Adversarial Machine Translation Inputs
Step 3:

Input sentence	machine translation output	correct translation output
你都老大不小了，还是这么毛躁	Your boss is not young anymore, still so frizzy.	You are not young anymore but still so irritable.
On Friday, President gave the order to the military to carry out the mission to take out Ben.	星期五，总统下令命令军方执行将本·本的任务。	星期五，总统下命令让军方将本除掉。
红豆生南国，春来发几枝	Red beans are born in the south, and spring comes to send a few branches.	Red beans are born in the south, and Several branches grow as spring coming.
人生得意须尽欢	Life is proud to be thoroughly enjoyed themselves	Enjoy life every chance you get
莫使金樽空对月	Don't make Jinzun empty to the moon	Don't hold a gold but empty cup under the moon.
快使用饿了么。现在下单你喜欢的菜品可享八折。	Are you hungry? Now you can get 20% off when you order your favorite dishes.	Use Ele.me now to get 20% off when you order your favorite dishes.
林尽水源，便得一山	If the forest is full of water, there is a mountain	At the end of forest and river, there is a mountain.
斯是陋室，惟吾德馨。	Si is a shabby room, but I am decent.	It is a shabby house, but I have good character in it.
谈笑有鸿儒，往来无白丁。	Talk and laugh with ru, no contacts STOMATOLOGY.	I talk and laugh with gentlemen, and do not contact with illiteracies.
唐宗宋祖，稍逊风骚。	Tang Zong and Song Zu, a little less coquettish.	Emperor Taizong of Tang and Emperor Taizu of Sang are not that elegant.
三尺微命，一介书生。	Sanchi Weiming, a scholar	I am not gentility, just a scholar.
The use of deep learning has significantly advanced translation quality.	深度学习的使用显着提高了翻译质量。	深度学习的使用显著提高了翻译质量。

Your task is to find and collect as many as possible foreign input sentences that Google translate will make errors on.	您的任务是查找并收集尽可能多的 Google 翻译会在其上出错的外来输入句子。	您的任务是查找并收集尽可能多的谷歌翻译出错句子。
总要有个好开始	Always have a good start	It should be a good beginning.
我爱的是不是你，你不知道吗？	Is it you who I love, don't you know?	I love you or not, don't you know?
他想吃披萨想吃到要晕倒	He wants to eat pizza until he faints.	He wants to eat pizza so much even going to be faint.
天上一天，人间一年。	One day in the sky, one year in the world.	A day in heaven, a year on earth.
你好	Hello there	Hello
我刚洗了碗，现在准备拖地。	I just washed the dishes and am now preparing to mop the floor.	I just washed the dishes, and am going to mop the floor now.
给岁月以文明，而不是给文明以岁月	Give time to civilization, not to civilization	Provide civilization to time, not provide time to civilization.

Step 4:

Type & Reason:

1/ Meaning error:

Polyseme: These words have different meanings in different context, and the machine sometimes makes mistakes.

Special words: Every languages have their own words or phrases that other languages can not translate as their literally meaning. Sometimes they’re from ancient languages which has some different with modern languages, and sometimes they are just proper nouns.

Special symbols

2/ Structure error:

Words order reversing: Sentences consist of words by grammar, and different languages have different grammars. Sometimes the results translating word by word will be weird in the other languages.

Words missing

Phrase splitting: Some phrases are split in a wrong way.

The machine split a sentence to each words and translate them, it learned meaning from a huge language database and obtain a most likely one. Firstly, The dataset may not big enough, so that it cannot have enough samples to study those flexible structure. As the result, it cannot split

sentences well and recognize polyseme well. Secondly, the original sentences contain some special symbols and words, the machine may consider them as invalid input and miss them.

Step 5:

I think the most direct way is to expand the datasets, so the machine will have more chance to meet the complexity of a language, whatever about the grammar or special words. The more data machine learns from, the less error it makes. At the same time, the network can change it structures. A more complex network with more layers or more units may not improve the Google translate that much, but we can add some new parts into the machine. For example, adding a context reinforcement system to the machine. Grabbing the information appears frequently on the Internet relative to each word to create its possible stronger context. So that the machine may have more chance to offer a precise meaning.

Attentional Neural Machine Translation

2/

```
def simple_model(input, vocab_size, trans_vocab_size):
    # Build the layers
    learning_rate = 1e-3
    rnn = GRU(64, return_sequences = True)(input)
    logits = TimeDistributed(Dense(trans_vocab_size))(rnn)
    model = Model(input_seq, Activation('softmax')(logits))
    model.compile(loss = sparse_categorical_crossentropy,
                  optimizer = Adam(learning_rate),
                  metrics = ['accuracy'])

    return model

# Train the neural network
simple_rnn_model = simple_model(
    Train[0],
    5208,
    5208)
simple_rnn_model.fit(Train[0], Train[1], batch_size=1000, epochs=10, validation_split=0.2)

# Print test accuracy(s)
print(logits_to_text(simple_rnn_model.predict(test[0], test[1])))
```

3/

```
1 Data = pd.read_table('data.txt',header=None,encoding = "utf-8",dtype = str,sep='\s')
2 Data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6510 entries, 0 to 6509
Data columns (total 2 columns):
0    6510 non-null object
1    6510 non-null object
dtypes: object(2)
memory usage: 101.8+ KB
```

```
1 Train, Test = train_test_split(Data, test_size=0.2)
2 Train.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 5208 entries, 4725 to 2120
Data columns (total 2 columns):
0    5208 non-null object
1    5208 non-null object
dtypes: object(2)
memory usage: 122.1+ KB
```

```
1 Test.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1302 entries, 3043 to 2872
Data columns (total 2 columns):
0    1302 non-null object
1    1302 non-null object
dtypes: object(2)
memory usage: 30.5+ KB
```

4/ Yes, it will have problem to translate the sentence. Because I choose a fix length of each character and change them to an index list, and a long sentence will beyond the length and can't be reshape.
5/ I think the LSTM RNN can work well.