CS 584-A: Natural Language Processing

Student information

- Full name: Thanapoom Phatthanaphan
- CWID: 20011296

Homework 3

Goals

The goal of HW3 is for you to get hands-on experience of utilizing Seq2Seq model for machine translation task. You will get a deeper understanding of how the input sequence is translated to the output sequence with the Seq2Seq model. The skills of you learnt in this homework will benefit your understanding of a wide range of NLP tasks beyond machine translation. Please feel free to use any packages or libraries in your implementation.

Similar to HW1 and HW2, all questions are open questions and there is no fixed solution. The difference in data selection and processing, parameter initialization, data split, etc., will lead to the differences in predictions and evaluation results. Therefore, during the grading, the specific values in the results are not required. It is important that you focus on implementing and setting up the pipelines of applying these models to solve the tasks.

Task: Machine Translation

1. Data preparation

```
# Install datasets library
!pip install datasets
!pip install transformers
!pip install torchtext
!pip install torch
!pip install tensorflow
!pip install sentencepiece
!pip install sacremoses
!pip install sacrebleu
!pip install ctranslate2
```

```
Requirement already satisfied: MarkupSafe>=2.1.1 in /usr/local/lib/python3.10/dist-packages (from werkzeug>=1.0.1->ten
Requirement already satisfied: pyasn1<0.6.0,>=0.4.6 in /usr/local/lib/python3.10/dist-packages (from pyasn1-modules>=0
Requirement already satisfied: oauthlib>=3.0.0 in /usr/local/lib/python3.10/dist-packages (from requests-oauthlib>=0.7
Requirement already satisfied: sentencepiece in /usr/local/lib/python3.10/dist-packages (0.1.99)
Requirement already satisfied: sacremoses in /usr/local/lib/python3.10/dist-packages (0.1.1)
Requirement already satisfied: regex in /usr/local/lib/python3.10/dist-packages (from sacremoses) (2023.6.3)
Requirement already satisfied: click in /usr/local/lib/python3.10/dist-packages (from sacremoses) (8.1.7)
Requirement already satisfied: joblib in /usr/local/lib/python3.10/dist-packages (from sacremoses) (1.3.2)
Requirement already satisfied: tqdm in /usr/local/lib/python3.10/dist-packages (from sacremoses) (4.66.1)
Requirement already satisfied: sacrebleu in /usr/local/lib/python3.10/dist-packages (2.3.2)
Requirement already satisfied: portalocker in /usr/local/lib/python3.10/dist-packages (from sacrebleu) (2.8.2)
Requirement already satisfied: regex in /usr/local/lib/python3.10/dist-packages (from sacrebleu) (2023.6.3)
