Another Look At Paraphrase Identification

Anonymous ACL submission

Abstract

In this paper, we take an updated look at the paraphrase identification task. We analyze commonly used English-language datasets such as MRPC, PAWS, and QQP. We study usage levels of these datasets, showing that dataset usage is heavily skewed towards MRPC. We also study and compare qualitative and quantitative characteristics of the datasets. We investigate the generalization performance of modern models trained on these datasets, showing that models do not generalize well across datasets. Lastly, we demonstrate methods to improve the generalization performance of models, showing that improved label consistency and MNLI pre-training are useful.

1 Introduction

Understanding paraphrasing and the related phenomenon is a foundational aspect of natural language understanding. In natural language, the same semantic meaning can often be conveyed using a variety of expressions, while similar expressions can convey different meanings. In education, students and learners are often encouraged to paraphrase ideas to test and reinforce the accuracy and completeness of their understanding (Hirvela and Du, 2013). Natural language processing (NLP) systems also need to handle paraphrases to achieve robust real-world performance. This has not been achieved even by cutting-edge NLP systems such as ChatGPT (OpenAI, 2022), publicly noted by its authors to be sensitive to input phrasing.

Paraphrase Identification is the task of determining if a pair of sentences are paraphrases of each other. Such a paraphrase identification system has many downstream applications where recognizing equivalent texts is important. For example, we may be required to evaluate if two generated textual summaries of a document are semantically equivalent, and not merely similar.

To identify paraphrases, a typical approach is to train a classifier model on a paraphrase identification dataset. Due to the high intrinsic performance of recent state-of-the-art NLP models, the community is adopting an increasingly data-centric view of how to improve performance on various NLP tasks. Thus, we would like to take an updated and closer look at datasets used to train such models for the paraphrase identification task and how they can be used more effectively.

In this paper, we will analyze some contemporary English language paraphrase identification datasets: Microsoft Research Paraphrase Corpus (MRPC), Quora Question Pairs (QQP) and Paraphrase Adversaries from Word Scrambling (PAWS). We look at the usage statistics of these datasets in the research community, showing how dataset usage is heavily skewed towards MRPC for the past decade. We then analyze the qualitative and quantitative characteristics of these datasets, showing similarities and differences between the datasets. We will also investigate the often poor generalization performance of models trained on the datasets and analyze some trends in the mistakes made by these models. Finally, we also investigate methods to improve the effectiveness of models trained on current paraphrase identification datasets, showing that we can improve generalization performance, without needing larger models or datasets, by improving improve label consistency and using models pre-trained on the MNLI task.

2 Related Work

There have been prior survey papers on the subject of paraphrase identification. However, many of them were published before 2020. Thus, they do not capture much of the modern NLP trends and the large increase in research activity after these previous survey papers had been written. In addition, most of the recent work focuses on surveying modelling approaches instead of datasets used.

In *On Paraphrase Identification Corpora* (Rus et al., 2014), the authors analyzed some paraphrase identification datasets. The two largest paraphrase identification datasets (consisting of paraphrase and non-paraphrase pairs) were MRPC and SemEval-2013 Task 7 Student Response analysis (SRA) (Dzikovska et al., 2015), of which SRA is no longer being used in a contemporary context. The authors made recommendations targeted at advancing our understanding of what a paraphrase is and developing future paraphrase datasets. We note that many of the recommendations have not been further explored, such as creating more precise definitions for paraphrases and unified annotation guidelines for consistent labelling of datasets.

In other survey papers, it is common to find a large focus on studying various modelling approaches. In *A survey on paraphrase recognition* (Magnolini, 2014), the authors focus primarily on studying the effectiveness of various methods of text classification applied to the paraphrase recognition task. Although they analyze some prior proposed definitions of paraphrases and how they are constructed, they do not perform an analysis of datasets, choosing to focus on the effectiveness of various contemporary models on the MRPC task. The model-centric focus is also true for more recent survey papers including *A survey on word embedding techniques and semantic similarity for paraphrase identification* (Kubal and Nimkar, 2019), *Corpus-based paraphrase detection experiments and review* (Vrbanec and Meštrović, 2020), and *Evaluation of state-of-the-art paraphrase identification and its application to automatic plagiarism detection* (Altheneyan and Menai, 2020).

