb{R}^n \tag{1}$$

where, $\langle \mathbf{A}, \mathbf{B} \rangle$ is the inner product of $\mathbf{A}$ and $\mathbf{B}$, $\|\mathbf{A}\|$ and $\|\mathbf{B}\|$ denotes the norms of $\mathbf{A}$ and $\mathbf{B}$ respectively, $\mathbb{R}$ is the set of real numbers and n is the predefined size of the embedding space.

Recall that a similarity of 0 means that there is no similarity between the two sentence embeddings, a similarity of 1 means the sentence embeddings are exactly the same (to a scaling value) and a similarity -1 means the sentence embeddings are exactly opposite (to a scaling value). Two sentences of different styles, such as the input and output of a style transfer system, carry approximately the same meaning if their cosine similarity is sufficiently close to 1.

In [3] and [4] cosine similarity is used by creating three different aggregated word embeddings, element-wise maximum, element-wise minimum and element-wise mean, then concatenating these into a single vector which has its similarity measured. In the case of [3] the `GloVe-wiki-gigaword-100` [9] embeddings are used and in the case of [4] the `word2vec` [8] embeddings are used.

The work of [15], where the use of cosine similarity in this way was popularised, the cosine similarity is directly used on the sentences that have been embedded with the `SentPiece`[1] embedder. The work of [6] takes the same approach as [15], with the caveats that the sentence embeddings are taken from the `[CLS]` vectors of a `RoBERTa` style classifier that they fine-tuned and that the similarity is later incorporated into an overall style transfer metric.

2.2 Disentanglement of Style and Content

Due to the discreteness of natural language, the fact that short sentences do not contain much style information and a severe lack of parallel corpora, it is not possible at present to fully disentangle the style and content of text, like what is possible in image processing [3,11,12]. Full disentanglement, however, may not be necessary. In the event that style and content cannot be fully disentangled, it should still be possible to attain a partial separation of style and content, such that it is possible to determine the similarity of content between two (or more) sentences.

It is to that end that much previous work has separated the measurement of correct style transfer and preservation of content, with much work going into the latter such as in [7] and [5]. The method of choice in many papers for measuring the content preservation between an input sentence and its transferred

[1] https://github.com/google/sentencepiece.

counterpart, likely chosen for its simplicity, is as previously mentioned, cosine similarity.

However, with the exception of the first authors to replace BLEU with cosine similarity, [15], which is a language translation task (so the overlap isn't perfect), not a style transfer task, no implementation of cosine similarity as the content preservation metric has, to the authors' knowledge, actually verified that the underlying sentence representations (at least partially) disentangle style and content and thus that, cosine similarity does what is intended.

2.3 The Style Invariant Embedding Assumption

More recent work, that which uses cosine similarity over older metrics such as BLEU [3,4,6,15], do not seem to make a mention of the assumption of their work relies upon, namely the one mentioned in Sect. 1:

> *The process of generating a sentence embedding implicitly removes non-semantic (i.e. stylistic) information, that is, it is style invariant.*

The reason it is important that the sentence embeddings are style invariant, is due to the following situations: (1) If the task is to transfer from style $\mathbb{A}$ to $\mathbb{B}$ and the style transfer system incorrectly outputs a sentence in style $\mathbb{A}$, then if the sentence embeddings are not style invariant, then the input/output pair of sentences will be evaluated as more similar to each other than is desired (especially in the case where style $\mathbb{A}$ has had a sentence output at random). (2) If the task is to transfer from style $\mathbb{A}$ to $\mathbb{B}$ and the style transfer system correctly outputs to a sentence in style $\mathbb{B}$, then as each style contains different sets of words that have the same meaning, the input/output pair of sentences will be evaluated as being less similar than is desired.

Fortunately, investigating whether the various sentence embedding methods are style invariant is rather straightforward. Let us assume for a moment that it is the case that the assumption is true. In a corpus of non-parallel sentences in multiple styles, then it should be the case that any given style within the corpus has an equal probability of being most similar to any other style, that it is equivalent to a single sample from a discrete uniform distribution over the number of styles (11 in the case of our data). If however the embedding methods are not style invariant then it is the case that multiple styles will demonstrate self-similarity (being most similar to itself).

