Howard University-AI4PC at SemEval-2025 Task 2: Improving Machine Translation With Context-Aware Entity-Only Pre-translations with GPT4o

Saurav K. Aryal and **Jabez Agyemang-Prempeh**
EECS, Howard University
Washington, DC 20059, USA
https://howard.edu/
saurav.aryal@howard.edu, jabez.agyemang-prem@bison.howard.edu

Abstract

This paper presents our work on a 3-Step GPT translation system developed for SemEval-2025 Task 2 to enhance the translation of named entities within machine translation. Our approach integrates (1) entity extraction via wikidata, (2) GPT-based refinement of entity translations, and (3) final context-aware GPT translation. Results from the original dataset of six languages show significant improvements in the handling of named entities compared to direct GPT-based translation baselines. We further discuss replicability, observed challenges, and outline future research directions.

1 Introduction

Modern translation technologies have enabled cross-cultural communication at scale. Additionally, machine translations are the initial step considered in tackling multilingual problems in natural language processing and understanding (Aryal et al., 2023). However, translations may lead to the loss of certain linguistic nuances and cultural information (Sapkota et al., 2023; Aryal and Adhikari, 2023), further contributing to the system's reduced effectiveness. In particular, Machine translation (MT) systems often struggle with named entities, particularly rare, ambiguous, or unknown to the translation model. Proper handling of names of people, organizations, locations, and products is crucial to maintaining correctness and cultural relevance across different languages.

SemEval-2025 Task 2 (Conia et al., 2025) challenges participants to improve named entity translation from English into multiple target languages. We propose a 3-Step GPT Translation pipeline that integrates external knowledge from wikidata to enrich named entity contexts, combined with carefully constructed GPT prompts. Our main contributions include:

1. A modular pipeline that leverages wikidata to retrieve accurate entity labels and descriptions for the target language.
2. A three-step approach, featuring (a) entity extraction from wikidata, (b) GPT-based refinement of entity translations, and (c) context-aware GPT translation of full sentences.

2 Task Description

SemEval-2025 Task 2 focuses on accurately translating sentences that contain named entities from English to a set of target languages. These include Italian (it), Spanish (es), French (fr), German (de), Arabic (ar), Japanese (ja), Chinese (zh), Korean (ko), Thai (th), and Turkish (tr). The task's official scoring metric is the harmonic mean of COMET and M-ETA.[1]

3 Related Work

Recent advances in retrieval-augmented machine translation further support our methodology. For instance, Conia et al. (2024) introduced KG-MT, a system that leverages knowledge graphs to incorporate structured external information into the translation process, while Zeng et al. (2023) proposed the "Extract and Attend" framework, which aligns entity representations with their surrounding context to improve named entity translation accuracy. In addition, work on entity pre-training, such as that by Hu et al. (2022), demonstrates that denoising strategies can boost translation accuracy for entities. Resources such as ParaNames (Sälevä and Lignos, 2022) have also established the value of extensive multilingual corpora derived from Wikidata. Unlike KG-MT, which integrates knowledge graph embeddings directly into the translation model, our approach leverages GPT's context-aware generative capabilities, systematically refining individual

[1]Final Score is defined as the harmonic mean of the M-ETA score (Manual Entity Translation Accuracy) and the COMET score.

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 1885–1889

Rank	Team	System	Uses Gold	Uses RAG	Uses LLM	LLM Name	Overall Final	Overall M-ETA	Overall COMET
14	Lunar	LLaMA-RAFT-Gold	True	True	True	Llama-3.1-8B-Instruct	86.76	82.12	92.60
15	SALT	Salt-Full-Pipeline + Gold	True	True	False	–	85.78	65.30	93.34
16	Howard University-AI4PC	DoubleGPT	True	True	True	gpt-4o-2024-08-06	84.44	77.93	93.63
17	SALT	Salt-Full-Pipeline	False	True	True	GPT-4o-mini	83.63	77.13	91.81
18	SALT	Salt-MT-Pipeline	False	True	False	–	80.42	71.66	92.52
19	FII-UAIC-SAI	Qwen2.5-Wiki-MT	False	False	True	–	78.17	68.24	91.64
20	Lunar	LLaMA-RAFT-Plus	False	True	True	Llama-3.1-8B-Instruct	74.26	62.90	91.82

Table 1: Non-metric system details and overall scores (Final, M-ETA, COMET) for leaderboard entries (ranks 14 to 20).