In our paper, we aim to provide a more updated data-centric survey, investigating paraphrase identification datasets and answering different questions, such as how we can improve the way we use the datasets to achieve better generalization performance.

2.1 Paraphrase Identification Task

2.2 Definition

Paraphrase identification is the task of identifying whether a pair of sentences are paraphrases. It is typically a binary classification task guided by the definition of a paraphrase, which will be discussed in greater detail in the next section.

Paraphrase identification can be an isolated task, or used in conjunction with other NLP tasks, such as detecting equivalent questions and answers as part of a question-answering task or detecting equivalent outputs in generative tasks such as translation or summarization.

There is a wide range of approaches for accomplishing this task. However, as with most tasks in NLP, deep learning models, in particular Transformer-based large language models, achieve state-of-the-art performance on various paraphrase identification benchmarks.

2.3 What is a paraphrase?

There is no universally accepted and precise definition of what constitutes a paraphrase. Differing definitions can be obtained from many sources (online sources, dictionaries and publications), and often some ambiguous aspects can differ depending on personal interpretation (Rus et al., 2014). This impacts the usefulness of paraphrase identification datasets as annotation guidelines and annotators' interpretation of those guidelines can vary significantly.

In the NLP research community, several definitions have been proposed:

1. Paraphrasing can be seen as bidirectional textual entailment (Androutsopoulos and Malakasiotis, 2010)

2. Paraphrases are differently worded texts with approximately the same content and have a symmetric relationship (Gold et al., 2019)

3. A sentence is a paraphrase of another sentence if they are not identical but share the same semantic meaning (Liu and Soh, 2022)

In our paper, we prefer the third definition as it captures the most important aspects of paraphrasing: we are looking at two non-identical sentences (different structure and/or different vocabulary) that express the same semantic meaning. However, the definitions are generally in agreement with each other except for the second definition. In this work, we do not consider "*approximately*" equivalent text to be equivalent for the purposes of paraphrase identification, and it introduces an additional aspect of ambiguity and subjectivity, namely how approximate or close enough the meaning has to be in order to be considered a paraphrase.

2.4 Usage Statistics

In our survey, we would like to perform in-depth investigations on the most commonly used datasets.

Hence, in this section, we investigate the usage statistics of major paraphrase identification datasets to identify the key datasets that are in use.

Based on citation counts from Google Scholar, there are 3 major English paraphrase identification datasets in modern use. They are:

1. Microsoft Research Paraphrase Corpus (MRPC) with 1074 citations[1]
2. Paraphrase Adversaries from Word Scrambling (PAWS) with 258 citations[2]
3. Quora Question Pairs (QQP) with 121 citations[3]

Of these three datasets, two datasets, MRPC and QQP, are part of the General Language Understanding Evaluation (GLUE) (Wang et al., 2018) benchmark suite, which itself has 3344 citations[4].

To visualize the usage of these datasets, we provide the following figures. In the figures below, the citation count is adjusted to account for overlaps in citations. For publications that cite both GLUE and another dataset, they are visualized as citing the other dataset instead. Thus, all the GLUE citations in the figure are of publications that **only** cite GLUE. This is because the majority of the publications that only cite GLUE without directly citing another paraphrase identification dataset predominantly do not explore paraphrase identification as a core aspect of the paper, but use the GLUE benchmark in various other ways. Thus, although they include the paraphrase identification task, they do not focus on it. This is in contrast to the papers that cite paraphrase identification datasets directly.

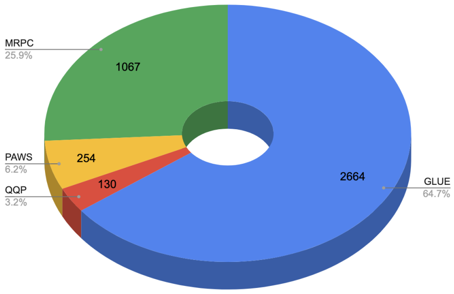


Figure 1: Total citation counts

As we can see in Figure 1 above, the majority of relevant publications cite **only** GLUE. We can also see that predominantly, the research community focuses on the MRPC task, where 25.0% of citations cite MRPC directly, and another 67.2% use MRPC as part of the GLUE benchmark. Thus, there is also a concern that paraphrase identification research is overly focused on MRPC.

