Advanced NLP Project Report

Topic Name: Contrastive Representation Learning for Exemplar-Guided Paraphrase Generation

Team Name: InBred

Team Members:

1. Harshit Gupta (2020114017)

2. Pratyaksh Gautam (2020114002)

Introducing the Problem Statement

The goal of Exemplar-Guided Paraphrase Generation (EGPG) is to produce a target sentence that matches the style of the provided exemplar while preserving the source sentence's content information.

In an effort to learn a better representation of the style and the substance, this study makes a novel approach suggestion. The recent success of contrastive learning, which has proven its effectiveness in unsupervised feature extraction tasks, is the key driving force behind this approach. Designing two contrastive losses with regard to the content and style while taking into account two problem features during training is the idea.

Dataset Characteristics

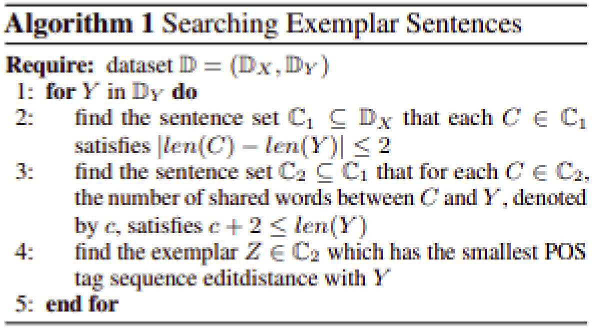
In order to train and assess our models, we use two datsets. As follows:

We

1. **ParaNMT Dataset:** Using back translation of the original English sentences from a different challenge, they were created automatically.
2. **QQPos:** Compared to the dataset above, the QQPos Dataset is more formal.

We use 93k sentences for training, 3k sentences for validation and 3k sentences for testing from both the datasets each.

Dataset Manipulation: While we get only the source and target sentence pairs from the above mentioned datasets, we **exemplar sentences** as well in order ot make our model. We come up with exemplar sentences on our own. To extract the sentences from the source-target pairs in the dataset, we take the following method.



```
Algorithm 1 Searching Exemplar Sentences
Require: dataset $\mathbb{D} = (\mathbb{D}_X, \mathbb{D}_Y)$
for $Y$ in $\mathbb{D}_Y$ do
    find the sentence set $\mathbb{C}_1 \subseteq \mathbb{D}_X$ that each $C \in \mathbb{C}_1$ satisfies $|len(C) - len(Y)| \leq 2$
    find the sentence set $\mathbb{C}_2 \subseteq \mathbb{C}_1$ that for each $C \in \mathbb{C}_2$, the number of shared words between $C$ and $Y$, denoted by $c$, satisfies $c + 2 \leq len(Y)$
    find the exemplar $Z \in \mathbb{C}_2$ which has the smallest POS tag sequence editdistance with $Y$
end for
```

The primary factor is that we look for appropriate exemplar pairings based on the edit distance between the source and target pairs' POS tagging sequences.

Modelling Approach

The model architecture can be summarized as:

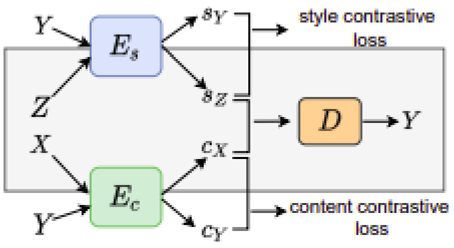


Figure 1: An Overview of Our Model

The notation here represents:

1. X, Y, Z as source, target and exemplar sentences repectively.
2. S_Y, S_Z as style encodings generated after passing them through an BERT based architecture to get the necessary feature representations. The dimensions of the style features is set to the standard `768`.
3. C_X, C_Y as content encodings that are generated after passing them through a bi-directional GRU architecture with hidden state size set to `512`.
4. S_Z, C_X are passed through the decoder after concatenating their feature representations and the loss is computed against the ground truth Y.

5. We use cross entropy loss simply because it is essentially NLL Loss + Softmax (which is done internally).

6. Along with that they have mentioned Style and Contrastive loss along with a mathematical formula. We have successfully vectorized it and made it more simpler and effecient to run using trasnformations and other relevant mathematical properties from linear algebra.

7. The Learning Rate is set to `1e-4` and the size od GloVe word Embeddings as `300`. Teacher forcing rate is set to `1.0` (total teacher forcing).

8. Regarding the 3 loss functions, they are cobined together using the following formula:

$$L = \sum_{i=1}^{n} L_i^{nll} + \lambda_1 L^{ccl} + \lambda_2 L^{scl}$$

9. In the above expression, λ_1 and λ_2 are set to `0.1` and τ is set to `0.5`.

We have implemented the same model architecture as explained above for this project.

