silp_nlp at SemEval-2025 Task 2: An Effect of Entity Awareness in Machine Translation Using LLM

Sumit Singh and Pankaj Kumar Goyal and Uma Shanker Tiwary
Indian Institute of Information Technology Allahabad, Allahabad
{sumitrsch, pankajgoyal02003}@gmail.com
ust@iiita.ac.in

Abstract

Our team, *silp_nlp*, participated in the SemEval-2025 Task 2: Entity-Aware Machine Translation (EA-MT) which is focused on translating from English to the target languages. We have utilized LLM like GPT-4o in our experiment in two cases. In the first case, we have performed translation with a straightforward prompt, and in the second one, first a named entity recognition (NER) model, Universal NER, was used for extracting the named entities of a source sentence, thereafter, with the added information of the named entities, the prompt is updated for machine translation to instruct the LLM. Our results show that the addition of named entities helps the LLM to generate better translation.

1 Introduction

Machine translation (MT) of named entities (NEs), such as person or place names, remains a major challenge even for advanced models. This is largely because NEs appear less frequently in training data compared to other words or phrases. Additionally, new and unseen NEs like organization or product names are constantly being created, and even common nouns can function as NEs in certain contexts.

On the other hand, named entity recognition (NER) has achieved reasonably high accuracy for many languages, with precision scores around 80–90% (Rikters and Miwa, 2024). However, since most MT models rely solely on parallel training data and contextual cues for translation, they often struggle with rare NEs that appear infrequently or not at all during training. In such cases, the models may "hallucinate," generating words or phrases that are statistically similar to the rare NE in embedding space but do not provide the correct translation. This can result in the creation of entirely new, unintended words instead of accurate translations.

In this translation work from the source language to the target language, we incorporated Entity Aware (EA) in the prompt to achieve better results. We utilized the Universal NER model (Mayhew et al., 2024) for entity extraction.

2 Related Work

(Ugawa et al., 2018) enhances named entity (NE) translation by encoding named NE tags alongside tokens and merging their embeddings. (Modrzejewski et al., 2020) investigates different ways to integrate NE annotations into MT models, showing that fine-grained NE annotations improve translation quality in English-German and English-Chinese MT on WMT 2019 test sets compared to baseline transformer models.

(Zeng et al., 2023) uses a dictionary-based approach, where translation candidates are retrieved and prepended to the decoder input. (Hu et al., 2022) improves NE handling by modifying pretraining data—replacing NEs in the target language, training the model to reconstruct original sentences, and applying multi-task fine-tuning for both reconstruction and MT.

(Conia et al., 2024) tackles cross-cultural translation by introducing XC-Translate, the first large-scale benchmark for translating culturally-nuanced entity names, and KG-MT, a novel method that integrates multilingual knowledge graphs into neural machine translation via dense retrieval. Experiments show that existing MT systems, including large language models, struggle with entity translation, while KG-MT significantly outperforms NLLB-200 and GPT-4, achieving 129% and 62% relative improvements, respectively.

3 data

This shared task focuses on translating from English to various target languages. Each target sentence in English is accompanied by a Wikidata ID, which represents the entity type. The goal of this task is to translate the source sentence into the tar-

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 2389–2394

get sentence, with a primary focus on accurately translating the entities corresponding to their respective types.

Training and validation data include source sentences along with their corresponding Wikidata IDs, which indicate the entity types present in the sentences and target sentences (Conia et al., 2025). The testing data consists of source sentences and their associated Wikidata IDs that correspond to the entity types found within those sentences.

Statistics for the datasets across all ten languages are presented in Table 1.

Language	Train	Valid	Test
en-ar	5350	722	4550
en-de	6680	731	5880
en-es	6150	739	5340
en-fr	6260	724	5470
en-it	5900	730	5100
en-ja	5900	723	5110
en-ko	5900	745	5080
en-th	4230	710	3450
en-tr	5280	732	4470

Table 1: Training, validation, and test set sizes for different language pairs

4 System Description

We have generated translations with LLMs (GPT-4o and GPT-4o-mini) (OpenAI et al., 2024) with the appropriate prompt, with and without entity aware. We have also fine-tuned the NLLB-200 model (Koishekenov et al., 2023) in both cases. Universal NER model is utilised for the extraction of NER. The NER results obtained using the Universal NER model on the training data are presented in Table 2. However, since gold annotations for the test data were not available, we assume that similar performance would be achieved on the test set.

