Example Based Machine Translation

Abstract

The project aims to implement an Example-Based Machine Translation (EBMT) system for translating sentences between English to Turkish and, Azerbaijani to Turkish. The code utilizes the NLTK library for tokenization and the Zemberek library for Turkish language processing. The code starts by reading parallel sentence corpora from text files for English, Azerbaijani, and Turkish. It preprocesses the corpora by tokenizing the sentences and normalizing the Turkish sentences using Zemberek. The main functionality of the code is the translate_example_based function, which performs the translation using the EBMT approach. It calculates the edit distance between the input sentence and the source corpus sentences and selects the most similar sentence from the target corpus. The code also includes a user interface that allows the user to select the translation direction and enter the input sentence. It then calls the translate_example_based function to obtain the translation and displays it to the user. Furthermore, the code includes an additional function, calculate_bleu_score, that calculates the BLEU score between a reference sentence and a translated sentence. This allows for evaluating the translation quality by measuring the similarity between the translated and reference sentences. Overall, the project provides a basic framework for performing example-based machine translation and includes functionality for user interaction and BLEU score evaluation.

Keywords: Example Based Machine Translation, NLTK, Zemberek, BLUE

EBMT Report

The purpose of this report is to provide an overview and evaluation of the Example-Based Machine Translation (EBMT) system implemented in the provided code. The EBMT system aims to translate sentences between English, Azerbaijani, and Turkish languages. The code utilizes the NLTK library for tokenization and the Zemberek library for Turkish language processing.

Text Generation For Corporas

English, Turkish, Azerbejaini

Corporas consist of two different sets parallel to each other along with 5007 texts per each, 10014 in parallel, and 15021 texts in total. There are various resources that are used to generate sentences to use in English corpus. These are as follows;

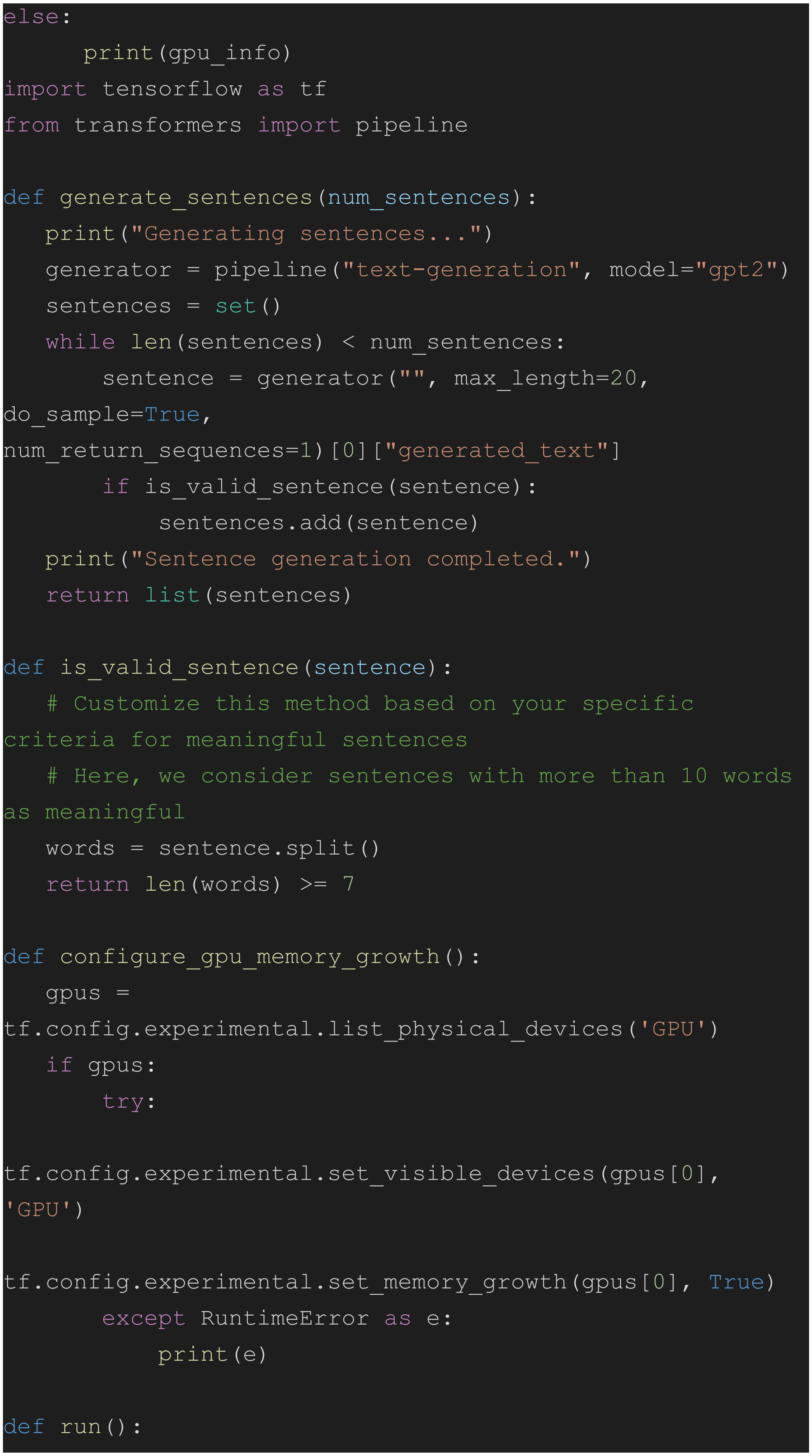
Manual Text Generating: Creating and implementing sentences by hand from various web resources such as, e-book, article, social media, newspaper, etc.