Rank	Team	ar_AE	de_DE	es_ES	fr_FR	it_IT	ja_JP	ko_KR	th_TH	tr_TR	zh_TW
14	Lunar	88.86	86.83	90.54	81.70	92.18	91.31	90.62	88.09	86.64	70.85
15	SALT	90.83	87.56	88.27	88.12	91.54	88.43	87.81	81.19	88.82	65.30
16	Howard University-AI4PC	89.30	84.55	89.73	85.28	87.25	89.90	90.15	88.25	82.20	57.84
17	SALT	87.29	83.04	87.49	85.11	86.14	85.77	85.97	82.59	85.11	67.82
18	SALT	87.09	82.02	83.01	82.43	84.77	81.30	82.56	76.11	84.76	60.19
19	FII-UAIC-SAI	76.91	77.27	81.22	80.52	83.40	78.11	77.14	75.16	77.77	74.19
20	Lunar	77.70	72.11	77.61	77.40	82.28	69.39	73.96	77.02	81.08	54.02

Table 2: Language-specific final scores for leaderboard entries (ranks 14 to 20) with team names.

entity translations through explicit prompts before final translation. This design choice prioritizes precise entity contextualization and improved handling of sparse or ambiguous data that might be inadequately captured by traditional embedding-based KG methods

4 System Overview: 3-Step GPT Translation

4.1 Step 1: Wikidata Entity Extraction

Our entity extraction process originally employed spaCy's `en_core_web_sm` model for named entity recognition. However, through experimentation, we found GPT-based entity recognition to be both faster (given our hardware and runtime environment constraints) and more accurate, especially in recognizing novel or domain-specific entities that spaCy struggled with. For instance, spaCy frequently failed to recognize rare or uniquely constructed proper nouns, whereas GPT succeeded due to its contextual reasoning capabilities. Consider a hypothetical example sentence: Consider the following hypothetical example sentence: "I recently visited Takunville to watch the grand opening performance by Awetu Tesfaye." Here, spaCy may fail to detect entities like "Takunville" or "Awetu Tesfaye," whereas GPT typically recognizes these from context. Consequently, we transitioned entirely to GPT for entity detection. We then query Wikidata using `wikidata.client` (a Python library for accessing wikidata) for entity metadata: short description, aliases (if present) and label in the target language. If Wikidata lacks relevant information for any of the entities identified for a given source sentence, our fallback strategy is to trust GPT to provide either transliteration or its best guess at an accurate translation from the context of the source sentence in the next step.

4.2 Step 2: GPT-Based Entity Translation Refinement

This step directly yields an entity translation result guided by context retrieved from wikidata information. Using this prompt, we make a query to obtain a refined entity translation:

```
You are an advanced translation service.
Translate the entity name in the input from English
to target_locale.
In the input, there's extra information to help you
translate the entity.
Input: label_info_input
Return only the entity translation.
```

Each entity translation refinement step results in a structured JSON-like object containing the following fields for every named entity:

- `label`: The entity name in the source language.
- `description`: A short description providing contextual information about the entity.
- `target_reference`: The refined translation of the entity into the target language.
- `entity_group`: A category abbreviation (e.g., "PER" for person, "LOC" for location).

We pass a list of these structured objects to the final GPT translation step (Section 4.3). This detailed structured representation ensures precise handling and consistent naming of entities in the translated output.