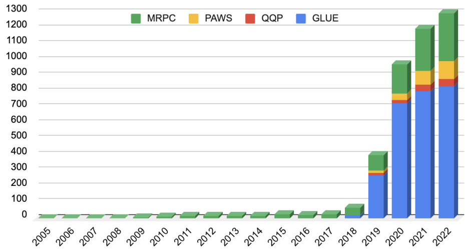


Figure 2: Citation counts per calendar year

In Figure 2, we can visualize the trend of dataset usage over time. We can see MRPC (including usage as part of GLUE) has been consistently a large majority of the usage, even after the introduction of newer datasets like PAWS. In addition, if we use citation counts as a proxy for the amount of research activity, we can see that research increased dramatically in 2020, after the previous survey papers have been published.

2.5 Overview of Datasets

2.5.1 MRPC

The Microsoft Research Paraphrase Corpus (MRPC) (Dolan and Brockett, 2005) dataset contains sentence pairs which were collected from various online news articles. Similar sentences are automatically mined from different articles and labelled by human annotators. Sentences in MRPC are often formal reporting and journalism-style text. This dataset is widely used, both independently and as part of the GLUE benchmark. MRPC contains 4076 training and 1725 test examples, with approximately 50% labelled as paraphrases.

2.5.2 QQP

The Quora Question Pairs (QQP) (Shankar et al., 2017) dataset contains 404,290 question pairs collected from the Quora platform. The questions contain a large variety of different content and textual styles written by social media users, and pairs of questions are labelled by human annotators. Approximately 40% of the data is annotated as a "duplicate" or paraphrase.

[1] View MRPC Google Scholar Page for latest statistics

[2] View PAWS Google Scholar Page for latest statistics

[3] Due to the lack of a officially provided citation, this dataset has been cited in varying ways. We document how we compute the total amount of citations in the Appendix

[4] View GLUE Google Scholar Page for latest statistics

2.5.3 PAWS

The Paraphrase Adversaries from Word Scrambling (PAWS) (Zhang et al., 2019) is a dataset containing sentence pairs extracted from Wikipedia. It consists of procedurally generated sentences created from sentences mined from Wikipedia and labelled by human annotators. The sentences are written factually and in a formal writing style. While it is less commonly used than MRPC, it is high-quality and much larger. PAWS contains approximately 45% paraphrases with 49,401 training, 8000 development and 8000 test examples.

2.6 Dataset differences

Each of the above datasets has different characteristics due to differences in domain, data collection methodology, and data annotation. In the overview, we have already provided some information on the different text domains and data collection methodology. In this section, we will focus on differences in data annotation and other characteristics.

2.6.1 Data annotation

All three datasets follow the same basic structure, where each example consists of a pair of sentences and a binary label indicating if they are a paraphrase. However, there are differences due to the inconsistencies in the annotation guidelines provided to annotators. However, such differences are difficult to quantify.

In MRPC, annotators were instructed to label two sentences as paraphrases if they "mean the same thing", with the interpretation of that instruction being left up to individual annotators. In addition, the "degree of mismatch allowed" before a sentence pair was disqualified as a paraphrase is also left to individual annotators. As such, there is great ambiguity in the labelling of MRPC. Sentences referring to the same subject but containing different information are often labelled as paraphrases, but sometimes not as well. This weakness is acknowledged by the authors of the dataset as well.

To illustrate the problem, we show the following sentence pair, which is labelled as a paraphrase in MRPC:

1. **Scientists** have figured out the complete genetic code of a **virulent pathogen** that has **killed** tens of thousands of California native oaks

2. The **East Bay-based Joint Genome Institute** said Thursday it has unraveled the genetic blueprint for the **diseases** that cause the **sudden death** of oak trees

Despite the clear information mismatch (marked in **bold**) and missing information (marked in red), this is labelled as a paraphrase.

In QQP, we do not have much information on the labelling process. According to the information provided via Quora (Shankar et al., 2017) and Kaggle[5] when the data was released, the question pairs are labelled by human experts, however, the process was acknowledged to be "noisy", with "inherently subjective" labels, and with reasonable possibility for disagreements. However, the authors believe that on a whole, the dataset can "represent a reasonable consensus". In our inspection of the data, we believe that the annotation is indeed done with reasonable consistency, although subjectivity remains.