3 Experiment

The purpose of this research is to determine if the various sentence embedding approaches (aggregated word embeddings and pre-trained sentence embedders) are style invariant. In this section, we evaluate the style invariance of three different sentence embedding approaches; two aggregated word embeddings and one pre-trained sentence embedding model, by computing the average cosine

similarity of each of eleven styles of the Corpus of Diverse Styles (CDS), and observing how often a style demonstrates self-similarity (that is, it is most similar to itself). If self-similarity is rare or non-existent when using a specific embedding method, then it holds that the sentence embedding method is style-invariant. We then use `t-SNE` to dimensionally reduce the sentence embeddings to determine if there is any clustering of the embeddings with respect to style of the sentence it represents.

It should be noted that this task is a challenge to solve without access to a corpus of sentences in different styles that are parallel with respect to content (i.e. they have the same content in a variety of styles.). The use of the CDS, which contains non-parallel sentences in multiple different styles will aid with task by providing sentences that can only be similar due to the style of the sentence and the use of common words across many or all styles (such as *the*, *and*, *of*, and *by*), and thus if a style shows self-similarity it must be due to one (or both) of those properties.

3.1 Dataset

The Corpus of Diverse Styles (CDS) [6] is a dataset that consists of non-parallel sentences in the following eleven styles, 1000 sentences each, collated from a variety of sources:

1. 1810–1820
2. 1890–1900
3. 1990–2000
4. African American English Tweets (AAE)
5. King James Bible
6. English Tweets
7. James Joyce
8. Song Lyrics
9. Romantic Poetry
10. Shakespeare
11. Switchboard

The data also contains generated sentences from the Style TRansfer as Paraphrasing (STRAP) of [6], but these are not used to ensure that our results are not biased by this model.

3.2 Procedure

The aim of the this experiment is to determine the style invariance of various sentence embedding techniques and therefore the appropriateness of each of these methods for determining the content preservation of various style transfer systems. To achieve this we undertake the following procedure. Each of the sentences in each of the styles were embedded using each of the following methods.

– `GloVe-wiki-gigaword-300` [9] concatenated with element-wise summation.

– `FastTest-Crawl-300d-2M` [1] concatenated with element-wise summation.
– `Distilroberta-base` [10]

In the cases where the embedding method is a sentence embedding model, no further action is undertaken, in the cases where the embedding method is a word embedding model, then a sentence embedding is generated by taking the element-wise sum of each of the embedded words, to create a sentence embedding with a dimensionality identical to that of the word embeddings.

For each of these sentence embeddings, the cosine similarity of that sentence embedding versus every sentence embedding in every style (including its own style) are calculated. Each sentence embedding in a style, which has associated similarities for each other sentence embedding it was compared with in each given style, has the mean of these similarities taken. These means, when taken together comprise the *sampling distribution* of style **A** when compared to style **B**.

Next, the mean of each of these sampling distributions is taken as the estimate of the input style's similarity to the target style. The tables that show all 121 pairwise comparisons for each of the three sentence embedding methods, is shown below in Sect. 4. Finally, for each style, we find which style it is the most similar to (i.e. has the highest average cosine similarity). This will allow us to determine if the embeddings are style invariant, by showing how close each of the style to style similarity comparison is to the result of pure chance.

4 Results and Discussion

Provided in Tables 1, 2 and 3 are the style similarity estimates for the eleven different styles in the CDS, as labelled in Sect. 3.1, using each of the different embedding methods, with the highest similarity in bold.

Table 1. Average Cosine Similarity of Each of the Eleven Styles in the CDS as labelled in Sect. 3.1 - `GloVe-wiki-gigaword-300`, bold indicates the highest level of similarity.