4.3 Step 3: Context-Aware GPT Translation

Finally, we combine:

1. The source sentence in English.
2. The target language (e.g., German).
3. The list of refined entity translations from Step 2.

The final translation prompt instructs GPT to incorporate the previously refined entity names into the resulting translation, ensuring consistency and accuracy:

```
You are an advanced translation service.
Given:
1. 'source': A sentence in English.
2. 'target_locale': The target language.
3. 'processed_wiki_entities': A list of refined
entity translations.
Task: Translate 'source' to target_locale using
'processed_wiki_entities'.
Return ONLY the translated sentence as a string.
```

By explicitly referencing the refined entity names, we mitigate GPT's tendency to guess or alter named entities.

5 Validation and Results

5.1 Validation Setup

We first conducted experiments on the initial 6 languages from the SemEval-2025 Task 2 validation dataset: Arabic (ar), German (de), Spanish (es), French (fr), Italian (it), and Japanese (ja). Using the harmonic mean of COMET and M-ETA as specified by the task organizers, we compared our proposed approach to both GPT-4o mini with no special entity handling or context beyond requesting a translation of the source text to the target language.

5.2 Results

The experimental results of our validation that are in Table 3 show that we outperformed our baseline models with no entity-associated context. This approach was then applied to the test set and submitted to the official leaderboard.

Table 1 presents non-metric system details and overall scores for leaderboard entries (ranks 14 to 20), while Table 2 summarizes the language-specific final scores for these entries.

On the official leaderboard by the organizers of SemEval, under the name Howard University-AI4PC, our system *DoubleGPT* was ranked 16th overall with an overall final score of 84.44, an overall M-ETA score of 77.93, and a COMET score of 93.63.[2] Notably, we achieved our highest per-language performance in Korean (90.15), Japanese (89.90), and Spanish (89.73), suggesting that our entity-refinement pipeline can excel in languages with strong tokenization or ample external resources. In contrast, our system struggled with Chinese (57.84), reflecting the need for further adaptation when wikidata coverage is sparse or script complexity is high. Despite this gap, our multi-step GPT approach remained competitive relative to single-pass LLM-based methods, demonstrating that targeted named entity handling and retrieval-augmented generation can yield robust improvements across a diverse range of languages.

5.3 Real-World Viability

While our multi-step GPT-based pipeline delivers notable improvements in entity translation, it introduces significant computational overhead due to multiple GPT interactions at both the entity-level and sentence-level. Each translated sentence requires multiple GPT calls, potentially causing latency and increased runtime costs in real-world or real-time translation scenarios. Moreover, our current prompt engineering strategy employs uniform prompt templates across all languages, potentially missing opportunities for language-specific optimizations that could further enhance performance, especially for languages that differ substantially in linguistic structures or available resources.

5.4 Omission of Additional Languages

We recognized that the official dataset was later updated to include Chinese, Korean, Thai, and Turkish. Although they were included in our official submission, we discovered these additions too late in our research cycle to fully evaluate or compute official metrics. We plan to incorporate these languages in a future version of this work.

6 Conclusion and Future Work

We presented a 3-Step GPT Translation system for SemEval-2025 Task 2, emphasizing named entity accuracy. By integrating external Wikidata and employing carefully engineered GPT prompts across two stages (entity translation refinement and final context-aware translation), our approach achieved notable improvements over baseline GPT-4o miniGPT-4o direct translations.

[2]Final Score is defined as the harmonic mean of the M-ETA score (Manual Entity Translation Accuracy) and the COMET score.

Despite these promising results, our method still exhibits certain limitations. Primarily, our system's performance heavily depends on Wikidata coverage; thus, entities without sufficient Wikidata entries risk incorrect or incomplete translations. Additionally, we identified notably lower performance for languages like Chinese due to sparse Wikidata coverage and unique script complexities. Future research will address these issues by exploring supplementary multilingual knowledge resources or custom lexical databases specifically targeted at improving performance in these languages.

