Paraphrase Identification

Team 4: Mehul Mathur, Manasvi Vaidyula, Sahil Bhatt, Srinath Nair

Github Repo Link -
https://github.com/manasvi-26/Paraphrase-Identification

Introduction

The major project allotted to our team required us to build a paraphrasing system that both detects paraphrased sentence-pairs and also generates paraphrased sentences when an input sentence is fed into the system. As we move to the second phase of the project, the primary aim was to test out various paraphrase identification tasks that have been happening in the research community. The intention was to complete building the paraphrase identification task after selecting a Transformer model that could help with the same.

In this document, we discuss the internal discussions we have had, the work that we have done since the last submission, progress tracking against the timeline we had submitted as part of phase 1, and key decisions made along with the reasoning behind the same.

The main task for this phase was **Paraphrase Identification**

Identifying the Dataset

Effort was put from our side to understand the datasets primarily used in paraphrase identification tasks. The following datasets were under consideration initially:

- Quora Question Pair (QQP)
- PAWS
- Microsoft Research Paraphrase Corpus (MRPC)

According to the PAWS paper -

- The best accuracy for the QQP dataset was - 95.3%
- The best accuracy for the PAWS_QQP dataset was - 89.9%
- The best accuracy for the PAWS_WiKi dataset was - 93.8%

We chose the PAWS dataset for the first round of understanding since PAWS is a much larger dataset.

Identifying a Model to Train

The paraphrase identification task is basically a binary classification task where we determine whether or not a given pair of sentences in the dataset are paraphrases of each other or not. We came across research works that have tried solving the paraphrase identification task using primarily the following transformer models or their modifications:

- BERT
- T5

- XLNet

Our best model was chosen to be **BERT - specifically the bert-base-uncased model**

Experimentation

Attempt 1:

The first attempt was to use a pre-trained T5 model from the Huggingface library, encode the sentences in the sentence pair and check for the cosine similarity. This gave very poor results and the accuracy was as low as 57%.

Method:

- We load the PAWS dataset
- We get a pre-trained **T5 model** and it's tokenizer from HuggingFace transformers library
- We put both sentences in each data point through the encoder in the model and got the sentence vectors from the last hidden state
- We calculated sklearn's cosine similarity on this and used a threshold (of about 0.95 to 0.98) to label it as a paraphrase

Findings

- Accuracy was 57%
- No Fine tuning was done on this
- This focused more on sentence semantic similarity and hence the PAWS dataset was not fully utilized explaining the lower accuracy

Attempt 2

After the first attempts , we decided to incorporate **Fine-Tuning** for our models. This was done using a framework called **Fast AI ,** which basically facilitates Transfer Learning on pre trained models.

Transfer learning is a machine learning method where a model developed for a task is reused as the starting point for a model on a second task. It basically means fine tuning and re training an already pre trained generic model for specific tasks to obtain much better results. The models and datasets used here were:

- PAWS and MRPC with BERT-base-uncased - (93.1% accuracy with PAWS and 82.84% accuracy with MRPC)
- PAWS with XLNET - (90% accuracy)

The Datasets were PAWS and MRPC

PAWS

	id	sentence1	sentence2	label
0	1	In Paris , in October 1560 , he secretly met t...	In October 1560 , he secretly met with the Eng...	0
1	2	The NBA season of 1975 -- 76 was the 30th seas...	The 1975 -- 76 season of the National Basketba...	1
2	3	There are also specific discussions , public p...	There are also public discussions , profile sp...	0
3	4	When comparable rates of flow can be maintaine...	The results are high when comparable flow rate...	1
4	5	It is the seat of Zerendi District in Akmola R...	It is the seat of the district of Zerendi in A...	1
...	...	...	...	...

The **label** column determines whether sentence1 and sentence2 are paraphrases of each other

This dataset had **49401** Training data points , and **2000** Testing data points

MRPC

	Quality	#1 ID	#2 ID	#1 String	#2 String
0	1	702876	702977	Amrozi accused his brother , whom he called " ...	Referring to him as only " the witness " , Amr...
1	0	2108705	2108831	Yucaipa owned Dominick 's before selling the c...	Yucaipa bought Dominick 's in 1995 for $ 693 m...
2	1	1330381	1330521	They had published an advertisement on the Int...	On June 10 , the ship 's owners had published ...
3	0	3344667	3344648	Around 0335 GMT , Tab shares were up 19 cents ...	Tab shares jumped 20 cents , or 4.6 % , to set...
4	1	1236820	1236712	The stock rose $ 2.11 , or about 11 percent , ...	PG & E Corp. shares jumped $ 1.63 or 8 percent...
...	...	...	...	...	...

