Towards Better Characterization of Paraphrases

Timothy Liu[1] **De Wen Soh**[1]
[1]Singapore University of Technology and Design
timothy_liu@mymail.sutd.edu.sg, dewen_soh@sutd.edu.sg

Abstract

To effectively characterize the nature of paraphrase pairs without expert human annotation, we propose two new metrics: word position deviation (WPD) and lexical deviation (LD). WPD measures the degree of structural alteration, while LD measures the difference in vocabulary used. We apply these metrics to better understand the commonly-used MRPC dataset and study how it differs from PAWS, another paraphrase identification dataset. We also perform a detailed study on MRPC and propose improvements to the dataset, showing that it improves generalizability of models trained on the dataset. Lastly, we apply our metrics to filter the output of a paraphrase generation model and show how it can be used to generate specific forms of paraphrases for data augmentation or robustness testing of NLP models.

1 Introduction

A robust understanding of semantic meaning, despite variances in sentence expression, is an integral part of natural language processing (NLP) tasks. However, many existing NLP models exhibit shortcomings in understanding real-world variations in natural language. These models are often over-reliant on learned spurious correlations resulting in poor generalization (Sanchez et al., 2018; McCoy et al., 2019). This problem is challenging to address since it is difficult to distinguish spurious correlations from useful features (Gardner et al., 2021).

One way of improving the performance and robustness of NLP model is to increase the size of the dataset (Hestness et al., 2017). It is possible to do so in an efficient manner through data augmentation, or the process of generating new data out of existing examples, thus creating more training data or test cases (Feng et al., 2021; Chen et al., 2021). This would also enhance the capability to detecting error in a wide range of NLP systems. We can also condition language models to generate paraphrases of input sentences (Witteveen and Andrews, 2019) through the use of large language models such as GPT (Radford et al., 2019). However, commonly used paraphrase datasets and paraphrase generation techniques that rely on such datasets suffer from several shortfalls, such as being noisy due to loose labelling in these datasets and lack of accurate, controllable generation. In this paper, we make three key contributions to address this issue.

Firstly, we propose two new metrics for better understanding of paraphrase pairs: word position deviation and lexical deviation. We show, with examples, how these metrics are more effective at quantitatively capturing the linguistic characteristics of paraphrase pair than existing methods such as ROGUE-L, SELF-BLEU and edit distance.

Secondly, we apply the proposed metrics to better understand the commonly used Microsoft Research Paraphrase Corpus (MRPC) (Dolan and Brockett, 2005) dataset. We also study how MRPC differs from Paraphrase Adversaries from Word Scrambling (PAWS) (Zhang et al., 2019), another paraphrase identification dataset. In the process, we perform a detailed study on MRPC and propose some revisions to the dataset. We demonstrate that this improves the quality of paraphrase identification models trained on MRPC, with higher transferability to other paraphrase identification datasets.

Lastly, we demonstrate the applicability of our proposed metrics. By applying our metrics to filter the output of a paraphrase generation model, we show how it can be used to generate specific forms of paraphrases, which can be used as training data for data augmentation purposes and to generate test cases for robustness testing of NLP models.

2 Related Work

There have been several survey papers done to better understand the task of paraphrase identification

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 8592 - 8601

and generation. *A Survey of Paraphrasing and Textual Entailment Methods* (Androutsopoulos and Malakasiotis, 2010) presented a comprehensive survey and review on the the aforementioned tasks. In this paper, the authors helped to properly define the tasks and identified some methods and their associated challenges. This was followed up by a more recent survey specifically on the task of paraphrase identification, *A Survey on Paraphrase Recognition* (Magnolini, 2014), where the focus of the survey was the performance of various statistical and non-deep learning approaches on paraphrase identification on the MRPC dataset. Additionally, in *On Paraphrase Identification Corpora* (Rus et al., 2014), the authors performed a survey of various paraphrase datasets, also highlighting several issues with paraphrase datasets, including MRPC, and providing some recommendations for improving the curation of paraphrase datasets.

There have also been previous work on the task of better quantifying various characteristics of paraphrase pairs. In *Texygen: A Benchmarking Platform for Text Generation Models* (Zhu et al., 2018), SELF-BLEU was proposed to measure the diversity in text generation. However, it suffers from limitations inherent to BLEU-style metrics: it captures the differences in presence of n-grams, but not their sequence, and is thus mostly limited to capturing the differences of vocabulary, but not the overall structure of a sentence. In *Paraphrasing with Large Language Models* (Witteveen and Andrews, 2019), ROUGE-L is used as a measurement of paraphrase diversity, where lower ROUGE-L scores correspond to greater diversity in paraphrasing generation. However, ROUGE-L mainly measures degree of similarity in sub-sequences, but not the order in which the sub-sequences occur, and thus cannot accurately capture the possible structural differences present in paraphrase pairs.