	1	2	3	4	5	6	7	8	9	10	11
1	**0.775**	0.765	0.730	0.558	0.768	0.591	0.588	0.653	0.572	0.654	0.758
2	**0.757**	0.749	0.718	0.550	0.750	0.581	0.579	0.643	0.561	0.641	0.745
3	**0.732**	0.724	0.696	0.541	0.720	0.568	0.558	0.629	0.542	0.622	0.727
4	0.594	0.591	0.569	0.521	0.594	0.512	0.466	0.575	0.456	0.563	**0.656**
5	0.766	0.753	0.717	0.554	**0.782**	0.582	0.583	0.647	0.572	0.657	0.740
6	0.624	0.617	0.597	0.511	0.620	0.520	0.485	0.579	0.471	0.570	**0.666**
7	**0.606**	0.601	0.576	0.452	0.602	0.475	0.471	0.527	0.456	0.525	0.605
8	0.706	0.699	0.673	0.570	0.702	0.578	0.547	0.646	0.533	0.638	**0.743**
9	0.601	0.592	0.565	0.448	**0.607**	0.467	0.464	0.522	0.464	0.529	0.592
10	0.691	0.681	0.650	0.545	0.696	0.559	0.533	0.622	0.523	0.632	**0.713**
11	0.756	0.748	0.718	0.606	0.743	0.617	0.581	0.687	0.562	0.676	**0.806**

Table 2. Average Cosine Similarity of Each of the Eleven Styles in the CDS as labelled in Sect. 3.1 - `FastText-Crawl-300d-2M`, bold indicates the highest level of similarity.

	1	2	3	4	5	6	7	8	9	10	11
1	0.591	0.580	0.559	0.463	0.617	0.464	0.399	0.550	0.470	0.535	**0.624**
2	0.569	0.564	0.546	0.461	0.590	0.457	0.389	0.540	0.455	0.517	**0.612**
3	0.552	0.549	0.537	0.457	0.569	0.449	0.380	0.530	0.442	0.500	**0.598**
4	0.466	0.472	0.466	0.478	0.493	0.440	0.338	0.510	0.390	0.473	**0.573**
5	0.586	0.572	0.549	0.466	**0.656**	0.460	0.400	0.549	0.481	0.549	0.612
6	0.472	0.472	0.464	0.448	0.492	0.431	0.334	0.496	0.387	0.465	**0.561**
7	0.409	0.409	0.396	0.355	0.436	0.344	0.296	0.400	0.341	0.391	**0.448**
8	0.508	0.506	0.495	0.467	0.534	0.444	0.355	0.532	0.411	0.492	**0.603**
9	0.490	0.483	0.468	0.398	**0.528**	0.392	0.339	0.464	0.415	0.461	0.510
10	0.497	0.490	0.473	0.439	0.544	0.421	0.347	0.498	0.410	0.501	**0.560**
11	0.560	0.558	0.543	0.506	0.577	0.490	0.390	0.581	0.444	0.535	**0.685**

Table 3. Average Cosine Similarity of Each of the Eleven Styles in the CDS as labelled in Sect. 3.1 - `Distilroberta-base`, bold indicates the highest level of similarity.

	1	2	3	4	5	6	7	8	9	10	11
1	**0.972**	**0.972**	0.968	0.967	0.960	0.970	0.966	0.966	0.963	0.964	0.963
2	0.972	**0.973**	0.970	0.966	0.956	0.969	0.964	0.966	0.960	0.962	0.960
3	0.968	0.970	**0.976**	0.964	0.953	0.964	0.960	0.965	0.957	0.958	0.955
4	0.967	0.965	0.964	**0.977**	0.969	0.965	0.972	0.973	0.969	0.966	0.969
5	0.960	0.957	0.955	0.970	**0.972**	0.961	0.967	0.968	0.970	0.970	0.970
6	**0.970**	0.969	0.964	0.966	0.961	**0.970**	0.965	0.964	0.963	0.964	0.964
7	0.968	0.966	0.963	**0.973**	0.967	0.967	0.970	0.969	0.967	0.965	0.968
8	0.966	0.966	0.966	**0.973**	0.967	0.964	0.969	0.972	0.968	0.965	0.967
9	0.962	0.959	0.957	0.970	0.969	0.962	0.967	0.968	**0.971**	0.969	0.968
10	0.965	0.962	0.960	0.966	0.967	0.965	0.964	0.965	0.968	**0.973**	0.967
11	0.963	0.960	0.957	**0.970**	**0.970**	0.964	0.968	0.968	0.970	0.969	0.970

Provided in Tables 4, 5 and 6 are each style (as labelled in Sect. 3.1), and which style it is most similar to. Any self similarity (a style being most similar to itself) is shown in bold. It is interesting to note that only `Distilroberta-base` had any occurrences of tied greatest similarity.

