Style-transfer and Paraphrase: Looking for a Sensible Semantic Similarity Metric

Ivan P. Yamshchikov
Max Planck Institute for Mathematics in the Sciences
Inselstraße 22
Leipzig, Germany 04103
ivan@yamshchikov.info

Viacheslav Shibaev
Ural Federal University
Ekaterinburg, Russia

Nikolay Khlebnikov
Ural Federal University
Ekaterinburg, Russia

Alexey Tikhonov
Yandex
Berlin, Germany

Abstract

The rapid development of such natural language processing tasks as style transfer, paraphrase, and machine translation often calls for the use of semantic preservation metrics. In recent years a lot of methods to control the semantic similarity of two short texts were developed. This paper provides a comprehensive analysis for more than a dozen of such methods. Using a new dataset of fourteen thousand sentence pairs human-labeled according to their semantic similarity, we demonstrate that none of the metrics widely used in the literature is close enough to human judgment to be used on its own in these tasks. The recently proposed Word Movers Distance (WMD), along with bilingual evaluation understudy (BLEU) and part-of-speech (POS) distance, seem to form a reasonable complex solution to measure semantic preservation in reformulated texts. We encourage the research community to use the ensemble of these metrics until a better solution is found.

1 Introduction

Style transfer and paraphrase are two tasks in Natural Language Processing (NLP). Both of them are centered around the problem of an automated reformulation. Given an input text, the system tries to produce a new rewrite that resembles the old text semantically. In the task of paraphrase, semantic similarity is the only parameter that one tries to control. Style transfer usually controls more aspects of the text and could, therefore, be regarded as an extension of a paraphrase. Intuitive understanding of style transfer problem is as follows: if an input text has some attribute A, say, politeness, a system generates new text similar to the input semantically but with attribute A changed to the target $\tilde{A}$. For example, given a polite sentence "could you be so kind, give me a hand" and a target "not polite" the system produces a rewrite "God damn, help me".

The significant part of current works perform style transfer via an encoder-decoder architecture with one or multiple style discriminators to learn disentangled representations (Hu et al., 2017). This basic architecture can have various extensions, for example, can control POS-distance between input and output (Tian et al., 2018), or have additional discriminator or an extra loss term to improve the quality of the latent representations (Yamshchikov et al., 2019). There are also other approaches to this problem that do not use ideas of disentangled latent representations but rather treat it as a machine translation problem; see, for example, (Subramanian et al., 2018). However, independently of a chosen architecture, one has to control the semantic component of the output text. It is expected to stay the same as the system changes the style of the input. This aspect makes the problem of style transfer naturally related to the problem of paraphrase (Prakash et al., 2016), (Gupta et al., 2018), (Roy and Grangier, 2019). It also raises the question of how one could automatically measure the semantic similarity of two texts in these problems.

As with every NLP task that is relatively new, the widely accepted baselines and evaluations metrics are still only emerging. There are ongoing discussions on which aspects of the texts are stylistic and could be changed by the style transfer system and which are semantic and therefore are technically out of the scope of the style transfer research (Tikhonov and Yamshchikov, 2018). This paper refrains from these discussions. It instead attempts to systematize existing methods of quality assessment for the tasks of style transfer that are used in different state of art research results. We also put these methods into the perspective of paraphrase tasks. To our knowledge, that was not done before. The contribution of the paper is four-fold:

- it compares more than a dozen of existing

semantic similarity metrics used by different researchers to measure the performance of different style transfer methods;

- using human assessment of 14 thousand pairs of sentences it demonstrates that there is still no optimal semantic-preservation metric that could be comparable with human judgment in context of paraphrase and textual style transfer, instead it also suggests to use an ensemble of three best available candidates;

- it proposes a simple necessary condition that a metric should comply with to be a valid semantic similarity metric for the task of style transfer;

- it shows that some metrics used in style transfer literature should not be used in the context of style transfer at all.