In our paper, we take a deeper look at some of the issues related to MRPC, proposing some useful improvements. We also build upon previous attempts to characterise paraphrases through the use of quantitative metrics, demonstrating how our proposed metrics can capture various different paraphrasing techniques better than previously proposed metrics.

3 What is a Paraphrase?

3.1 Definition of Paraphrase

To facilitate more precise discussions in our paper, we clearly define a paraphrase as follows:

Definition 1 (Paraphrase)**.** A sentence is a paraphrase of another sentence if they are not identical but share the same semantic meaning.

Therefore, there are two distinct criteria in order to fulfill the definition of being a paraphrase pair:

1. The two sentences must have the *same* meaning: it is impossible to derive different information from a paraphrase of a sentence. Where two sentences are not certain to have exactly the same meaning, a common interpretation of both sentences should be the same in order for it to be a *reasonable* paraphrase. This also implies that both sentences in a paraphrase pair necessarily entail each other.

2. The two sentences must not be identical, for example having lexical differences (differences in vocabulary) or structural differences (differences in word order, punctuation and syntax).

In *A Survey of Paraphrasing and Textual Entailment Methods* (Androutsopoulos and Malakasiotis, 2010), the following example is provided, which we shall discuss:

> 1. Wonderworks Ltd. constructed the new bridge.
> 2. The new bridge was constructed by Wonderworks Ltd.
> 3. Wonderworks Ltd. is the constructor of the new bridge.

It is argued that sentence 3 is not a precise paraphrase of sentences 1 and 2 as it is not stated precisely in sentence 3 that the bridge was completed. For the purposes of our discussion, we would consider sentence 3 a *reasonable* paraphrase as well as it is very likely that all three sentences would be interpreted in the same way, and thus share the same semantic meaning based on the most common interpretation of the sentences.

These examples illustrate that it is non-trivial to precisely define what is a paraphrase pair, as there is some variance (depending on subjective interpretation) on what would be a precise paraphrase. This problem is observed to have caused issues due to the imprecise definitions used while creating paraphrase datasets, such as the MRPC dataset which is very widely used. By adhering strictly to the definition of a paraphrase as detailed in this section, we hope to better facilitate discussion throughout the paper.

3.2 Paraphrase Datasets

In this paper, we will utilize and compare two commonly used paraphrase datasets, MRPC and PAWS.

3.2.1 Microsoft Research Paraphrase Corpus (MRPC)

The Microsoft Research Paraphrase Corpus (MRPC) is a corpus consists of sentence pairs collected from web news articles (Dolan and Brockett, 2005). This dataset is widely used as a benchmark for the paraphrase identification task. It can be used directly or indirectly as part of the GLUE benchmark (Wang et al., 2019). In particular, as part of the GLUE benchmark, the dataset has been used for training and evaluation in more than 50 research papers as can be determined from the GLUE leaderboard[1]. It is also less commonly used as a paraphrase generation dataset, in works such as (Huang and Chang, 2021). MRPC contains 4076 training and 1725 test examples.

3.2.2 Paraphrase Adversaries from Word Scrambling (PAWS)

The Paraphrase Adversaries from Word Scrambling (PAWS) is a dataset contains sentence pairs extracted from Wikipedia and the Quora Question Pairs (QQP) dataset (Zhang et al., 2019). While it is less commonly used than MRPC, it is a high quality and larger dataset, and is used in a number of papers such as (Yu and Ettinger, 2021), (Tu et al., 2020) and (Chen et al., 2020) for the purpose of paraphrase identification. PAWS contains 49,401 training, 8000 development and 8000 test examples.

4 Proposed Metrics

4.1 Objectives

Our objective is to comprehensively evaluate the diverse linguistic phenomena involved in paraphrasing, which can include techniques such as synonym substitution, negation, diathesis alternation, coordination changes and more. We can broadly classify these techniques into the use of *structural* alternations and *lexical* alternations to achieve paraphrasing.

Thus, to better characterise a paraphrase pair, we propose two metrics: *word position deviation* and *lexical deviation*. These two metrics are introduced so as to provide a quantitative understanding on what type of paraphrase it is along the two types of changes. A key design consideration of these metrics is the need to be able to capture the extents of structural and lexical alterations in an efficient manner, without resorting to costly human annotation or large amounts of computation. We will use these metrics to provide a good understanding of the characteristics of paraphrase pairs both at a individual (paraphrase pair) level and at an aggregate level over the whole dataset. In addition, we apply these metrics to filter outputs from paraphrase generation systems to select for specific types of paraphrases.

