Zero at SemEval-2025 Task 2: Entity-Aware Machine Translation: Fine-Tuning NLLB for Improved Named Entity Translation

Revanth Gundam
IIIT Hyderabad
revanth.gundam@research.iiit.ac.in

Abhinav Marri
IIIT Hyderabad
abhinav.marri@research.iiit.ac.in

Advaith Malladi
IIIT Hyderabad
advaith.malladi@research.iiit.ac.in

Radhika Mamidi
IIIT Hyderabad
radhika.mamidi@iiit.ac.in

Abstract

Machine Translation (MT) is an essential tool for communication among people across different cultures, yet Named Entity (NE) translation remains a major challenge due to its rarity in occurrence and ambiguity. Traditional approaches, like using lexicons or parallel corpora, often fail to generalize to unseen entities and, hence, do not perform well. To address this, we create a silver dataset using the Google Translate API and fine-tune the facebook/nllb-200-distilled-600M model with LoRA (Low-Rank Adaptation) to enhance translation accuracy while also maintaining efficient memory use. Evaluated with metrics such as BLEU, COMET, and M-ETA, our results show that fine-tuning a specialized MT model improves NE translation without having to rely on large-scale general-purpose models.

1 Introduction

Machine Translation (MT) has proven to be significant at enabling cross-cultural and cross-border communication. It is essential in multilingual content creation, border business interaction, and real-time translation communication services. While the contemporary Neural Machine Translation (NMT) models have achieved high fluency and accuracy, they struggle with certain aspects of translation. One of the most difficult parts is the translation of named entities (NEs) which are handled comparatively poorly. NEs include proper names of people, places, organizations and other cultural references. They pose difficulties due to their rarity in occurrence in natural language, ambiguity, and due to the language-specific variations.

Entity-Aware Machine Translation (EA-MT) seeks to resolve the complexities faced when translating sentences with named entities. Unlike generic MT that depends on the patterns of language and context, EA-MT focuses on entity recognition, retention, and accurate translation. The importance of EA-MT is very evident in real-world applications such as **news translation, medical and legal document translation, and localization of entertainment content**, where small mistakes in NE translation can lead to misinformation or loss of meaning, which would cause problems.

Handling named entities in translation is a difficult task because entities might not have direct equivalents across all languages. For example, consider the English sentence:

"*Elon Musk announced an exciting new feature for X (formerly Twitter) in an interview with CNBC today.*"

An MT model might struggle to translate the above sentence due to several possible reasons:

- "**X (formerly Twitter)**" might be translated poorly if the model fails to recognize that both "X" and “Twitter” here refer to social-media platforms.
- "**CNBC**" could be wrongly translated if the model assumes it to be a random acronym instead of recognizing it to be the media organization.
- "**Elon Musk**" should in an ideal situation remain unchanged since it is the name of an individual, but certain translation models might fail and attempt transliteration, changing the original intended meaning of the sentence.

To handle such issues as discussed above, a variety of techniques have been used in MT such as dictionary-based approaches, parallel corpus training, and using external knowledge, in addition to other approaches. Traditional approaches utilize large bilingual lexicons or pre-aligned corpora in order to guarantee the correct entity mappings. Novel techniques are now using knowledge graphs, entity linking, or explicit annotation to assist models in differentiating between named entities and regular phrases.

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 1187–1191
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

In this paper, we propose an approach that leverages a **silver dataset generated using Google Translate** and fine-tunes the **NLLB model** to enhance entity-aware translation. Such a method allows the model to learn important patterns in entity translation. The approach used in this paper also provides flexibility which helps the model to be able to generalize and tackle samples with unseen entities and still have a high accuracy.