Requirement already satisfied: tabulate>=0.8.9 in /usr/local/lib/python3.10/dist-packages (from sacrebleu) (0.9.0)
Requirement already satisfied: numpy>=1.17 in /usr/local/lib/python3.10/dist-packages (from sacrebleu) (1.23.5)
Requirement already satisfied: colorama in /usr/local/lib/python3.10/dist-packages (from sacrebleu) (0.4.6)
Requirement already satisfied: lxml in /usr/local/lib/python3.10/dist-packages (from sacrebleu) (4.9.3)
Requirement already satisfied: ctranslate2 in /usr/local/lib/python3.10/dist-packages (3.21.0)
Requirement already satisfied: numpy in /usr/local/lib/python3.10/dist-packages (from ctranslate2) (1.23.5)
```

```
# Download the Flores200 dataset
from datasets import load_dataset
dataset = load_dataset("Muennighoff/flores200", 'all')
```

```
# Import random library for randomly selecting sentence pairs
import random

# Select Thai as source language and English as target language for translation
source_language = dataset['dev']['sentence_tha_Thai']
target_language = dataset['dev']['sentence_eng_Latn']

# Combine source and target sentences into pairs
sentence_pairs = list(zip(source_language, target_language))

# Randomly select 100 sentence pairs
random.seed(50)
selected_pairs = random.sample(sentence_pairs, 100)

# Define the list for source and target sentences
source_sentences = []
target_sentences = []

# Display the selected sentence pairs
for idx, (source, target) in enumerate(selected_pairs):
    source_sentences.append(source)
    target_sentences.append(target)
    print(f"Pair {idx + 1} - Source: {source}\n         Target: {target}\n")
```

```
Pair 95 - Source: กีฬาฟันดาบสมัยใหม่มีการแข่งขันกันในหลายระดับ ตั้งแต่ระดับนักศึกษามหาวิทยาลัยไปจนถึงระดับมืออาชีพและในกีฬาโอลิมปิก
          Target: The modern sport of fencing is played at many levels, from students learning at a university to profe

Pair 96 - Source: เขาเสียชีวิตที่โอซาก้าในวันอังคาร
          Target: He died in Osaka on Tuesday.

Pair 97 - Source: มีการจัดให้มีการสอบสวนเพื่อสืบสวน
          Target: An inquiry was established to investigate.

Pair 98 - Source: โรคติดต่อคือโรคชนิดหนึ่งที่แพร่เชื้อได้โดยง่ายเมื่ออยู่ใกล้ชิดกับผู้ติดเชื้อ
          Target: A contagious disease is a disease which is easily transmitted by being in the vicinity of an infected

Pair 99 - Source: ก๊าซจะเบาบางลงเมื่อเราเคลื่อนที่ไกลออกมาจากศูนย์กลางของดวงอาทิตย์
          Target: The gas becomes thinner as you go farther from the center of the Sun.

Pair 100 - Source: ปกติแล้วกิจกรรมเหล่านี้จะกินเวลาระหว่างสามถึงหกเดือนและจัดขึ้นในพื้นที่ซึ่งมีขนาดไม่เกิน 50 เฮกตาร์
          Target: These events normally last anywhere between three and six months, and are held on sites no smaller th
```