4.2 Key Definitions

In this section, we define some terms that will be used across various metrics computations. Let $s1$ and $s2$ denote two sentences. We will also refer to the pair of sentences (s_1, s_2) as a paraphrase pair.

Definition 4.1 (Set of common words)**.** The set of common words $\mathcal{C}_{(s1,s2)}$ of a paraphrase pair is the set of words, in uncased lemmatized form, which occurs in both s_1 and s_2.

Definition 4.2 (Set of all words)**.** The set of all words $\mathcal{A}_{(s1,s2)}$ of a paraphrase pair is the complete set of words, in uncased lemmatized form, which occurs in either or both sentences s_1 and s_2.

Thus, given two sentences:

- s_1: Yesterday, Bob met Tom at the store.
- s_2: Tom met Bob yesterday while they were at the store.
- $\mathcal{C}_{(s1,s2)}$:
  { yesterday, bob, meet, tom, at, the, store }
- $\mathcal{A}_{(s1,s2)}$:
  { yesterday, bob, meet, tom, at, the, store, while, they, be }

We will also use the notation $N_{\mathcal{C}_{(s1,s2)}}$ to refer to the size of set $\mathcal{C}_{(s1,s2)}$ and $N_{\mathcal{A}_{(s1,s2)}}$ to refer to the size of set $\mathcal{A}_{(s1,s2)}$. We use $N_{\mathcal{C}}$ and $N_{\mathcal{A}}$ for short when it is obvious which statements s_1 and s_2 we are referring to. For a word W and a sentence s, we denote by $N_s(W)$ the number of times that the word W appears in the sentence s.

4.3 Word Position Deviation (WPD)

We propose the *word position deviation* (WPD) of a paraphrase pair as a metric that effectively captures the degree of deviation in the structure of paraphrased sentences by looking at changes in word

[1] https://gluebenchmark.com/leaderboard

positions. WPD can be intuitively understood as the mean of how much words shift in position after a paraphrase. We find that this proposed metric is effective in identifying the amount of structural alterations present in paraphrase pairs.

To properly define WPD, we first introduce the concept of *normalized word position* in a paraphrase pair.

Definition 4.3 (Normalized Word Position)**.** Let s be a sentence and W be a word. For $1 \leq n \leq N_s(W)$, the normalized word position $\rho_{s,n}(W)$ of n-th appearance of W in s is its index divided by the index of the last word. Thus, a normalized word position value ranges from the first word in the sentence having a value of 0.0 and last word having value of 1.0. For example, if the second appearance of W has index a and the last word has index b in the sentence s, then $\rho_{s,2}(W) = a/b$.

In WPD, we consider the mean differences between the normalized word positions. For any given word that is common in both sentences in a paraphrase pair (s_1, s_2), we can calculate the *relative position shift* as the difference in normalized word position.

Definition 4.4 (Relative Position Shift)**.** The relative position shift of a word W with respect to sentence s_1 in paraphrase pair (s_1, s_2) is denoted as $\delta_{s1,s2}(W)$, only defined for words in $\mathcal{C}_{(s1,s2)}$, and has the expression

$$\delta_{s1,s2}(W) = \sum_{n=1}^{N_{s_1}(W)} \min_{1 \leq k \leq N_{s_2}(W)} \frac{|\rho_{s_1,n}(W) - \rho_{s_2,k}(W)|}{N_{s_1}(W)}. \quad (1)$$

For each occurrence of W in s_1, we calculate the smallest difference between its normalized word position and that of the occurrences of W in s_2. We then average these smallest differences over all occurrence of W in s_1 to get the relative position shift of W with respect to s_1 in paraphrase pair (s_1, s_2).

In a simple case with only one occurrence of W in each sentence, this reduces to the distance between $\rho_{s_1,1}(W)$ and $\rho_{s_2,1}(W)$, which is

$$\delta_{s1,s2}(W) = |\rho_{s_1,1}(W) - \rho_{s_2,1}(W)|. \quad (2)$$

To the concepts described above, a simple example is provided in Figure 1 below.

We can see that if we had a word W is near the start of s_1 and near the end of s_2, $\delta_{s1,s2}(W)$ is close to 1.0. Conversely, if the word W is near the start of s_1 and near the start of the s_2, $\delta_{s1,s2}(W)$ is close to 1.0.

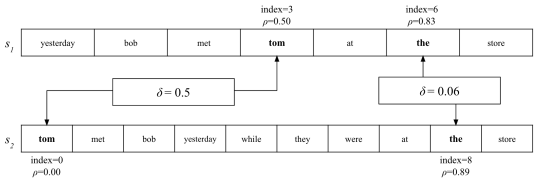

Figure 1: Illustration of individual words' relative position shifts.