Incorporating ScalarDotAttention

How to apply and use attention is not specifically addressed in the study. To our framework, we, nonetheless, applied the same. Scalar Dot Product Attention is the type of attention applied. A sort of similarity score between the query and key vectors is calculated using the dot product. We treat the Content Encoder and Decoder as the true Encoder Decoder framework and concentrate our attention just on the layers of the Content Encoder. However, we combine the content and style encoders' hidden states. This makes sure that style representations are included in the model training as well and affect the likelihood scores appropriately.

```
class ScaledDotProductAttention(nn.Module):
    def __init__(self, query_dim, key_dim, value_dim):
        super().__init__()
        self.scale = torch.sqrt(torch.FloatTensor([query_dim])).to(device) #dimensions: (32,qd)
        self.query_linear = nn.Linear(query_dim, query_dim)
        self.key_linear = nn.Linear(key_dim, query_dim)
        self.value_linear = nn.Linear(value_dim, query_dim)
    def forward(self, query, key, value, mask=None):
        query = self.query_linear(query) #dimensions: (1, batch_size, query_dim)
        key = self.key_linear(key) #dimensions: (batch_size, src_len, query_dim)
        value = self.value_linear(value) #dimensions: (batch_size, src_len, query_dim)
        attn_weights = torch.bmm(query.permute(1,0,2), key.permute(0,2,1)).to(device)
#dimensions: (b, 1, s)
        attn_weights = attn_weights / self.scale
        if mask is not None:
            attn_weights = attn_weights.masked_fill(mask == 0, -1e10)
        attn_weights = F.softmax(attn_weights, dim=2) #dimensions: (b,1,s)
        context = torch.bmm(attn_weights, value).to(device) #dimensions: (b,1,q)
        return context, attn_weights
```

Results and Analysis

Based on the model architecture above, we experimented and generated the following results:

1. Model with only NLL Loss

2. Model with NLL Loss+Both Contrastive Losses

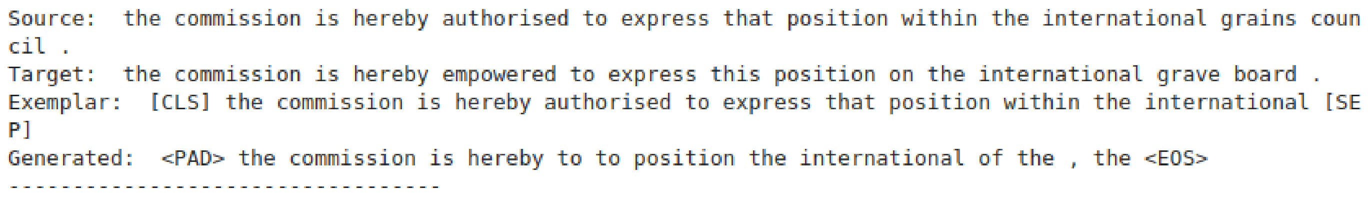
Metric	QQPos	ParaNMT
ROUGE1: Precision	0.4004	0.5416
ROUGE1: Recall	0.3754	0.5179
ROUGE1: F-Measure	0.3861	0.5271
ROUGE2: Precision	0.1711	0.3116
ROUGE2: Recall	0.1640	0.3014
ROUGE2: F-Measure	0.1670	0.3052
ROUGEL: Precision	0.3442	0.4890
ROUGEL: Recall	0.3246	0.4696
ROUGEL: F-Measure	0.3329	0.4770
BLEU	0.0995	0.1196
METEOR	0.3234	0.4813

Examples Of Generated Output

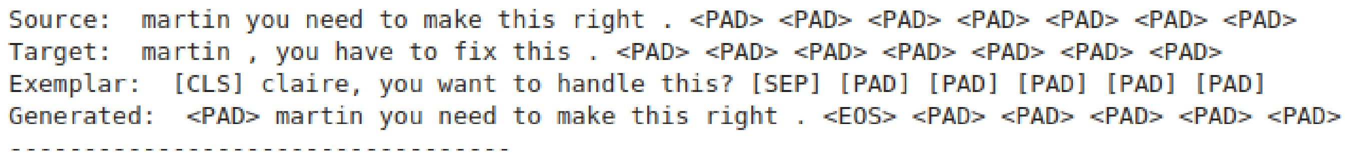
Observation: We can see from the generated sentences that the model both catches the information and makes sure the exemplar style is preserved. Using precise and accessible substitutes, several crucial words have been replaced in intriguing ways that highlight how well learned embeddings and representations work. The sentence contains some syntactic errors, however this is due to the target sentences' attempts to mimic the syntactic structure of the exemplar sentences, which causes some awkward syntax. We do see, however, that the overall meaning and content has been retained quite well.