2. Machine Translation with Seq2Seq model

2.1 OPUS-MT

```
# Import neccessary libraries for OPUS-MT
from transformers import AutoTokenizer, AutoModelForSeq2SeqLM


# load a pre-trained model and tokenizer for translation from Thai to English
opus_tokenizer = AutoTokenizer.from_pretrained('Helsinki-NLP/opus-mt-th-en')
opus_model = AutoModelForSeq2SeqLM.from_pretrained('Helsinki-NLP/opus-mt-th-en')


# Function for translation
def opus_translate_sentence(sentence, model, tokenizer):
    inputs = tokenizer(sentence, return_tensors="pt")
    outputs = model.generate(**inputs)
    translated_sentence = tokenizer.batch_decode(outputs, skip_special_tokens=True)[0]
    return translated_sentence

# Define the list to store the translated sentences
opus_translated_sentences_list = []

# Perform translations on the 100 sentence pairs
for idx, (source_language, target_language) in enumerate(selected_pairs):
    # Translate from source to target
    opus_translated_sentence = opus_translate_sentence(source_language, opus_model, opus_tokenizer)

    # Add the sentence into the list
    opus_translated_sentences_list.append(opus_translated_sentence)

    # Display the results
    print(f"Pair {idx + 1}")
    print(f"Source: {source_language}")
    print(f"Target: {target_language}")
    print(f"Translated: {opus_translated_sentence}\n")
```

```
Source: กีฬาฟันดาบสมัยใหม่มีการแข่งขันกันในหลายระดับ ตั้งแต่ระดับนักศึกษามหาวิทยาลัยไปจนถึงระดับมืออาชีพและในกีฬาโอลิมปิก
Target: The modern sport of fencing is played at many levels, from students learning at a university to professional a
Translated: Modern swordfights have many levels of competition, from university to professional and in the Olympics.

Pair 96
Source: เขาเสียชีวิตที่โอซาก้าในวันอังคาร
Target: He died in Osaka on Tuesday.
Translated: He died in Osaka on Tuesday.

Pair 97
Source: มีการจัดให้มีการสอบสวนเพื่อสืบสวน
Target: An inquiry was established to investigate.
Translated: There's been an investigation.

Pair 98
Source: โรคติดต่อคือโรคชนิดหนึ่งที่แพร่เชื้อได้โดยง่ายเมื่ออยู่ใกล้ชิดกับผู้ติดเชื้อ
Target: A contagious disease is a disease which is easily transmitted by being in the vicinity of an infected person.
Translated: Infection is a disease that can easily spread when it's close to the infected.

Pair 99
Source: ก๊าซจะเบาบางลงเมื่อเราเคลื่อนที่ไกลออกมาจากศูนย์กลางของดวงอาทิตย์
Target: The gas becomes thinner as you go farther from the center of the Sun.
Translated: The gas will be thinr as we move far from the center of the Sun.

Pair 100
Source: ปกติแล้วกิจกรรมเหล่านี้จะกินเวลาระหว่างสามถึงหกเดือนและจัดขึ้นในพื้นที่ซึ่งมีขนาดไม่เกิน 50 เฮกตาร์
Target: These events normally last anywhere between three and six months, and are held on sites no smaller than 50 hec
Translated: These activities usually take between three and six months and hold in an area that is no more than 50 hag
```

```
# This is the code that I just practice to use Ctranslate2, I will set the codes as comments but not run it to show the resul
# However, you can set it as a code and try to run to see the result
# Note: The final result will focus on the above method.
# !wget https://object.pouta.csc.fi/OPUS-MT-models/th-en/opus-2020-01-16.zip
# !unzip opus-2020-01-16.zip
# !ct2-opus-mt-converter --model_dir . --output_dir then_ctranslate2

# import ctranslate2
# import sentencepiece as spm

# sp = spm.SentencePieceProcessor()
# sp.load("source.spm")

# ctranslate2_list = []

# for idx, (source_language, target_language) in enumerate(selected_pairs):
#     source = sp.encode(source_language, out_type=str)
#     translator = ctranslate2.Translator("then_ctranslate2")
#     results = translator.translate_batch([source])
#     output = sp.decode(results[0].hypotheses[0]).replace("_", " ")

#     # Add the sentence into the list
#     ctranslate2_list.append(output)

#     # Display the results
#     print(f"Pair {idx + 1}")
#     print(f"Source: {source_language}")
#     print(f"Target: {target_language}")
#     print(f"Translated: {output}\n")
```