2 Measuring semantic preservation

Style transfer, as well as a paraphrase, naturally demands the preservation of the semantic component as the input sentence is transformed into the desired output. Different researchers use different methods to measure this preservation of semantics.

Despite its disadvantages (Larsson et al., 2017), one of the most widely used semantic similarity metrics is BLEU. (Tikhonov et al., 2019) show that it could be manipulated in a way that the system would show higher values of BLEU on average, producing sentences that are completely detached from the input semantically. However, BLEU is easy to calculate and is broadly accepted for various NLP tasks that demand semantic preservation (Vaswani et al., 2017), (Hu et al., 2017), (Cohn-Gordon and Goodman, 2019). Alongside BLEU, there are other, less broadly accepted metrics for semantic preservation. For example, (Zhang et al., 2018) work with different versions of ROUGE.

(Fu et al., 2018), (John et al., 2018) or (Romanov et al., 2018) compute a sentence embedding by concatenating the min, max, and mean of its word embeddings and use the cosine similarity between the source and generated sentence embeddings as an indicator of content preservation. (Tian et al., 2018) uses POS-distance alongside with BLEU and BLEU between human-written reformulations and the actual output of the system.

One of the most recent contributions in this area (Mir et al., 2019) evaluates several of the metrics mentioned above as well as METEOR (Banerjee and Lavie, 2005) and Word Mover's Distance (WMD). This metric is calculated as the minimum "distance" between word embeddings of input and output (Kusner et al., 2015).

In this paper, we use these metrics of content-preservation listed above alongside with several others that are used for semantic similarity in other NLP tasks recently. We put all these metrics into the context of paraphrase and style transfer. These metrics are:

- POS-distance that looks for nouns in the input and output and is calculated as a pairwise distance between the embeddings of the found nouns;

- Word overlap calculated as a number of words that occur in both texts;

- chrF (Popović, 2015) – a character n-gram F-score that measures number of n-grams that coincide in input and output;

- cosine similarity calculated in line with (Fu et al., 2018) with pre-trained embeddings by GloVe (Pennington et al., 2014);

- cosine similarity calculated similarly but using FastText word embeddings (Joulin et al., 2016);

- L2 distance based on ELMo (Peters et al., 2018)

- WMD (Kusner et al., 2015) that defines the distance between two documents as an optimal transport problem between the embedded words;

- BLEU (Papineni et al., 2002);

- ROUGE-1 (Lin and Hovy, 2000) compares any text to any other (typically human-generated) summary using a recall-oriented approach and unigrams;

- ROUGE-2 that uses bigrams;

- ROUGE-L (Lin and Och, 2004) that identifies longest co-occurring in sequence n-grams;

- Meteor (Banerjee and Lavie, 2005) metric that is based on a harmonic mean of unigram precision and recall, with recall weighted higher than precision and some additional features, such as stemming and synonymy matching;

- and the most novel BERT score proposed in (Zhang et al., 2019) for the estimation of the generated texts.

All these metrics are known to vary from dataset to dataset but show consistent results within one data collection. In the next section, we try to come up with a set of various paraphrases and style transfer datasets that would allow us to see qualitative differences between these metrics of semantic similarity.

3 Data

The task of paraphrasing a given sentence is better formalized than the task of style transfer. However, to our knowledge, there were no attempts to look at these two tasks in one context. Here we intend to work with the metrics listed in the previous section and calculate them over three paraphrase and two style transfer datasets that are often used for these two NLP tasks. The paraphrase datasets include:

- different versions of English Bibles (Carlson et al., 2017);
- English Paralex dataset[1];
- English Paraphrase dataset[2].

The style transfer datasets are:

- Dataset of politeness introduced in (Rao and Tetreault, 2018) that we in line with the original naming given by the authors refer to as GYAFC later on;
- Yelp! Reviews[3] enhanced with human written reviews with opposite sentiment provided by (Tian et al., 2018).

We suggest to work with these datasets, since they are frequently used for baseline measurements in paraphrase and style transfer literature.

