Pointwise Paraphrase Appraisal is Potentially Problematic

Hannah Chen, Yangfeng Ji, David Evans
Department of Computer Science
University of Virginia
Charlottesville, VA 22904
{yc4dx,yangfeng,evans}@virginia.edu

Abstract

The prevailing approach for training and evaluating paraphrase identification models is constructed as a binary classification problem: the model is given a pair of sentences, and is judged by how accurately it classifies pairs as either paraphrases or non-paraphrases. This pointwise-based evaluation method does not match well the objective of most real world applications, so the goal of our work is to understand how models which perform well under pointwise evaluation may fail in practice and find better methods for evaluating paraphrase identification models. As a first step towards that goal, we show that although the standard way of fine-tuning BERT for paraphrase identification by pairing two sentences as one sequence results in a model with state-of-the-art performance, that model may perform poorly on simple tasks like identifying pairs with two identical sentences. Moreover, we show that these models may even predict a pair of randomly-selected sentences with higher paraphrase score than a pair of identical ones.

1 Introduction

Paraphrase identification is a well-studied sentence pair modeling task that refers to the problem of determining whether two sentences are semantically equivalent. Detecting paraphrases can be very useful for many NLP applications such as machine translation (MT), question answering (QA), and information retrieval (IR). In a QA system, we would like to find the most probable question paraphrases from a database of question answer pairs for a given input question (Rinaldi et al., 2003; Dong et al., 2017). In a MT model, we would like to obtain the best translation by comparing the target sentence to a list of translated sentences. Even though pretrained language models have reached state-of-the-art performance on paraphrase identification tasks, the current problem setup is insufficient to produce models with consistent and robust performance on unseen samples and real world problems.

The typical current problem setup for paraphrase identification is different from intended uses in real world applications. They often involve finding best paraphrases from a group of documents given a particular query, rather than just determining whether two sentences are paraphrases of each other. Besides, getting the order and identifying the most relevant documents is usually more important than getting the binary decision of a pair of sentences (Zuccon et al., 2012). However, to make the task simpler, current methods and existing datasets such as Quora Question Pairs (QQP) (Iyer et al., 2017) and Microsoft Research Paraphrase Corpus (MRPC) (Lan et al., 2017) are all framed as a binary classification problem at the sentence pair level.

Contributions As a first step to improve the way paraphrase identification is evaluated for ranking tasks, we analyze some of the anomalies found in the current pointwise task setting. We first demonstrate the standard way of fine-tuning BERT for pointwise paraphrase evaluation makes the model sometimes fails on simple problems including recognizing two identical sentences and reversing the order of two sentences in a pair (Section 3). We find that it performs worse than a bag-of-words model due to its asymmetrical model architecture. Lastly, we show that the model may fail to capture the correct relative order of two sentence pairs using the pointwise approach, sometimes even predicting a pair of random sentences with a higher paraphrase score than a pair of identical ones (Section 4).

2 Background

This section provides background on the paraphrase identification task, evaluation methods, and the datasets and models we use in our experiments.

2.1 Paraphrase Identification

We consider the general definition of paraphrase as sentences having the same meaning. In addition, paraphrase requires a symmetric relation. Paraphrase identification originates from the real-world applications such as machine translation (Dolan et al., 2004; Quirk et al., 2004) and document summarization (Barzilay and McKeown, 2001), where an essential task is to evaluate the semantic relatedness of translated sentences or generated texts.

2.2 Evaluation Methods

The current problem setting for paraphrase identification is similar to the pointwise method for learning-to-rank problems in information retrieval (Li, 2011). There are three types of approaches to solve learning-to-rank: pointwise, pairwise, and listwise (Liu, 2009). The *pointwise* approach learns to predict a binary relevance judgement for a single document given a specific query. It retrieves the most relevant document by computing the relevance score between each candidate document and the query and returning the document with the maximum score. The *pairwise* approach learns to predict the relative order of a pair of documents, (d_1, d_2), for a given query q. This is closer to the nature of ranking than the pointwise approach. However, both the pointwise and pairwise approaches neglect the fact that some documents are related to the same query. The *listwise* approach directly optimizes the model on the permutations of a list of documents $D = \{d_1, d_2, ..., d_n\}$ (Cao et al., 2007), and hence it most closely matches the objective of ranking.

2.3 Datasets

For our experiments, we use four datasets designed for evaluating paraphrase identification models.