2.2 M2M-100

```
# Import necessary library for M2M-100
from transformers import M2M100ForConditionalGeneration, M2M100Tokenizer


# load a pre-trained model and tokenizer for the M2M100 model
m2m_model = M2M100ForConditionalGeneration.from_pretrained("facebook/m2m100_418M")
m2m_tokenizer = M2M100Tokenizer.from_pretrained("facebook/m2m100_418M")


# Function for translation (Thai to English)
def m2m_translate_sentence(sentence, model, tokenizer):
    tokenizer.src_lang = "th"
    inputs = tokenizer(sentence, return_tensors="pt")
    outputs = model.generate(**inputs, forced_bos_token_id=tokenizer.get_lang_id("en"))
    translated_sentence = tokenizer.batch_decode(outputs, skip_special_tokens=True)[0]
    return translated_sentence

# Define the list to store the length of the translated sentences
m2m_translated_sentences_list = []

# Perform translations on the 100 sentence pairs
```

```
for idx, (source_language, target_language) in enumerate(selected_pairs):
    # Translate from source to target
    m2m_translated_sentence = m2m_translate_sentence(source_language, m2m_model, m2m_tokenizer)

    # Add the sentence into the list
    m2m_translated_sentences_list.append(m2m_translated_sentence)

    # Display the results
    print(f"Pair {idx + 1}")
    print(f"Source: {source_language}")
    print(f"Target: {target_language}")
    print(f"Translated: {m2m_translated_sentence}\n")
```

```
Translated: Many buildings are pretty beautiful and visible. And a view from a height or from a smartly placed window

Pair 90
Source: บรรยากาศมืดสลัวในบริเวณวัดและทิวทัศน์เหนือทะเลสาบโตนเลสาบทำให้การปีนขึ้นสู่เนินเขาเป็นเรื่องที่คุ้มค่า
Target: The gloomy atmosphere of the temple and the view over the Tonle Sap lake make the climb to the hill worthwhile
Translated: The dark atmosphere in the area of the temple and the views on the north of the Boton Lake make climbing t

Pair 91
Source: งานรวมศิลปินครั้งนี้ที่ยังเป็นส่วนหนึ่งของแคมเปญที่จัดโดยศาลาว่าการกรุงบูคาเรสต์ ในการพยายามสร้างภาพลักษณ์ของเมืองหลวงของโรมาเนียให้เป็นมห
Target: The artistic event is also part of a campaign by the Bucharest City Hall that seeks to relaunch the image of t
Translated: This group of artists is also part of a campaign organized by the devil as the Bucharest. In an attempt to

Pair 92
Source: เด็กเหล่านี้มักเผชิญกับปัญหามากมายเพราะพวกเขา "เข้าไปพัวพันกับพฤติกรรมเสี่ยง การต่อยตี และท้าทายเจ้าหน้าที่" เพื่อกระตุ้นสมองของตนเอง เนื่อ
Target: These children tend to get into a lot of trouble, because they "engage in risky behaviors, get into fights, an
Translated: These children are often faced with a lot of problems because they are "in the middle of risk behavior, st

Pair 93
Source: มนุษย์ได้ผลิตและใช้เลนส์เพื่อประโยชน์ในการขยายภาพมานานหลายพันปีแล้ว
Target: Humans have been making and using lenses for magnification for thousands and thousands of years.
Translated: Humans have produced and used lenses for the benefit of image expansion for thousands of years.

Pair 94
Source: การขว้างบูมเมอแรงคือทักษะยอดนิยมที่นักท่องเที่ยวหลายต่อหลายคนต้องการมี
Target: Boomerang throwing is a popular skill that many tourists want to acquire.
Translated: Strong booming is a popular skill that many tourists want to have.

Pair 95
Source: กีฬาฟันดาบสมัยใหม่มีการแข่งขันกันในหลายระดับ ตั้งแต่ระดับนักศึกษามหาวิทยาลัยไปจนถึงระดับมืออาชีพและในกีฬาโอลิมปิก
Target: The modern sport of fencing is played at many levels, from students learning at a university to professional a
Translated: Modern shoe sports are competing on several levels. From college level to professional level and in the Ol

Pair 96
Source: เขาเสียชีวิตที่โอซาก้าในวันอังคาร
Target: He died in Osaka on Tuesday.
Translated: He died in Osaka on Tuesday.

Pair 97
Source: มีการจัดให้มีการสอบสวนเพื่อสืบสวน
Target: An inquiry was established to investigate.
Translated: There is an investigation to explore.

Pair 98
Source: โรคติดต่อคือโรคชนิดหนึ่งที่แพร่เชื้อได้โดยง่ายเมื่ออยู่ใกล้ชิดกับผู้ติดเชื้อ
Target: A contagious disease is a disease which is easily transmitted by being in the vicinity of an infected person.
Translated: Contact disease is a type of disease that is easily transmitted when close to an infected person.

Pair 99
Source: ก๊าซจะเบาบางลงเมื่อเราเคลื่อนที่ไกลออกมาจากศูนย์กลางของดวงอาทิตย์
Target: The gas becomes thinner as you go farther from the center of the Sun.
Translated: Gas becomes thin when we move far from the center of the sun.

Pair 100
Source: ปกติแล้วกิจกรรมเหล่านี้จะกินเวลาระหว่างสามถึงหกเดือนและจัดขึ้นในพื้นที่ซึ่งมีขนาดไม่เกิน 50 เฮกตาร์
Target: These events normally last anywhere between three and six months, and are held on sites no smaller than 50 hec
Translated: Normally, these activities take a period of three to six months and take place in an area of no more than
```