Random Text Generator: A website called https://randomwordgenerator.com that gives you random 50 texts including questions, sentences, phrases, etc.

Using Machine Learning Techniques: English corpora is used to use ML techniques due to volume and number of English corpora.

- We have used a pre-trained model called gpt2 model along with transformers which are accessible by Hugging Face cloud models.
- Gpt2 model used (By it is own) pre-trained from web resources as well mostly social media entries and newspapers

```
gpu_info = !nvidia-smi
gpu_info = '\n'.join(gpu_info)
if gpu_info.find('failed') >= 0:
      print('Not connected to a GPU')
```



```
else:
      print(gpu_info)
import tensorflow as tf
from transformers import pipeline

def generate_sentences(num_sentences):
   print("Generating sentences...")
   generator = pipeline("text-generation", model="gpt2")
   sentences = set()
   while len(sentences) < num_sentences:
       sentence = generator("", max_length=20,
do_sample=True,
num_return_sequences=1)[0]["generated_text"]
       if is_valid_sentence(sentence):
           sentences.add(sentence)
   print("Sentence generation completed.")
   return list(sentences)

def is_valid_sentence(sentence):
   # Customize this method based on your specific
criteria for meaningful sentences
   # Here, we consider sentences with more than 10 words
as meaningful
   words = sentence.split()
   return len(words) >= 7

def configure_gpu_memory_growth():
   gpus =
tf.config.experimental.list_physical_devices('GPU')
   if gpus:
       try:

tf.config.experimental.set_visible_devices(gpus[0],
'GPU')

tf.config.experimental.set_memory_growth(gpus[0], True)
       except RuntimeError as e:
           print(e)

def run():
   configure_gpu_memory_growth()
   num_sentences = 2500
   generated_sentences =
generate_sentences(num_sentences)
   for sentence in generated_sentences:
       print(f'"{sentence}",')

if __name__ == "__main__":
   run()
```

- The given code above uses the Transformers library to generate sentences using the GPT-2 model. It defines a function to generate a specified number of sentences by iteratively sampling from the model until the desired number is reached. The code also includes a function to define criteria for meaningful sentences, such as the minimum number of words. Additionally, it sets up GPU memory growth using TensorFlow's configuration. The main function calls these functions to generate and print the sentences. Overall, the code demonstrates how to generate sentences with the GPT-2 model, customize the criteria for meaningful sentences, and configure GPU memory growth.

After 5007 English texts we used Google Cloud Translation API to get the translated data of Turkish and Azerbaijani corpuses as well with the same amount of number per each.

Script For Azerbaijani - Turkish:

```
    const fs = require("fs");
    const { Translate } =
require("@google-cloud/translate").v2;
```