From Para-NMT Dataset

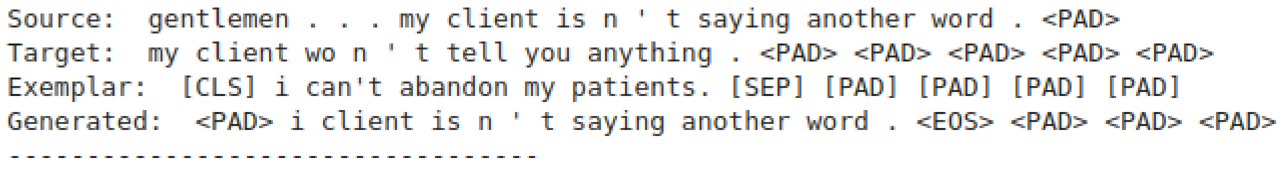
```
Source:  chemical and physical compatibility has been demonstrated with the following <UNK> <PAD> <PAD> <PAD> <PAD>
Target:  chemical and physical stability has been demonstrated with the following solvents <PAD> <PAD> <PAD> <PAD>
Exemplar:  [CLS] chemical and physical compatibility has been demonstrated with the following diluents [SEP]
Generated:  <PAD> chemical and quality has been demonstrated with <EOS> <EOS> <PAD> <PAD> <PAD> <PAD> <PAD>
```



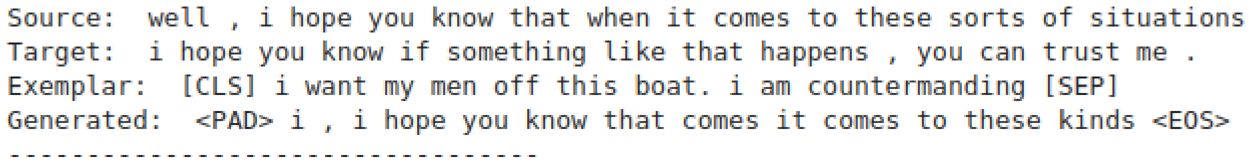
```
Source:  the commission is hereby authorised to express that position within the international grains coun
cil .
Target:  the commission is hereby empowered to express this position on the international grave board .
Exemplar:  [CLS] the commission is hereby authorised to express that position within the international [SE
P]
Generated:  <PAD> the commission is hereby to to position the international of the , the <EOS>
---------------------------------
```



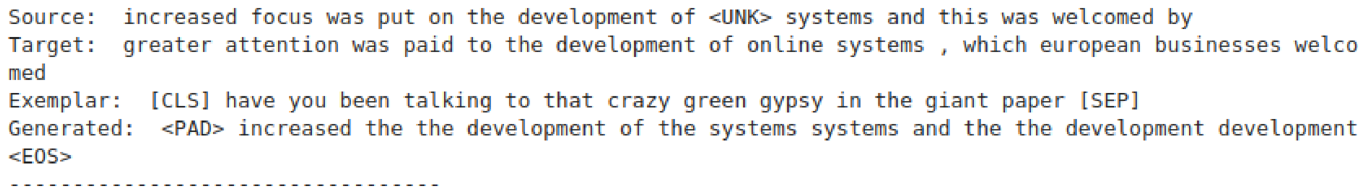
```
Source:  martin you need to make this right . <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD>
Target:  martin , you have to fix this . <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD>
Exemplar:  [CLS] claire, you want to handle this? [SEP] [PAD] [PAD] [PAD] [PAD] [PAD]
Generated:  <PAD> martin you need to make this right . <EOS> <PAD> <PAD> <PAD> <PAD> <PAD>
---------------------------------
```



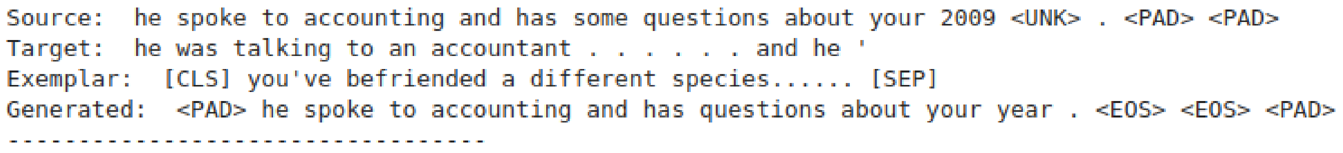
```
Source:  gentlemen . . . my client is n ' t saying another word . <PAD>
Target:  my client wo n ' t tell you anything . <PAD> <PAD> <PAD> <PAD> <PAD>
Exemplar:  [CLS] i can't abandon my patients. [SEP] [PAD] [PAD] [PAD] [PAD]
Generated:  <PAD> i client is n ' t saying another word . <EOS> <PAD> <PAD> <PAD>
---------------------------------
```