2.3 MBART-50

```
# Import necessary libraries for MBART-50
from transformers import MBartForConditionalGeneration, MBart50TokenizerFast


# load a pre-trained model and tokenizer for the MBART-50 model
mbart_model = MBartForConditionalGeneration.from_pretrained("facebook/mbart-large-50-many-to-many-mmt")
mbart_tokenizer = MBart50TokenizerFast.from_pretrained("facebook/mbart-large-50-many-to-many-mmt")


# Function for translation (Thai to English)
def mbart_translate_sentence(sentence, model, tokenizer):
    tokenizer.src_lang = "th_TH"
    inputs = tokenizer(sentence, return_tensors="pt")
```

```
    outputs = model.generate(**inputs, forced_bos_token_id=tokenizer.lang_code_to_id["en_XX"])
    translated_sentence = tokenizer.batch_decode(outputs, skip_special_tokens=True)[0]
    return translated_sentence

# Define the list to store the length of the translated sentences
mbart_translated_sentences_list = []

# Perform translations on the 100 sentence pairs
for idx, (source_language, target_language) in enumerate(selected_pairs):
    # Translate from source to target
    mbart_translated_sentence = mbart_translate_sentence(source_language, mbart_model, mbart_tokenizer)

    # Add the sentence into the list
    mbart_translated_sentences_list.append(mbart_translated_sentence)

    # Display the results
    print(f"Pair {idx + 1}")
    print(f"Source: {source_language}")
    print(f"Target: {target_language}")
    print(f"Translated: {mbart_translated_sentence}\n")
```

```
Translated: Growing a lot of beautiful landscapes and viewing from high-rise buildings or from smart architecture, may

Pair 90
Source: บรรยากาศมืดสลัวในบริเวณวัดและทิวทัศน์เหนือทะเลสาบโตนเลสาบทำให้การปีนขึ้นสู่เนินเขาเป็นเรื่องที่คุ้มค่า
Target: The gloomy atmosphere of the temple and the view over the Tonle Sap lake make the climb to the hill worthwhile
Translated: The dark atmosphere in the temple area and the views over Lake Tanganyika, the mountain climbing, the moun

Pair 91
Source: งานรวมศิลปินครั้งนี้ที่ยังเป็นส่วนหนึ่งของแคมเปญที่จัดโดยศาลาว่าการกรุงบูคาเรสต์ ในการพยายามสร้างภาพลักษณ์ของเมืองหลวงของโรมาเนียให้เป็นมห
Target: The artistic event is also part of a campaign by the Bucharest City Hall that seeks to relaunch the image of t
Translated: It's a congress of artists from the capital city, organized by the city of Bucharest, to try to paint a po

Pair 92
Source: เด็กเหล่านี้มักเผชิญกับปัญหามากมายเพราะพวกเขา "เข้าไปพัวพันกับพฤติกรรมเสี่ยง การต่อยตี และท้าทายเจ้าหน้าที่" เพื่อกระตุ้นสมองของตนเอง เนื่อ
Target: These children tend to get into a lot of trouble, because they "engage in risky behaviors, get into fights, an
Translated: The councils often have a lot of problems because they're dealing with their own behavior, their own fight

Pair 93
Source: มนุษย์ได้ผลิตและใช้เลนส์เพื่อประโยชน์ในการขยายภาพมานานหลายพันปีแล้ว
Target: Humans have been making and using lenses for magnification for thousands and thousands of years.
Translated: The industry has been producing and expanding for thousands of years.

Pair 94
Source: การขว้างบูมเมอแรงคือทักษะยอดนิยมที่นักท่องเที่ยวหลายต่อหลายคนต้องการมี
Target: Boomerang throwing is a popular skill that many tourists want to acquire.
Translated: Strong bowel movement is a very popular skill among many people.

Pair 95
Source: กีฬาฟันดาบสมัยใหม่มีการแข่งขันกันในหลายระดับ ตั้งแต่ระดับนักศึกษามหาวิทยาลัยไปจนถึงระดับมืออาชีพและในกีฬาโอลิมปิก
Target: The modern sport of fencing is played at many levels, from students learning at a university to professional a
Translated: Ice hockey in the old days was played at a variety of levels, from university level to professional level

Pair 96
Source: เขาเสียชีวิตที่โอซาก้าในวันอังคาร
Target: He died in Osaka on Tuesday.
Translated: He died in Ossetia on Tuesday.

Pair 97
Source: มีการจัดให้มีการสอบสวนเพื่อสืบสวน
Target: An inquiry was established to investigate.
Translated: There was an investigation.

Pair 98
Source: โรคติดต่อคือโรคชนิดหนึ่งที่แพร่เชื้อได้โดยง่ายเมื่ออยู่ใกล้ชิดกับผู้ติดเชื้อ
Target: A contagious disease is a disease which is easily transmitted by being in the vicinity of an infected person.
Translated: Diseases are diseases that are caused by a combination of diseases and diseases.

Pair 99
Source: ก๊าซจะเบาบางลงเมื่อเราเคลื่อนที่ไกลออกมาจากศูนย์กลางของดวงอาทิตย์
Target: The gas becomes thinner as you go farther from the center of the Sun.
Translated: It will lighten us away from the center of the sun.

Pair 100
Source: ปกติแล้วกิจกรรมเหล่านี้จะกินเวลาระหว่างสามถึงหกเดือนและจัดขึ้นในพื้นที่ซึ่งมีขนาดไม่เกิน 50 เฮกตาร์
Target: These events normally last anywhere between three and six months, and are held on sites no smaller than 50 hec
Translated: Typically, the conferences are three to six months long, and the conferences are 50 hectares in size.
```

