Paraphrase Identification Datasets: Usage Survey and Generalization Patterns

Anonymous ACL submission

Abstract

We perform a survey to identify the most commonly used paraphrase identification datasets. We then look deeper at the top three English datasets containing sentential paraphrases, comparing various qualitative and quantitative characteristics of the datasets. In addition, we investigate the generalization performance of modern models trained on these datasets, showing that models do not generalize well across datasets, showing a weakness in real-world generalisation ability. Lastly, we test some methods to improve generalisation ability, showing that MNLI pre-training and improved label consistency 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 (Kletzien, 2009; 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 examine how they can be employed more effectively.

A variety of different paraphrase identification datasets exist. In Section 3, we look at the usage levels of various openly available datasets, finding that usage is skewed towards the MRPC dataset. In Section 4, we analyse the top high-quality English-language datasets containing sentential paraphrases, the Microsoft Research Paraphrase Corpus (Dolan and Brockett, 2005), Quora Question Pairs (Shankar et al., 2017) and Paraphrase Adversaries from Word Scrambling (Zhang et al., 2019), showing various similarities and differences between them. In addition, in Section 5 we investigate the often poor generalization performance of models trained on the datasets. Lastly, in Section 6, we investigate methods to improve the generalization ability of models trained on current paraphrase identification datasets. We show that we can improve generalization performance, without needing larger models or datasets, by performing MNLI pre-training and enhancing label consistency of the datasets.

2 Related Work

There is some prior work in this area in the form of survey papers. In our paper, we aim to provide a more updated data-centric investigation of the most commonly used paraphrase identification datasets and their efficacy for training modern paraphrase identification models.

In *On Paraphrase Identification Corpora* (Rus et al., 2014), the authors analyzed some paraphrase

identification datasets. The two largest paraphrase identification datasets analyzed 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 several 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).

2.1 Paraphrase Identification Task

2.2 Task 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.

2.3 What is a paraphrase?

There is no universally accepted and precise definition of what constitutes a paraphrase (Zhou et al., 2022). Differing definitions can be obtained from many sources such as online sources, dictionaries and various publications. Often, there can be disagreements on which sentences are paraphrases due to subjective differences in personal opinions (Roig, 2001) or interpretations (Rus et al., 2014). One key element of subjectivity is how "close" or "precise" the meaning of two sentences need to be in order to be considered a paraphrase. This impacts the usefulness of current paraphrase identification datasets as annotation guidelines and annotators' interpretation of those guidelines can vary significantly. Thus, there is a need to have a less subjective framework to more precisely define what is considered a paraphrase.

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.

3 Datasets Survey

3.1 Overview of English-language Datasets

Microsoft Research Paraphrase Corpus (MRPC) The 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.

Quora Question Pairs (QQP) The 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.

Paraphrase Adversaries from Word Scrambling (PAWS) The PAWS dataset ((Zhang et al., 2019)) contains 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.

Paraphrase Database (PPDB) The PPDB dataset proposed in (Ganitkevitch et al., 2013) contains over 220 million paraphrase pairs. Each paraphrase pair contains a set of associated scores including paraphrase probabilities and monolingual distributional similarity scores. Despite its size and variety, this dataset only contains phrasal and lexical paraphrases without any sentence paraphrases. Thus, it is not commonly used as it is not appropriate to be used as training or testing data for sentential paraphrases, which are the dominant type of paraphrases encountered.

Twitter URL The Twitter URL dataset ((Lan et al., 2017)) is constructed by collecting large-scale sentential paraphrases from Twitter by linking tweets through shared URLs. Due to the nature of how the dataset is collected, the text is usually short and of extremely varying qualities. The annotation of the dataset is also noisy when even high-confidence annotations have a large amount of subjectivity.

ParaNMT ParaNMT ((Wieting and Gimpel, 2017)) is a dataset of more than 50 million uncased sentential paraphrase pairs. The pairs were generated automatically by using back-translation to translate the non-English side of a large Czech-English parallel corpus. Due to the relatively low quality of the generated text, this dataset is not suitable to be used without extensive cleaning and post-processing.

TaPaCo TaPaCo ((Scherrer, 2020)) is a paraphrase corpus extracted from the Tatoeba database. Sentences in this database are simple sentences geared towards language learners. The paraphrase corpus is created by populating a graph with Tatoeba sentences and equivalence links between sentences "meaning the same thing". This graph is then traversed to extract sets of paraphrases. A manual evaluation performed on three languages shows that between half and three-quarters of inferred paraphrases are correct and that most remaining ones are either correct but trivial or "near-paraphrases". The corpus contains a total of 200k–250k sentences per language. However, due to its highly simplistic nature and lack of consistent annotation, this dataset is not very useful as well.