Every dataset is preprocessed as follows. We align every paraphrase or rewrite pairwise. These aligned datasets are then used to sample sets of pairs over which we average the metric of semantic similarity. We also estimate the standard deviation for every metric. Then we pair sentences in every dataset randomly, forming a new dataset of aligned pairs. With random pairing, there is hardly any semantic similarity between two texts. Still, both sentences could probably have similar style characteristics. These random datasets are also useful to estimate 'zero' for every semantic similarity metric and see how noisy it is.

All the metrics that we include in this paper already have undergone validation. These metrics hardly depend on the size of the random data sample provided it is large enough. Also, they are known to vary from one dataset to another. However, due to the laborious nature of this project, we do not know of any attempts to characterize these differences across various datasets.

4 Assessment

This paper is focused on the applications of semantic similarity to the tasks of style transfer and paraphrase, however there are more NLP tasks that depend on semantic similarity measures. We believe that the reasoning and measurements presented in this paper are general enough to be transferred to other NLP tasks that depend upon a semantic similarity metric.

Table 1 and Table 2 show the results for five datasets and thirteen metrics as well as the results of the human evaluation of semantic similarity. It is essential to mention that (Rao and Tetreault, 2018) provide different reformulations of the same text both in an informal and formal style. That allows us to use the GYAFC dataset not only as a style transfer dataset but also as a paraphrase dataset, and, therefore, extend the number of datasets in the experiment. To stimulate further research of semantic similarity measurements, we publish[4] our dataset that consists of 14 000 different pairs of sentences alongside with average semantic similarity score. The semantic similarity scores were given by 300+ English native speakers with at least three scores for every sentence pair. Each sentence was annotated by at least three humans independently. The annotator was presented with two parallel sentences and was asked to assess how similar is their meaning. We used AmazonTurk with several restrictions on the turkers: these should be native speakers of English in the top quintile of the internal rating). Humans were to assess "how similar is the meaning of these two sentences" on a scale from 1 to 5. Across all pairs of sentences, annotators agree with the variance 0.96. This variance

[1] http://knowitall.cs.washington.edu/paralex/
[2] http://paraphrase.org
[3] https://www.yelp.com/dataset
[4] https://drive.google.com/file/d/1pUJ03NWFuJpqiwfVjtiJDF6ZXNuVjdZO

differs on various datasets but never exceeds 1.3. We publish all scores that were provided by the annotators to enable further methodologic research. We hope that this dataset could be further used for a deeper understanding of semantic similarity.

5 Discussion

Let us briefly discuss the desired properties of a hypothetical ideal content preservation metric. We do understand that this metric can be noisy and differ from dataset to dataset. However, there are two basic principles with which such metrics should comply. First, every content preservation metric that is aligned with actual ground truth semantic similarity should induce similar order on any given set of datasets. Indeed, if we have two metrics M_1 and M_2 both of which claim to measure content preservation in two given parallel datasets D_a and D_b, then if in terms of the order induced by M_1 the following holds

$$M_1(D_a) \leq M_1(D_b),$$

one should expect that in terms of the order induced by M_2 the following would be true as well

$$M_2(D_a) \leq M_2(D_b).$$

Since style is a vague notion it is hard to intuitively predict what would be the relative ranking of style transfer pairs of sentences D_s, and paraphrase pairs D_p. However, under order induced by an ideal semantic preservation metric one expects to see both these datasets to be ranked above the dataset D_r that consists of random pairs

$$M(D_r) \leq M(D_s); \quad M(D_r) \leq M(D_p). \quad (1)$$

As we have mentioned before, style is a relatively new and vague notion, so one can not demand that under a metric-induced order the style-transfer data would be clearly distinguished from the paraphrase data $M(D_s) < M(D_p)$. However, it seems more than natural to disqualify any metric that induces such an order that a randomized dataset ends up above the paraphrase dataset, equivalently one can demand that $M(D_r) < M(D_p)$.

