

NMT System with Automatic Evaluation

Project Report: Part 1

1. Application Overview

This application is a specialized Neural Machine Translation (NMT) tool designed for the healthcare domain. It leverages the **Transformer architecture** to translate medical text from English to French and provides real-time quality evaluation using the **BLEU (Bilingual Evaluation Understudy)** metric.

2. Design Choices & Integration

2.1 Model Selection

- **Architecture:** Transformer-based **MarianMT** (via [Helsinki-NLP](#)).
- **Reasoning:** Unlike older RNN/LSTM models, the Transformer's **Self-Attention** mechanism allows it to maintain the context of long medical sentences, ensuring that subjects and their respective clinical findings remain linked across the translation.
- **Library:** Built using [transformers](#) for the backend and [streamlit](#) for a responsive, browser-based UI.

2.2 Evaluation Strategy

- **Metric:** [sacrebleu](#) was integrated to provide a standardized BLEU score.
 - **N-Gram Precision:** The system calculates 1-gram (word accuracy) through 4-gram (sentence fluency) scores to give a granular view of translation quality.
 - **Multiple References:** The application supports multiple reference inputs to account for medical synonyms (e.g., "Physician" vs. "Doctor").
-

3. Implementation Flow

1. **Input:** User enters medical source text and one or more reference translations.
2. **Tokenization:** Text is broken into subwords using **SentencePiece**, handling complex medical terms.
3. **Inference:** The Transformer model generates the target translation.

4. **Scoring:** The output is compared to the reference(s), calculating the **Brevity Penalty** and **N-gram precision**.
-

4. Challenges Faced

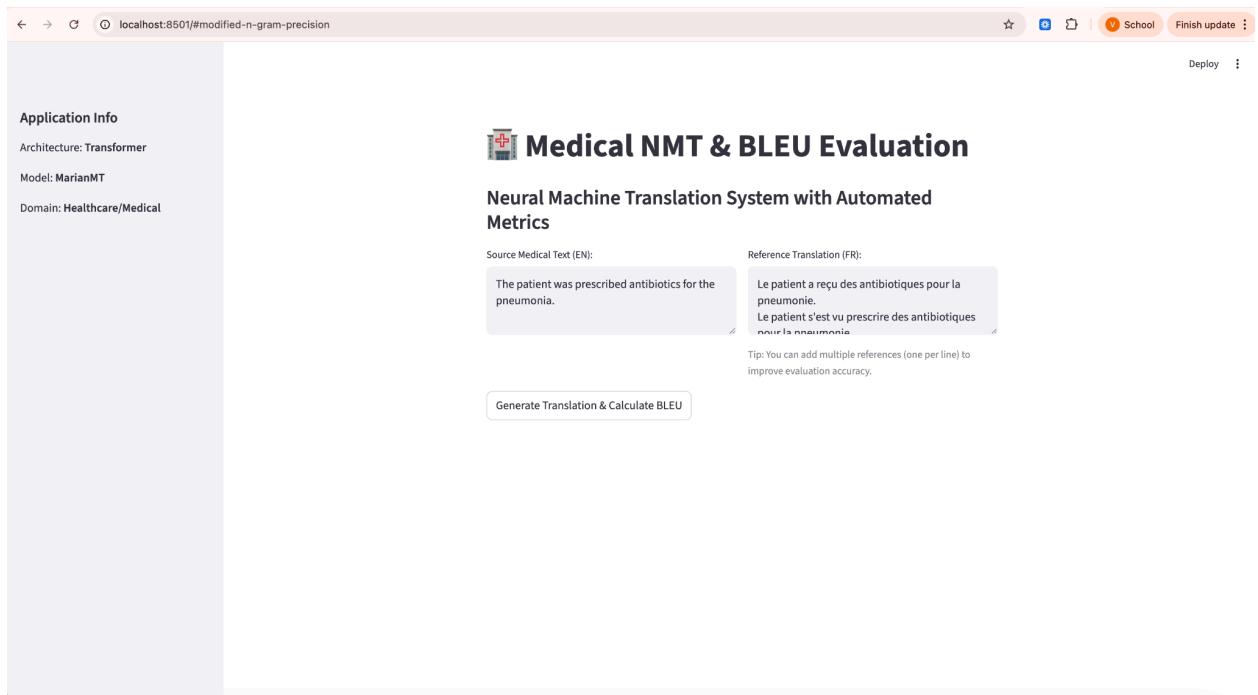
- **Medical Jargon:** Handling specialized terms like “*myocardial infarction*” required a model with robust subword tokenization to avoid “Unknown” token errors.
 - **Lexical Sensitivity:** BLEU penalizes synonyms. To solve this, the UI was designed to accept multiple valid reference strings to ensure a fair evaluation of the NMT engine.
-

5. Application Screenshots & Results

5.1 User Interface Layout

[Screenshot 1: Main UI]

Description: This screenshot shows the input fields for source text and the multi-line reference box.



5.2 Translation & BLEU Score

[Screenshot 2: Output and Metric]

Description: Displays the generated French translation and the final BLEU score metric.

The screenshot shows a web-based application for medical NMT and BLEU evaluation. On the left, there's a sidebar titled "Application Info" with details: Architecture: Transformer, Model: MarianMT, and Domain: Healthcare/Medical. At the top right, there are "Deploy" and three-dot buttons. The main area has a title "Medical NMT & BLEU Evaluation" with a red cross icon. Below it is a sub-section titled "Neural Machine Translation System with Automated Metrics". It contains two text input fields: "Source Medical Text (EN)" with the text "The patient was prescribed antibiotics for the pneumonia.", and "Reference Translation (FR)" with the text "Le patient a reçu des antibiotiques pour la pneumonie." followed by "Le patient s'est vu prescrire des antibiotiques pour la pneumonie.". A tip below says "Tip: You can add multiple references (one per line) to improve evaluation accuracy." A "Generate Translation & Calculate BLEU" button is at the bottom of this section. Below this is a "Results" section with a blue header bar containing "NMT Output: Le patient a reçu des antibiotiques pour la pneumonie.". Under "Results", the "Total BLEU Score" is listed as "100.00".

5.3 N-Gram Precision Table

[Screenshot 3: Evaluation Table]

Description: A detailed view of the modified n-gram precision (1-gram to 4-gram), showing the accuracy of individual words versus full phrases.

This screenshot shows the same application interface as Screenshot 2, but with a focus on the "Modified N-Gram Precision" table. The sidebar and top sections are identical. The "Results" section now includes a table titled "Modified N-Gram Precision" with four rows. The table has columns for "N-Gram Type" and "Precision (%)".

	N-Gram Type	Precision (%)
0	1-Gram (Unigram)	100.00%
1	2-Gram (Bigram)	100.00%
2	3-Gram (Trigram)	100.00%
3	4-Gram (4-gram)	100.00%

At the bottom of the table is a link "View Evaluation Details (Brevity Penalty & Stats)".