In a generalised case where there can be multiple occurrences of W can be present in s_1 or s_2, the mean distance between one occurrence and the nearest occurrence in the other sentence is considered. However, such instances are much rarer. We illustrate the handling of using a real example, showing how the word `his` occurs twice, resulting in a mean $\delta(\text{"}his\text{"})$ of 0.263.

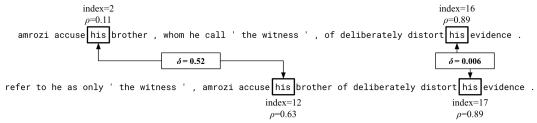

Figure 2: Illustration of a special case where multiple instances of a word occurs.

Thus, we can now define WPD.

Definition 4.5 (Word Position Deviation)**.** Let (s_1, s_2) be a paraphrase pair. The WPD of (s_1, s_2), denoted as $\sigma_{\text{pos}}(s1, s2)$, is the mean of all the relative position shifts of all the words in the set $\mathcal{C}_{(s1,s2)}$, namely,

$$\sigma_{\text{pos}}(s1, s2) = \frac{1}{N_{\mathcal{C}}} \sum_{W \in \mathcal{C}} \max\{\delta_{s1,s2}(W), \delta_{s2,s1}(W)\}. \quad (3)$$

Below are additional examples of the WPD computation on paraphrases in the MRPC dataset. To aid visualization of what the metric measures, the common words are underlined and coloured to aid comparison.

4.4 Lexical Deviation (LD)

We propose *lexical deviation* (LD), a metric that effectively captures the degree of deviation in the vocabulary used between the sentences in a paraphrase pair. We find that the proposed metric is effective in identifying and ranking paraphrase pairs

from various datasets according to meaningful differences in their usage of lexical changes to perform paraphrasing.

Definition 4.6 (Lexical Deviation)**.** Let (s_1, s_2) be a paraphrase pair. The lexical deviation $\sigma_{\text{lex}}(s_1, s_2)$ for a paraphrase pair (s_1, s_2) is defined by

$$\sigma_{\text{lex}}(s_1, s_2) = 1 - \frac{N_{\mathcal{C}}}{N_{\mathcal{A}}}. \tag{4}$$

For a case where there is complete reuse of words (in other words, $N_{\mathcal{C}} = N_{\mathcal{A}}$), the metric will compute to 0.0. Likewise, in a case where there is no reuse of words, the metric computes to 1.0.

For the purpose of computing the total set of words and the set of common words, we consider words that are the same after lemmatization (ignoring capitalization) to be the same word. Therefore, we do not consider words that are of different forms (e.g. tense) and capitalization to be different words. This allows our metric to more accurately capture the range of vocabulary used. As word forms tend to vary when used as part of different sentence structures, we do not wish to capture that in this metric, which focuses on the diversity of vocabulary (using of different words), and not the grammatical usage of a word. In addition, we consider changes in capitalization a trivial paraphrase, and hence do not consider it in this metric.

5 Application of Metrics

To demonstrate the applicability of our proposed metrics of WPD and LD, we compare them against other metrics with similar purposes: ROGUE-L (Lin, 2004), SELF-BLEU (Zhu et al., 2018) and Damerau–Levenshtein edit distance (Levenshtein, 1965). In the examples below, we show that with WPD and LD, we can effectively distinguish between different types of paraphrases that have similar scores via various other metrics.

S1: A conviction could bring a maximum penalty of 10 years in prison and a $250,000 fine. S2: If convicted, he faces a maximum penalty of 10 years in prison and a $250,000 fine.		
WPD: 0.03	LD: 0.39	ROUGE-L: 0.76
S1: The top rate will go to 4.45 percent for all residents with taxable incomes above $500,000. S2: For residents with incomes above $500,000, the income-tax rate will increase to 4.45 percent.		
WPD: 0.50	LD: 0.33	ROUGE-L: 0.75

Figure 3: Example Pair 1

In Example Pair 1 (Figure 3), we show that two paraphrases can have very similar ROGUE-L scores of 0.76 and 0.75, where ROUGE-L primarily measures the degree of sub-string similarity (longest common sub-strings). However, with WPD, we are able to additionally distinguish the degree in which the similar sub-strings have been shuffled in position, which is a structural alteration to the sentence.