Quora Question Pairs (QQP) consists of 400k question pairs from Quora (Iyer et al., 2017). The goal is to reduce the number of duplicate questions on the platform. Each question pair is either labeled as duplicate or non-duplicate. Recently, it has been shown to have selection bias, where models can simply rely on the frequency of the sentences or the intersection of the neighbor sentences to make predictions (Zhang et al., 2019a).

Paraphrase Adversaries from Word Scrambling (PAWS) contains two datasets constructed from Wikipedia and QQP (Zhang et al., 2019b). To

Dataset	BERT		BOW	
	Acc	F1	Acc	F1
QQP	90.10	86.7	64.75	51.56
QQP+PAWS$_{QQP}$	90.69	87.48	64.13	51.28
MRPC	83.65	87.97	68.12	79.45
Twitter URL	89.98	76.75	84.32	50.44

Table 1: Model accuracy and F1 scores trained on different datasets. Both metrics are scaled by 100. QQP + PAWS$_{QQP}$ indicates models are trained and evaluated on both datasets.

compare with the original QQP dataset, we only tested PAWS$_{QQP}$. The sentence pairs are created by swapping words that have the same part-of-speech or named entity tags to construct higher lexical overlap sentences. The training set contains 11,988 sentence pairs, and the testing set contains 667.

Microsoft Research Paraphrase Corpus (MRPC) contains 5801 sentence pairs extracted from online news articles (Dolan and Brockett, 2005). The sentence pairs are created with very similar syntactic features and high *n*-gram overlap causing the model to make skewed decisions based on these shallow features (Das and Smith, 2009).

Twitter URL Paraphrase Corpus is extracted from tweets posted by 22 English news accounts on Twitter (Lan et al., 2017). Relevant tweets are paired up based on the same embedded URLs, and each pair is then labeled by 6 human annotators. After discarding sentence pairs with neutral decisions (3 out of 6 annotators labeled it as paraphrase), the dataset consists of 42k sentence pairs for training and 9k pairs for testing.

2.4 Models

We fine-tuned the BERT$_{BASE}$ model on different paraphrase datasets with the default configuration (Devlin et al., 2018). We also implemented early stopping during the training process. For baseline comparison, we trained a bag-of-words (BOW) model with unigram and bigram encodings. The model makes predictions based on the consine similarity between the encodings of the two sentences. A consine similarity value above 0.5 is considered a paraphrase. The performance of both models for each task is shown in Table 1.

We include the results for testing QQP model on its adversarial set, PAWS$_{QQP}$, in Table 2, and it shows BERT performing as poorly as BOW of this

Models	$QQP \rightarrow PAWS_{QQP}$	
	Acc	F1
BERT	32.94	42.68
BOW	28.21	44.01

Table 2: Model accuracy and F1 score tested in the adversarial setting, where models are trained on QQP and evaluated on $PAWS_{QQP}$ development set.

dataset. We also report the results of models that trained and tested on a concatenated set of QQP and $PAWS_{QQP}$ in Table 1.

3 Asymmetry

For semantic matching tasks, the BERT paraphrase identification model considers two sentences (s_1, s_2) as a single sequence by concatenating them with a separator token. However, due to this asymmetrical approach, the sequence representations before the final classification layer would be entirely different if we permute the order of the two input sentences. We explore two implications of this method for identifying paraphrases: sensitivity to input order (Section 3.1) and possibility of considering identical sentences non-paraphrases (Section 3.2).

3.1 Sensitivity to Sentence Order

In the original datasets, each sentence pair is only concatenated in one way as (s_1, s_2) and a label y will be predicted by the model. We constructed new sentence pairs in the reverse order as (s_2, s_1), and tested the model on these sentence pairs and got their predicted labels y'. To find out how much it would affect the prediction results, we computed the ratio of sentence pairs that are predicted with a different label ($y \neq y'$). The results for BERT and BOW models are shown in the second and third column of Table 3.

In normal setting (model is trained and evaluated on the the same dataset), there are more than 3% of sentence pairs that are predicted with an opposite label by BERT. The ratio decreases on $PAWS_{QQP}$, but it increases when the model includes adversarial examples in the training data. The percentages are even higher on MRPC and Twitter corpus. BOW, trivially, has zero disagreement since the order does not effect the bag-of-words model.

We reproduced the same experiment in section 3.1 on the RoBERTa$_{BASE}$ model (Liu et al., 2019), and found that the model also has inherent

asymmetry issue as BERT. The ratio of sentence pairs from the QQP development set with opposite labels is around 4.7% (But it performs well on identifying identical sentences with an error rate less than 1%). We further tried fine-tuning BERT on the augmented QQP dataset that includes sentence pairs in both original and reverse order. Although the ratio of sentence pairs with opposite predicted labels decreases about half, the asymmetrical issue is not completely eliminated. These results suggest that these pre-trained language model do not really understand the symmetric relation within paraphrases. One possible reason is combining two sentences as a single input encourages the model to learn paraphrase as an asymmetric relation.