Let us now look closely at Table 1 and Table 2 to see if any of the examined metrics comply with these criteria. Table 3 summarizes order induced on the set of the paraphrase, style transfer, and randomized datasets. It is important to note here that we have also gathered human assessment of semantic similarity for one thousand pairs out of every datasets. One can see that humans, indeed, rank random pairs as less semantically similar than paraphrases or style-transfer rewrites. Generally, human ranking corresponds to the intuition described in Equations 1.

What is particularly interesting is that humans assess GYAFC reformulations (the sentences with supposedly similar semantic but varying level of politeness) as the most semantically similar, however Yelp! rewrites that contain the same review of a restaurant but with a different sentiment are ranked as the least similar. A close examination of Table 3 shows that several metrics work quite well in terms of semantic preservation measurement, yet diverge from human assessment when working with Yelp! rewrites. This illustrates the argument made in (Tikhonov and Yamshchikov, 2018) that sentiment is perceived as an aspect of semantics rather than style. Indeed, humans treat sentiment as a semantic attribute of a text. Therefore, addressing the sentiment transfer problem as an example of the style transfer problem can cause systemic errors in terms of semantic similarity assessment. Unfortunately this often happens in modern style transfer research and should be corrected.

Closely examining Table 3 one can make several conclusions. First of all, cosine similarity metrics do not seem to be useful as metrics of semantic preservation since they state that randomized pairs of sentences are closer semantically than parallel paraphrases. All the other metrics induce similar order on the set of the datasets. Indeed, Table 4 shows pairwise correlations between human-induced order and the orders that other metrics induce as well as the correlation of the absolute metric values with human assessments.

One can also introduce several scoring systems to estimate how well every metric performs in terms of Inequalities 1. For example, we can calculate, how many datasets get the same rank from the metric-induced order as from the human-induced one. Another possible score could be a number of swaps needed to produce the human-induced order out of the metric-induced one. Table 5 shows these scores for the the semantic similarity metrics in questions.

Looking at the results listed above we can recommend the following. First of all, one has to conclude that there is no ”silver bullet” for seman-

Dataset	Human Labeling	POS-distance	Word overlap	chrF	Cosine Similarity Word2Vec	Cosine Similarity FastText	WMD
Bibles	3.54 ± 0.72	2.39 ± 3.55	0.47 ± 0.18	0.54 ± 0.18	0.04 ± 0.04	0.04 ± 0.02	0.57 ± 0.29
Paralex	3.28 ± 0.8	2.91 ± 4.28	0.43 ± 0.18	0.48 ± 0.18	0.13 ± 0.09	0.09 ± 0.04	0.62 ± 0.3
Paraphrase	3.6 ± 0.79	2.29 ± 2.85	0.31 ± 0.2	0.41 ± 0.23	0.29 ± 0.17	0.21 ± 0.12	0.77 ± 0.34
GYAFC formal	3.63 ± 0.75	2.27 ± 3.97	0.5 ± 0.22	0.53 ± 0.22	0.06 ± 0.04	0.05 ± 0.03	0.57 ± 0.35
GYAFC informal	3.41 ± 0.78	3.79 ± 4.54	0.32 ± 0.17	0.34 ± 0.17	0.09 ± 0.05	0.09 ± 0.04	0.76 ± 0.31
Yelp! rewrite	2.68 ± 0.83	1.11 ± 2.34	0.45 ± 0.25	0.51 ± 0.23	0.08 ± 0.06	0.08 ± 0.06	0.61 ± 0.31
GYAFC rewrite	3.83 ± 0.75	2.32 ± 3.91	0.47 ± 0.21	0.53 ± 0.22	0.06 ± 0.04	0.06 ± 0.04	0.54 ± 0.35
GYAFC random informal	2.13 ± 0.73	10.61 ± 7.2	0.05 ± 0.05	0.13 ± 0.04	0.15 ± 0.05	0.15 ± 0.04	1.24 ± 0.08
GYAFC random formal	2.12 ± 0.74	10.82 ± 8.64	0.08 ± 0.05	0.14 ± 0.04	0.15 ± 0.04	0.14 ± 0.03	1.26 ± 0.07
GYAFC random rewrite	2.07 ± 0.7	10.58 ± 8.03	0.06 ± 0.05	0.13 ± 0.04	0.15 ± 0.04	0.14 ± 0.03	1.25 ± 0.07
Yelp! random rewrite	2.14 ± 0.79	8.97 ± 4.35	0.06 ± 0.06	0.14 ± 0.04	0.19 ± 0.06	0.17 ± 0.05	1.26 ± 0.08