2.4 Data statistics

Displaying the data statistics of the 100 sampled sentences

```
# Get the minimum length of the sentences
def min_length(sentences_list):
```

```
    min_length = len(sentences_list[0])

    for sentence in sentences_list:
        if min_length > len(sentence):
            min_length = len(sentence)

    return min_length

# Get the maximum length of the sentences
def max_length(sentences_list):

    max_length = len(sentences_list[0])

    for sentence in sentences_list:
        if max_length < len(sentence):
            max_length = len(sentence)

    return max_length

# Get the average length of the sentences
def avg_length(sentences_list):

    sum = 0

    for sentence in sentences_list:
        sum += len(sentence)

    avg_length = sum // len(sentences_list)

    return avg_length

import pandas as pd

# Define the columns and rows of the table containing the statistics of the source and target sentences
source_target_sentences_length_table = {
    'Sentences' : ['Source sentences',
                   'Target sentences'],
    'Minimum length of the sentences' : [min_length(source_sentences),
                                         min_length(target_sentences)],
    'Average length of the sentences' : [avg_length(source_sentences),
                                         avg_length(target_sentences)],
    'Maximum length of the sentences' : [max_length(source_sentences),
                                         max_length(target_sentences)],
}

# Create the dataframe of the statistics of the source and target sentences
source_target_sentences_statistics_table = pd.DataFrame(source_target_sentences_length_table)

# Define the columns and rows of the table containing the statistics of the translated sentences
translated_sentences_length_table = {
    'Models' : ['OPUS_MT',
                'M2M-100',
                'MBART-50'],
    'Minimum length of the translated sentences' : [min_length(opus_translated_sentences_list),
                                                    min_length(m2m_translated_sentences_list),
                                                    min_length(mbart_translated_sentences_list)],
    'Average length of the translated sentences' : [avg_length(opus_translated_sentences_list),
                                                    avg_length(m2m_translated_sentences_list),
                                                    avg_length(mbart_translated_sentences_list)],
    'Maximum length of the translated sentences' : [max_length(opus_translated_sentences_list),
                                                    max_length(m2m_translated_sentences_list),
                                                    max_length(mbart_translated_sentences_list)],
}

# Create the dataframe of the statistics of the translated sentences
translated_sentences_statistics_table = pd.DataFrame(translated_sentences_length_table)


source_target_sentences_statistics_table
```

	Sentences	Minimum length of the sentences	Average length of the sentences	Maximum length of the sentences
0	Source sentences	33	126	287

```
translated_sentences_statistics_table
```

Models	Minimum length of the translated sentences	Average length of the translated sentences	Maximum length of the translated sentences

3. Results analysis and evaluation

MBART-

3.1 Results of each model

```
import nltk
nltk.download('punkt')
from nltk.tokenize import word_tokenize
from nltk.translate import bleu_score

target_ref = [[word_tokenize(sentence) for sentence in target_sentences] for reference in target_sentences]

# Perform evaluation for OPUS-MT model
# Tokenize the sentences using NLTK's word_tokenize
opus_predictions = [word_tokenize(sentence) for sentence in opus_translated_sentences_list]
# Compute BLEU scores
opus_bleu = bleu_score.corpus_bleu(target_ref, opus_predictions)
print("The BLEU score of OPUS-MT model:")
print(opus_bleu)

# Perform evaluation for M2M-100 model
m2m_predictions = [word_tokenize(sentence) for sentence in m2m_translated_sentences_list]
m2m_bleu = bleu_score.corpus_bleu(target_ref, m2m_predictions)
print("\nThe BLEU score of M2M-100 model:")
print(m2m_bleu)

# Perform evaluation for MBART-50 model
mbart_predictions = [word_tokenize(sentence) for sentence in mbart_translated_sentences_list]
mbart_bleu = bleu_score.corpus_bleu(target_ref, mbart_predictions)
print("\nThe BLEU score of MBART-50 model:")
print(mbart_bleu)
```

```
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
The BLEU score of OPUS-MT model:
0.25161329155896006

The BLEU score of M2M-100 model:
0.24422825751240484

The BLEU score of MBART-50 model:
0.13185801454648585
```