Table 4. Each Style and its Most Similar Style as labelled in Sect. 3.1 - `GloVe-wiki-gigaword-300`, bold indicates self-similarity.

Style	1	2	3	4	5	6	7	8	9	10	11
Is most similar to	**1**	1	1	11	**5**	11	1	11	5	11	**11**

Table 5. Each Style and its Most Similar Style as labelled in Sect. 3.1 - `FastTest-Crawl-300d-2M`, bold indicates self-similarity.

Style	1	2	3	4	5	6	7	8	9	10	11
Is most similar to	11	11	11	11	**5**	11	11	11	5	11	**11**

Table 6. Each Style and its Most Similar Style as labelled in Sect. 3.1 - `Distilroberta-base`, bold indicates self-similarity.

Style	1	2	3	4	5	6	7	8	9	10	11
Is most similar to	**1**, 2	**2**	**3**	**4**	**5**	1, **6**	4	4	9	10	4, 5

In Table 7, it is shown how many times each embedding method results in self similarity.

Table 7. Self Similarity Counts and Proportions for Each of the Three Embedding Methods

	`GloVe-wiki-gigaword-300`	`FastTest-Crawl-300d-2M`	`Distilroberta-base`
Count	3	2	6
Percentage	27.3	18.2	54.5

As there are eleven styles in the CDS, if the various sentence embeddings were indeed fully style invariant, then each style has an equal probability of being as similar to itself as any other style (i.e. the style similarity is discretely uniform over the number of styles), thus it would not be unreasonable to see approximately $\frac{1}{11} \approx 9\%$ occurrence of self-similarity. Although, as stated in Sect. 2.2, style and content are (at least at present) not fully separable from one another, 9% provides a maximum lower bound that sentence embedding methods can strive towards.

Table 7 shows that `Distilroberta-base` performs quite poorly with over half of the styles as embedded showing self-similarity. `GloVe-wiki-gigaword-300` does much better with only three styles showing self-similarity and `FastTest-Crawl-300d-2M` does better again with only two styles showing self-similarity.

From these results it can be concluded that `Distilroberta-base` performs quite poorly at being style invariant, with `GloVe-wiki-gigaword-300` perform-

ing twice as well and `FastTest-Crawl-300d-2M` performing the best with only two occurrences of self-similarity and is the closest to the target of 9% ($\frac{1}{11}$).

Shown in the Figs. 1, 2 and 3 are the `t-SNE` plots for the sentence embeddings normalised to length 1.

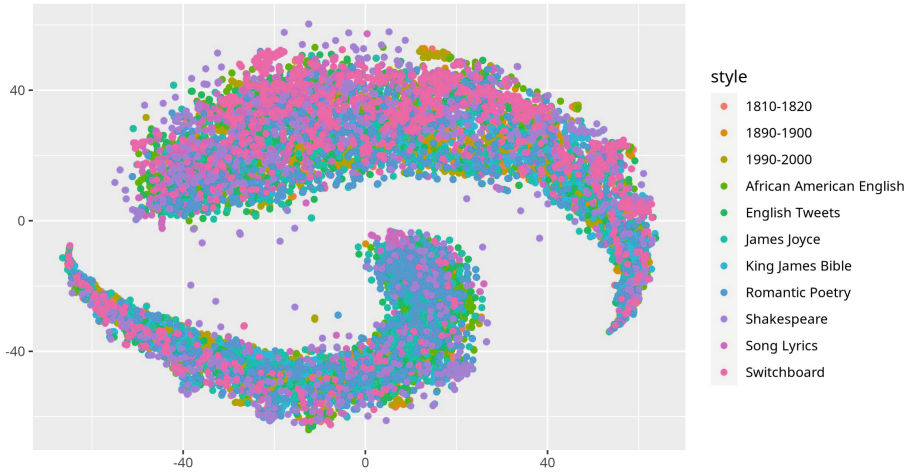

Fig. 1. t-SNE Plot - GloVe aggregated word embeddings

As can be seen in Table 7, both types of aggregated word embeddings have only 3 and 2 cases of self-similarity (GloVe and FastText, respectively) compared to the comparatively high 6 in the case of the dedicated sentence embedding model.