Language	**Accuracy**
ar	0.19
de	0.17
es	0.15
fr	0.18
it	0.18
ja	0.19

Table 2: NER accuracy of six languages with the Universal NER on training data.

4.1 Translation with GPT-4o and GPT-4o-mini

We have utilized GPT-4o and GPT-4o-mini LLMs to generate machine translation in two cases. In the first case, a source sentence is provided directly to the model along with the instruction to translate it into the target language. In the second case, sentences are provided to the language model with extracted entities for translation. We extracted the named entities using Universal NER (Mayhew et al., 2024). Universal NER is the most comprehensive named entity recognition model for the English language, covering 13,020 distinct entity types. Overall architecture of translation with GPT-4o illustrated in Fig. 1.

1. **Prompt without the extracted named entities** Translate the following sentences from `{source_lang}` to `{target_lang}`.

   For each sentence, transliterate the entities into the target language, then translate the sentence while ensuring the entities are correctly placed.

2. **Prompt with the extracted named entities**

   Translate the following sentences from `{source_lang}` to `{target_lang}`, ensuring entity awareness. Entities are enclosed within XML-like tags (e.g., `<PER>Henry I</PER>`), and they should be correctly transliterated and placed in the translated sentence. Additionally, use the following entity categorization to enhance translation accuracy:

   Entity Categories: `{tag_to_category}`

   Instructions:

   (a) Preserve the XML-like tags for entities in the translated sentence.

   (b) Transliterate entities (e.g., names, locations) appropriately for the target language.

   (c) Ensure the translated sentence is grammatically correct and contextually accurate.

   Output Format: The output should be a list of dictionaries (in JSON format), where each dictionary contains:

   - `"source_sentence"`: The original sentence in `{source_lang}`.

- `"translated_sentence"`: The translated sentence in `{target_lang}`.

Sentences to Translate: `{batch}`

4.2 NLLB-200 Fine-tune (Team et al., 2022; Koishekenov et al., 2023)

NLLB-200 (No Language Left Behind) is a state-of-the-art massively multilingual machine translation model developed by Meta AI, designed to support translation across 202 languages, including many low-resource languages. The largest variant of NLLB-200 (Koishekenov et al., 2023) adopts a Mixture-of-Experts (MoE) architecture with 54.5 billion parameters, where each input token is routed through a small subset of specialized experts. This design enables both scalability and high translation quality, achieving superior performance on multilingual benchmarks such as FLORES-200 (Team et al., 2022). However, the full model requires at least four 32GB GPUs for inference, limiting its deployment. Recent advancements demonstrate that language-specific expert pruning can reduce memory usage significantly (up to 80%) without sacrificing translation quality, enabling single-GPU deployment while preserving the benefits of expert specialization.

We have fine-tuned NLLB-200 using both entity-aware and non-entity-aware methods. In the entity-aware approach, we incorporated entities into sentences along with their corresponding entity tags. The extraction of entities is done using Universal NER. For example, a source sentence like "What is the main goal of the Confederacy of Independent Systems?" includes the entity information in a specified format:

What is the main goal of the <FIC>Confederacy of Independent Systems</FIC>?

Here FIC is the entity type of "Confederacy of Independent Systems" entity.

We fine-tuned the pretrained model using the Hugging Face Transformers framework for five epochs. A batch size of 32 was used throughout the training process. After experimenting with various hyperparameters, we found that a learning rate of 5e-05 yielded the best results. The model was evaluated on a validation set after each epoch, and the checkpoint with the lowest validation loss was selected as the best-performing model. This approach ensured stable convergence and helped prevent overfitting. The Hugging Face framework enabled efficient integration of tokenization, batching, and model checkpointing, making the training process streamlined and reproducible. These settings were chosen to balance computational efficiency with effective learning in low-resource conditions.

5 Evaluation Metric

The evaluation of this task is based on three metrics for the comprehensive assessment:

1. **M-ETA Score** Measures the accuracy of named entity translation, ensuring proper preservation and correctness of entities.

2. **COMET Score** Evaluates the overall quality of translation at the sentence level.

3. **Overall Score** Evaluates overall sentence-level translation quality.