The **Quality** column determines whether #1 String and #2 String are paraphrases or not

This dataset had **3668** training points and **1725** testing points

Attempt 2a - Bert-base-uncased with PAWS and MRPC:

PAWS Data

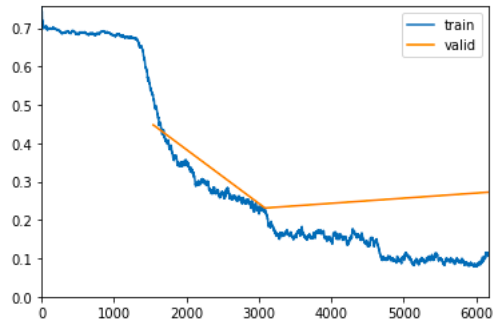
Method:

- Loading the PAWS dataset and pre processing
- Initializing the Learner Fast ai object
- Preparing a Databunch for the learner
- Training for **4** epochs
  - We trained the model on a Colab GPU.
- Saving the model

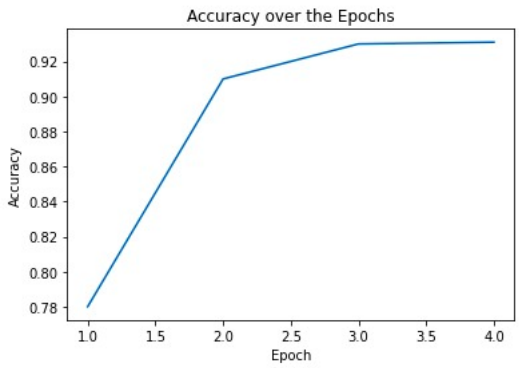
Findings

- Initial accuracy without any fine tuning was **53.6%**
- The accuracy values at each epoch

epoch	train_loss	valid_loss	accuracy
0	0.522695	0.447943	0.788000
1	0.230631	0.231232	0.918500
2	0.150596	0.251808	0.931000
3	0.112169	0.272477	0.931000



Decrease of Training loss and Validation loss over training iterations



- Final accuracy value was - 93.1%. This was more than the paper about models trained on QQP and PAWS which was 87%
- Final F1 score - *92.65%*
- Some sample predictions:

```
SENTENCE 1 -> The exception was between late 2005 and 2009 when he played in Sweden with Carlstad United BK , Serbia with FK Borac Čača
SENTENCE 2 -> The exception was between late 2005 and 2009 , when he played in Sweden with Carlstad United BK , Serbia with FK Borac Ča
PREDICTED =  1 ACTUAL =  1
-------

SENTENCE 1 -> The Tabaci River is a tributary of the River Leurda in Romania .
SENTENCE 2 -> The Leurda River is a tributary of the River Tabaci in Romania .
PREDICTED =  0 ACTUAL =  0
-------

SENTENCE 1 -> He played with the A-level Kane County Cougars in 1993 and the AA Portland Sea Dogs .
SENTENCE 2 -> He played in 1993 with the A - Level Portland Sea Dogs and the AA Kane County Cougars .
PREDICTED =  0 ACTUAL =  0
-------

SENTENCE 1 -> Winarsky is a member of the IEEE , Phi Beta Kappa , the ACM and Sigma Xi .
SENTENCE 2 -> Winarsky is a member of ACM , the IEEE , the Phi Beta Kappa and the Sigma Xi .
PREDICTED =  1 ACTUAL =  1
-------

SENTENCE 1 -> In 1938 he became the government anthropologist of the anglo-Egyptian Sudan and led fieldwork with the Nuba .
SENTENCE 2 -> In 1938 he became the Government Anthropologist of the Egyptian-Anglo Sudan and conducted fieldwork with the Nuba .
PREDICTED =  0 ACTUAL =  0
```

MRPC Data

Method

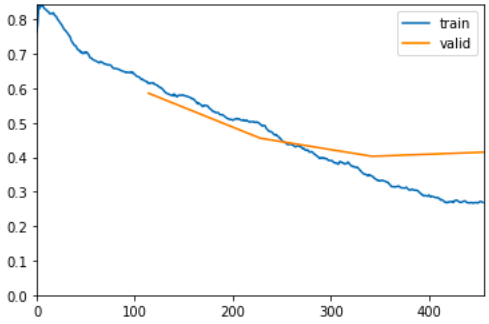
- Loading the MRPC dataset and pre processing

- Initializing the Learner Fast ai object
- Preparing a Databunch for the learner
- Training for **4** epochs
    - We trained the model on a Colab GPU.

Findings

- Initial accuracy without any fine tuning was **33.85%**
- Accuracy at each epoch:

epoch	train_loss	valid_loss	accuracy
0	0.615572	0.585727	0.704348
1	0.498354	0.455152	0.805797
2	0.346862	0.402653	0.822029
3	0.268923	0.414614	0.828406



Decrease of Training loss and Validation loss over training iterations

- Final Accuracy score = 82.84%
- Final F1 score = 87.2%

Attempt 2b - XLNet with PAWS:

In this attempt, the model architecture used was XLNet.