PAWS has the most rigorous labelling process of all 3 datasets. Each sentence pair is presented to five annotators with an extremely high agreement of above 90% on average. Therefore, we have the highest confidence in the consistency and quality of labelling in PAWS, which is confirmed by our own inspections.

2.6.2 Data characteristics

The combination of different domains, data collection and annotation methods results in differing data characteristics. To quantify these differences, we explore using the metrics proposed in Liu and Soh (2022). We use the word position deviation (WPD) and lexical deviation (LD) metrics to analyze the different characteristics of these datasets.

First, we compute and visualise WPD for each of the datasets: MRPC, QQP and PAWS.

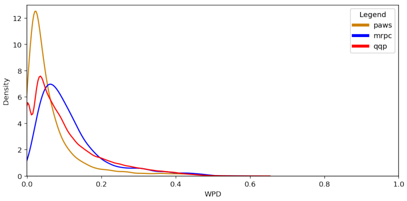


Figure 3: Distribution of WPD in each dataset

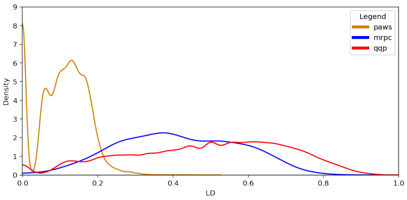


Figure 4: Distribution of LD in each dataset

[5] Kaggle: QQP Dataset Description

From the comparison, we can see that each of the datasets has a remarkably similar distribution of WPD, but PAWS has a different distribution of LD as compared to MRPC and QQP: PAWS has relatively low LD while MRPC and QQP are much higher. By considering the above characteristics, we can come to several preliminary conclusions:

1. We expect the datasets to contain similar levels of structural variations in the paraphrases. Hence, there is limited benefit to combining the datasets in an attempt to increase the diversity of structural paraphrases due to the lack of structural paraphrases in the datasets. Additionally, this also means that for structural paraphrases, all datasets would likely perform similarly.

2. The main difference between the datasets is in terms of the vocabulary, since PAWS has the least amount of LD, followed by QQP and MRPC. Based on what we know of MRPC and PAWS, we can make the following hypothesis that MRPC and PAWS will be challenging in terms of vocabulary, but in different ways. MRPC will be more challenging based on its diversity of vocabulary. On the other hand, PAWS will be more challenging as the classifier cannot rely on recognising similar words, since similar words are present in both paraphrase and non-paraphrase pairs.

3. There is a reasonable chance the much higher LD in MRPC and QQP compared to PAWS is a side effect of a less rigorous annotation process, leading to less semantic equivalence for sentences labelled as paraphrases.

3 Experiments

In this section, we will perform three experiments.

1. We will take each of the three datasets (MRPC, QQP and PAWS) and train a model on them, evaluating on the other two datasets. We term this the **generalization experiment** (Section 3.1). We use this to measure how generalizable each dataset is as a training dataset.

2. We will create a combined version of all three datasets and evaluate a model trained on them on each individual dataset. We term this the **combined dataset experiment** (Section 3.2). We use this to test if combining the datasets is effective in improving the performance of the model.

3. We use the method proposed in *Towards Better Characterization of Paraphrases* (Liu and Soh, 2022) to correct the labelling in MRPC and QQP, and re-run the above experiments to measure the differences when the labelling is made more consistent. We term this as the **rectified dataset experiment** (Section 3.3).

For comparison within our experiments, the main metric of comparison will be the Macro F1 score on the respective test sets, as the different datasets have different proportions of examples labelled as paraphrases. Thus, the Macro F1 score will let us evaluate the datasets more holistically as the score will not be affected by the proportion of paraphrases in the test set. The implementation we use is from the Scikit-learn (Pedregosa et al., 2011) `sklearn.metrics` package.

3.1 Model and Training

For all our experiments, we used a state-of-the-art DeBERTa-Large (He et al., 2020) pre-trained language model, intended for English language sequence classification tasks. We performed the training using the HuggingFace Transformers library and PyTorch. We used a learning rate of 5e-6, the Adam optimizer, a batch size of 128, and training for up to 10 epochs. We use validation scores to select optimal checkpoints for evaluation on the held-out test set.