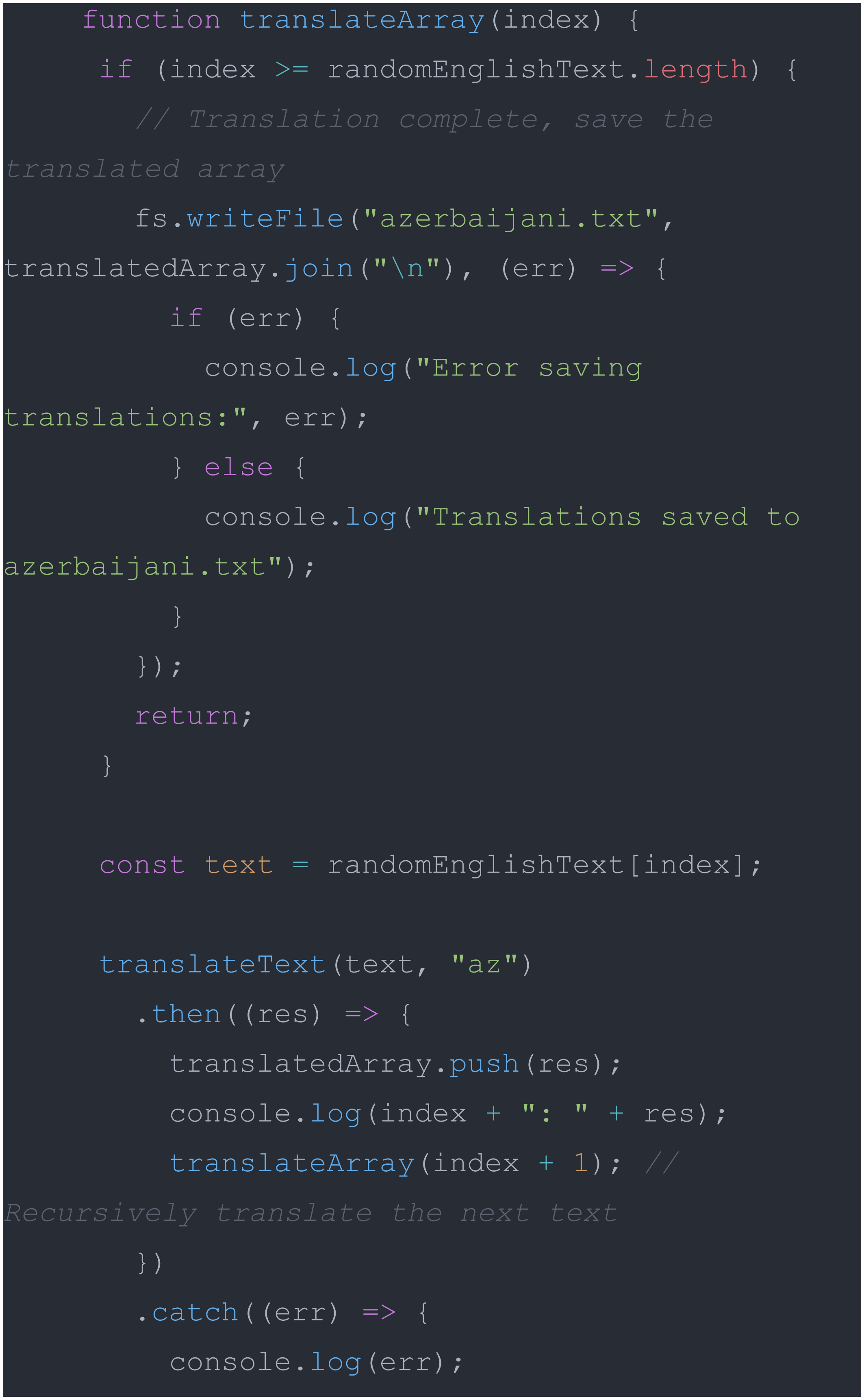
```
    const { randomEnglishText } =
require("./english");

    // Your credentials
    const CREDENTIALS =
JSON.parse(process.env.CREDENTIALS);

    // Configuration for the client
    const translate = new Translate({
     credentials: CREDENTIALS,
     projectId: CREDENTIALS.project_id,
    });

    const translateText = async (text,
targetLanguage) => {
     try {
       let [response] = await
translate.translate(text, targetLanguage);
       return response;
     } catch (error) {
       console.log(`Error at translateText -->
${error}`);
       return 0;
     }
    };

    const translatedArray = [];
```



```
    function translateArray(index) {
     if (index >= randomEnglishText.length) {
       // Translation complete, save the translated array
       fs.writeFile("azerbaijani.txt", translatedArray.join("\n"), (err) => {
         if (err) {
           console.log("Error saving translations:", err);
         } else {
           console.log("Translations saved to azerbaijani.txt");
         }
       });
       return;
     }

     const text = randomEnglishText[index];

     translateText(text, "az")
       .then((res) => {
         translatedArray.push(res);
         console.log(index + ": " + res);
         translateArray(index + 1); // Recursively translate the next text
       })
       .catch((err) => {
         console.log(err);
```

```
          translateArray(index + 1); // Move to
the next text even if translation fails
        });
     }

     translateArray(0); // Start translation from
the beginning of the array
```