Table 1: Various metrics of content preservation calculated across three paraphrase datasets, datasets of rewrites and various randomized datasets. GYAFC Formal and Informal correspond to the content preservation scores for GYAFC data treated as paraphrases in a formal or informal mode respectively. GYAFC and Yelp! rewrite correspond to the score between an input and a human-written reformulation in a different style. GYAFC and Yelp! random stand for the scores calculated on samples of random pairs from the respective dataset.

Dataset	ELMo L2	ROUGE-1	ROUGE-2	ROUGE-L	BLEU	Meteor	BERT score
Bibles	3.71 ± 1.18	0.61 ± 0.17	0.38 ± 0.22	0.58 ± 0.19	0.28 ± 0.24	0.6 ± 0.2	0.93 ± 0.03
Paralex	5.74 ± 1.41	0.58 ± 0.17	0.24 ± 0.2	0.52 ± 0.18	0.07 ± 0.17	0.49 ± 0.22	0.91 ± 0.03
Paraphrase	6.79 ± 1.87	0.43 ± 0.24	0.13 ± 0.22	0.41 ± 0.24	0.01 ± 0.09	0.4 ± 0.27	0.91 ± 0.05
GYAFC informal	5.56 ± 1.31	0.45 ± 0.2	0.22 ± 0.19	0.39 ± 0.19	0.1 ± 0.17	0.4 ± 0.21	0.89 ± 0.04
GYAFC formal	4.17 ± 1.49	0.61 ± 0.21	0.4 ± 0.26	0.57 ± 0.22	0.27 ± 0.28	0.6 ± 0.23	0.93 ± 0.04
Yelp! rewrite	4.89 ± 1.8	0.57 ± 0.24	0.37 ± 0.27	0.54 ± 0.26	0.22 ± 0.28	0.54 ± 0.28	0.92 ± 0.04
GYAFC rewrite	4.57 ± 1.54	0.61 ± 0.21	0.39 ± 0.25	0.56 ± 0.22	0.25 ± 0.27	0.57 ± 0.23	0.92 ± 0.04
GYAFC random informal	7.67 ± 1.02	0.08 ± 0.08	0.01 ± 0.03	0.06 ± 0.07	0.01 ± 0.01	0.06 ± 0.07	0.82 ± 0.02
GYAFC random formal	7.55 ± 0.92	0.08 ± 0.08	0.01 ± 0.03	0.07 ± 0.07	0.000 ± 0.01	0.08 ± 0.06	0.84 ± 0.02
GYAFC random rewrite	7.68 ± 1.03	0.07 ± 0.07	0.01 ± 0.02	0.06 ± 0.06	0.000 ± 0.000	0.06 ± 0.05	0.83 ± 0.02
Yelp! random rewrite	8.19 ± 0.9	0.08 ± 0.09	0.002 ± 0.02	0.07 ± 0.08	0.000 ± 0.000	0.06 ± 0.06	0.85 ± 0.02

Table 2: Various metrics of content preservation calculated across three paraphrase datasets, datasets of rewrites and various randomized datasets. GYAFC Formal and Informal correspond to the content preservation scores for GYAFC data treated as paraphrases in a formal or informal mode respectively. GYAFC and Yelp! rewrite correspond to the score between an input and a human-written reformulation in a different style. GYAFC and Yelp! random stand for the scores calculated on samples of random pairs from the respective dataset.