S1: However, prosecutors have declined to take criminal action against guards, though Fine said his inquiry is not finished. S2: Prosecutors have declined to take criminal action against corrections officers, although Fine said his inquiry was not finished.		
WPD: 0.06	LD: 0.29	SELF-BLEU: 0.56
S1: In trading on the New York Stock Exchange, Kraft shares fell 25 cents to close at $32.30. S2: Kraft's shares fell 25 cents to close at $32.30 yesterday on the New York Stock Exchange.		
WPD: 0.44	LD: 0.21	SELF-BLEU: 0.57

Figure 4: Example Pair 2

In Example Pair 2 (Figure 4), we again show that two paraphrases can have very similar SELF-BLEU scores of 0.60 and 0.59, where SELF-BLEU primarily measures the degree n-gram overlap. However, similar to Example Pair 1, in one of the paraphrases, the two "halves" of the sentence has been swapped in position, and this structural alteration is captured by the WPD score.

S1: An attempt last month in the Senate to keep the fund open for another year fell flat. S2: An attempt to keep the fund open for another year fell flat in the Senate last month.		
WPD: 0.31	LD: 0.00	Edit Distance: 0.59
S1: Prisoners were tortured and executed -- their ears and scalps severed for souvenirs. S2: They frequently tortured and shot prisoners, severing ears and scalps for souvenirs.		
WPD: 0.09	**LD: 0.42**	Edit Distance: 0.51

Figure 5: Example Pair 3

Lastly, in Example Pair 3 (Figure 5), we show that two paraphrases can have very similar Damerau–Levenshtein edit distance, but feature two completely different types of paraphrasing method.

5.1 Comparing MRPC and PAWS

5.1.1 Degree of Structural Paraphrasing

Using WPD, we are able to obtain an aggregate view of both the MRPC and PAWS datasets. We see that both datasets feature similar distributions of structural paraphrasing, where the average amount of structural paraphrasing is fairly low and MRPC features more structural paraphrasing compared to

PAWS. A visualization is provided in Figure 6 below. Hence, we would expect the MRPC dataset to be somewhat more diverse in structural paraphrases as compared to PAWS.

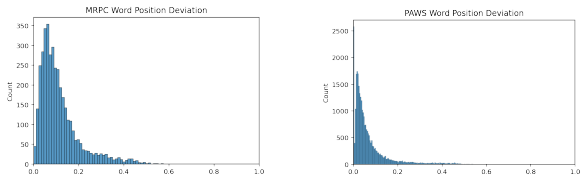

Figure 6: Visualization of WPD in MRPC and PAWS.

5.1.2 Degree of Lexical Paraphrasing

Using LD, we are able to obtain an aggregate view of both the MRPC and PAWS datasets to see that both datasets feature a very different distribution of lexical paraphrasing. A visualization is provided in Figure 7 below. MRPC features a large amount of lexical paraphrasing, in contrast to PAWS where lexical paraphrasing is almost absent. Hence, we would expect the MRPC dataset to be substantially more diverse in having different examples of lexical paraphrases as compared to PAWS.

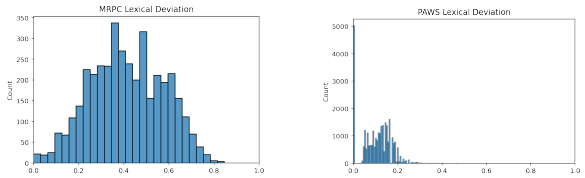

Figure 7: Visualization of LD in MRPC and PAWS.

We investigated the source of high LD in MRPC and determined that the reason is due to large inconsistencies in entities, such as named entities and quantities, present in MRPC paraphrase pairs. We can see that many of the examples at the high-end of lexical deviation are not reasonable paraphrases of each other as they contain extremely different information in each sentence.

s1:	But Secretary of State Colin Powell brushed off ***this possibility Wednesday***.
s2:	Secretary of State Colin Powell ***last week*** ruled out ***a non-aggression treaty***.
s1:	October gasoline prices settled ***1.47 cents*** lower at ***78.70 cents*** a gallon.
s2:	October heating oil ended down ***0.41 cent*** to ***70.74 cents*** a gallon.

Figure 8: Some problematic sentence pairs from MRPC that are not reasonable paraphrases.

When used as training data for paraphrase identification or generation tasks, this can introduce undesired behaviour into models. For example, this can make paraphrase generation models more prone to "hallucinating" additional information in paraphrases, while paraphrase identification models are less able to detect such inconsistencies.

Hence, this motivates us to more closely inspect the quality and consistency of labels in the MRPC dataset, and then propose improvements.

5.2 Evaluation of MRPC Label Quality

Despite its wide usage as a benchmark for paraphrase identification, the labels in the MRPC dataset are not of a consistently high quality. This is a result of the annotation process used to create the MRPC dataset.

