Approach and Results

LLM-Based Translation Pipeline with Glossary Retrieval

Author: Shivaji Gaikwad

Date: October 2025

1. Introduction

Machine translation powered by Large Language Models (LLMs) provides high-quality translations, but ensuring domain-specific term consistency remains a challenge. This project demonstrates a glossary-enhanced translation pipeline that uses a glossary embedded in a vector store for retrieval and application during translation. The pipeline compares baseline translations with glossary-enhanced translations to measure term adherence and translation quality.

2. Objective

- Implement a translation pipeline that integrates glossary retrieval.
- Compare translation quality with and without glossary application.
- Evaluate results across three target languages: French (FR), Italian (IT), and Japanese (JP).
- Demonstrate improvements in term consistency and translation accuracy.

3. Approach

3.1 Glossary Preparation

- A glossary file (glossary.csv) was created containing terms, translations, target languages, notes, and tags.
- Example entries include proper nouns (DNT – Do Not Translate), roles, tools, processes, and strategies.

3.2 Embedding & Vector Store

- Glossary terms were embedded using OpenAI embeddings and stored in ChromaDB.
- This enables semantic retrieval of relevant glossary terms given an input segment.

3.3 Translation Process

- Step 1 — Baseline Translation

A plain LLM prompt translating an English sentence to the target language without glossary context.

- Step 2 — Glossary Retrieval
Relevant glossary entries retrieved from ChromaDB using the source sentence as query.
Retrieved entries filtered by target language.

- Step 3 — Glossary-Enhanced Translation
LLM prompt extended with retrieved glossary terms and specific rules:
1. Apply glossary terms exactly.
2. Do not translate DNT entries.
3. Preserve grammar, punctuation, and casing.

4. Results

Example Output Table:

Source Sentence	Baseline Translation	Glossary-Enhanced Translation	Retrieved Terms
The localization engineer prepared a translation kit.	Le ingénieur de localisation a préparé un kit de traduction.	L'ingénieur de localisation a préparé un kit de traduction.	- "translation kit" → "kit de traduction"
Our vendor strategy ensures quality and efficiency.	Notre stratégie de fournisseurs garantit la qualité.	Notre stratégie fournisseur garantit qualité et efficacité.	- "vendor strategy" → "stratégie fournisseur" - "hybrid model" → "modèle hybride"
We measure performance with KPIs.	Nous mesurons la performance avec des KPI.	Nous mesurons la performance avec des indicateurs clé de performance.	- "KPI" → "indicateur clé de performance" - "parser" → "analyseur"

5. Observations

- Glossary-enhanced translations consistently adhered to term definitions.
- Proper nouns and domain-specific terms were preserved accurately.
- Fluency and adequacy were maintained or improved.
- The glossary mechanism adds minimal computational overhead while significantly improving consistency.

6. Conclusion

The LLM-Based Glossary Retrieval Pipeline demonstrated the value of combining large language models with domain-specific term retrieval. This approach enhances translation consistency, particularly in specialized domains, without sacrificing fluency or meaning.

7. Future Work

- Automate glossary updates from domain corpora.
- Implement automatic term-adherence scoring (BLEU, COMET).
- Add a web interface for live glossary-enhanced translation.
- Expand language support.

8. Deliverables

- translation_pipeline.ipynb — Jupyter notebook implementing the pipeline.
- glossary.csv — Glossary dataset.
- translation_results.csv — Output results.
- README.md — Documentation.
- Approach_and_Results.pdf — This document.
- demo_video.mp4 — Demonstration.