6 Results and Analysis

Tables 3 and 4 present the comparative results of various machine translation systems under entity-aware (EA) and non-entity-aware configurations. The evaluation is based on three key metrics: M-ETA, COMET, and a composite Overall score, averaged across all languages. Additionally, Table 4 provides a language-wise breakdown of the M-ETA scores for ten representative languages.

The results in Table 3 demonstrate that incorporating entity information substantially improves the performance of large language models, particularly GPT-4o. The M-ETA score for GPT-4o improves from 13.52 to 23.26, while the COMET score rises from 77.6 to 88.59 when EA is included, resulting in a 77.6% relative gain in the Overall score (from 20.72 to 36.83). A similar performance boost is observed for GPT-4o-mini, with the Overall score increasing from 20.26 to 35.27 upon adding entity annotations. These results confirm the significance of explicit entity marking in enhancing translation fidelity, especially for models that rely on contextual semantics.

In contrast, the fine-tuned NLLB-200 model exhibits relatively stable performance across both settings. The improvement in the Overall score from 20.44 to 20.50 with EA is minimal, despite a small rise in COMET (from 85.2 to 86.0) and M-ETA (from 11.61 to 11.64). This suggests that NLLB-200, while effective in standard translation tasks, is less sensitive to explicit entity markup, possibly due to its architecture and training design, which

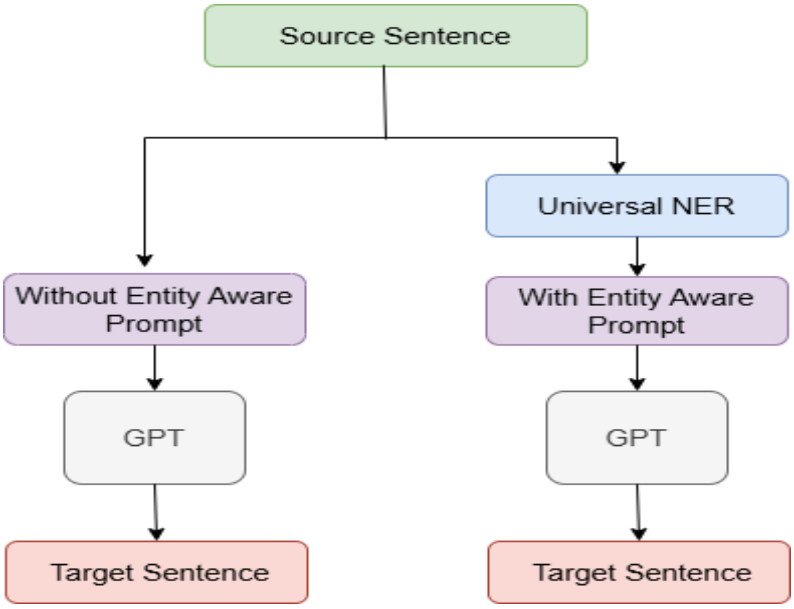

Figure 1: Architecture for Target Language Generation from Source Language with and without Entity Awareness with GPT.

	Average across all languages		
System	**M-ETA**	**Comet**	**Overall**
GPT-4o (without EA)	13.52	77.6	20.72
GPT-4o (with EA)	23.26	88.59	36.83
GPT-4o-mini (without EA)	13.18	77.67	20.26
GPT-4o-mini (with EA)	22.12	87.23	35.27
Fine-tuned NLLB-200 (without EA)	11.61	85.2	20.44
Fine-tuned NLLB-200 (with EA)	11.64	86.0	20.50

Table 3: Table shows the average score across all languages of different models.

may already capture entity-level semantics to some extent.

The language-specific results in Table 4 provide deeper insights into how entity awareness affects individual language directions. For instance, GPT-4o with EA shows surprising gains for German (DE), Spanish (ES), and French (FR), languages that frequently include multi-token named entities, indicating better preservation and contextual translation of named entities. The M-ETA scores for Spanish and French jump from 2.21 to 34.54 and from 1.5 to 27.89, respectively. Similarly, entity awareness improves Arabic (AR) and Turkish (TR) translations. However, languages such as Chinese (ZH) and Thai (TH) show little to no gain, likely due to tokenization challenges or limitations in the NER model used for entity extraction.

For fine-tuned NLLB-200, while the M-ETA scores are generally lower, entity awareness does show marginal improvements across Italian, Arabic, and French, though the gains are not consistent across all languages.