The annotation process used for MRPC, as described in the paper *Automatically constructing a corpus of sentential paraphrases* (Dolan and Brockett, 2005), is as follows: a collection of news articles is collected from the web over a 2-year period, and candidates for paraphrase pairs are extracted using automated approaches, followed by human evaluation used to determine if two similar sentences are paraphrases. However, the instructions given to the human annotators of the pairs were "ill-defined". Compounding the issue is that several classes of named entities in the text were replaced by generic tags, introducing large amounts of ambiguity. As a result, the annotators labelled sentences with very inconsistent entities as valid paraphrases, leading to a relatively large number of sentences inside that are not in fact reasonable paraphrases, despite being labelled as such. Thus, models that perform well on MRPC may not able to correctly identify paraphrases in a precise manner. We can show this in Section 5.2.2, where a state-of-the-art language model that performs well on MRPC has nearly random performance on PAWS, despite both being paraphrase identification datasets.

To illustrate this issue, we use an example of a sentence pair, labelled as a paraphrase, from the MRPC dataset:

1. The stock rose $2.11, or about 11 percent, to close Friday at $21.51 on the New York Stock Exchange.

2. PG&E Corp. shares jumped $1.63 or 8 percent to $21.03 on the New York Stock Exchange on Friday.

In this example, which is labelled as a paraphrase-pair, there are a total of 9 entities across the paraphrase pair, but only 2 ("the New York

Stock Exchange" and "Friday") are present across the two. In other words, there is a great inconsistency in the entities present between each of the paraphrase pairs. In this case, this results in a large discrepancy in the information contained in each sentence, and thus the two sentences are not in fact paraphrases despite being labelled as such in MRPC. In MRPC, there are a total of 3900 paraphrase pairs. Of those, 3016 (77%) have at least 1 inconsistent entity. Thus, this is a common issue in MRPC.

5.2.1 Proposed Amendments to MRPC

With the aim to improve the precision of sentence pairs labelled as paraphrases in MRPC, we proposed some amendments to MRPC, including the following specific objectives:

1. Automatically correcting the inconsistency in entities;

2. Rectifying the labels where automated correction is not possible.

Our process to achieve this has two main steps.

First, we search for inconsistent examples where the inconsistency is limited to singular instances of any type of quantity. For example, one instance of a monetary value that differs between two sentences in a paraphrase pair.

Next, when a match is found, we proceed with a to correct the paraphrase. In this specific scenario, as we know that both values share the same type, we can correct one of the values to be identical to the instance in the other sentence, making it a more precise paraphrase. In order to avoid being overly zealous in this replacement, we inspect the most frequent replacements to ensure that no unintended replacements occur.

Of the 3016 inconsistent paraphrase pairs in MRPC, 476 (16%) can be corrected using our approach. For the rest of the paraphrase pairs that we cannot correct, we label them as non-paraphrases. After the corrections, 2064 (53%) out of the original 3900 paraphrase pairs are relabelled as non-paraphrases. This also changes the ratio of paraphrase:non-paraphrase in MRPC from approximately 8:5 to approximately 4:8. We term this revised version of MRPC as MRPC-R1.

To illustrate the corrections to text performed during the creation of MRPC-R1, a few examples are shown in the table below:

s1: The findings are being published ***today*** in the Annals of Internal Medicine
s2: The findings are published in the ***July 1st*** issue of the Annals of Internal Medicine.

Correct **DATE** to either **today** or **July 1st**

s1: American has laid off ***6,500*** of its flight attendants since ***Dec. 31***.
s2: Since ***October 2001***, American has laid off ***6,149*** flight attendants.

Correct **CARDINAL** to either **6,500** or **6,149**
Correct **DATE** to either **Dec. 31** or **October 2001**

Figure 9: Correcting some examples from MRPC

5.2.2 Evaluating Changes to MRPC

In order to evaluate the differences in quality of the datasets, we compare the transferability of a model trained on MRPC and MRPC-R1 to the PAWS test set.

Our training setup is as follows: We used a state-of-the-art DeBERTa (He et al., 2021) pretrained langauage model and fine-tuned it on each of the following: MRPC training set, MRPC-R1 training set, and lastly for a baseline, the PAWS training set). We performed the training using the HuggingFace Transformers library (Wolf et al., 2020) and PyTorch (Paszke et al., 2019), learning rate of 1e-5, and the Adam optimizer (Kingma and Ba, 2015). For MRPC and MRPC-R1, we use a batch size of 32, and for PAWS, which has a much larger training set, we use a batch size of 128. We did not perform extensive hyper-parameter tuning. We tested two variations of the DeBERTa model: DeBERTa-base (140M parameters) and DeBERTa-large (400M parameters) Each of the models are evaluated every 50 steps on the PAWS development set, and the best model checkpoint is evaluated against the PAWS test set. We report the results below (median from 5 runs).