In addition, we also test the same model that has been fine-tuned on a text entailment task, MNLI (Williams et al., 2018). Some previous works have suggestion such models can perform better on paraphrase identification tasks. In addition, we hypothesize that DeBERTa-Large-MNLI would require less data, and thus perform better on smaller datasets. Thus, we seek to validate if MNLI pre-training would be effective in improving the model's performance and generalization abilities.

3.2 Generalization experiment setup

Each of the three component datasets is separated beforehand into a fixed training, validation and test dataset.

PAWS has a predetermined dataset split for training, validation and test sets. Hence, we use it in our experiments.

MRPC has a predetermined test set but does not have a predetermined validation set. We split the original training set into a training set (90%) and a validation set (10%), keeping the proportion of the labels in the original training set.

QQP does not have a publicly labelled test set, nor does it have a predetermined validation set. We split the original training set into a training set (80%), validation set (10%) and test set(10%), keeping the proportion of the labels constant.

3.3 Combined dataset experiment setup

Since all the datasets follow the same basic structure (a pair of sentences and a binary label), it is a reasonable assumption that these datasets should all be interoperable. For example, we should be able to combine all datasets to create a more effective paraphrase identification dataset. In this experiment, we will test this hypothesis.

In this experiment, we train on all three datasets simultaneously, instead of only training on one dataset. After training, we evaluate on each individual evaluation set. We maintain the train-valid-test splits from the previous experiments.

3.4 Rectified dataset experiments setup

In this experiment, we test the impact of attempting to improve the labeling consistency between the three datasets using the method proposed in Liu and Soh (2022). We run the automated correction on the MRPC and QQP datasets. Following that, we repeat the generalization experiment, as well as the combined dataset experiment, keeping all other factors the same. We will then compare the results between the original datasets and the rectified datasets.

4 Results

4.0.1 Generalization experiment

First, we report the performance of the trained DeBERTa-Large and DeBERTa-Large-MNLI models in terms of the Test Macro F1 score on each of the various datasets. The table indicates which dataset the model is trained on, and how it performs when evaluated on the other datasets.

		Test Macro F1		
Model	**Training**	**MRPC**	**QQP**	**PAWS**
DeBERTa	MRPC	85.53	72.06	32.89
	QQP	67.16	**91.10**	45.49
	PAWS	68.51	76.49	94.83
DeBERTa-MNLI	MRPC	**88.37**	77.41	55.21
	QQP	69.50	90.88	66.40
	PAWS	70.40	79.15	**94.91**

Next, we report some aggregated statistics to show the average performance of each model when trained and evaluated on the same dataset (Same Task), as well as the average performance when evaluated on a different dataset from which it is trained (Other Tasks).

	Test Macro F1 (Mean)	
Model	**Same Task**	**Other Tasks**
DeBERTa	90.42	60.43
DeBERTa-MNLI	**91.39**	**69.69**

Unsurprisingly, the model performs best when it is evaluated on the test set from the dataset it is trained on. When evaluated on another dataset, the average performance suffers significantly.

4.0.2 Combined dataset experiment

	Test Macro F1 (Mean)		
Model	**MRPC**	**QQP**	**PAWS**
DeBERTa	85.46	**91.29**	93.95
DeBERTa-MNLI	**86.44**	91.12	**94.69**

We observe that the overall performance is very similar to when we train and evaluate on a single dataset. This is an interesting result since it shows that training on more data from other datasets did not result in any significant improvement. In fact, it can introduce some slight regression. Green indicates improvement and red indicates regression.

4.0.3 Rectified dataset experiments

First, we report the performance of the trained DeBERTa-Large and DeBERTa-Large-MNLI models in terms of the Test F1 score on each of the various rectified datasets, along with PAWS.

In the table below, we use the following colors to mark the significant changes of **at least 5.0** Test Macro F1 score. Green indicates an improvement and red indicates a regression when compared to training on the original datasets.

		Test Macro F1		
Model	**Training**	**MRPC-R1**	**QQP-R1**	**PAWS**
DeBERTa	MRPC-R1	88.14	75.98	56.10
	QQP-R1	85.46	89.66	61.73
	PAWS	61.41	73.54	94.83
DeBERTa-MNLI	MRPC-R1	**89.38**	78.83	76.62
	QQP-R1	87.87	**89.88**	75.86
	PAWS	68.92	75.58	**94.91**

As can be seen in the table, when evaluated on the same task only, the performance did not change significantly: MRPC shows a slight improvement (mean of 1.81 F1), while QQP shows a slight regression (mean of 1.22 F1). However, there was a significant improvement for 6 out of the 12 generalization experiments, with another 2 showing slight improvements. There was one significant regression (trained on PAWS and evaluated on MRPC-R1), which was an unexpected result.

