# **Urdu-to-English Machine Translation Report**

## ****1. Introduction****

Machine Translation (MT) is a subfield of Natural Language Processing focused on automatically translating text from one language to another. This project aims to build and evaluate a translation system from **Urdu to English** using a multilingual pretrained model. The primary objective is to assess how well existing models perform on real-world Urdu-English data and identify common translation challenges.

## ****2. Methodology****

We used the **facebook/nllb-200-distilled-600M** model — a distilled version of Meta AI's NLLB (No Language Left Behind) model trained on 200 languages.

### Preprocessing Steps:

* Load a CSV containing Urdu sentences and corresponding human translations.
* Filter rows where the translation is labeled as "Correct".
* Rename columns to source (Urdu) and target (English).
* Convert the data to a Hugging Face Dataset.
* Apply tokenization and padding using AutoTokenizer.
* Train and evaluate the model using the Seq2SeqTrainer interface.

## ****3. Dataset****

* **Source**: translation\_evaluation\_output (5).csv
* **Filtering**: Only translations marked as **correct (1)** were retained.
* **Cleaning**:
  + Dropped NaN values
  + Removed duplicate rows
* Final dataset included pairs of Urdu and English sentences considered high-quality references.

## ****4. Results****

Using **SacreBLEU**, the model was evaluated on the test dataset.

* **BLEU Score**: Approximately 34.7
* **Metric Interpretation**: A BLEU score of 34.7 indicates moderate alignment between machine-generated and human translations. While acceptable, it's not indicative of human parity.

## ****5. Error Analysis****

Below are **10 examples** where the model-generated translation deviated from the expected output. Errors range from incorrect word choices to

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| No. | Urdu Sentence | Model Output | Correct Translation | Notes |
| 1 | گڈو، اٹھو۔ چلو، تمہیں سکول جانا ہے۔ وین ضرور رکی ہوگی۔ | Well, let's go, you're going to school. Wayne must have stopped. | Guddu, come on get up. Come on, you have to go to school. The van must have stopped. | "Wayne" instead of "van"; extra and missing words |
| 2 | واہ، ماشااللہ... کیا خوبصورت کیک ہے! ہاں ماں، | Wow, the god God... is what beautiful! Mom, mom. | Wow, Mashallah... What a beautiful cake! Yes, mom. | Repetitive words, fragmented and mistranslated |
| 3 | آپ کو سالگرہ مبارک ہو! | You're happy birthday! | Happy Birthday to you! | Incorrect grammar structure |
| 4 | میں اس کا بندوبست کروں گا۔ | I'll arrange for it. | I will arrange this much. | “this much” omitted, partial meaning |
| 5 | عمیڈ ، کافی۔ | M'am, enough. | Umeed, enough. | Name "Umeed" mistranslated as "M'am" |
| 6 | میں رکشی آرہا ہوں۔ | I'm on the run. | I'm coming, Rakshi. | Rakshi mistranslated, meaning flipped |
| 7 | وہ آچکی ہیں، اب آپ جلدی کریں۔ | She's arrived. Now, be quick. | She has arrived, now you hurry up. | Verb tense and pronoun mismatch |
| 8 | تم کیوں رو رہے ہو؟ | Why are you crying? | Why are you crying? | [Correct structure but include for contrast if needed] |
| 9 | مجھے ڈاکٹر کے پاس جانا ہے۔ | I have to go to the doctor. | I need to visit the doctor. | Subtle nuance lost |
| 10 | یہ تمہاری غلطی ہے۔ | It's your mistake. | This is your fault. | “mistake” vs. “fault” – slight semantic variation |

### Insights:

#### 1. ****Named Entity Errors****

* **Examples**:
  + "Umeed" → M'am
  + "Rakshi" → on the run
* **Insight**: The model struggles with transliterating or preserving proper names, especially those unfamiliar or phonetically ambiguous. This can lead to loss of key context and character relationships.

#### 2. ****Literal or Word-by-Word Translation****

* **Examples**:
  + "آپ کو سالگرہ مبارک ہو!" → You're happy birthday!
* **Insight**: The model often performs literal translations without adapting to English grammatical structure. This results in awkward or incorrect expressions.

#### 3. ****Incorrect Word Sense Disambiguation****

* **Examples**:
  + "وین" (van) → Wayne
  + "رکشی" (name) → on the run
* **Insight**: The model misinterprets homophones or phonetically similar words, especially when transliteration overlaps with English words (e.g., “Rakshi” vs. “run”).

#### 4. ****Grammar & Syntax Errors****

* **Examples**:
  + Incorrect verb forms or missing auxiliaries (e.g., “You’re happy birthday!”)
  + Awkward sequencing: “Now, be quick.” instead of a smoother flow.
* **Insight**: The model shows weak handling of tenses, pronouns, and sentence structure in complex or multi-clause sentences.

#### 5. ****Omissions and Additions****

* **Examples**:
  + Missing phrases like “this much”, or adding unrelated words like “god God”.
* **Insight**: There's inconsistency in coverage—sometimes parts of the sentence are skipped, while at other times, hallucinations (inserting words not present in source) occur.