Datasets	Reverse Order		Identical	
	BERT	BOW	BERT	BOW
QQP→QQP	3.70	0.0	2.40	0.0
QQP→PAWS$_{QQP}$	2.66	0.0	7.36	0.0
QQP+PAWS$_{QQP}$	4.0	0.0	0.54	0.0
MRPC	8.46	0.0	0.0	0.0
Twitter URL	7.08	0.0	0.0	0.0

Table 3: The percentage (%) of sentence pairs with asymmetrical prediction results. Reverse Order: sentence pairs predicted with different labels when reversing the order of the sentences. Identical: identical pairs that are predicted as non-paraphrases. (Please see Section 2.3 for actual data sizes.)

3.2 Inability to Recognize Identical Sentences

We would like to know if the asymmetrical structure also affect BERT's ability to identify identical sentences as paraphrases. We collected distinct sentences for each dataset and constructed a new set of sentence pairs by pairing each one with itself. Each pair is labeled as paraphrase. We calculated the ratio of pairs that are predicted as non-paraphrase by the model. As shown in Table 3, BERT trained on QQP recognizes 2.4% of identical pairs as non-paraphrases and the ratio increases about 5% when tested on $PAWS_{QQP}$. BOW trivially achieves perfect accuracy on pairs of identical sentences, since they have exactly the same bags of words.

The models trained on MRPC and Twitter corpus do recognize all the identical pairs as paraphrases. This may be the fact that many sentences appear in Twitter corpus multiple times pairing with different sentence each time. Thus, the model may better capture the difference between a variety of sentences. As for MRPC, many sentence pairs look

	QQP → QQP	QQP → PAWS$_{QQP}$	QQP + PAWS$_{QQP}$	MRPC	Twitter URL
Paraphrase > Identical	30.51	41.88	21.27	4.18	1.76
Avg Score Difference	5.09	3.07	2.58	0.12	0.60
Non-paraphrase > Identical	0.97	43.21	0.41	0.0	0.03
Avg Score Difference	6.28	2.53	3.01	0.0	1.52

Table 4: Percentage of paraphrase and non-paraphrase pairs with higher paraphrase score (%) than a pair of identical sentences given the same query sentence. Avg Score Differences: average score difference between paraphrase/non-paraphrase and identical pairs. (Only pairs with higher scores than the identical ones are included.)

quite alike, and hence the model can better identify small differences between sentences even though most sentences only appear once. Since PAWS$_{QQP}$ contains higher lexical overlap sentence pairs, the model trained on both QQP and PAWS$_{QQP}$ decreases the error rate to less than 1%.

We also fine-tuned a BERT model on the augmented QQP training set with identical sentence pairs, and it can correctly identify every identical pairs as paraphrases. This suggests that the amount of lexical overlap in the dataset would affect the model's ability to identify identical sentences.

4 Problems with Pointwise Evaluation

For a given query sentence, we assume that a well-generalized paraphrase identification model should output a higher paraphrase score to the query sentence itself than a randomly-selected sentence. However, models trained with pointwise evaluation cannot learn the relative order based on the degree of semantic equivalence. We test this by considering how often models recognize a random sentence as more similar than the query sentence itself, and looking at the distribution of paraphrase scores across a dataset.

4.1 Random Sentences

We augmented the original datasets with sentence pairs concatenated in opposite order, as in Section 3.1, and labeled them same as their original pairs. We then compared each sentence pair, (s, s'), to a pair of identical sentences, (s, s), given the same query sentence s. We fine-tuned BERT on each dataset to learn a paraphrase score function f, and computed the fraction of tests where a randomly-selected pair gets a higher paraphrase score than an identical pair, $f(s, s') > f(s, s)$. Table 4 shows the results, revealing a similar pattern as in Section 3.2. The model trained on QQP considers more than 30% of randomly-selected paraphrase sentence pairs to be more similar than the

identical pairs, but the ratio decreases to 21% when adding the adversarial set into training. For MRPC and Twitter URL corpus, less than 5% of paraphrase pairs are considered to be more similar than the identical pairs.