3.2 Results discussion

- **OPUS-MT Model**
  BLEU Score: 0.25, the OPUS-MT model provided the highest BLEU score among the three models, indicating better alignment with the reference translations. This model may have been well-trained on the specific language pair, resulting in more accurate translations.
- **M2M-100 Model**
  BLEU Score: 0.24, the M2M-100 model also performed well but slightly lower than OPUS-MT. This model may have faced challenges in certain language nuances or domain-specific differences compared to the evaluation set.
- **MBART-50 Model**
  BLEU Score: 0.13, the MBART-50 model provided the lowest BLEU score. This model may not be as well-suited for this specific language pair or may require additional fine-tuning.

3.3 Examples of two data samples

```
# Gett the data samples
source_two_samples = source_sentences[0:2]
target_two_samples = target_sentences[0:2]
opus_two_samples = opus_translated_sentences_list[0:2]
m2m_two_samples = m2m_translated_sentences_list[0:2]
mbart_two_samples = mbart_translated_sentences_list[0:2]

# Display the sampled sentences
print("The first sample")
print("Source:", source_two_samples[0])
print("Target:", target_two_samples[0])
print("OPUS-MT translated:", opus_two_samples[0])
print("M2M-100 translated:", m2m_two_samples[0])
print("MBART-50 translated:", mbart_two_samples[0])
```

```
print("\nThe second sample")
print("Source:", source_two_samples[1])
print("Target:", target_two_samples[1])
print("OPUS-MT translated:", opus_two_samples[1])
print("M2M-100 translated:", m2m_two_samples[1])
print("MBART-50 translated:", mbart_two_samples[1])
```

```
The first sample
Source: นี่ไม่ใช่กลุ่มอาการขาดทักษะด้านการเรียนรู้ แต่เป็นความผิดปกติทางการเรียนรู้ โรคนี้ "ส่งผลกระทบต่อเด็กร้อยละ 3 ถึง 5 ซึ่งอาจเป็นเด็กอเมริกันมากถึง
Target: It is not a learning disability, it is a learning disorder; it "affects 3 to 5 percent of all children, perhaps
OPUS-MT translated: This isn't a group of learning disability, it's a learning disorder, it's an effect on children from
M2M-100 translated: This is not a group of learning failure, but a learning disorder. This disease "affects 3 to 5 per c
MBART-50 translated: We're talking about skills shortages, diseases, diseases that affect three to five percent of Ameri

The second sample
Source: ผู้ที่ไม่คุ้นเคยกับศัพท์เฉพาะทางการแพทย์คิดว่า คำว่าติดเชื้อและโรคติดต่อนั้นมีความหมายต่างกัน
Target: For those unfamiliar with medical jargon, the words infectious and contagious have distinct meanings.
OPUS-MT translated: Those who aren't familiar with medical terms think that infection and infectious diseases have diffe
M2M-100 translated: People who are not familiar with the specific medical term think that the word infection and contact
MBART-50 translated: It used to be synonymous with specialized medicines, and diseases were synonymous.
```

From the output of each model translating the two sampled sentences,

Example 1:

- OPUS-MT: The translation is accurate but includes some inaccuracies in conveying the precise meaning of the source. It mentions "an effect on children from three to five" instead of the intended "affects 3 to 5 percent of all children."
- M2M-100: The translation captures the essence but introduces a few errors, such as "learning failure" instead of "learning disability" and "3 to 5 per child" instead of "3 to 5 percent of all children."
- MBART-50: The translation is quite different from the reference and introduces unrelated terms like "skills shortages." It fails to accurately convey the intended meaning.

Example 2:

- OPUS-MT: The translation provides a relatively accurate translation, capturing the intended meaning effectively.
- M2M-100: The translation provides minor errors but conveys the main idea effectively.
- MBART-50: The translation provides a translation that deviates significantly from the target meaning and provides unrelated terms.

Conclusion

- OPUS-MT: Generally provides more accurate translations but may have some inaccuracies.
- M2M-100: Performs reasonably well, capturing the main ideas with occasional errors.
- MBART-50: Shows more significant divergence from the intended meaning and introduces unrelated terms.