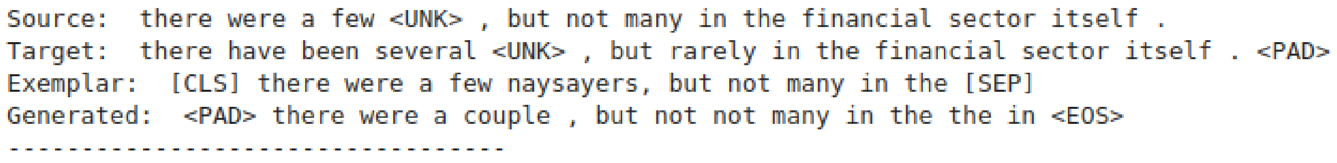
```
Source:  well , i hope you know that when it comes to these sorts of situations
Target:  i hope you know if something like that happens , you can trust me .
Exemplar:  [CLS] i want my men off this boat. i am countermanding [SEP]
Generated:  <PAD> i , i hope you know that comes it comes to these kinds <EOS>
---------------------------------
```



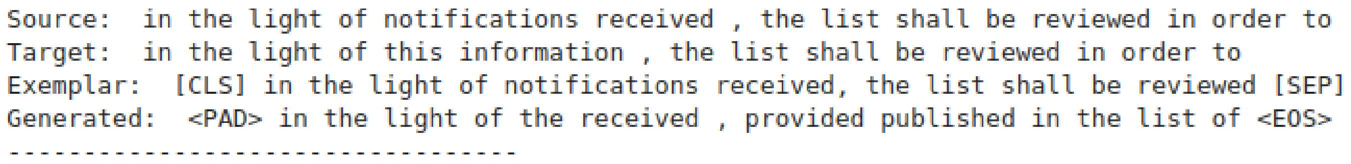
```
Source:  increased focus was put on the development of <UNK> systems and this was welcomed by
Target:  greater attention was paid to the development of online systems , which european businesses welco
med
Exemplar:  [CLS] have you been talking to that crazy green gypsy in the giant paper [SEP]
Generated:  <PAD> increased the the development of the systems systems and the the development development
<EOS>
---------------------------------
```



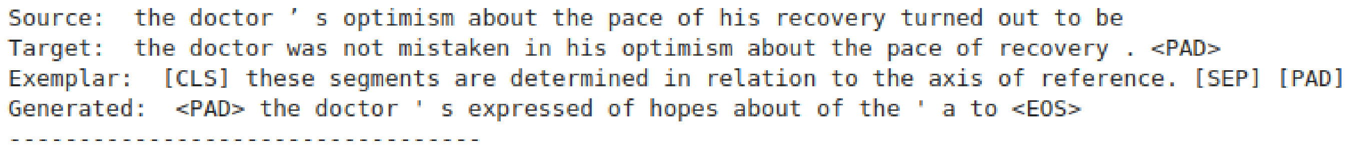
```
Source:  he spoke to accounting and has some questions about your 2009 <UNK> . <PAD> <PAD>
Target:  he was talking to an accountant . . . . . . and he '
Exemplar:  [CLS] you've befriended a different species...... [SEP]
Generated:  <PAD> he spoke to accounting and has questions about your year . <EOS> <EOS> <PAD>
---------------------------------
```



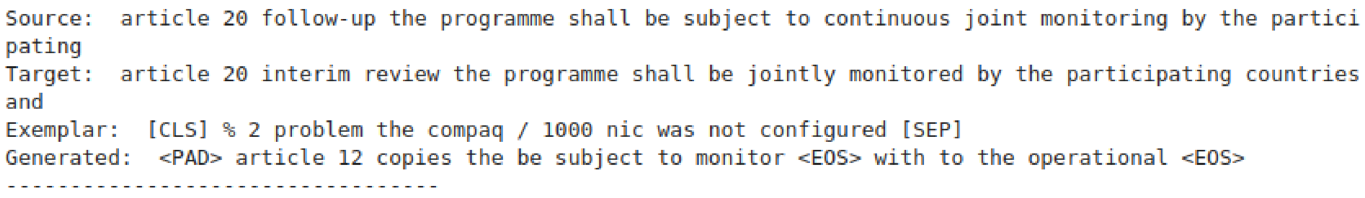
```
Source:  there were a few <UNK> , but not many in the financial sector itself .
Target:  there have been several <UNK> , but rarely in the financial sector itself . <PAD>
Exemplar:  [CLS] there were a few naysayers, but not many in the [SEP]
Generated:  <PAD> there were a couple , but not not many in the the in <EOS>
---------------------------------
```



```
Source:  in the light of notifications received , the list shall be reviewed in order to
Target:  in the light of this information , the list shall be reviewed in order to
Exemplar:  [CLS] in the light of notifications received, the list shall be reviewed [SEP]
Generated:  <PAD> in the light of the received , provided published in the list of <EOS>
---------------------------------
```