In summary, the results highlight the effectiveness of entity-aware translation in improving multilingual translation quality, particularly in transformer-based generative models like GPT-4o. While both models benefit from entity incorporation, the gains are significantly more pronounced in autoregressive generative models than in encoder-decoder models like NLLB-200. This underscores the importance of integrating named entity tagging as an auxiliary signal, especially in settings where entity fidelity and factual grounding are critical.

7 Conclusion

This study explored the impact of entity-aware translation on multilingual machine translation performance using both GPT-based models (GPT-4o

System	AR	DE	ES	FR	IT	JA	KO	TH	TR	ZH	Avg
GPT-4o (without EA)	28.24	0.92	2.21	1.5	32.97	31.15	28.85	0.09	9.21	0.12	13.52
GPT-4o (with EA)	29.29	24.03	34.54	27.89	26.09	30.15	25.9	6.12	28.55	0.12	23.26
Fine-tuned NLLB-200 (without EA)	15.6	21.77	22.32	27.03	22.32	8.89	-	-	-	-	11.61
Fine-tuned NLLB-200 (with EA)	19.39	19.87	21.02	21.20	26.26	8.73	-	-	-	-	11.64

Table 4: M-ETA score of different methods on the task across different languages. Language codes: Arabic (AR), German (DE), Spanish (ES), French (FR), Italian (IT), Japanese (JA), Korean (KO), Thai (TH), Turkish (TR), and Chinese (ZH).

and GPT-4o-mini) and the fine-tuned NLLB-200 model. Our results demonstrate that incorporating explicit entity information significantly enhances translation quality in GPT models. The models exhibited particularly strong gains in languages with rich named entity structures, such as Spanish, French, and German. In contrast, the NLLB-200 model showed only marginal improvements, indicating that while it may implicitly learn entity representations, it does not benefit substantially from direct entity injection using the current architecture.

These findings highlight the suitability of GPT-style autoregressive models for prompt-based entity-aware enhancements, whereas encoder-decoder architectures like NLLB-200 may require structural changes to leverage entity information more effectively.

For GPT-based models, future research could explore advanced prompting strategies, such as multi-turn dialogue prompts, chain-of-thought reasoning, or contextual expansion around named entities. Additionally, integrating retrieval-augmented generation (RAG) or few-shot in-context learning using entity-rich examples could further boost translation fidelity.

For NLLB-200 and similar encoder-decoder models, future work may focus on architectural enhancements, such as entity-aware attention mechanisms, adapter layers for entity embedding, or multi-task learning frameworks that jointly train on translation and NER objectives. In conclusion, this work emphasizes the importance of integrating structured entity knowledge into multilingual translation systems and opens up promising directions for enhancing translation quality in low-resource and entity-rich contexts.

References

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.

Yeskendir Koishekenov, Alexandre Berard, and Vassilina Nikoulina. 2023. Memory-efficient NLLB-200: Language-specific expert pruning of a massively multilingual machine translation model. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3567–3585, Toronto, Canada. Association for Computational Linguistics.

Stephen Mayhew, Terra Blevins, Shuheng Liu, Marek Suppa, Hila Gonen, Joseph Marvin Imperial, Börje Karlsson, Peiqin Lin, Nikola Ljubešić, Lester James Miranda, Barbara Plank, Arij Riabi, and Yuval Pinter. 2024. Universal NER: A gold-standard multilingual named entity recognition benchmark. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4322–4337, Mexico City, Mexico. Association for Computational Linguistics.

Maciej Modrzejewski, Miriam Exel, Bianka Buschbeck, Thanh-Le Ha, and Alexander Waibel. 2020. Incorporating external annotation to improve named entity translation in NMT. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 45–51, Lisboa, Portugal. European Association for Machine Translation.

OpenAI, :, Aaron Hurst, Adam Lerer, and Adam P. Goucher ... 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Matiss Rikters and Makoto Miwa. 2024. Entity-aware multi-task training helps rare word machine translation. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 47–54, Tokyo, Japan. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Arata Ugawa, Akihiro Tamura, Takashi Ninomiya, Hiroya Takamura, and Manabu Okumura. 2018. Neural machine translation incorporating named entity. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3240–3250, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Zixin Zeng, Rui Wang, Yichong Leng, Junliang Guo, Shufang Xie, Xu Tan, Tao Qin, and Tie-Yan Liu. 2023. Extract and attend: Improving entity translation in neural machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1697–1710, Toronto, Canada. Association for Computational Linguistics.