3.2 Appropriateness of Image Captioning Datasets

MSCOCO, proposed in (Lin et al., 2014), was originally described as a large-scale object detection dataset. It additionally contains human-annotated captions of over 120K images, and each image is associated with five captions from five different annotators. In most cases, annotators describe the most prominent object or action in an image. MSCOCO's image captioning data is a common source of paraphrase data for tasks such as paraphrase generation. However, in almost all cases, the contents of the captions vary widely with different features of the image described. As such, this dataset is not appropriate for most paraphrasing-related tasks.

3.3 Usage Levels

We use the openly available citation counts from Google Scholar as a proxy for measuring the usage of various paraphrase identification datasets. Another statistic, dataset usage counts on PapersWithCode, are also based on citation counts and exhibit the same trends. However, we did not use the PapersWithCode data as we were not able to obtain the raw data for dataset usage counts. We summarize the statistics that we collected in Table 1 and visualized in Figure 1 (next page).

Dataset	Size	Sentential?	Citations
MRPC	6k	Yes	1624
PPDB	220m	No	945
PAWS	65k	Yes	457
ParaNMT	50m	Uncased	332
QQP	405k	Yes	179
TwitterURL	2.9m	Yes	168
TaPaCo	250k	Yes	47

Table 1: Summary comparison of the major paraphrase datasets

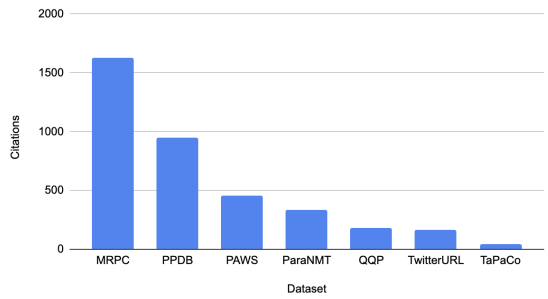

Figure 1: Citation counts of the top datasets containing paraphrases

3.4 Most Notable Datasets

Based on the citation counts (up to end of May 2024), there are 3 major English paraphrase identification datasets in modern use. They are:

1. Microsoft Research Paraphrase Corpus (MRPC) with 1624 citations[1]

2. Paraphrase Adversaries from Word Scrambling (PAWS) with 457 citations[2]

3. Quora Question Pairs (QQP) with 179 citations[3]

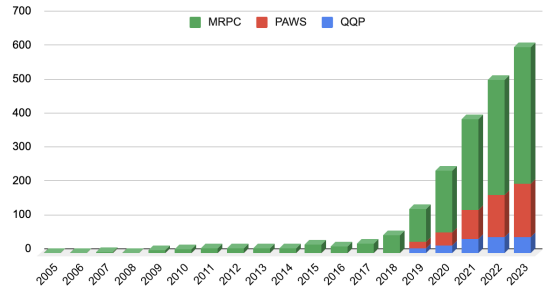

Figure 2: Citation counts per calendar year

[1]View MRPC Google Scholar Page for latest statistics
[2]View PAWS Google Scholar Page for latest statistics
[3]Due to the lack of an officially provided citation, this dataset has been cited in varying ways. We document how we compute the total amount of citations in Appendix A.

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 and a steep increase in research activity.

4 Comparisons between Notable Datasets

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.

4.1 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[4] 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. However, some element of subjectivity can still exist, highlighting the challenge of precise definitions. For example, in the below sentence pair, labelled as a non-paraphrase, it is challenging to outline the differences in meaning, which is visualised in Figure 3.

1. John Barrow Island is a member of the Queen Elizabeth Islands and the Canadian Arctic Archipelago in the territory of Nunavut

2. John Barrow Island is a member of the Canadian Arctic Archipelago and the Queen Elizabeth Islands in the Nunavut area

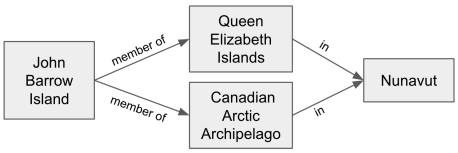

Figure 3: Visualisation of meaning in sentence pair

4.2 Data characteristics

The combination of different domains, data collection and annotation methods results in differing data characteristics. We would like to use quantifiable metrics to analyze the different characteristics of these datasets.