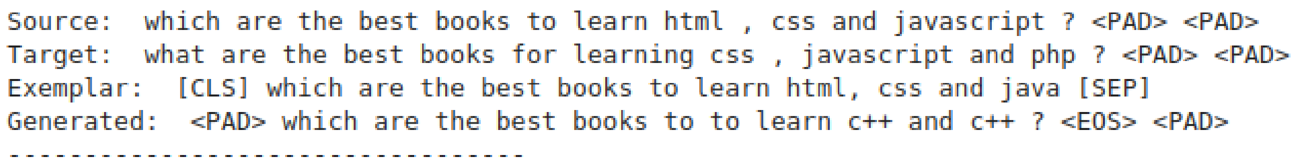


```
Source:  the doctor ’ s optimism about the pace of his recovery turned out to be
Target:  the doctor was not mistaken in his optimism about the pace of recovery . <PAD>
Exemplar:  [CLS] these segments are determined in relation to the axis of reference. [SEP] [PAD]
Generated:  <PAD> the doctor ' s expressed of hopes about of the ' a to <EOS>
---------------------------------
```

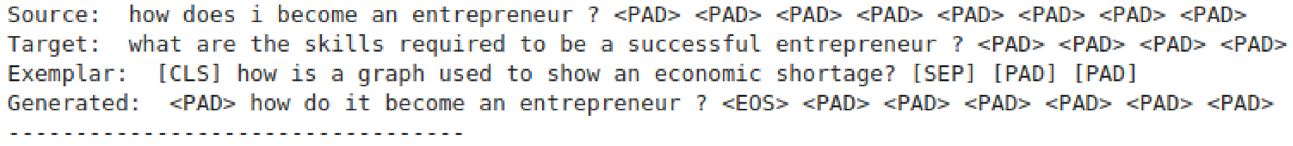


```
Source:  article 20 follow-up the programme shall be subject to continuous joint monitoring by the partici
pating
Target:  article 20 interim review the programme shall be jointly monitored by the participating countries
and
Exemplar:  [CLS] % 2 problem the compaq / 1000 nic was not configured [SEP]
Generated:  <PAD> article 12 copies the be subject to monitor <EOS> with to the operational <EOS>
---------------------------------
```

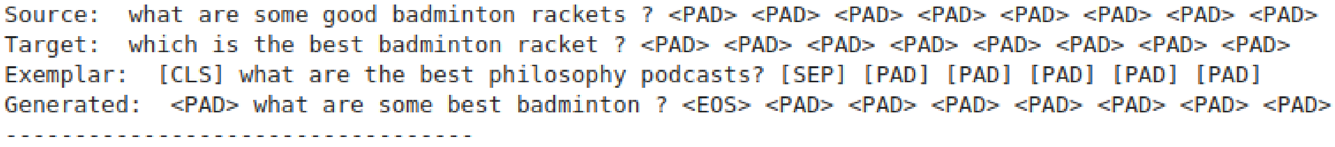
From QQPos Dataset



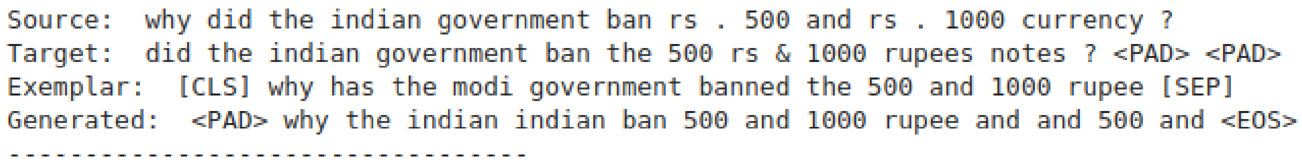
```
Source:  which are the best books to learn html , css and javascript ? <PAD> <PAD>
Target:  what are the best books for learning css , javascript and php ? <PAD> <PAD>
Exemplar:  [CLS] which are the best books to learn html, css and java [SEP]
Generated:  <PAD> which are the best books to to learn c++ and c++ ? <EOS> <PAD>
---------------------------------
```



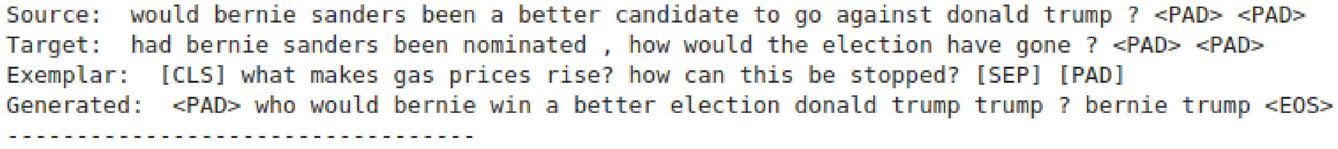
```
Source:  how does i become an entrepreneur ? <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD>
Target:  what are the skills required to be a successful entrepreneur ? <PAD> <PAD> <PAD> <PAD>
Exemplar:  [CLS] how is a graph used to show an economic shortage? [SEP] [PAD] [PAD]
Generated:  <PAD> how do it become an entrepreneur ? <EOS> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD>
---------------------------------
```