Overall, the mean Test Macro F1 score increased by 5.89 for the DeBERTa-Large model and 5.06 for the DeBERTa-Large-MNLI model.

Below, we report some aggregated statistics to compare the mean generalization (transfer) performance before and after the dataset rectification, showing the differences in generalization characteristics of the resulting models.

	Test Macro F1 (Transfer)	
Model	**Before**	**After**
DeBERTa	60.43	69.04
DeBERTa-MNLI	**69.69**	**77.28**

We see that the mean Macro Test F1 generalization performance increased by approximately 8.60 F1 for the DeBERTa-Large model and 7.59 F1 for the DeBERTa-Large-MNLI model. This is much higher than the overall increase in performance, since the performance in the individual datasets did not change much.

Lastly, we look at the performance of the model trained on the combined dataset after rectification.

	Test Macro F1 (Mean)		
Model	**MRPC-R1**	**QQP-R1**	**PAWS**
DeBERTa	87.22	**89.88**	93.96
DeBERTa-MNLI	**89.33**	89.83	**94.56**

There was a notable improvement for MRPC-R1 over MRPC (+2.89 F1), a regression for QQP (-1.41 F1) and the PAWS scores remain approximately the same.

5 Analysis

5.1 Generalization experiment results

5.1.1 Effect of MNLI pretraining

We can see that overall, the DeBERTa-Large-MNLI model performs the best. It is better in almost every scenario, including being evaluated on a different dataset from which it is trained. The only exception is when it is trained and evaluated on the QQP dataset, which is an interesting point that warrants further investigation.

The performance benefit of the DeBERTa-Large-MNLI model is evident in the generalization test scenario, where the model is trained and evaluated on different datasets. In particular, the benefit is largest when evaluated on PAWS. The improvement is also relatively large when the model is trained and evaluated on MRPC. We believe that the initial MNLI fine-tuning helps to overcome the small dataset size of MRPC.

In general, we believe that the initial MNLI fine-tuning is very beneficial for paraphrase recognition in general as it is a related task, while also aiding possibly reducing the tendency of the model to over-fit on smaller datasets. In addition, the benefit is the largest for a benchmark PAWS as it requires the highest "precision" in recognising paraphrases, and the level of "precision" is very similar to that involved in the MNLI task.

5.1.2 Trends in generalization performance

We get an interesting observation when considering the trends in generalization performance, where a dataset is trained on one dataset and evaluated on another. The MRPC dataset provides the least generalization performance, likely due to the large amount of inconsistency in annotation combined with the small number of examples. On the other hand, PAWS provides the greatest generalization performance, likely due to its labelling consistency and larger size.

5.1.3 Analysis of Model Mistakes

We are interested to analyze the model's error rates in different categories of paraphrases. Thus, we define three different categories of paraphrases to analyze for the three datasets:

1. sentences are very similar to each other
2. sentences are only different structurally (low LD, high WPD)
3. sentences are only different lexically (high LD, low WPD)

The model trained on MRPC performs best when evaluated on lexically different QQP examples and performs worst on lexically different PAWS examples. This aligns with what we expect from the known characteristics of these datasets.

The model trained on QQP performs best when evaluated on lexically different MRPC examples. This is a reciprocal relationship and falls within our expectations. The model also performs the worst on PAWS examples which are different structurally.

The model trained on PAWS performs best on lexically different QQP examples. It performs worst on MRPC examples which are close, which is expected since it is where the label inconsistency will be the greatest.

The full table of raw results is included in Appendix A.3.

5.2 Combined dataset experiment results

There are two main observations from the results of the combined dataset experiment:

Firstly there is no real benefit to training with a combined dataset. Despite having more ostensibly similar data for the same type of task, MRPC and

PAWS see minor reductions in performance (more significantly for MRPC), while QQP sees a mostly negligible performance improvement.

Secondly, the trend remains that DeBERTa-Large-MNLI is the better performing model, except when evaluated on QQP.

