Entity Aware Machine Translation SemEval 2025

Vincenzo Catalano
Politecnico di Torino
s326954@studenti.polito.it

Andrea Zenotto
Politecnico di Torino
s327473@studenti.polito.it

Ruben Tetamo
Politecnico di Torino
s317569@studenti.polito.it

Abstract

Machine translation plays a crucial role in bridging language barriers worldwide, yet accurately translating named entities remains a significant challenge, especially in specialized domains. Traditional translation models often overlook the nuanced treatment of entities, leading to inaccuracies that can compromise the integrity of the translated content. This work aims to address this gap by developing an entity-aware machine translation system that integrates Named Entity Recognition (NER) to detect entities in the source text with a Retrieval Augmented Generation (RAG) module to fetch the most contextually appropriate translation from a reliable source. Evaluated on datasets tailored for entity translation, our study examines the effectiveness of this combined approach in preserving semantic accuracy and context. Preliminary results demonstrate notable improvements in handling entity-specific translations, underscoring the potential of this method for enhancing machine translation in technical and professional settings.

GitHub repository

1 Introduction

The adoption of lightweight language models in natural language processing presents unique challenges, particularly when handling specialized content such as named entities in multilingual contexts. Ensuring that these entities are accurately recognized and translated is critical for maintaining semantic integrity and providing reliable communication across languages. Traditional machine translation systems often struggle with this task, leading to potential misinterpretations and errors in contexts that demand high precision.

This research addresses the current shortcoming in entity-specific translation by leveraging a lightweight Large Language Model (LLM) in combination with a Named Entity Recognition (NER) module and a Retrieval Augmented Generation (RAG) component. The integrated approach enhances translation performance by first detecting entities within the source text and then retrieving the most contextually appropriate translations from a reliable source. Our translator supports 10 languages: Italian (it), Spanish (es), French (fr), German (de), Arabic (ar), Japanese (ja), Chinese (zh), Korean (ko), Thai (th), and Turkish (tr). We evaluate the proposed system on the Mintaka dataset, a multilingual question answering dataset based on Wikidata, which provides a robust environment for testing entity-specific translation capabilities.

The effectiveness of our approach is assessed using advanced evaluation metrics such as COMET and M-ETA. These metrics offer a comprehensive view of translation quality and semantic alignment, ensuring that the system not only produces fluent output but also preserves the precise meaning of named entities. By optimizing the synergy between the lightweight LLM, NER, and RAG modules, our study demonstrates significant improvements in translation performance, paving the way for more accurate and context-aware machine translation solutions.

2 Related Works

Early neural machine translation (NMT) models often struggled with accurately translating named entities, frequently resulting in misinterpretations in multilingual contexts. Sennrich et al. (13) addressed the issue of rare word translation using subword units, which indirectly alleviated some challenges related to infrequent or out-of-vocabulary entities.

More recently, retrieval-augmented frameworks have gained prominence. Lewis et al. (8) introduced the Retrieval Augmented Generation (RAG) framework to enhance knowledge-intensive tasks by integrating external document retrieval into the generation process. Building on this idea, Khandelwal et al. (7) proposed a nearest neighbor machine translation approach that leverages a large datastore to improve translation accuracy, particularly

for out-of-distribution terms and rare entities.

In parallel, several studies have focused specifically on entity-aware translation. For instance, Li et al. (4) explored integrating Named Entity Recognition (NER) into the translation pipeline, isolating and processing entities separately to improve translation precision. However, many existing approaches either rely solely on internal model knowledge or treat entity translation as a secondary task without fully exploiting external contextual retrieval.

Our work differentiates itself by integrating a lightweight Large Language Model (LLM) with both a NER module and a RAG component to specifically target the accurate translation of named entities across 10 languages. Evaluated using advanced metrics such as COMET and M-ETA on the Mintaka dataset, a multilingual question answering dataset based on Wikidata, our approach aims to bridge the gap left by previous methods by ensuring that both general linguistic content and specialized entity information are translated with high semantic fidelity.