For a randomly-selected sentence pair, (s, s'), and a pair of identical sentences, (s, s), given the same query s sentence, we computed the score difference as $f(s, s') - f(s, s)$. The distributions of the score differences are shown in Figure 1a. We filtered out the pairs that have lower paraphrase score than the identical pairs, and report the average score difference in Table 4. In Figure 1a, the model trained on QQP has the largest score difference between paraphrase and identical pairs. After augmenting the training set with PAWS$_{QQP}$, the right tail of the distribution for paraphrase pairs diminishes. This indicates that the model considers fewer non-identical sentences as more similar to the query sentence than itself.

4.2 Paraphrase Score Distribution

To better understand how the scores are distributed, we plot the histograms of paraphrase score for random, paraphrase, non-paraphrase, and identical sentence pairs in Figure 1b. In the normal setting, there are two peaks in the distributions of randomly-selected pairs since they include both paraphrase and non-paraphrase pairs. On the other hand, the sentence pairs from PAWS$_{QQP}$ all seem very similar to the model. The distributions clearly show the model cannot distinguish them. Compared with the distribution for the Twitter corpus, the distribution of paraphrase pairs from QQP is more spread out, and it has slightly larger gap between the distribution of paraphrase and identical pairs.

5 Discussion

Defining Paraphrases. Our experiments assume that the "best" paraphrase for a given sentence s
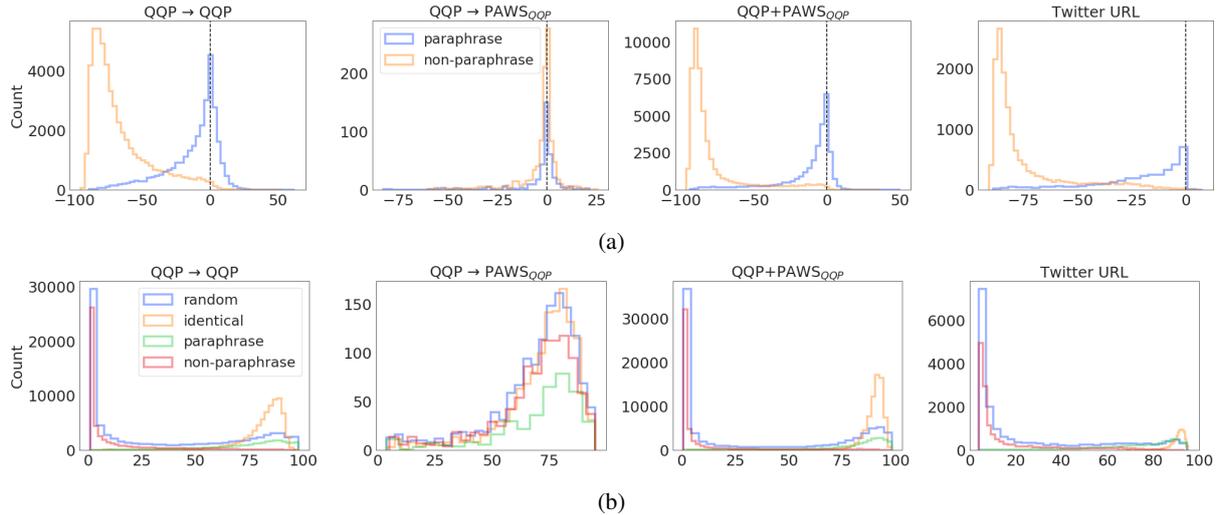
Figure 1: Histograms of (a) the score difference between randomly-selected and identical pairs and (b) paraphrase score for sentence pairs. Randomly-selected pairs contain the sentence pairs their original and reverse order. (We do not include the plot for MRPC since most paraphrase pairs from the dataset look alike, and it is hard to distinguish the distributions from the graph.)

is s itself. This assumes an equivalent in meaning definition of paraphrase, but other definitions may be appropriate. Bhagat and Hovy (2013) defined paraphrases as "sentences that convey the same meaning using different wording". By this definition, identical sentences are not paraphrases. Of course, we do not need a complex model to identify identical sentences when a simple equality test will do. However, when considering paraphrase detection as a test for how well language models can understand meaning, it would be counterproductive to consider identical sentences non-paraphrases, and require a trivial modification to consider them perfect paraphrases. Thus, we would expect a model to be able to identify sentence pairs with the same meaning as paraphrases regardless of whether they are the same in their surface forms.

Our experiments also assume that the paraphrase relationship should be symmetrical. This is consistent with the notion that the paraphrase identification task is meant to identify sentences with similar meaning, but not consistent with the purpose of many uses of paraphrase identification (e.g., in some real world question retrieval tasks, finding questions that contain the query, or that have the opposite meaning, would still be useful). This suggests the importance of a clear notion of what a paraphrase is, in both constructing test datasets and in determining how a given application can use a paraphrase detection model.