Given that some of the similarities are quite close (as seen in Tables 1, 2 and 3), it is not unreasonable to believe that at least some of these are due to chance.

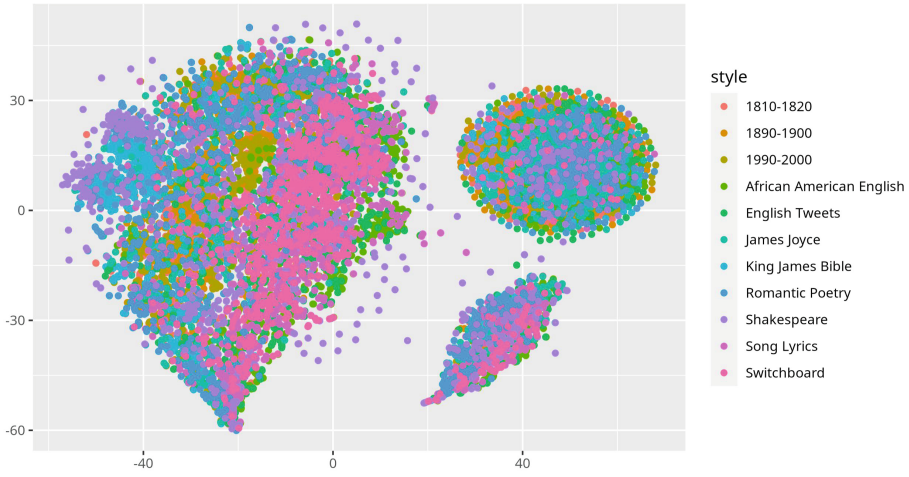

Fig. 2. t-SNE Plot - Fast text aggregated word embeddings

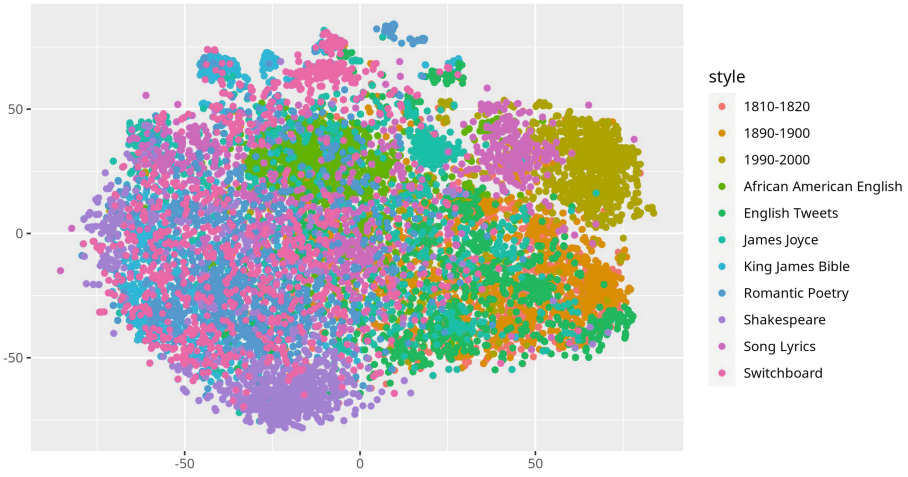

Fig. 3. t-SNE Plot - Distilroberta sentence embeddings

It is hence desirable to visualise a lower dimensional projection of the points to see if there is indeed any clustering.

Figure 1, the `t-SNE` plot for the aggregated `GloVe` word embeddings, shows a classic 'Swiss Roll' shape. Figure 2, the corresponding plot for `FastText`, shows three distinct groups of points. Figure 3, the corresponding plot for `Distilroberta` shows a relatively even spread of points, with one style clustering on the right hand side.

