Exercise 3

Q1: How would you design a data pipeline for a Machine Translation system? (e.g. necessary steps, main challenges, etc.)

Answer: Designing a data pipeline for a Machine Translation system involves major steps like below: The main challenge is to ensure the quality of the training data. Here are the necessary steps in more detail:

(i) Data acquisition: Collecting parallel data corpus from various sources, such as publicly available datasets, web crawling, and professional translators and transcribing the audio recordings.

(ii) Data Cleaning: This helps in Removing the noisy data sentences with too many errors or untranslated words, and filtering out unreliable sources.

(iii) Sentence Alignment: Aligning the sentences in the source and target languages so that each source sentence corresponds to its translation in the target language using some tools like giza++.

(iv) Data Pre-processing: Tokenizing, lowercasing, stopword removal, stemming, lemmatization and normalizing the text to make it suitable for the neural network model.

(v) Training: Training the neural network model (like Encoder-Decoder Transformer model, pre-trained models like MT5, mBART etc.) on the preprocessed data to learn the mapping between the source and target languages.

(vi) Evaluation: Evaluating the performance of the system on a test dataset to measure its accuracy, fluency, and other metrics.

Major challenges: The major challenge is to ensure the quality of the training data, which directly affects the quality of the translation output. To address this challenge, it is important to have a diverse and representative set of training data that covers a wide range of domains and language varieties. The data collection and pre-process should be done in a very keen manner.

Q2: What would you do to collect new and good quality data from the web? Assume you want to use them to train a neural model for Machine Translation.

Answer: To collect the good and high quality data from the web, several techniques can be used, including web crawling, scraping, and crowd-sourcing. It is important to ensure the quality of the data by applying several filters and quality checks as below:

(i) Identify the websites and domains that contain high-quality parallel text for a particular domain and language.

(ii) Keenly scrap the web pages to extract the text and align the translations with a corresponding sentence accordingly.

(iii) Apply the detailed methods of quality check, such as checking if the parallel sentence corpora is syntactically and semantically correct(grammaticality), fluency, and consistency with other translations.

(iv) Applying machine learning and statistical models to filter out low-quality data and improve the accuracy of the final dataset.

(v) Finally, use the collected and pre-processed and quality checked data to train a neural machine translation model, and evaluate its performance on the test set.

Q3: Suppose that low quality translations created by the system are post-edited by professional translators. How would you use this process to monitor the quality of the Machine Translation system?

Answer: Some of the steps and usage of this process to monitor the quality of the Machine Translation system.

(i) Collect the output of the machine translation system and give it to professional translators (annotators) for post-editing.

(ii) Measure the time and effort required for post-editing and compare it with the effort required for translating from scratch.

(iii) Evaluate the quality of the post-edited translations using several metrics, such as BLEU, TER, WER or METEOR.

1. TER: Translated Error Rate determines the amount of Post-Editing required for machine translation jobs. An automatic evaluation metric that calculates the number of edits required to change a machine translation output into one of the references.

2. WER: Word Error Rate computes the minimum Edit Distance between the human-generated sentence and the machine-predicted sentence.

(iv) Analyze the errors made by the system and identify the patterns that need to be improved.

(v) Use this feedback to retrain the machine translation system on the corrected data and improve its quality.