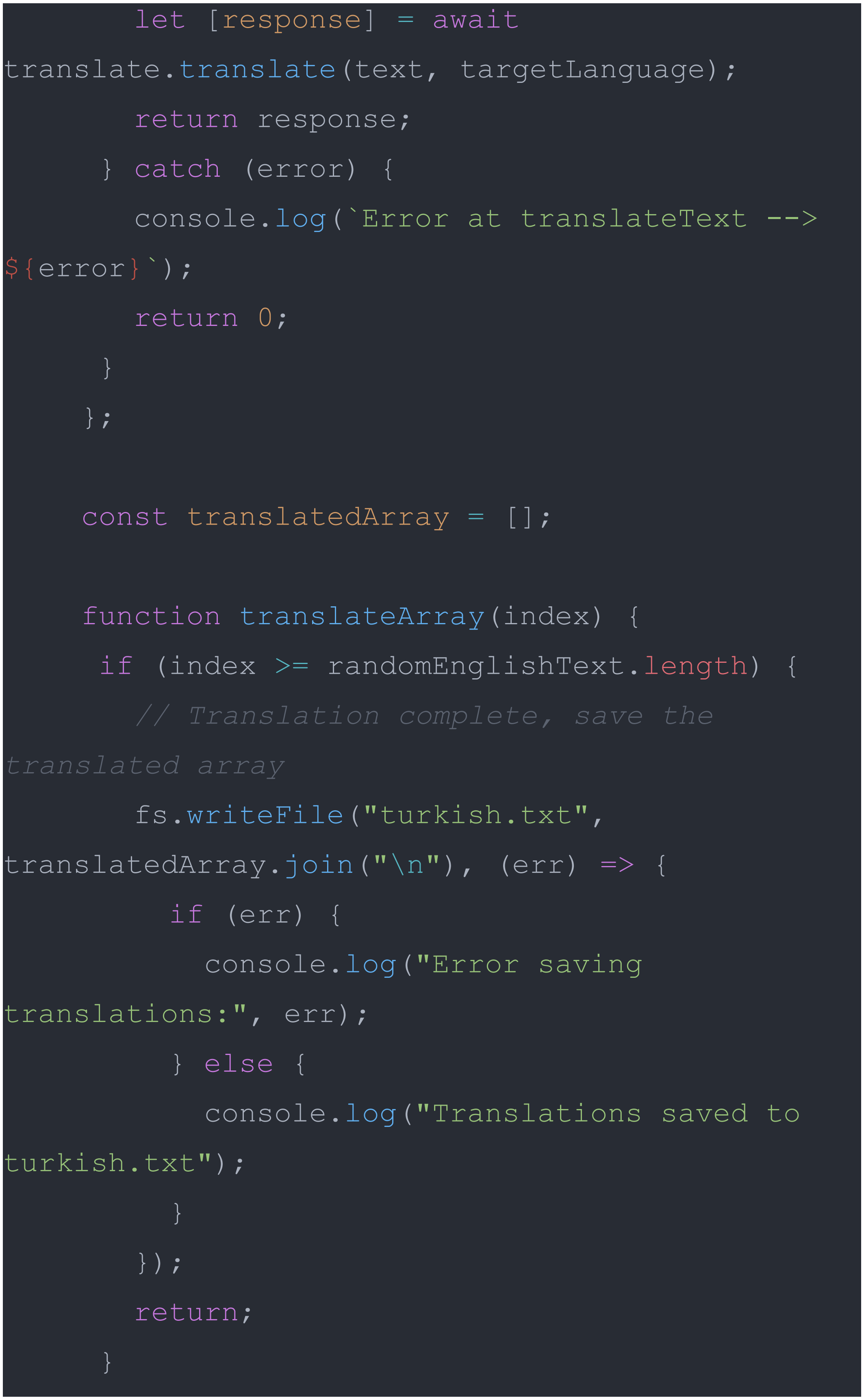
Script for English - Turkish

```
     const fs = require("fs");
     const { Translate } =
require("@google-cloud/translate").v2;
     const { randomEnglishText } =
require("./english");

     // Your credentials
     const CREDENTIALS =
JSON.parse(process.env.CREDENTIALS);

     // Configuration for the client
     const translate = new Translate({
      credentials: CREDENTIALS,
      projectId: CREDENTIALS.project_id,
     });

     const translateText = async (text,
targetLanguage) => {
      try {
```



```
       let [response] = await
translate.translate(text, targetLanguage);
       return response;
     } catch (error) {
       console.log(`Error at translateText -->
${error}`);
       return 0;
     }
    };

    const translatedArray = [];

    function translateArray(index) {
     if (index >= randomEnglishText.length) {
       // Translation complete, save the
translated array
       fs.writeFile("turkish.txt",
translatedArray.join("\n"), (err) => {
         if (err) {
           console.log("Error saving
translations:", err);
         } else {
           console.log("Translations saved to
turkish.txt");
         }
       });
       return;
     }
```

```
      const text = randomEnglishText[index];

      translateText(text, "tr")
        .then((res) => {
          translatedArray.push(res);
          console.log(index + ": " + res);
          translateArray(index + 1); // Recursively translate the next text
        })
        .catch((err) => {
          console.log(err);
          translateArray(index + 1); // Move to the next text even if translation fails
        });
    }

    translateArray(0); // Start translation from the beginning of the array

    // Save the original English text to english.txt
    fs.writeFile("english.txt", randomEnglishText.join("\n"), (err) => {
      if (err) {
        console.log("Error saving English text:", err);
      } else {
        console.log("English text saved to
english.txt");
      }
     });
```

Code Overview

The code consists of several components:

- Data Loading: The code reads parallel sentence corpora from text files for English, Azerbaijani, and Turkish languages.
- Preprocessing: The corpora are preprocessed by tokenizing the sentences and normalizing the Turkish sentences using Zemberek.
- Translation: The main functionality is the translate_example_based function, which performs the translation using the EBMT approach. It calculates the edit distance between the input sentence and the source corpus sentences and selects the most similar sentence from the target corpus.
- User Interface: The code includes a user interface that allows users to select the translation direction and enter the input sentence. It then calls the translation function and displays the translated sentence.
- BLEU Score Calculation: The code incorporates a function to calculate the BLEU score, which measures the similarity between the translated and reference sentences. The BLEU score provides an evaluation metric for translation quality.

Performance Evaluation

To evaluate the performance of the EBMT system, several aspects can be considered:

- Translation Accuracy: The system's accuracy can be assessed by manually inspecting the translated sentences and comparing them with the expected translations. This evaluation can be performed for different language pairs (e.g., English to Turkish, Azerbaijani to Turkish).
- Speed and Efficiency: The code does not explicitly measure translation speed or efficiency. However, these factors can be assessed by measuring the time taken to translate sentences of varying lengths and evaluating the code's resource usage.
- BLEU Score Analysis: The BLEU scores can be calculated for a set of reference sentences and their translations. The scores can provide an indication of the system's overall translation quality and can be compared across different language pairs and corpus sizes.

Limitations and Future Improvements

The implemented EBMT system has a few limitations and areas for potential improvement:

- Limited Language Support: The system currently supports translation between English, Azerbaijani, and Turkish only. Extending the system to include additional languages would enhance its usefulness.
- Dependency on Edit Distance: The current translation approach relies solely on edit distance for similarity measurement. Exploring other techniques, such as word alignment models or neural machine translation, could improve translation accuracy.