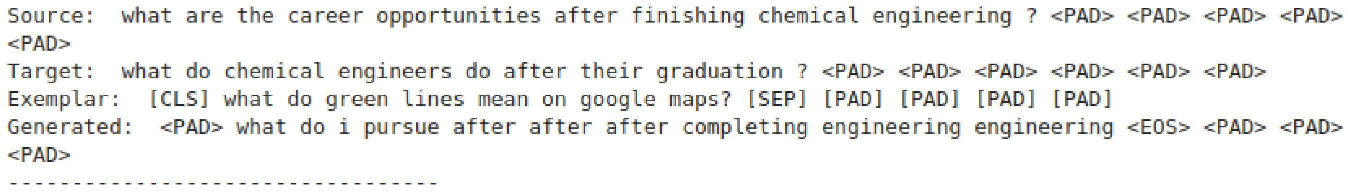
```
Source:  what are some good badminton rackets ? <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD>
Target:  which is the best badminton racket ? <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD>
Exemplar:  [CLS] what are the best philosophy podcasts? [SEP] [PAD] [PAD] [PAD] [PAD] [PAD]
Generated:  <PAD> what are some best badminton ? <EOS> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD>
---------------------------------
```



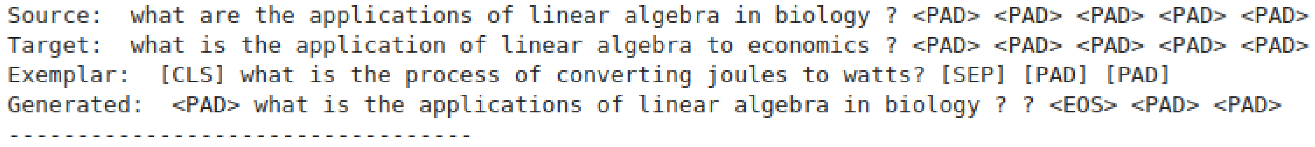
```
Source:  why did the indian government ban rs . 500 and rs . 1000 currency ?
Target:  did the indian government ban the 500 rs & 1000 rupees notes ? <PAD> <PAD>
Exemplar:  [CLS] why has the modi government banned the 500 and 1000 rupee [SEP]
Generated:  <PAD> why the indian indian ban 500 and 1000 rupee and and 500 and <EOS>
---------------------------------
```



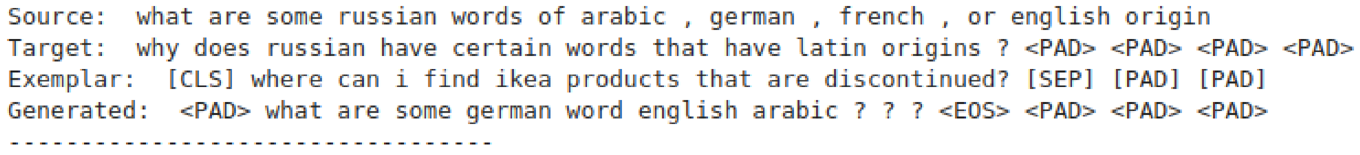
```
Source:  would bernie sanders been a better candidate to go against donald trump ? <PAD> <PAD>
Target:  had bernie sanders been nominated , how would the election have gone ? <PAD> <PAD>
Exemplar:  [CLS] what makes gas prices rise? how can this be stopped? [SEP] [PAD]
Generated:  <PAD> who would bernie win a better election donald trump trump ? bernie trump <EOS>
---------------------------------
```



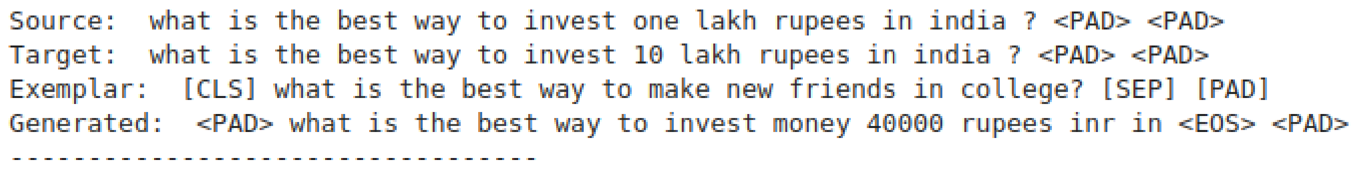
```
Source:  what are the career opportunities after finishing chemical engineering ? <PAD> <PAD> <PAD> <PAD>
<PAD>
Target:  what do chemical engineers do after their graduation ? <PAD> <PAD> <PAD> <PAD> <PAD> <PAD>
Exemplar:  [CLS] what do green lines mean on google maps? [SEP] [PAD] [PAD] [PAD] [PAD]
Generated:  <PAD> what do i pursue after after after completing engineering engineering <EOS> <PAD> <PAD>
<PAD>
---------------------------------
```