Model	Training Data	Dev F1	Test F1
DeBERTa-base	PAWS (baseline)	92.77	91.98
DeBERTa-large	PAWS (baseline)	95.12	94.32
DeBERTa-base	MRPC	35.67	35.07
DeBERTa-large	MRPC	30.80	30.80
DeBERTa-base	MRPC-R1	52.25	50.83
DeBERTa-large	MRPC-R1	**56.04**	**55.22**

From our results, we can see that training on MRPC-R1 results in much better scores on the PAWS test set for both models. Additionally, if we use the more powerful DeBERTa-large model, the model overfits more on MRPC training data. Thus, DeBERTa-large scores lower than DeBERTa-base on the PAWS test set. However, DeBERTa-large performs better than DeBERTa-base when trained on MRPC-R1, showing that more powerful

models benefit more from MRPC-R1. Thus, we can see that that MRPC-R1 has greater transferability to the PAWS test set. These results demonstrate that we have increased the generalization ability of the trained model through the improving the consistency and quality of the labels in MRPC.

5.3 Evaluation and Filtering for Paraphrase Generation

To demonstrate the applicability of our metrics to filter and thus control the output from a paraphrase generation model, we combine the paraphrase pairs from MRPC-R1 and PAWS to form a corpus to train a sequence-to-sequence T5 (Raffel et al., 2020) transformer language model to generate paraphrases. We performed the training using the HuggingFace Transformers library and PyTorch, using the the pretrained T5-large model (770M parameters). We performed training for a total of 10 epochs with a batch size of 16, learning rate of 1e-5, the Adam optimizer and did not perform extensive hyper-parameter tuning. By using WPD and LD, we are able to effectively filter for specific types of paraphrases.

In the following example, we pass "*I keep a glass of water next to my bed when I sleep.*" as an input to be paraphrased by the model. Some of the outputs are sampled and ranked below according to WPD, showing how WPD can be used to select paraphrases with varying extents of structural paraphrases, and the results can be seen in the table below:

Generated Paraphrase	**WPD**
I keep a glass of water beside my bed when I sleep.	0.02
A glass of water is kept next to my bed when I sleep.	0.10
When I sleep, I always keep a glass of water near my bed.	0.37

We can also do the same for LD, where we can see that the lower the the extent of word overlap between the original and paraphrase, the greater the LD value. Words are marked with *italics* to visually indicate words that have changed from the source sentence. The results can be seen in the table below:

Generated Paraphrase	**LD**
When I sleep I keep a glass of water next to my bed.	0.00
I keep a glass of water *beside* my bed when I sleep.	0.23
During the night, I keep a glass of water next to my bed.	0.33

Thus, we can use WPD, LD, or a combination of both to select specific types of paraphrases, therefore efficiently obtaining specific variations of data for data augmentation or robustness testing purposes.

6 Ethical Considerations

To the best of our knowledge, we do not introduce any ethical concerns in this work. Our work is based on the existing MRPC and PAWS datasets, which are sampled from online news articles as well as Wikipedia. Hence we expect our findings to generalize well to other English datasets in the general domain. Generalization of our work to domains where usage of language is markedly different (for example, in some forms of technical writing) is not certain. When our proposed metrics are used in conjunction with other technology (such as large generative language models), it does not affect the existing ethical considerations of using those technology.

7 Conclusions and Future Work

In our paper, we have proposed two new metrics to better understand paraphrase pairs: word position deviation (WPD) and lexical deviation (LD). We have applied these metrics to better understand the MRPC and PAWS datasets, and also to filter the output of a paraphrase generation model to obtain specific forms of paraphrases. However, our metrics still have some limitations, which can be address in future work. Although we are able to measure the extent of structural and lexical alterations, we cannot determine the fine-grained type of alterations that is being made, for example, a specific form of structural alteration or word substitution. We anticipate that improvements in this area would be valuable to improve our ability to effectively characterize various properties of paraphrases, leading to better data augmentation and robustness testing approaches that eventually resulting in better performing NLP systems.

References

I. Androutsopoulos and P. Malakasiotis. 2010. A survey of paraphrasing and textual entailment methods. *Journal of Artificial Intelligence Research*, 38:135–187.

Hannah Chen, Yangfeng Ji, and David Evans. 2020. Pointwise paraphrase appraisal is potentially problematic.

Jiaao Chen, Derek Tam, Colin Raffel, Mohit Bansal, and Diyi Yang. 2021. An empirical survey of data augmentation for limited data learning in nlp.