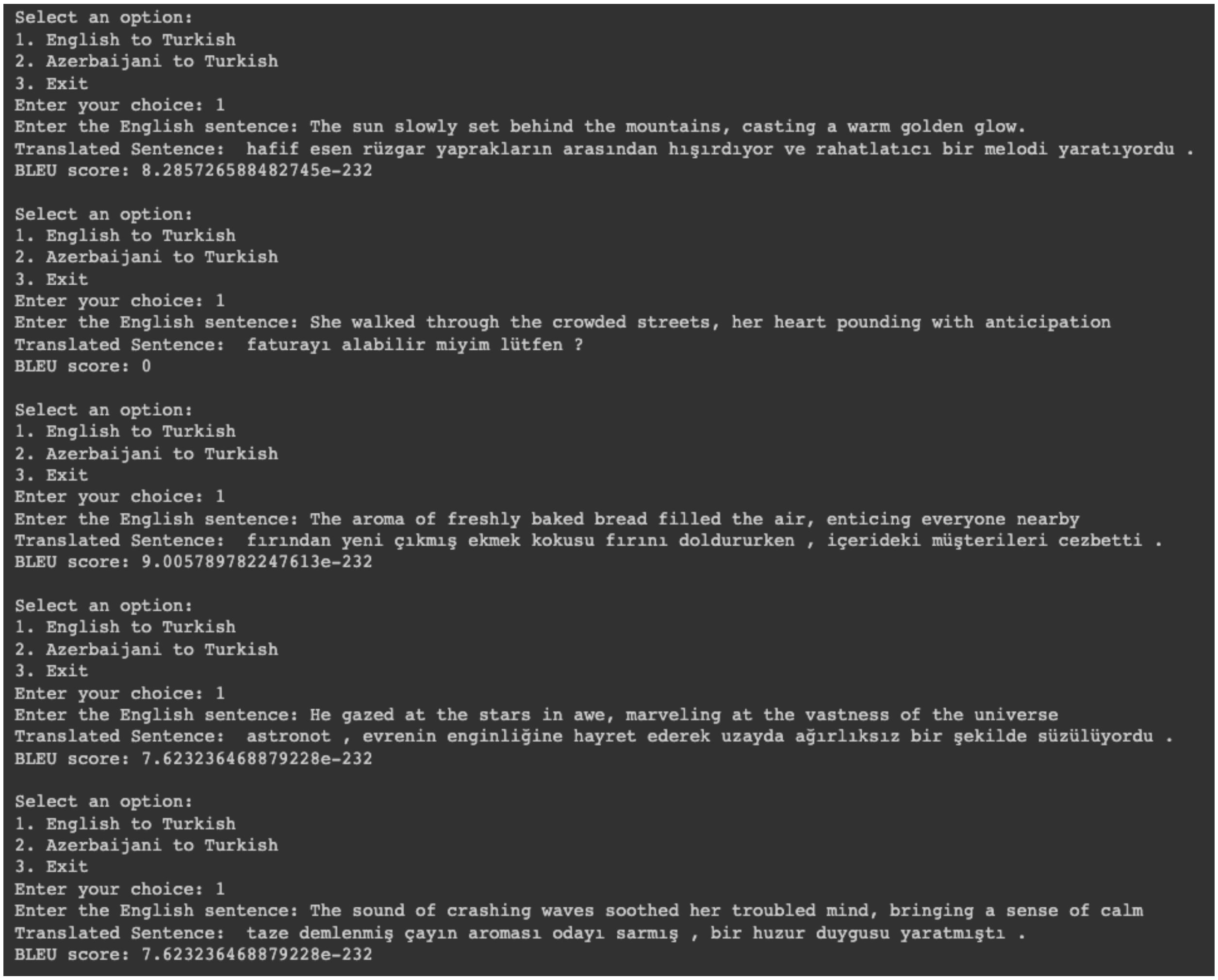
- Lack of Model Training: The system does not involve training any translation models. Integrating machine learning techniques, such as training on parallel corpora, could lead to more accurate and context-aware translations.
- Performance Optimization: The code could be optimized for speed and efficiency, especially for large corpora. Implementing parallel processing or utilizing more efficient algorithms for edit distance calculation could enhance the system's performance.

Conclusion

The Example-Based Machine Translation system implemented in the provided code offers a basic framework for translating sentences between English, Azerbaijani, and Turkish languages. The code performs tokenization, sentence normalization, and utilizes edit distance for similarity calculation. The system's performance can be evaluated based on translation accuracy, speed, and the calculated BLEU scores. While the system has limitations, such as limited language support and reliance on edit distance, it serves as a starting point for further development and improvement in the field of machine translation.

Results

Outputs are given below in Screenshots taken from Google Colab User Interface. The short sentences are providing poor results because probably mostly long sentences which are greater or equal than ten words have been used in English to Turkish corpora. Nevertheless, Azerbaijan BLUE score and translations are poor due to lack of Morphological Analysis tool for Azerbaijani NLP approaches. See below for results.