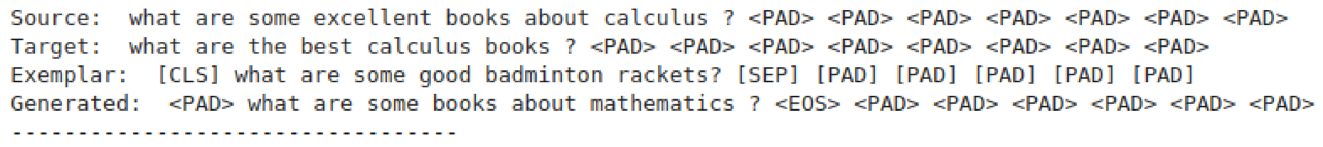
```
Source:  what are the applications of linear algebra in biology ? <PAD> <PAD> <PAD> <PAD> <PAD>
Target:  what is the application of linear algebra to economics ? <PAD> <PAD> <PAD> <PAD> <PAD>
Exemplar:  [CLS] what is the process of converting joules to watts? [SEP] [PAD] [PAD]
Generated:  <PAD> what is the applications of linear algebra in biology ? ? <EOS> <PAD> <PAD>
---------------------------------
```



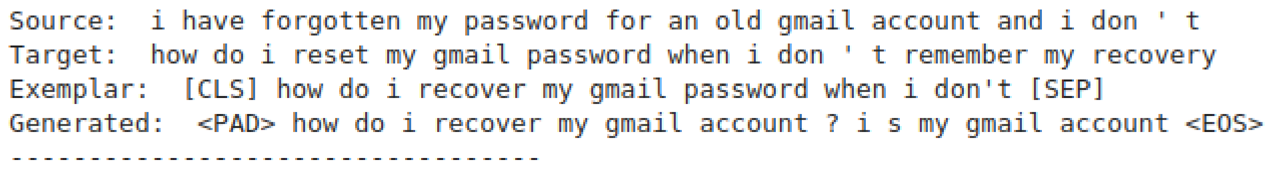
```
Source:  what are some russian words of arabic , german , french , or english origin
Target:  why does russian have certain words that have latin origins ? <PAD> <PAD> <PAD> <PAD>
Exemplar:  [CLS] where can i find ikea products that are discontinued? [SEP] [PAD] [PAD]
Generated:  <PAD> what are some german word english arabic ? ? ? <EOS> <PAD> <PAD> <PAD>
---------------------------------
```



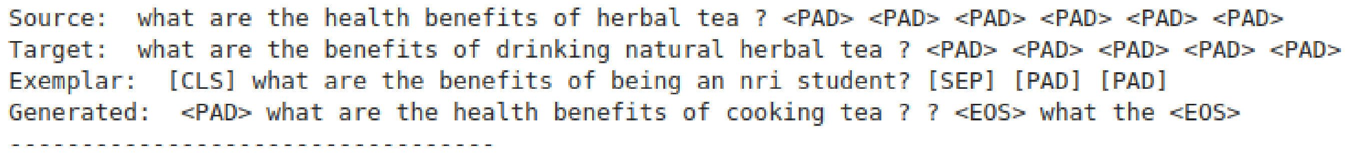
```
Source:  what is the best way to invest one lakh rupees in india ? <PAD> <PAD>
Target:  what is the best way to invest 10 lakh rupees in india ? <PAD> <PAD>
Exemplar:  [CLS] what is the best way to make new friends in college? [SEP] [PAD]
Generated:  <PAD> what is the best way to invest money 40000 rupees inr in <EOS> <PAD>
----------------------------------
```



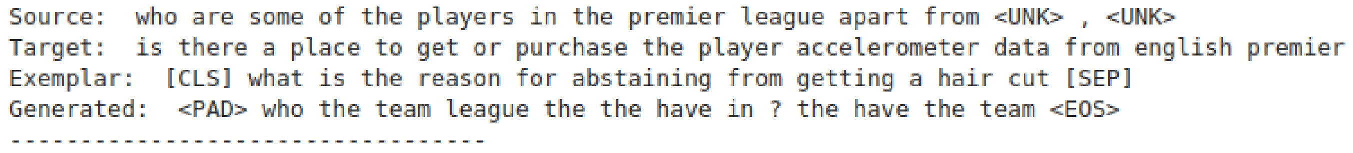
```
Source:  what are some excellent books about calculus ? <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD>
Target:  what are the best calculus books ? <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD>
Exemplar:  [CLS] what are some good badminton rackets? [SEP] [PAD] [PAD] [PAD] [PAD] [PAD]
Generated:  <PAD> what are some books about mathematics ? <EOS> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD>
---------------------------------
```