Metric	Yelp! random rewrite	GYAFC random rewrite	GYAFC random informal	GYAFC random formal	Yelp! rewrite	GYAFC rewrite	GYAFC informal	GYAFC formal	Bibles	Paralex	Paraphrase
POS	8	9	10	11	1	4	7	2	5	6	3
Word overlap	9	10	11	8	4	3	6	1	2	5	7
chrF	9	10	11	8	4	2	7	3	1	5	6
Word2Vec	10	7	9	8	4	2	5	3	1	6	11
FastText	10	8	9	7	4	3	6	2	1	5	11
WMD	10	9	8	11	4	1	6	3	2	5	7
ELMo L2	11	10	9	8	4	3	5	2	1	6	7
ROUGE-1	9	11	10	8	5	3	6	1	2	4	7
ROUGE-2	11	10	8	9	4	2	6	1	3	5	7
ROUGE-L	9	11	10	8	4	3	7	2	1	5	6
BLEU	10	11	8	9	4	3	5	2	1	6	7
Meteor	9	10	11	8	4	3	7	2	1	5	6
BERT score	8	10	11	9	3	4	7	1	2	5	6
Human Labeling	8	11	9	10	7	1	5	2	4	6	3

Table 3: Different semantic similarity metrics sort the paraphrase datasets differently. Cosine similarity calculated with Word2Vec or FastText embeddings do not comply with Inequality $M(D_r) < M(D_p)$. All other metrics clearly distinguish randomized texts from style transfers and paraphrases and are in line with Inequalities 1. However, none of the metrics is completely in line with human evaluation.

tic similarity yet. Every metric that is used for semantic similarity assessment at the moment fails to be in line with human understanding of semantic similarity. Second, judging by Table 4 and Table 5 there are three metrics that seem to be the most apt instruments for the task. These are: WMD that shows a high correlation with human-induced order as well as the highest correlation with human assessment values; POS-distance and BLEU both have lower results in terms of correlation with human-induced order but assign same ranks as humans do to four datasets and are only eleven swaps

Metric	Correlation of the induced orders with human-induced order	Correlation of the metric with human evaluation
POS	0.75	0.87
Word overlap	0.79	0.89
chrF	0.8	0.9
Word2Vec	0.5	0.46
FastText	0.5	0.52
WMD	**0.81**	**0.92**
ELMo L2	0.76	0.82
ROUGE-1	**0.82**	0.9
ROUGE-2	**0.81**	0.84
ROUGE-L	**0.81**	0.89
BLEU	0.8	0.72
Meteor	0.79	**0.91**
BERT score	0.77	0.89

Table 4: Pairwise correlations of the orders induced by different semantic similarity metrics with human-induced semantic similarity alongside with pairwise correlations of the semantic similarity metric with human assessment similarity scores. It is interesting that some metrics show high correlation with human assessment yet produce erroneous ranking of the datasets, particularly failing to reproduce human assessment or Yelp! rewrites.

Metric	Number of ranks coinciding with human-induced ranking	Number of swaps needed to reconstruct human-induced ranking
POS	**4**	**11**
Word overlap	0	12
chrF	0	**11**
Word2Vec	**3**	18
FastText	2	19
WMD	1	**11**
ELMo L2	**4**	12
ROUGE-1	1	**10**
ROUGE-2	0	**11**
ROUGE-L	2	**11**
BLEU	**4**	**11**
Meteor	1	12
BERT score	1	**11**

Table 5: Scores for the orders induced by different semantic similarity metrics. The best scores are marked with asterisks.

away from a human-induced order.

One should also emphasize that the Word2Vec and FastText metrics should not be used for the semantic similarity assessment.

Finally, let us look at Figure 1. There is a clear correlation between all orders induced by the other metrics listed in Table 4. This correlation of induced orders is not only a consistent result that shows that the majority of semantic preservation metrics are aligned to a certain extent. This correlation could also be regarded as a justification of an order theory inspired methodology that we propose here for comparative analysis of metrics.