Method:

- A pre-trained XLNet model imported from huggingface library
- preprocessing the PAWS dataset - Created a custom Tokenizer and numericalizer
- Preparing the databunch
- Fast Ai learner: Adam Optimizer imported from huggingface was used.
- Training: The model was trained for a total of 3 epochs

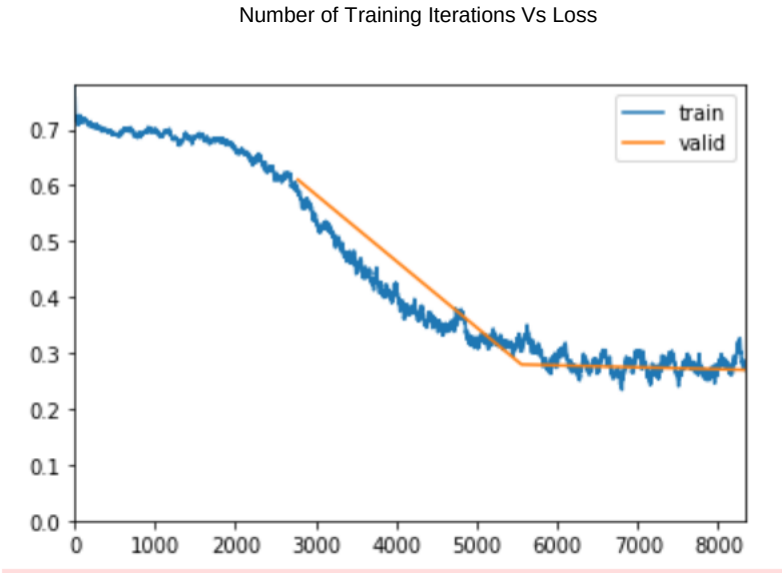
Findings:

Final accuracy: 90%

Below is the accuracy value for each epoch -

epoch	train_loss	valid_loss	accuracy	error_rate	time
0	0.586868	0.610249	0.669028	0.330972	25:40
1	0.307644	0.279444	0.891093	0.108907	26:37
2	0.272397	0.269873	0.900000	0.100000	26:12

Below is the training loss and validation loss -



Some Sample Paraphrase results -

```
SENTENCE 1 -> In the reviews below will be the highest rating for the show in red ,
              and the lowest evaluation for the show will be in blue episode .
SENTENCE 2 -> In the reviews below will be the lowest rating for the show in red ,
              and the highest rating for the show will be in blue sequence .
PREDICTED =  0 ACTUAL =  0
-------

SENTENCE 1 -> From the merger of the Four Rivers Council and the Audubon Council , the Shawnee Trails Council was born .
SENTENCE 2 -> Shawnee Trails Council was formed from the merger of the Four Rivers Council and the Audubon Council .
PREDICTED =  1 ACTUAL =  1
-------

SENTENCE 1 -> The family moved to Camp Hill in 1972 , where he attended Trinity High School in Harrisburg , Pennsylvania .
SENTENCE 2 -> In 1972 , the family moved to Camp Hill , where he visited the Trinity High School in Harrisburg , Pennsylvania .
PREDICTED =  1 ACTUAL =  1
-------

SENTENCE 1 -> Components of elastic potential systems store mechanical energy if
              they are deformed when forces are applied to the system .
SENTENCE 2 -> Components of elastic potential systems store mechanical energy if
              they are deformed to the system when applied to forces .
PREDICTED =  1 ACTUAL =  1
-------

SENTENCE 1 -> Aamir Khan has agreed to play immediately after reading Mehra 's script in `` Rang De Basanti '' .
SENTENCE 2 -> Mehra agreed to act in `` Rang De Basanti '' immediately after reading Aamir Khan 's script .
PREDICTED =  0 ACTUAL =  0
-------
```

So our final scores were from our best saved model which was:
bert-base-cased fine tuned on PAWS dataset

Accuracy = 93.1%

F1 score = 92.65%

These scores are comparable to that of the PAWS paper mentioned above. It is more than the score they got on PAWS_QQP dataset (89%).

Agreement with Proposed Timeline

Our proposed timeline expected a report on the paraphrase generation task as well. However, after much deliberation with everyone including the TA, it was decided to keep information to a minimum on the generation task. The paraphrase generation task was slightly explored until the point where we decided to focus on the task at hand.

The first attempt at the paraphrase identification task was done with the help of the T5 model that was trained on the PAWS dataset.

Plans for Next Phase

Our current classification F1 and accuracy scores are close to the baseline. In the next phase, the primary aim would be to generate paraphrased sentences. Based on the completion of the paraphrased sentence generation task, attempts will be made to 'beat' the deadline.

Learnings from Attempt 1:

The major learning was the fact that we can not very likely proceed with the idea of "off the shelf" models. And hence, we would have to look at means of training the model, irrespective of what we select finally.