Thus, our main conclusion is that if our goal is to improve performance on a particular dataset, simply adding more data from another dataset does not help, even if they are all paraphrase identification datasets. In fact, there is a slight overall regression even with a powerful classifier. This informs us that each of these datasets is likely "testing" the classifier on slightly different aspects of the tasks, and thus do not generalize well to each other, reinforcing our results from the generalization experiments. The differences in annotation criteria and rigour are also likely to be detrimental. This hypothesis is partially supported by our results where after rectification, the model train on the combined dataset has notably higher performance on MRPC-R1.

5.3 Rectified dataset experiment results

Our first observation is that the dataset rectification results in largely positive improvements. Across all the 12 generalization experiments, 8 see improvements, with 6 having a significant improvement of more than 5.0 F1 on the test set and another 2 experiments showing a slight improvement. 3 experiments show some regression (1.5-3.6 F1), and 1 experiment (train on PAWS and test on MRPC) shows a significant regression of 7.1 F1. However, overall the generalization performance increases by a significant margin.

For the combined experiment, we see that the general trends from the non-rectified experiment carry over. However, there is an improvement for MRPC. Previously the performance dropped for the DeBERTa-Large-MNLI model by 1.93 F1. After rectification, the performance drop is much less at 0.05 F1. We note that an interesting performance trend remains, where QQP performance is slightly better on DeBERTa-Large compared to DeBERTa-Large MNLI. Another interesting observation is QQP performance decreases, but the gap between DeBERTa-Large and DeBERTa-Large-MNLI is greatly reduced. However, the reason for this is unclear. This can be a subject of future study.

In general, the higher-performing DeBERTa-Large-MNLI model improved less during this experiment. We hypothesize that it is likely due to its already higher performance baseline, in addition to the pre-trained MNLI task offsetting some of the benefits of the rectified dataset.

6 Limitations and Future Work

Due to limitations on computing resources and the already large number of existing experiments, we did not do multiple runs of the experiments nor extensive hyper-parameter searches to tune the results for both individual experiments and the entire set of experiments. Instead, we stuck to established hyper-parameters that are known to provide reasonably good results. We also did not repeat the set of experiments across different pre-trained models. We believe that the same trends in results would hold for different combinations of hyper-parameters and pre-trained models, although model performance might vary slightly. In future work, more experiments can be conducted to further validate our results with multiple sets of hyper-parameters and different pre-trained models.

7 Ethical Considerations

To the best of our knowledge, we do not introduce any ethical concerns or risks in this work.

8 Conclusion

In this paper, we took another look at the paraphrase identification task. We looked at usage trends and took a deep dive into commonly used English-language datasets for this task. We highlighted some issues, including inconsistent standards used to label these datasets, as well as interesting similarities and differences in dataset characteristics. We also studied how well models trained on these datasets performed when evaluated on other datasets, showing that generalization performance is relatively low. We conclude that current paraphrase identification datasets have various shortcomings that can be improved with better annotation processes. In addition, we demonstrated that better generalization performance can be achieved by improving labelling consistency and using a model pre-trained on the MNLI task, while other strategies such as combining existing datasets have limited utility.

References

Alaa Altheneyan and Mohamed El Bachir Menai. 2020. Evaluation of state-of-the-art paraphrase identification and its application to automatic plagiarism detection. *International Journal of Pattern Recognition and Artificial Intelligence*, 34(04):2053004.

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

Zihang Chen, Hongbo Zhang, Xiaoji Zhang, and Leqi Zhao. 2017. Quora question pairs.

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

Myroslava Dzikovska, Rodney Nielsen, and Claudia Leacock. 2015. The joint student response analysis and recognizing textual entailment challenge: making sense of student responses in educational applications. *Language Resources and Evaluation*, 50.

Darina Gold, Venelin Kovatchev, and Torsten Zesch. 2019. Annotating and analyzing the interactions between meaning relations. In *Proceedings of the 13th Linguistic Annotation Workshop*, pages 26–36, Florence, Italy. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention.

Alan Hirvela and Qian Du. 2013. “why am i paraphrasing?”: Undergraduate esl writers’ engagement with source-based academic writing and reading. *Journal of English for Academic Purposes*, 12(2):87–98.

Divesh R Kubal and Anant V Nimkar. 2019. A survey on word embedding techniques and semantic similarity for paraphrase identification. *International Journal of Computational Systems Engineering*, 5(1):36–52.

