Paraphrase Detection on PAWS Dataset

Shantanu Purandare
spuranda@asu.edu

Vaibhavi Kardale
vkardale@asu.edu

Ishani Bhatt
ibhatt1@asu.edu

Daniel Mathew
dgmathe@asu.edu

Mohammed Sauban Mussaddique
mmussadd@asu.edu

Abstract

The task of paraphrase detection is given a pair of sentences, is one the paraphrase of the other. This is, in essence, a binary classification problem. A Lot of the state-of-the-art models that performed well on other relatively trivial datasets, performed rather poorly on the PAWS dataset, showing nearly 60% drop in accuracy in the case of BERT. We, as a group try to explore the architecture of BERT and how we can make some adaptations to improve the capability of these models. We do this in a twofold way: 1) By adapting the model architecture and hyperparameter finetuning. Through this experiment we hope to achieve better benchmark results on the PAWS dataset. 2) By augmenting the PAWS dataset with more difficult pairs of sentences.

1 Introduction

Paraphrases are a pair of phrases in a language that essentially have the same meaning. There are many datasets and models [Zhang et al.,2019] proposed to automate the identification of a pair of sentences as paraphrases or not. Most of these models lack sentence pairs with high lexical overlap and are not paraphrases. It is found that even the state of the art models performs poorly when categorizing sentence pairs with high Bag of Word(BOW) overlap.[Zhang et al.,2019]

Most of the paraphrase adversary pair datasets draw their negative examples from related sentences that lack a high BOW overlap. Training on such datasets would undersample the training for the pairs with high BOW overlap. Even when there are such pairs with high word overlap, very few tend to be non-paraphrase pairs. As a result, models built on such a dataset would be biased towards categorizing pairs with high BOWs as Paraphrases. The Quora Question Pairs(QQP) dataset [Iyer et al., 2017] contains only 1000 of 400k pairs with high word overlap, of which only 20% are negative examples. This unequal distribution of examples leads to models not having sufficient information to encode the importance of word order.

PAWS (Paraphrase Adversaries from Word Scrambling) [Zhang et al.,2019] addresses this problem by introducing a dataset that has sentence pairs with high word overlap but with equal distribution of negative and positive example pairs. Sentences are pooled from sentences in Quora and Wikipedia.

2 Dataset Description

Most of the datasets lacked either a high bag-of-words(BOW) overlap or non-paraphrase pairs even if there was a BOW overlap. Models trained or evaluated with only this data may not perform well on real world tasks where such sensitivity is important. To address this, Paraphrase Adversaries from Word Scrambling data set is created where challenging pairs are generated by controlled word swapping and back translation, followed by fluency and paraphrase judgments by human raters.

The automatic generation method is based on two ideas. The first swaps words to generate a sentence pair with the same BOW, controlled by a language model. The second uses back translation to generate paraphrases with high BOW overlap but different word order. PAWS effectively measures sensitivity of models to word order and structure.

The PAWS dataset consists of two corpora viz. Quora Question Pairs (QQP) and Wikipedia. The QQP corpus consists of 11,988 training examples and 677 validation and testing examples. The Wikipedia corpus consists of three subsets- 1) PAWS-Wiki Labeled Final 2) PAWS-Wiki Labeled Swap-Only and 3) PAWS-Wiki Unlabeled

Sentence 1	Sentence 2	Label
Can a bad person become good?	Can a good person become bad?	1
Although interchangeable, the body pieces on the 2 cars are not similar.	Although similar, the body parts are not interchangeable on the 2 cars.	0
Katz was born in Sweden in 1947 and moved to New York City at the age of 1.	Katz was born in 1947 in Sweden and moved to New York at the age of one.	1
The team also toured in Australia in 1953	In 1953, the team also toured in Australia	1

Table 1: Paraphrase and Non-paraphrase examples from PAWS dataset.

Final

For phase 1, we have used the **PAWS-Wiki Labeled Final.** It contains pairs that are generated from both word swapping and back translation methods. All pairs have human judgements on both paraphrasing and fluency and they are split into Train/Dev/Test containing 49,401/8000/8000 examples respectively with 44.2% of examples being paraphrases. The training as well as test data sets contain 2 sentences with label 0 and 1, determining if the sentences have different semantic meaning or if the pair is a paraphrase. Table 1 shows examples of Paraphrase and Non-paraphrase pairs generated by the PAWS dataset using the above data generation techniques.

3 Methods/Implementation

3.1 BERT

The Bidirectional Encoder Representations from Transformers is a technique for NLP pre-training developed by Google. It achieved a state-of-the-art performance in 11 NLU tasks including Question Answering and Natural Language Inference. BERT involves pre-training a transformer encoder on a large corpus with over three billion words. It is extremely approachable and can be quickly fine-tuned with ease for variety of tasks.

Data Size	Accuracy(%)	AUC(%)
10,000	83	84
49,401	92	92

Table 2: Effect of training dataset size on Accuracy and AUC scores.

3.2 Experimental Setup

In our experiment we train and evaluate our model on the Paws-Wiki Labeled (Final) dataset. We use **BertModelForSequenceClassification** for this task. It is a BERT model transformer with a sequence classification head on top (a linear layer on top of the pooled output). After hyperparameter fine-tuning, we decided to use the following configurations for BERT: Pre-trained model- 'bert-base-cased' and maximum length of the sequences- 128 (due to GPU memory limitations) and we pad the sequences to maximum length, learning rate- 2e-5. We use two metrics to judge our model's performance: accuracy and AUC scores.

3.3 Results

To analyze how the training data size affects our results we train our model on subsets of the dataset. Table 2 shows the performance of the models with a change in the training dataset size. The model improves by 8% AUC scores using the complete dataset vs a subset of 10k examples. Furthermore, BERT is developed to be trained on large datasets, so it would likely still benefit from more PAWS training examples.

4 Conclusion

The PAWS datasets provide a new resource for training and evaluating paraphrase identifiers. We show using state-of-the-art models like BERT on the PAWS datasets provides a great performance on challenging examples and makes them more robust to real world examples. We also demonstrate that using more examples from the PAWS dataset would likely help us gain a higher accuracy on the paraphrase identification task.

References

[Zhang et al.,2019] Yuan Zhang, Jason Baldridge and Luheng He. 2019. PAWS: Paraphrase Adversaries from Word Scrambling. *arXiv:1904.00130v1*

[Iyer et al.,2017] Shankar Iyer, Nikhil Dandekar, and Kornel Csernai. 2017. *First quora dataset release: Question pairs.*