First Semester 2024-25: CS F429 (NLP) Assignment 2
Deadline: 17th November, 2024, 11:55 PM

Problem Statement

The objective of this assignment is to develop a machine translation system capable of translating between multiple languages **without using parallel corpora (datasets where there are source and target sentence pairs) for training**. The system must be designed to take in a monolingual corpora (any dataset with just a single language) in various languages as input and learn to translate between language pairs. The system will be evaluated using parallel corpora, but parallel corpora cannot be used in the training process.

The assignment shall be done in groups of **at most 3 people**.

Requirements

1. **Input:**
    a. The training data consists exclusively of **monolingual corpora** in multiple languages. These corpora must be independently found online and should not contain any sentences from parallel source-target pairs (i.e., in your report you should mention where you got the corpora from).
    b. Students are responsible for sourcing monolingual data in at least three different languages, ensuring that no sentence-level alignments exist between these corpora.
    c. Choice of languages is up to the group
2. **Testing:**
    a. A separate **parallel corpus** for the chosen language pairs must be found and used **only for testing**.
    b. This test set will be used to report evaluation metrics like **BLEU, METEOR, or TER scores** to quantify the translation performance of the system.
3. **Constraints:**
    a. The use of pre-trained models or fine-tuning on pre-existing models is allowed, but the final system must include new models trained using the provided monolingual corpora. Simple prompt engineering or single-model fine-tuning is insufficient.
    b. Pre-trained embeddings (such as word embeddings from multilingual models) may be used to initialize the models, but no parallel data can be used for the actual training process.
4. **Submission Requirements:**
    a. A **command-line interface (CLI)** capable of taking monolingual corpora in different languages and training models that can translate between the given languages (and/or) being able to run inference in the command line itself.
    b. A detailed **report** outlining the methods used, experiments conducted, and results obtained, including BLEU scores or other relevant metrics.
    c. A description of how the parallel test sets were obtained and a demonstration that they were independently sourced.
    d. **Code and documentation** for running the system, including instructions on how to input new corpora and evaluate translations.

Evaluation

1. **Evaluation Criteria:**
    a. **Effectiveness of the Translation System**: How well the system performs on translating between language pairs using metrics like BLEU, METEOR, TER etc.
    b. **Reproducibility**: Whether the system can be easily executed using new monolingual corpora inputs.
2. **Marks Distribution:**
    a. **Total:** 50M
    b. Split:
        i. Training procedure (code and explanation/viva): 22M
        ii. CLI tool (code, features and usability): 8M
        iii. Effectiveness of tool (translation metrics and real time inference): 10M (5+5)
        iv. Report and project documentation: 10M