2 Related Work

(Conia et al., 2025) is the task description article and (Conia et al., 2024) is the work that led to the creation of this task and initial dataset. Past research has explored different strategies to try and improve Named Entity translation in Neural Machine Translation. (Modrzejewski et al., 2020) uses external annotations for the NMT models, which shows that using samples which are explicitly entity labelled can enhance translation quality. The article (Li et al., 2021) proposes a unique approach which is lexicon-based to ensure consistency in the translations of the model, but such an approach would end up lacking the ability to translate any unseen entities. In contrast to this, a fine-tuning approach would learn entity mappings from the sample data rather than simply relying on predefined lexicons. (Awadallah et al., 2016) uses an alternative approach which improves translation quality by aligning entities across comparable and parallel corpora. The approach in (Jiang et al.) employs strategies such as web mining and transliteration to extract bilingual named entities in an attempt to handle unknown entities. The methodology in (Huang and Vogel, 2002) focuses on statistical NE extraction and entity disambiguation, which is similar to the goal of this paper of improving entity representation in machine translation. These studies provide some key insights into the challenges in Named Entity Machine Translation, which this paper tries to build upon by generating a silver dataset and fine-tuning a transformer model for a more adaptable and accurate entity-aware translation system.

3 Dataset

The dataset (Sen et al., 2022) provided for the task is a collection of English text data translated into various languages such as Italian, Spanish, French, etc. The data is present in the JSONL format. Fig 1 depicts a sample of evaluation data. The sample has an id, a wikidata_id, a list of entity types present in the sentence, the source language, the target language, the source text in English and the translated target sentence.

```
{
  "id": "Q850522_0",
  "wikidata_id": "Q850522",
  "entity_types": [
    "Movie"
  ],
  "source": "Who are the main characters
       in the movie Little Women?",
  "targets": [
    {
      "translation": "  Quines   son los
           personajes principales de la
          pel cula Mujercitas?",
      "mention": "Mujercitas"
    }
  ],
  "source_locale": "en",
  "target_locale": "es"
}
```

Figure 1: Evaluation data sample

The training data provided has a slightly different format. An example is shown in Fig 2. The training sample contains the source text in English, the target translation in the required language, list of Wikidata IDs for entities present in the source text and the source of the data sample. Prediction data is also provided, which contains predictions by GPT-4o and GPT-4o-mini, which can be used to analyze the performance of proposed systems.

```
{
    "source": "Did Gone With The Wind
        come out before 1940?",
    "target": "Via col vento    uscito
        prima del 1940?",
    "entities": [
        "Q2875"
    ],
    "source_locale": "en",
    "target_locale": "it",
    "instance_id": "826528e6",
    "from": "mintaka"
}
```

Figure 2: Training data sample

4 System Description

4.1 Silver Dataset Creation

As we have mentioned in the previous sections, the provided dataset contains predictions generated using GPT 4o and GPT 4o-mini. However, we wanted to create a different dataset using an expert machine translation system as opposed to using the predictions from a general purpose large language model. This is because of the fact that GPT-4o and GPT-4o-mini are optimized for language modeling over a vast corpus rather than being trained specifically for machine translation.

To ensure we have good quality silver predictions which stem from a model specifically trained for machine translation, we made use of the Google Translate API to translate the sentences in the dataset from the source language to the target language. We call this newly created dataset our “Silver Dataset”.

4.2 Model

For this task, we have decided to fine-tune a smaller pre-trained machine translation model on our newly created silver dataset. We propose using a smaller expert model as opposed to using a general-purpose large language model which fits all tasks because we believe in training smaller expert models which specialize in specific tasks instead of having generic models.

To this end, we chose to fine-tune the **Facebook / nllb-200-distilled-600M** (Costa-Jussà et al., 2022) model. This model was selected because its base pre-trained variant supports all languages present in the task dataset. We use LoRA (Low-Rank Adaptation) (Hu et al., 2022) to fine-tune the NLLB model for each language individually. LoRA was chosen because of its very minimal memory requirements compared to full fine-tuning.

After fine-tuning the base model on each language individually, we obtained the predictions for the test set. For the fine-tuning procedure, we made use of 4 RTX 3080 Ti graphics cards.

5 Results and Analysis

Our machine translation system was individually fine-tuned for each of the target languages, and its performance was evaluated using BLEU scores and the harmonic mean of COMET and M-ETA scores. The results across ten languages are summarized in Table 1 and Table 2 and can also be seen in Fig 3 and Fig 4.