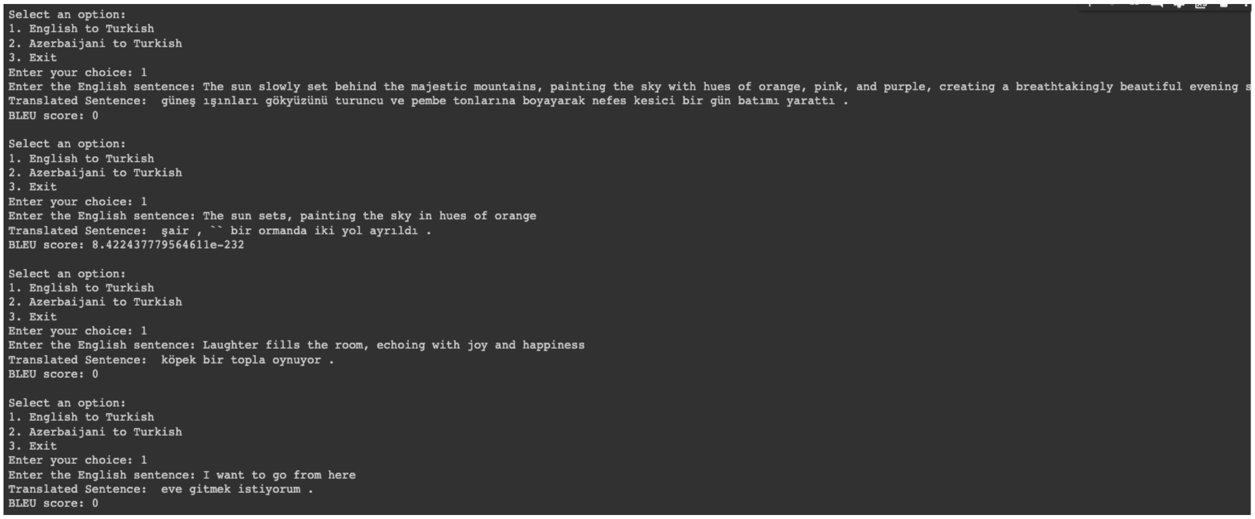
```
Select an option:
1. English to Turkish
2. Azerbaijani to Turkish
3. Exit
Enter your choice: 1
Enter the English sentence: The sun slowly set behind the mountains, casting a warm golden glow.
Translated Sentence:  hafif esen rüzgar yaprakların arasından hışırdıyor ve rahatlatıcı bir melodi yaratıyordu .
BLEU score: 8.285726588482745e-232

Select an option:
1. English to Turkish
2. Azerbaijani to Turkish
3. Exit
Enter your choice: 1
Enter the English sentence: She walked through the crowded streets, her heart pounding with anticipation
Translated Sentence:  faturayı alabilir miyim lütfen ?
BLEU score: 0

Select an option:
1. English to Turkish
2. Azerbaijani to Turkish
3. Exit
Enter your choice: 1
Enter the English sentence: The aroma of freshly baked bread filled the air, enticing everyone nearby
Translated Sentence:  fırından yeni çıkmış ekmek kokusu fırını doldururken , içerideki müşterileri cezbetti .
BLEU score: 9.005789782247613e-232

Select an option:
1. English to Turkish
2. Azerbaijani to Turkish
3. Exit
Enter your choice: 1
Enter the English sentence: He gazed at the stars in awe, marveling at the vastness of the universe
Translated Sentence:  astronot , evrenin enginliğine hayret ederek uzayda ağırlıksız bir şekilde süzülüyordu .
BLEU score: 7.623236468879228e-232

Select an option:
1. English to Turkish
2. Azerbaijani to Turkish
3. Exit
Enter your choice: 1
Enter the English sentence: The sound of crashing waves soothed her troubled mind, bringing a sense of calm
Translated Sentence:  taze demlenmiş çayın aroması odayı sarmış , bir huzur duygusu yaratmıştı .
BLEU score: 7.623236468879228e-232
```