Timothy Liu and De Wen Soh. 2022. Towards better characterization of paraphrases. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8592–8601, Dublin, Ireland. Association for Computational Linguistics.

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

OpenAI. 2022. https://openai.com/blog/chatgpt/.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

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).

Iyer Shankar, Dandekar Nikhil, and Csernai Kornel. 2017. First quora dataset release: question pairs (2017). *URL https://www. quora. com/q/quoradata/First-Quora-Dataset-Release-Question-Pairs*.

Tedo Vrbanec and Ana Meštrović. 2020. Corpus-based paraphrase detection experiments and review. *Information*, 11(5):241.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. 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 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. PAWS: Paraphrase adversaries from word scrambling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1298–1308, Minneapolis, Minnesota. Association for Computational Linguistics.

A Appendix

A.1 Quora Question Pairs citations

A large number of publications (102) cite *Quora question pairs* (Chen et al., 2017). However, this is not correct, since this is not the paper that introduced the QQP dataset, but an early paper that demonstrates some techniques to tackle the dataset. The dataset was first introduced in *First Quora Dataset Release: Question Pairs* (Shankar et al., 2017), which is a blog post on the Data@Quora blog.

Therefore, we aggregate the total number of QQP citations as the sum of citations of the above paper and the blog post, which are referenced with three differing titles. The four Google scholar URLs are as follows:

1. https://scholar.google.com/scholar?cluster=3336862162093221896&hl=en&as_sdt=2005&sciodt=0,5

2. https://scholar.google.com/scholar?cluster=5155042585544784702&hl=en&as_sdt=2005&sciodt=0,5

3. https://scholar.google.com/scholar?cluster=11073074702727464584&hl=en&as_sdt=2005&sciodt=0,5

4. https://scholar.google.com/scholar?cluster=5249091588465214420&hl=en&as_sdt=2005&sciodt=0,5

A.2 Pre-trained Models used

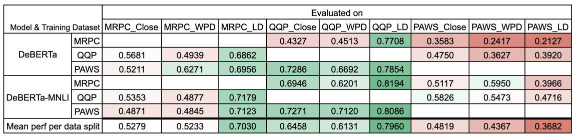
We used two pre-trained models in our experiments.

1. DeBERTa-Large, a 350M-parameter pretrained language by Microsoft proposed in *DeBERTa: Decoding-enhanced BERT with Disentangled Attention* (He et al., 2020). The model is available on the HuggingFace Hub at microsoft/deberta-large.

2. DeBERTa-Large-MNLI, the DeBERTa-Large model fine-tuned on MNLI by Microsoft. The benchmark results are as reported in the DeBERTa paper. The model is available on the HuggingFace Hub at microsoft/deberta-large-mnli.

A.3 Full set of results for Section 5.1.3

In the graphic below, we provide the full visualization of results that accompany the analysis in Section 5.1.3 "Analysis of Model Mistakes". Green indicates higher performance, and red indicates poorer performance.



Model & Training Dataset		Evaluated on								
		MRPC_Close	MRPC_WPD	MRPC_LD	QQP_Close	QQP_WPD	QQP_LD	PAWS_Close	PAWS_WPD	PAWS_LD
DeBERTa	MRPC				0.4327	0.4513	0.7708	0.3583	0.2417	0.2127
	QQP	0.5681	0.4939	0.6862				0.4750	0.3627	0.3920
	PAWS	0.5211	0.6271	0.6956	0.7286	0.6692	0.7854			
DeBERTa-MNLI	MRPC				0.6946	0.6201	0.8194	0.5117	0.5950	0.3966
	QQP	0.5353	0.4877	0.7179				0.5826	0.5473	0.4716
	PAWS	0.4871	0.4845	0.7123	0.7271	0.7120	0.8086			
Mean perf per data split		0.5279	0.5233	0.7030	0.6458	0.6131	0.7960	0.4819	0.4367	0.3682

A.4 Hardware used

All the training was done on an single NVIDIA RTX 3090 with 24GB of VRAM. Training was done in automatic mixed precision mode with mix FP32 and FP16 computations. The total estimated GPU hours taken for the full set of experiments (19 × 2 experiments) is approximately 100 hours.

A.5 Code and Raw Data

After the review period, the code and data will be available publicly on GitHub.