5.1 BLEU Scores Analysis

BLEU (Bilingual Evaluation Understudy) is a widely used metric in machine translation that evaluates the quality of translation by comparing n-grams in the predicted translation with the N-grams in the reference translation. The BLEU scores show significant variations in performance across the different languages. The best performance was noticed in the case of Spanish (59.14), which was closely followed by Italian (54.52), French (49.17), and Turkish (49.38). Arabic, German, and Thai showed mostly moderate scores, with Arabic (47.24) and German (44.98) showing equally competitive results. However, the model seems to have struggled a lot with Japanese (1.71), Thai (4.93), and Chinese (Traditional) (0.11).

Language	BLEU Score
Arabic	47.24
German	44.98
Spanish	59.14
French	49.17
Italian	54.52
Japanese	1.71
Korean	28.19
Thai	4.93
Turkish	49.38
Chinese (Traditional)	0.11

Table 1: BLEU Scores for Different Languages

5.2 COMET and M-ETA Scores Analysis

We use a combined metric based on **COMET** (Rei et al., 2020) and **M-ETA**. **COMET (Cross-lingual Optimized Metric for Evaluation of Translation)** is a model-based metric that compares the machine translation output with a human reference translation, leveraging pre-trained embeddings to capture semantic similarity. **M-ETA (Manual Entity Translation Accuracy)**, on the other hand, measures how well named entities are translated by

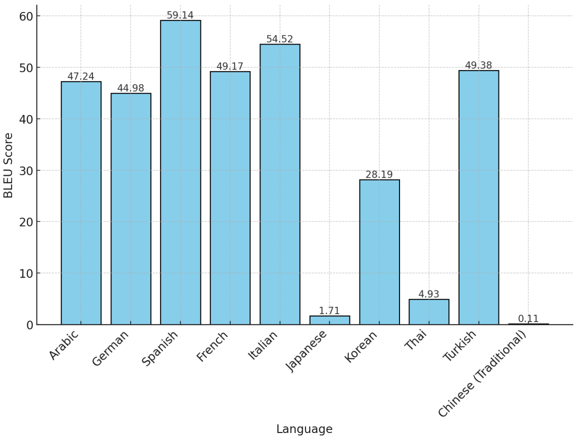

Figure 3: BLEU Scores for Different Languages.

Language	Overall Score
Arabic	37.50
German	40.32
Spanish	46.46
French	33.16
Italian	39.37
Japanese	35.28
Korean	35.97
Thai	13.75
Turkish	46.50
Chinese (Traditional)	8.41

Table 2: Harmonic mean of COMET and M-ETA Scores for Different Languages

calculating the proportion of correctly translated entities. The final **composite score** is computed as:

$$Score = 2 \times \frac{(COMET \times M - ETA)}{(COMET + M - ETA)}$$

While Spanish (46.46) and Turkish (46.50) still performed well, Japanese (35.28) and Korean (35.97) saw considerable improvement compared to their BLEU scores, which suggests that while exact word matching is poor, most of the semantic content is relatively preserved. Chinese (Traditional) (8.41) and Thai (13.75) are still the lowest-performing languages, showing the difficulty of translation in these languages.

5.3 Language-Specific Observations

- **High BLEU and M-ETA Scores:** Spanish, Italian, and Turkish performed well across both of the above metrics.

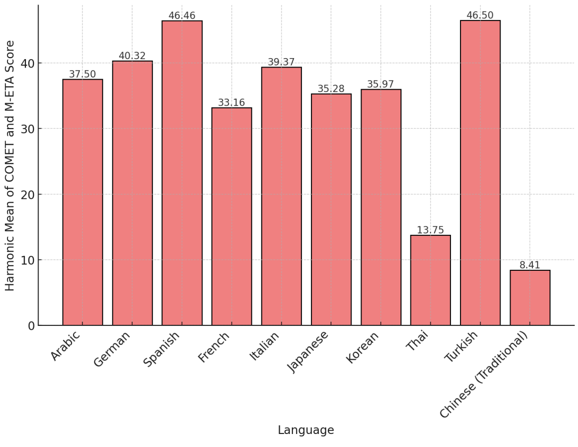

Figure 4: Harmonic mean of COMET and M-ETA Scores for Different Languages.