We also recognize missed opportunities for prompt customization tailored explicitly to linguistic nuances and specific entity types. Therefore, future work will involve developing and testing language-specific and entity-type-specific prompts, aiming to further enhance translation accuracy.

Finally, addressing computational efficiency remains a critical component of future improvements. We plan to investigate optimization strategies such as single-pass GPT prompts or selectively triggered GPT calls, dynamically invoking GPT only when entity translations are uncertain or when comprehensive Wikidata coverage is lacking. These optimizations will aim to sustain high accuracy while significantly reducing computational resources and translation latency.

Ethics Statement

Our work depends on pretrained LLMs (GPT-4o mini / GPT-4o) and a publicly available knowledge base (wikidata). While these technologies can improve named entity translation, we note that wiki data entries might be incomplete or skewed toward certain languages or cultures, leading to uneven performances. Mistranslations of personal or place names have cultural and ethical implications. Users should verify correctness when translating culturally sensitive content.

Acknowledgements

This research project was supported in part by the Office of Naval Research grant N00014-22-1-2714. The work is solely the responsibility of the authors and does not necessarily represent the official view of the Office of Naval Research.

References

Saurav K Aryal, Howard Prioleau, Gloria Washington, and Legand Burge. 2023. Evaluating ensembled transformers for multilingual code-switched sentiment analysis. In *2023 International Conference on Computational Science and Computational Intelligence (CSCI)*, pages 165–173. IEEE.

Saurav Keshari Aryal and Gaurav Adhikari. 2023. Evaluating impact of emoticons and pre-processing on sentiment classification of translated african tweets. *ICLR Tiny Papers*.

Simone Conia, Daniel Lee, Min Li, Umar Farooq Minhas, Saloni Potdar, and Yunyao Li. 2024. Towards cross-cultural machine translation with retrieval-augmented generation from multilingual knowledge graphs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16343–16360, Miami, Florida, USA. Association for Computational Linguistics.

Simone Conia, Min Li, Roberto Navigli, and Saloni Potdar. 2025. SemEval-2025 task 2: Entity-aware machine translation. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*. Association for Computational Linguistics.

Junjie Hu, Hiroaki Hayashi, Kyunghyun Cho, and Graham Neubig. 2022. DEEP: DEnoising entity pre-training for neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1753–1766, Dublin, Ireland. Association for Computational Linguistics.

Jonne Sälevä and Constantine Lignos. 2022. ParaNames: A massively multilingual entity name corpus. In *Proceedings of the 4th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 103–105, Seattle, Washington. Association for Computational Linguistics.

Hrishav Sapkota, Saurav Keshari Aryal, and Howard Prioleau. 2023. Zero-shot classification reveals potential positive sentiment bias in african languages translations. *ICLR Tiny Papers*.

Zixin Zeng, Rui Wang, Yichong Leng, Junliang Guo, Xu Tan, Tao Qin, and Tie yan Liu. 2023. Extract and attend: Improving entity translation in neural machine translation.

Appendix

Table 3 illustrates a rough estimation of the M-ETA metric from our internal validation experiments. This analysis suggests that our refined entity-focused translation pipeline achieves approximately 1.5–2× improvement in the final score compared to direct GPT-4 or GPT-4o translation

baselines. This approximation underscores the significant contribution of explicit entity translation refinement and contextual GPT prompting in improving overall translation quality.

Language	3-Step	GPT-4o mini	GPT-4o
Arabic	0.44	0.28	0.36
German	0.41	0.29	0.36
Spanish	0.51	0.33	0.37
French	0.51	0.31	0.36
Italian	0.47	0.30	0.36
Japanese	0.48	0.29	0.36

Table 3: Comparison of estimated M-ETA scores between our entity-focused GPT pipeline (3-Step) and direct GPT-4o mini/GPT-4o translation baselines.