Thus, we explore using two metrics, word position deviation and lexical deviation (Liu and Soh, 2022), for our analysis. Word position deviation (WPD) is a measure of the difference in sentence structure. On the other hand, lexical deviation (LD) measures the difference in the vocabulary used. This allows us to obtain a more holistic view of differences in the sentence pairs.

[4]Kaggle: QQP Dataset Description

First, we compute WPD and LD for each of the datasets: MRPC, QQP and PAWS and visualize them in Figures 4 and 5.

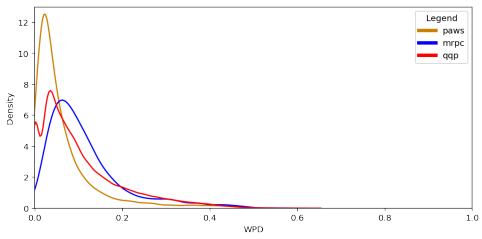

Figure 4: Distribution of WPD in each dataset

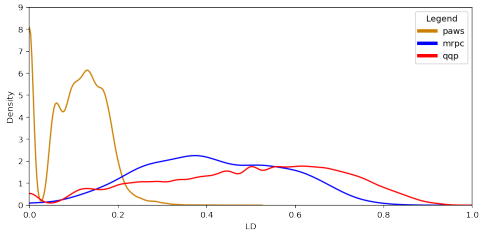

Figure 5: Distribution of LD in each dataset

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.

Firstly, 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.

Next, the main difference between the datasets is in terms of 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.

Lastly, 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.

5 Generalisation Testing

In this section, we will perform experiments to test the generalisation ability of models. Our method of doing so is to train a model on one dataset, and then evaluating on another. For example, we can train a model on the MRPC training dataset and evaluate it on the PAWS test set.

5.1 Experiment Setup

For all our experiments, we used a modern DeBERTa-Large (He et al., 2020) pre-trained language model, with strong performance for English language sequence classification tasks. We performed the training using the HuggingFace Transformers library (Wolf et al., 2020) and PyTorch (Paszke et al., 2019). We used a learning rate of 5e-6, the Adam optimizer (Kingma and Ba, 2017), 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.

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.

Train-test splits Each of the three component datasets is separated beforehand into a fixed training, validation and test dataset. This split is maintained across all the experiments in the paper. PAWS has a predetermined dataset split for training, validation and test sets which we use for 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. Lastly, 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.

5.2 Results

		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

Table 2: Results from generalization experiment

As summarized in Table 2 When trained and evaluated on the same task, the models exhibit very good performance, scoring a range of between 85.53 and 94.83 Macro F1 score on the test set. However, when evaluated on other test sets, the performance drops drastically, falling to between 32.89 to 76.49. In general, the generalization ability of MRPC and QQP is especially poor.

6 Improving Generalisation

In this section, we test three ideas for improving the generalisation performance: performing pretraining in the MNLI task, combining the datasets, and improving labelling consistency in the datasets.

6.1 MNLI pretraining

In this section, we test the same DeBERTa model which has been fine-tuned on a text entailment task, MNLI (Williams et al., 2018) beforehand. Some previous works (Ko and Choi, 2020; Arase and Tsujii, 2021) have suggested that 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 pretraining would be effective in improving the model's performance and generalization abilities on the datasets.

In Table 3 provide a summary table below to show the overall performance with and without the MNLI pretraining. The full set of results is available in Appendix A.3.

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

Table 3: Aggregated results showing the performance difference with and without the MNLI pretraining

6.2 Combining Datasets

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

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 by training on all three datasets simultaneously, instead of only training on one dataset. After training, we evaluate each individual evaluation set. We maintain the existing train-valid-test splits.

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

Table 4: Results from combined dataset experiment

Our results are summarized in Table 4. Green indicates improvement and red indicates regression when compared to training and evaluating on individual datasets. We can make one key observation: Combining datasets does not improve the individual task performances for 2 out of 3 tasks, despite the larger dataset size and increased diversity of data. In fact, there is a slight regression in performance on 2 tasks (MRPC and PAWS), even though the original training data is included. This also leads us to expect the resulting model will continue to generalise poorly when tested on data it is not trained on.

6.3 Improving Label Consistency

We use the method proposed in *Towards Better Characterization of Paraphrases* (Liu and Soh, 2022) to rectify the labelling in MRPC and QQP and re-run the above experiments to measure the differences when the labelling is made more consistent.

In this experiment, we test the impact of improving the labelling consistency between the three datasets using the method proposed in Liu and Soh (2022), running the automated correction procedure 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 and rectified datasets.