- **Low BLEU, Higher M-ETA:** Japanese and Korean exhibited low BLEU scores but higher M-ETA and COMET scores, suggesting that BLEU may not fully capture translation adequacy in morphologically complex languages.

- **Extremely Low Scores:** Chinese (Traditional) performed the worst across both metrics, indicating significant model limitations in handling the language's complex structure and large vocabulary space.

6 Conclusion

This paper explored the challenge of Named Entity translation in Machine Translation, a task where the generic models often fall short. To address this, we created a **silver dataset** using **Google Translate** and fine-tuned the **facebook/nllb-200-distilled-600M** model with **LoRA (Low-Rank Adaptation)**, enabling a more efficient and specialized approach to tackle the task of entity-aware translation.

Our evaluation using **BLEU, COMET, and M-ETA** metrics demonstrated the effectiveness of fine-tuning to improve NE translation quality without the need to use generalized large language models. While **Spanish and Turkish** achieved high scores across both general translation and entity accuracy, languages like **Japanese and Korean** displayed weaker BLEU scores but better semantic preservation, which can be seen from COMET and M-ETA scores. Overall, our approach shows the strengths of fine-tuning models for named entity machine translation.

Limitations

Entity Awareness in Fine-Tuning

While our approach successfully fine-tunes a specialized model for named entity translation, it does not explicitly enforce fine-tuning on the entity awareness aspect of each sample. The model learns these entity translation patterns indirectly from the silver dataset, but there is no direct, specific mechanism to ensure that these named entities are treated differently from all other words. Another key limitation of our approach is the reliance on the silver dataset as discussed above. While Google Translate usually provides high-quality translations, it might not always ensure accurate named entity translations. Some of those entities may be falsely transliterated or replaced with the wrong words, which could introduce noise into the training data, causing the model to perform comparatively poorly.

Limitations of BLEU for Entity Translation

The BLEU score primarily measures N-gram overlap; hence it might not be a great way to measure the quality of named entity translation. It does not account for semantic accuracy and often fails to penalize incorrect entity translations effectively. The following examples illustrate these shortcomings:

Incorrect Entity Translation with a High BLEU Score

Source: Who starred in the 1972 film Taming of the Fire?
Predicted: Qui a joué dans le film Taming of the Fire de 1972 ?
Reference: Qui a joué dans le film de 1972 Dompter le feu ?
BLEU Score: 44.08

Here, even though the named entity *"Taming of the Fire"* was incorrectly translated, the BLEU score still remains considerably high because the rest of the predicted sentence aligns with the reference sentence. This shows that BLEU does not effectively penalize named entity translation errors.

Correct Entity Translation with an Average BLEU Score

Source: How old is Emmaus Monastery in Prague?
Predicted: Quel âge a le monastère d'Emmaüs à Prague ?
Reference: Quel âge a le cloître d'Emmaüs à Prague ?
BLEU Score: 43.16

In this case, the entity *"Emmaus Monastery"* is correctly translated in the predicted sentence as *"monastère d'Emmaüs"*, but still the BLEU score remains average due to small structural differences in the remaining words of the sentence. This clearly shows that BLEU alone is not sufficient for evaluating the quality of entity-aware translation and hence COMET and M-ETA scores were also used in this paper.

References

Ahmed Hassan Awadallah, H. S. Fahmy, and Hany Hassan Awadalla. 2016. Improving named entity translation by exploiting comparable and parallel corpora.

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.

Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Fei Huang and S. Vogel. 2002. Improved named entity translation and bilingual named entity extraction. In *Proceedings. Fourth IEEE International Conference on Multimodal Interfaces*, pages 253–258.

Long Jiang, Ming Zhou, Lee-Feng Chien, and Cheng Niu. Named entity translation with web mining and transliteration.

Panpan Li, Mengxiang Wang, and Jian Wang. 2021. Named entity translation method based on machine translation lexicon. *Neural Computing and Applications*, 33:3977–3985.

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.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Priyanka Sen, Alham Fikri Aji, and Amir Saffari. 2022. Mintaka: A complex, natural, and multilingual dataset for end-to-end question answering. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1604–1619, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.