Selection Bias in the Pointwise Setting. Previ-

ous studies have addressed the problem of selection bias when constructing the task as a pointwise learning problem (Wang et al., 2016; Zadrozny, 2004). Datasets tend to have inconsistent frequency of sentences causing the model biased towards the dominating sentences. For instance, we found that some sentences from the Twitter corpus are repeated almost a hundred times as the first input sentence. This is part of the reason that the model gets more asymmetrical prediction results for sentences in reverse order (Table 3).

6   Conclusion

Although the state-of-art paraphrase identification models can achieve impressive performance under the pointwise evaluation method, they cannot handle real-world problems and unseen data well and even have worse results than a BOW model on simple tasks. We show that the asymmetry in BERT can produce inconsistent prediction results when reversing the order of the two sentences. We examined the relation of semantic equivalence learned by models trained with pointwise approach, and found that they may consider a random sentence as more similar to the query sentence itself. This suggests future work to reconsider how to match the training and evaluation to the actual objective of downstream applications, and thus create more reliable evaluation metrics and benchmarks.

References

Regina Barzilay and Kathleen McKeown. 2001. Extracting paraphrases from a parallel corpus. In *Proceedings of the 39th annual meeting of the Association for Computational Linguistics*, pages 50–57.

Rahul Bhagat and Eduard Hovy. 2013. What is a paraphrase? *Computational Linguistics*, 39(3):463–472.

Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. 2007. Learning to rank: From pairwise approach to listwise approach. In *Proceedings of the 24th International Conference on Machine Learning*, ICML 07, page 129136, New York, NY, USA. Association for Computing Machinery.

Dipanjan Das and Noah Smith. 2009. Paraphrase identification as probabilistic quasi-synchronous recognition. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 468–476. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Bill Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Third International Workshop on Paraphrasing (IWP2005)*. Asia Federation of Natural Language Processing.

Bill Dolan, Chris Quirk, and Chris Brockett. 2004. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *Proceedings of the 20th International Conference on Computational Linguistics*, COLING 04, page 350es, USA. Association for Computational Linguistics.

Li Dong, Jonathan Mallinson, Siva Reddy, and Mirella Lapata. 2017. Learning to paraphrase for question answering. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 875–886, Copenhagen, Denmark. Association for Computational Linguistics.

Shankar Iyer, Nikhil Dandekar, and Kornl Csernai. 2017. First quora dataset release: Question pairs. https://www.quora.com/q/quoradata/First-Quora-Dataset-Release-Question-Pairs.

Wuwei Lan, Siyu Qiu, Hua He, and Wei Xu. 2017. A continuously growing dataset of sentential paraphrases. *CoRR*, abs/1708.00391.

Hang Li. 2011. A short introduction to learning to rank. *IEICE Transactions*, 94-D:1854–1862.

Tie-Yan Liu. 2009. Learning to rank for information retrieval. *Found. Trends Inf. Retr.*, 3(3):225331.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Chris Quirk, Chris Brockett, and William B Dolan. 2004. Monolingual machine translation for paraphrase generation. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 142–149.

Fabio Rinaldi, James Dowdall, Kaarel Kaljurand, Michael Hess, and Diego Mollá. 2003. Exploiting paraphrases in a question answering system. In *Proceedings of the Second International Workshop on Paraphrasing - Volume 16*, PARAPHRASE 03, page 2532, USA. Association for Computational Linguistics.

Xuanhui Wang, Michael Bendersky, Donald Metzler, and Marc Najork. 2016. Learning to rank with selection bias in personal search. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR 16, page 115124, New York, NY, USA. Association for Computing Machinery.

Bianca Zadrozny. 2004. Learning and evaluating classifiers under sample selection bias. In *Proceedings of the Twenty-First International Conference on Machine Learning*, ICML 04, page 114, New York, NY, USA. Association for Computing Machinery.

Guanhua Zhang, Bing Bai, Jian Liang, Kun Bai, Shiyu Chang, Mo Yu, Conghui Zhu, and Tiejun Zhao. 2019a. Selection bias explorations and debias methods for natural language sentence matching datasets. *CoRR*, abs/1905.06221.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019b. PAWS: paraphrase adversaries from word scrambling. *CoRR*, abs/1904.01130.

Guido Zuccon, Leif Azzopardi, Dell Zhang, and Jun Wang. 2012. Top-k retrieval using facility location analysis. In *Proceedings of the 34th European Conference on Advances in Information Retrieval*, ECIR12, page 305316, Berlin, Heidelberg. Springer-Verlag.