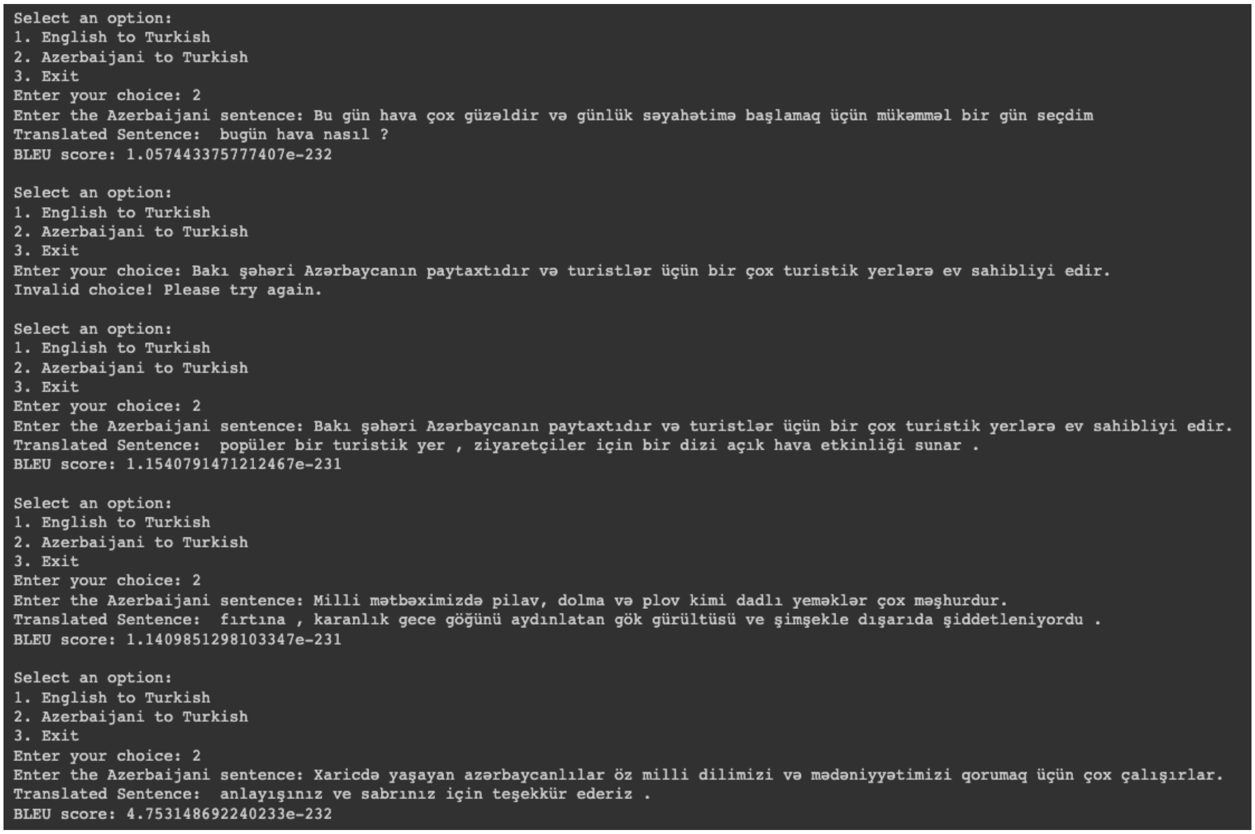


```
Select an option:
1. English to Turkish
2. Azerbaijani to Turkish
3. Exit
Enter your choice: 1
Enter the English sentence: The sun slowly set behind the majestic mountains, painting the sky with hues of orange, pink, and purple, creating a breathtakingly beautiful evening s
Translated Sentence:  güneş ışınları gökyüzünü turuncu ve pembe tonlarına boyayarak nefes kesici bir gün batımı yarattı .
BLEU score: 0

Select an option:
1. English to Turkish
2. Azerbaijani to Turkish
3. Exit
Enter your choice: 1
Enter the English sentence: The sun sets, painting the sky in hues of orange
Translated Sentence:  şair , `` bir ormanda iki yol ayrıldı .
BLEU score: 8.422437779564611e-232

Select an option:
1. English to Turkish
2. Azerbaijani to Turkish
3. Exit
Enter your choice: 1
Enter the English sentence: Laughter fills the room, echoing with joy and happiness
Translated Sentence:  köpek bir topla oynuyor .
BLEU score: 0

Select an option:
1. English to Turkish
2. Azerbaijani to Turkish
3. Exit
Enter your choice: 1
Enter the English sentence: I want to go from here
Translated Sentence:  eve gitmek istiyorum .
BLEU score: 0
```