First, looking at the case of `Distilroberta`, it is clear to see that, although there is a good deal of overlap between each of the styles, several styles, most notably 1990–2000, do have a relatively clear, well-defined centre, so although some level of style disentanglement has been achieved, ideally something much better can be done. Next, looking at both Figs. 1 and 2, it can be seen that there are distinct clusters that form within both plots, however it can be seen that these clusters are not with respect to the sentence style. Thus aggregated word embeddings (at least compared to trained sentence embedding approaches), particularly those of `FastTest-Crawl-300d-2M` can be seen as having a much better disentanglement of style.

5 Conclusion

Transferring the style of text from one style to another, requires the measurement of two separate, but entangled properties of text; the style of the original sentence and its transferred version, to determine if the style was changed correctly and content preservation of the transfer, to determine if the two sentences still mean the same thing.

The measurement of content preservation currently utilises taking the cosine similarity of the sentence embeddings of the input and output sentences of

the style transfer system. This however relies on the assumption that sentence embedding methods are style invariant.

In this work, using previously existing data, we estimated the average similarity of multiple different textual styles to each other when embedded using various sentence embedding methods and found that sentence embeddings made from aggregated word embeddings have acceptable levels of style invariance, but that pre-trained sentence embeddings do not. We further also generated and plotted, dimensionally reduced points to investigate the nature of any of any clustering.

Possible future work could be undertaken along two paths. First is to investigate why dedicated sentence embedding models such as `Distilroberta-base` are worse at removing the stylistic properties than aggregated word embeddings. Second is to investigate if it is possible to explicitly train a sentence embedding model to remove as much style from a sentence as possible.

References

1. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Trans. Assoc. Comput. Linguist. **5**, 135–146 (2017)
2. Callison-Burch, C., Osborne, M., Koehn, P.: Re-evaluating the role of bleu in machine translation research. In: 11th Conference of the European Chapter of the Association for Computational Linguistics, pp. 249–256 (2006)
3. Fu, Z., Tan, X., Peng, N., Zhao, D., Yan, R.: Style transfer in text: exploration and evaluation. arXiv preprint arXiv:1711.06861 (2017)
4. Gong, H., Bhat, S., Wu, L., Xiong, J., Hwu, W.: Reinforcement learning based text style transfer without parallel training corpus. arXiv preprint arXiv:1903.10671 (2019)
5. Jin, D., Jin, Z., Hu, Z., Vechtomova, O., Mihalcea, R.: Deep learning for text style transfer: a survey. Comput. Linguist. **48**(1), 155–205 (2022)
6. Krishna, K., Wieting, J., Iyyer, M.: Reformulating unsupervised style transfer as paraphrase generation. arXiv preprint arXiv:2010.05700 (2020)
7. Li, J., Jia, R., He, H., Liang, P.: Delete, retrieve, generate: a simple approach to sentiment and style transfer. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, New Orleans, Louisiana, Volume 1 (Long Papers), pp. 1865–1874. Association for Computational Linguistics (2018). https://doi.org/10.18653/v1/N18-1169. https://www.aclweb.org/anthology/N18-1169
8. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems, vol. 26 (2013)
9. Pennington, J., Socher, R., Manning, C.D.: Glove: global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543 (2014)
10. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv abs/1910.01108 (2019)
11. Shen, T., Lei, T., Barzilay, R., Jaakkola, T.: Style transfer from non-parallel text by cross-alignment. In: Advances in Neural Information Processing Systems, pp. 6830–6841 (2017)

12. Subramanian, S., Lample, G., Smith, E.M., Denoyer, L., Ranzato, M., Boureau, Y.L.: Multiple-attribute text style transfer. arXiv preprint arXiv:1811.00552 (2018)
13. Tikhonov, A., Shibaev, V., Nagaev, A., Nugmanova, A., Yamshchikov, I.P.: Style transfer for texts: retrain, report errors, compare with rewrites. arXiv preprint arXiv:1908.06809 (2019)
14. Wang, A., Cho, K., Lewis, M.: Asking and answering questions to evaluate the factual consistency of summaries. arXiv preprint arXiv:2004.04228 (2020)
15. Wieting, J., Berg-Kirkpatrick, T., Gimpel, K., Neubig, G.: Beyond bleu: training neural machine translation with semantic similarity. arXiv preprint arXiv:1909.06694 (2019)