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 Table 5, we use the following colours 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**

Table 5: Results from rectified dataset experiment

When evaluated on the same task, the performance did not change significantly: MRPC shows a slight improvement, while QQP shows a slight regression. However, there was a significant improvement (>5.0 F1) for 6 out of the 12 generalization experiments. 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.

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

Table 6: Aggregated results

In Table 6, we report some aggregated statistics to compare the mean generalization (transfer) performance before and after the dataset rectification. 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.

	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**

Table 7: Results using combined rectified dataset

In Table 7, we look at the performance of the model trained on the combined dataset after rectification. There was a notable improvement for MRPC-R1 over MRPC (+2.89 F1), and a regression for QQP (-1.41 F1).

7 Discussion

7.1 Impact of MNLI fine-tuning

In our experiments, we chose to test the DeBERTa-Large model with and without MNLI pre-training, looking at the impact of this factor on task performance.

Overall, the DeBERTa-Large-MNLI model is the better-performing model across most tasks. The model trained on MRPC benefits the most from the MNLI pretraining while exhibiting the weakest original baseline performance. There is one exception where the normal DeBERTa-Large model performs better, which is when the model is trained and evaluated on QQP. Currently, we do not have a hypothesis as to why this is the case. Despite that, this indicates that MNLI pre-training is likely beneficial for improved paraphrase recognition performance. This performance improvement is also consistent even when combined with other approaches, such as combining datasets and improving the label consistency.

7.2 Impact of labelling consistency

One of the key issues we hope to learn more about is the impact of current levels of label consistency in paraphrase classification datasets. Our results show that not only is good label consistency key to having a useful dataset, but it can be more crucial than simply having a larger dataset.

In our experiments, 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 due to its labelling consistency and larger size. Finally, when we attempt to improve the consistency of the labels, we see improvements across 8 out of 12 different experiments.

While having a larger dataset is theoretically useful, there is no benefit if the labelling is not consistent. In our combined dataset experiments, we show that although we can use a larger combined dataset, we see mostly a minor reduction in performance in individual task evaluation. Thus, simply having a larger dataset is not useful.

Our results also highlight the need for a more standardized and less subjective annotation framework for paraphrase recognition tasks. With a better annotation framework, it would be possible to collect more consistent labels to create a larger and more diverse paraphrase corpus that works better than the current approach of combining existing datasets.

8 Limitations and Future Work

Due to limitations on computing resources and the already large number of existing experiments, we only performed our experiments on the DeBERTa-V3-Large model. We believe that the same trends in results would hold for different combinations of hyper-parameters and pretrained large language models, although the exact performance may vary. In future work, more experiments can be conducted to further validate our results with multiple sets of hyper-parameters and different pretrained models.

9 Ethical Considerations

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

10 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 pretrained 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.

Yuki Arase and Junichi Tsujii. 2021. Transfer fine-tuning of bert with phrasal paraphrases. *Computer Speech and Language*, 66:101164.

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.

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. Ppdb: The paraphrase database. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 758–764.

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.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization.

Sharon B Kletzien. 2009. Paraphrasing: An effective comprehension strategy. *The reading teacher*, 63(1):73–77.

Bowon Ko and Ho-Jin Choi. 2020. Twice fine-tuning deep neural networks for paraphrase identification. *Electronics Letters*, 56(9):444–447.

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.

Wuwei Lan, Siyu Qiu, Hua He, and Wei Xu. 2017. A continuously growing dataset of sentential paraphrases. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1224–1234, Copenhagen, Denmark. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

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/.

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 *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

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.

Miguel Roig. 2001. Plagiarism and paraphrasing criteria of college and university professors. *Ethics & behavior*, 11(3):307–323.

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).

Yves Scherrer. 2020. Tapaco: A corpus of sentential paraphrases for 73 languages. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6868–6873.

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.

John Wieting and Kevin Gimpel. 2017. Paranmt-50m: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. *arXiv preprint arXiv:1711.05732*.

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.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi 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 M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. PAWS: Paraphrase Adversaries from Word Scrambling. In *Proc. of NAACL*.

Chao Zhou, Cheng Qiu, and Daniel E. Acuna. 2022. Paraphrase identification with deep learning: A review of datasets and methods.

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 pre-trained 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 MNLI Experiments Results (Section 6.1)

		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**

Table 8: Results from dataset generalization experiment

A.4 Hardware used

All the training was done on a single NVIDIA RTX 3090 with 24GB of VRAM. The training was done in automatic mixed-precision mode with mixed FP32 and FP16 computations. The total estimated GPU hours taken for the full set of experiments (19×2 experiments) is approximately 120 hours.

A.5 Code and Raw Data

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