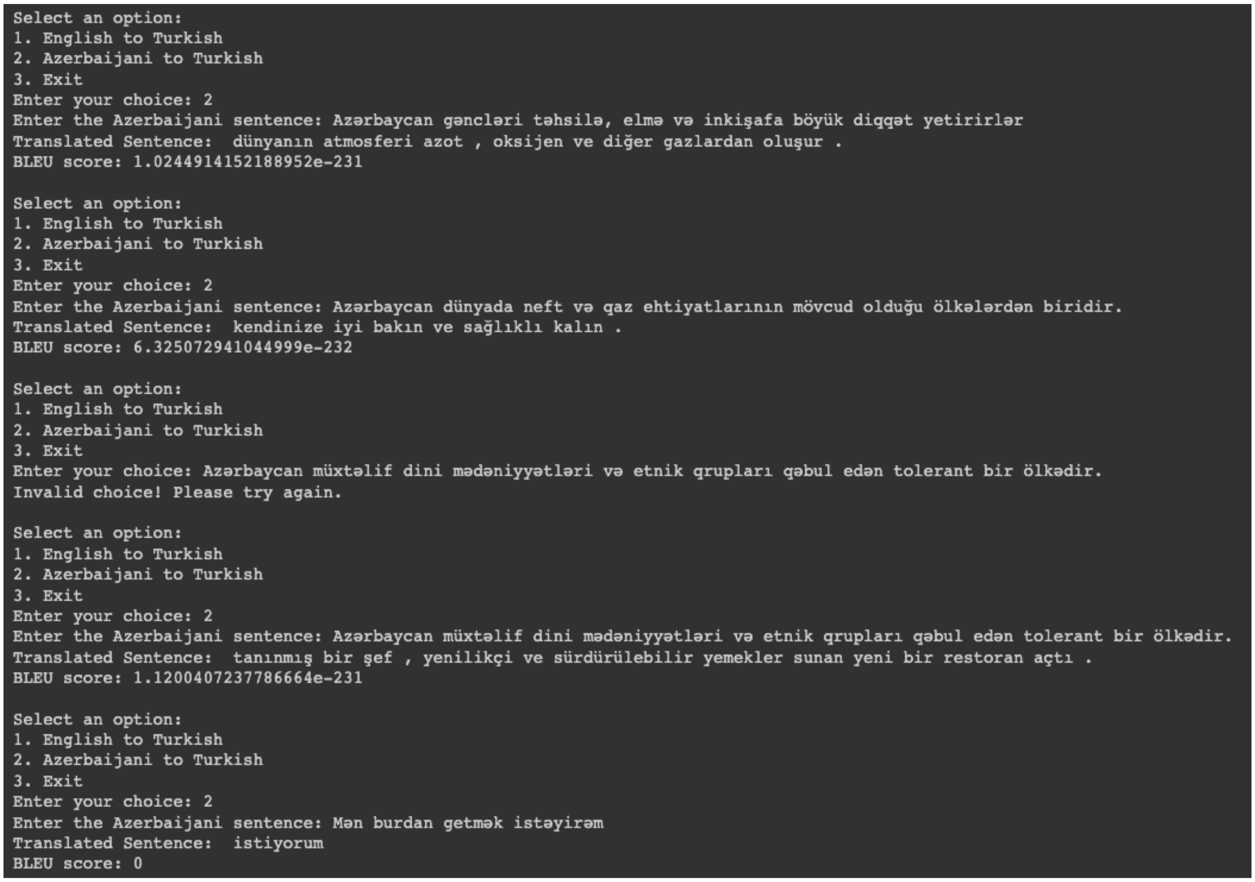
```
Select an option:
1. English to Turkish
2. Azerbaijani to Turkish
3. Exit
Enter your choice: 2
Enter the Azerbaijani sentence: Bu gün hava çox gözəldir və günlük səyahətimə başlamaq üçün mükəmməl bir gün seçdim
Translated Sentence:  bugün hava nasıl ?
BLEU score: 1.057443375777407e-232

Select an option:
1. English to Turkish
2. Azerbaijani to Turkish
3. Exit
Enter your choice: Bakı şəhəri Azərbaycanın paytaxtıdır və turistlər üçün bir çox turistik yerlərə ev sahibliyi edir.
Invalid choice! Please try again.

Select an option:
1. English to Turkish
2. Azerbaijani to Turkish
3. Exit
Enter your choice: 2
Enter the Azerbaijani sentence: Bakı şəhəri Azərbaycanın paytaxtıdır və turistlər üçün bir çox turistik yerlərə ev sahibliyi edir.
Translated Sentence:  popüler bir turistik yer , ziyaretçiler için bir dizi açık hava etkinliği sunar .
BLEU score: 1.1540791471212467e-231

Select an option:
1. English to Turkish
2. Azerbaijani to Turkish
3. Exit
Enter your choice: 2
Enter the Azerbaijani sentence: Milli mətbəximizdə pilav, dolma və plov kimi dadlı yeməklər çox məşhurdur.
Translated Sentence:  fırtına , karanlık gece göğünü aydınlatan gök gürültüsü ve şimşekle dışarıda şiddetleniyordu .
BLEU score: 1.1409851298103347e-231

Select an option:
1. English to Turkish
2. Azerbaijani to Turkish
3. Exit
Enter your choice: 2
Enter the Azerbaijani sentence: Xaricdə yaşayan azərbaycanlılar öz milli dilimizi və mədəniyyətimizi qorumaq üçün çox çalışırlar.
Translated Sentence:  anlayışınız ve sabrınız için teşekkür ederiz .
BLEU score: 4.753148692240233e-232
```



```
Select an option:
1. English to Turkish
2. Azerbaijani to Turkish
3. Exit
Enter your choice: 2
Enter the Azerbaijani sentence: Azərbaycan gəncləri təhsilə, elmə və inkişafa böyük diqqət yetirirlər
Translated Sentence:  dünyanın atmosferi azot , oksijen ve diğer gazlardan oluşur .
BLEU score: 1.0244914152188952e-231

Select an option:
1. English to Turkish
2. Azerbaijani to Turkish
3. Exit
Enter your choice: 2
Enter the Azerbaijani sentence: Azərbaycan dünyada neft və qaz ehtiyatlarının mövcud olduğu ölkələrdən biridir.
Translated Sentence:  kendinize iyi bakın ve sağlıklı kalın .
BLEU score: 6.325072941044999e-232

Select an option:
1. English to Turkish
2. Azerbaijani to Turkish
3. Exit
Enter your choice: Azərbaycan müxtəlif dini mədəniyyətləri və etnik qrupları qəbul edən tolerant bir ölkədir.
Invalid choice! Please try again.

Select an option:
1. English to Turkish
2. Azerbaijani to Turkish
3. Exit
Enter your choice: 2
Enter the Azerbaijani sentence: Azərbaycan müxtəlif dini mədəniyyətləri və etnik qrupları qəbul edən tolerant bir ölkədir.
Translated Sentence:  tanınmış bir şef , yenilikçi ve sürdürülebilir yemekler sunan yeni bir restoran açtı .
BLEU score: 1.1200407237786664e-231

Select an option:
1. English to Turkish
2. Azerbaijani to Turkish
3. Exit
Enter your choice: 2
Enter the Azerbaijani sentence: Mən burdan getmək istəyirəm
Translated Sentence:  istiyorum
BLEU score: 0
```