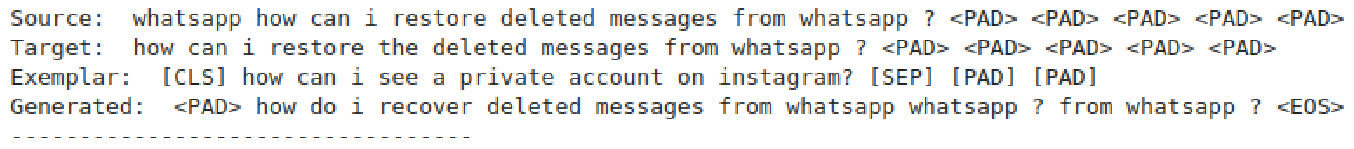
```
Source:  i have forgotten my password for an old gmail account and i don ' t
Target:  how do i reset my gmail password when i don ' t remember my recovery
Exemplar:  [CLS] how do i recover my gmail password when i don't [SEP]
Generated:  <PAD> how do i recover my gmail account ? i s my gmail account <EOS>
----------------------------------
```



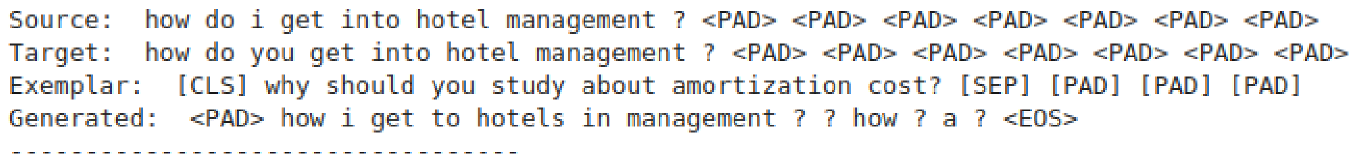
```
Source:  what are the health benefits of herbal tea ? <PAD> <PAD> <PAD> <PAD> <PAD> <PAD>
Target:  what are the benefits of drinking natural herbal tea ? <PAD> <PAD> <PAD> <PAD> <PAD>
Exemplar:  [CLS] what are the benefits of being an nri student? [SEP] [PAD] [PAD]
Generated:  <PAD> what are the health benefits of cooking tea ? ? <EOS> what the <EOS>
---------------------------------
```



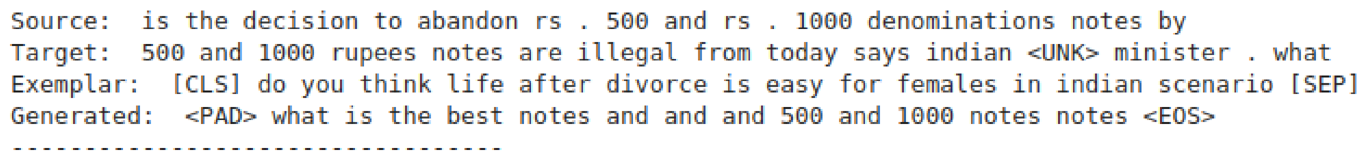
```
Source:  who are some of the players in the premier league apart from <UNK> , <UNK>
Target:  is there a place to get or purchase the player accelerometer data from english premier
Exemplar:  [CLS] what is the reason for abstaining from getting a hair cut [SEP]
Generated:  <PAD> who the team league the the have in ? the have the team <EOS>
---------------------------------
```



```
Source:  whatsapp how can i restore deleted messages from whatsapp ? <PAD> <PAD> <PAD> <PAD> <PAD>
Target:  how can i restore the deleted messages from whatsapp ? <PAD> <PAD> <PAD> <PAD> <PAD>
Exemplar:  [CLS] how can i see a private account on instagram? [SEP] [PAD] [PAD]
Generated:  <PAD> how do i recover deleted messages from whatsapp whatsapp ? from whatsapp ? <EOS>
---------------------------------
```



```
Source:  how do i get into hotel management ? <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD>
Target:  how do you get into hotel management ? <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD>
Exemplar:  [CLS] why should you study about amortization cost? [SEP] [PAD] [PAD] [PAD]
Generated:  <PAD> how i get to hotels in management ? ? how ? a ? <EOS>
----------------------------------
```



```
Source:  is the decision to abandon rs . 500 and rs . 1000 denominations notes by
Target:  500 and 1000 rupees notes are illegal from today says indian <UNK> minister . what
Exemplar:  [CLS] do you think life after divorce is easy for females in indian scenario [SEP]
Generated:  <PAD> what is the best notes and and and 500 and 1000 notes notes <EOS>
---------------------------------
```