Looking at Table 5 one should also mention that POS-distance seems to be less aligned with every other metric that was tested, yet shows relatively good results in terms of the correspondence between the induced order and the order based on human evaluation. This could mean the POS-distance captures certain semantic features that other methods to not look at yet that seem to be used by humans when they assess semantic similarity.

The observed correlation of the induced orders gives hope that there is a universal measure of semantic similarity for texts and that all these metrics proxy this potential metric to certain extent. However it is clear that none of them could model human judgement. There are several reasons that account for that. One is the phenomenal recent success of the semantic extraction methods that are based on local rather than global context that made local information-based metrics dominate NLP in recent years. Humans clearly operate in a non-local semantic context yet even state of art models in NLP can not account for it. The fact that BERT score that theoretically could model inner non-local semantics still does not reproduce human semantic similarity estimations is a proof for that. Second reason is the absence of rigorous, universally accepted definition for the problem of style transfer. We hope further research of disentangled semantic representations would allow to designate semantic information in NLP in a more rigorous way, especially in context of several recent attempts to come up with unified notion of semantic information, see for example (Kolchinsky and Wolpert, 2018).

6 Conclusion

In this paper, we examine more than a dozen metrics for semantic similarity in the context of NLP tasks of style transfer and paraphrase. We publish human assessment for semantic similarity of fourteen thousand short text pairs and hope that will facilitate further reserach of semantic similarity metrics. Using very general order theory reasoning, we demonstrate that cosine similarity based metrics should not be used in this context. We also show that the majority of the metrics that occur in style transfer literature induce similar order on the sets of data. This is not only to be expected but also justifies the proposed order-theory methodology. POS-distance is somehow less aligned with this general semantic similarity order but induces order that mimics certain aspects of human judgement more successfully than other metrics. The combination of POS-distance, BLEU and WMD score seems to be the best semantic similarity solution that could be used for style transfer problems as well as problems of paraphrase at the mo-

	POS-distance	Word overlap	chrF	Word2Vec	FastText	WMD	ELMO L2	ROUGE-1	ROUGE-2	ROUGE-L	BLEU	Meteor	BERT score	Human score
POS-distance	100%	77%	76%	47%	48%	75%	70%	72%	75%	77%	72%	78%	86%	75%
Word overlap	77%	100%	96%	83%	87%	87%	95%	98%	93%	97%	93%	98%	97%	79%
chrF	76%	96%	100%	80%	83%	89%	93%	95%	89%	98%	91%	99%	95%	80%
Word2Vec	47%	83%	80%	100%	97%	85%	87%	79%	83%	77%	84%	79%	74%	51%
FastText	48%	87%	83%	97%	100%	82%	89%	85%	85%	83%	85%	84%	79%	50%
WMD	75%	87%	89%	85%	82%	100%	91%	87%	95%	88%	93%	87%	85%	81%
ELMO L2	70%	95%	93%	87%	89%	91%	100%	94%	95%	95%	98%	94%	89%	76%
ROUGE-1	72%	98%	95%	79%	85%	87%	94%	100%	94%	97%	94%	96%	95%	82%
ROUGE-2	75%	93%	89%	83%	85%	95%	95%	94%	100%	92%	95%	90%	88%	81%
ROUGE-L	77%	97%	98%	77%	83%	88%	95%	97%	92%	100%	95%	99%	96%	81%
BLEU	72%	93%	91%	84%	85%	93%	98%	94%	95%	95%	100%	92%	89%	80%
Meteor	78%	98%	99%	79%	84%	87%	94%	96%	90%	99%	92%	100%	97%	79%
BERT score	86%	97%	95%	74%	79%	85%	89%	95%	88%	96%	89%	97%	100%	77%
Human score	75%	79%	80%	51%	50%	81%	76%	82%	81%	81%	80%	79%	77%	100%

Figure 1: Pairwise correlations of the orders induced by the metrics of semantic similarity.

ment. There is still no metric that could distinguish paraphrases form style transfers definitively. This fact is essential in the context of future style transfer research. To put that problem in the context of paraphrase, such semantic similarity metric is needed. Until it is found combination of WMD score, BLEU and POS-distance seems to be the optimal option for comparative analysis of style transfer systems.