William B Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 968–988, Online. Association for Computational Linguistics.

Matt Gardner, William Merrill, Jesse Dodge, Matthew E. Peters, Alexis Ross, Sameer Singh, and Noah Smith. 2021. Competency problems: On finding and removing artifacts in language data.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention.

Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md. Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. 2017. Deep learning scaling is predictable, empirically.

Kuan-Hao Huang and Kai-Wei Chang. 2021. Generating syntactically controlled paraphrases without using annotated parallel pairs.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

V. Levenshtein. 1965. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics. Doklady*, 10:707–710.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Simone Magnolini. 2014. A survey on paraphrase recognition. In *DWAI@AI*IA*.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Vasile Rus, Rajendra Banjade, and Mihai Lintean. 2014. On paraphrase identification corpora. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2422–2429, Reykjavik, Iceland. European Language Resources Association (ELRA).

Ivan Sanchez, Jeff Mitchell, and Sebastian Riedel. 2018. Behavior analysis of NLI models: Uncovering the influence of three factors on robustness. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1975–1985, New Orleans, Louisiana. Association for Computational Linguistics.

Lifu Tu, Garima Lalwani, Spandana Gella, and He He. 2020. An empirical study on robustness to spurious correlations using pre-trained language models.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. Glue: A multi-task benchmark and analysis platform for natural language understanding.

Sam Witteveen and Martin Andrews. 2019. Paraphrasing with large language models. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 215–220, Hong Kong. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen,

Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Lang Yu and Allyson Ettinger. 2021. On the interplay between fine-tuning and composition in transformers.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. Paws: Paraphrase adversaries from word scrambling.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models.

A Appendix

A.1 Models Checkpoints, Data, Hardware

This section lists down the specific pretrained model checkpoints and data used for various purposes in this paper.

- Lemmatization (used for WPD, LD computation): SpaCy Lemmatizer, which uses the spacy-lookups-data package.
- Named Entity Recognition (used for MRPC dataset corrections): SpaCy en_core_web_trf model.
- Paraphrase classification models: Fine-tuned from the microsoft/deberta-base and microsoft/deberta-large checkpoints.
- Paraphrase generation model: Fine-tuned from the t5-large checkpoint.

For model fine-tuning, a single RTX 3090 was used. Automatic mixed precision and TF32 is enabled.

A.2 Code Implementation

The code relevant to this paper can be found in this GitHub repository: `https://github.com/tlkh/paraphrase-metrics`

A.3 Additional Examples from MRPC

This section contains additional examples of WPD and LD applied to data from the MRPC training set.

s1: However, the talk was downplayed by PBL which said it would focus only on smaller purchases that were immediately earnings and cash flow-accretive. *s2:* The talk, however, has been downplayed by PBL which said it would focus only on smaller purchases that were immediately earnings and cash flow-accretive.	
WPD: 0.04	LD: 0.04

s1: With an estimated net worth of $1.7 billion, Mrs. Kroc ranked No. 121 on Forbes magazine's latest list of the nation's wealthiest people. *s2:* Kroc ranked No. 121 on Forbes magazine's latest list of the nation's wealthiest people, with an estimated net worth of $1.7 billion.	
WPD: 0.45	LD: 0.04

s1: As a result, Nelson now faces up to a 10 year jail term instead of life. *s2:* The verdict means Nelson faces up to 10 years in prison rather than a life sentence.	
WPD: 0.14	LD: 0.65

s1: Federal Emergency Management Administration designated $20 million to establish the registry. *s2:* The registry was launched with $20 million from the Federal Emergency Management Agency.	
WPD: 0.41	LD: 0.56

A.4 Additional Examples from PAWS

This section contains additional examples of WPD and LD applied to data from the PAWS training set.

s1: Brockton is approximately 25 miles northeast of Providence, Rhode Island, and 30 miles south of Boston. *s2:* Brockton is located approximately 25 miles northeast of Providence, Rhode Island and 30 miles south of Boston.	
WPD: 0.03	LD: 0.06

s1: Wollstonecraft arrived in Grenada on board the ship 'Sydney' on 31 August 1819. *s2:* Wollstonecraft arrived on August 31, 1819 on board the ship 'Sydney' in Grenada.	
WPD: 0.30	LD: 0.00

s1: Based on the city of Baltimore, only mentioned, never visited in the show. *s2:* Based on the city of Baltimore, only mentioned, has never visited in the show.	
WPD: 0.02	LD: 0.08

s1: The dividends have increased the total return on the average equity to double, approximately 3.2 %. *s2:* The dividends increased the real 'total return' of the average equity to double, about 3.2 % .	
WPD: 0.03	LD: 0.38