References

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Keith Carlson, Allen Riddell, and Daniel Rockmore. 2017. Zero-shot style transfer in text using recurrent neural networks. *arXiv preprint arXiv:1711.04731*.

Reuben Cohn-Gordon and Noah Goodman. 2019. Lost in machine translation: A method to reduce meaning loss. *arXiv preprint arXiv:1902.09514*.

Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2018. Style transfer in text: Exploration and evaluation. *AAAI*.

Ankush Gupta, Arvind Agarwal, Prawaan Singh, and Piyush Rai. 2018. A deep generative framework for paraphrase generation. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

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.

Vineet John, Lili Mou, Hareesh Bahuleyan, and Olga Vechtomova. 2018. Disentangled representation learning for text style transfer. In *arXiv preprint*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. Fasttext. zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Artemy Kolchinsky and David H Wolpert. 2018. Semantic information, autonomous agency and non-equilibrium statistical physics. *Interface focus*, 8(6):20180041.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *International conference on machine learning*, pages 957–966.

Maria Larsson, Amanda Nilsson, and Mikael Kågebäck. 2017. Disentangled representations for manipulation of sentiment in text. *arXiv preprint arXiv:1712.10066*.

Chin-Yew Lin and Eduard Hovy. 2000. The automated acquisition of topic signatures for text summarization. In *Proceedings of the 18th conference on Computational linguistics-Volume 1*, pages 495–501. Association for Computational Linguistics.

Chin-Yew Lin and Franz Josef Och. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 605. Association for Computational Linguistics.

Remi Mir, Bjarke Felbo, Nick Obradovich, and Iyad Rahwan. 2019. Evaluating style transfer for text. *arXiv preprint arXiv:1904.02295*.

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.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference*

on empirical methods in natural language processing (EMNLP), pages 1532–1543.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Maja Popović. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395.

Aaditya Prakash, Sadid A. Hasan, Kathy Lee, Vivek Datla, Ashequl Qadir, Joey Liu, and Oladimeji Farri. 2016. Neural paraphrase generation with stacked residual lstm networks. In *arXiv preprint*.

Sudha Rao and Joel Tetreault. 2018. Dear sir or madam, may i introduce the gyafc dataset: Corpus, benchmarks and metrics for formality style transfer. *arXiv preprint arXiv:1803.06535*.

Alexey Romanov, Anna Rumshisky, Anna Rogers, and David Donahue. 2018. Adversarial decomposition of text representation. In *arXiv preprint*.

Aurko Roy and David Grangier. 2019. Unsupervised paraphrasing without translation. In *arXiv preprint*.

Sandeep Subramanian, Guillaume Lample, Eric Michael Smith, Ludovic Denoyer, Marc'Aurelio Ranzato, and Y-Lan Boureau. 2018. Multiple-attribute text style transfer. *arXiv preprint arXiv:1811.00552*.

Youzhi Tian, Zhiting Hu, and Zhou Yu. 2018. Structured content preservation for unsupervised text style transfer. In *arXiv preprint*.

Alexey Tikhonov, Viacheslav Shibaev, Aleksander Nagaev, Aigul Nugmanova, and Ivan P. Yamshchikov. 2019. Style transfer for texts: Retrain, report errors, compare with rewrites.

Alexey Tikhonov and Ivan P. Yamshchikov. 2018. What is wrong with style transfer for texts? In *arXiv preprint*.

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.

Ivan P Yamshchikov, Viacheslav Shibaev, Aleksander Nagaev, Jürgen Jost, and Alexey Tikhonov. 2019. Decomposing textual information for style transfer. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 128–137.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Ye Zhang, Nan Ding, and Radu Soricut. 2018. Shaped: Shared-private encoder-decoder for text style adaptation. *arXiv preprint arXiv:1804.04093*.