3 Methods

Entity-aware machine translation seeks to overcome limitations in conventional translation systems by ensuring that named entities are accurately recognized and translated. Traditional neural machine translation models, despite their successes on large-scale parallel corpora, often struggle to maintain the semantic fidelity of specialized entity names, particularly in multilingual contexts where subtle nuances are critical.

In our approach, a lightweight Large Language Model (LLM) is augmented with a dedicated Named Entity Recognition (NER) module and a Retrieval Augmented Generation (RAG) component. The NER module first identifies entity mentions within the source text, while the RAG component retrieves the most contextually appropriate translations from reliable external sources. This multi-stage process not only addresses general linguistic translation but also specifically enhances the handling of named entities across the supported 10 languages.

3.1 Baseline Machine Translation with State-of-the-Art Models

To establish strong baselines for our translation pipeline, we evaluate a range of state-of-the-art machine translation models with varying parameter sizes and architectures. Specifically, we test `m2m100_418M` and `m2m100_1.2B`, two versions of Meta's `M2M100` multilingual translation model (2) designed to handle direct translation between 100 languages without relying on English as a pivot. We also evaluate `LLaMA 3.1_8B-Instruct` (3), a large language model fine-tuned for instruction following, which has demonstrated robust general-purpose capabilities, including machine translation. Furthermore, we assess two versions of `Qwen 2.5-Instruct` (10), a family of instruction-tuned models optimized for conversational and task-oriented scenarios, namely `Qwen2.5_3B-Instruct` and `Qwen2.5_7B-Instruct`, developed by Alibaba. Finally, we incorporate `Gemma 2-9B-it` (14), a model from Google fine-tuned for Italian, representing a high-capacity option tailored for downstream natural language understanding and generation tasks. These models serve as the foundation upon which we introduce entity-aware enhancements to improve named entity translation accuracy.

3.2 Integrating Named Entity Recognition

Subsequently, we integrate a dedicated Named Entity Recognition (NER) module into our translation pipeline. NER is a sequence labeling task in Natural Language Processing (NLP) that involves identifying and classifying spans of text representing named entities, such as persons, organizations, or locations. Modern NER systems often rely on pre-trained transformer-based models, such as BERT-base and BERT-large (9), which have shown state-of-the-art performance across various NER benchmarks. Additionally, we evaluate CamemBERT (6), a BERT-based model pre-trained on a large corpus of French text, to analyze performance differences in a multilingual setting. Instead of relying on manually annotated ground truth entity mentions, the NER module automatically detects entities in the source text. The identified entities are subsequently used to query the RAG component for accurate translations. This experimental setup allows us to assess the impact of automatic entity detection on translation quality by comparing the performance metrics against the baseline system that utilizes ground truth annotations.

3.3 Entity Linking using WikiData Resources

To enhance the handling of named entities in machine translation, we integrated an Entity Linking (EL) component into our translation pipeline, leveraging Wikidata as a reference knowledge base (1). Wikidata is a free and collaborative knowledge graph that provides structured data on various topics, including entities such as persons, organizations, and locations, with multilingual labels. In our approach, we first apply Named Entity Recognition (NER) to identify named entities in the source text. Then, we perform entity linking to disambiguate the detected entities and obtain their unique Wikidata identifiers (IDs). Once an entity is linked to its corresponding Wikidata ID, we retrieve its label in the target language from the entity's Wikidata page. This translation is then included in the prompt passed to the Large Language Model (LLM), guiding it to ensure that the named entity is correctly translated and preserving its identity across languages. This method allows us to reduce errors in entity translation, which is crucial for maintaining the factual accuracy and coherence of the output.

4 Experiments

In this section, we describe the evaluation of our proposed translation pipeline, focusing on the integration of Entity Linking (EL) and its impact on translation quality. We investigate two configurations: one leveraging gold entity IDs provided in the dataset, and another relying on entity IDs extracted through a Named Entity Recognition (NER) model.

The evaluation is conducted using two metrics: M-ETA (Manual Entity Translation Accuracy) (5) and COMET (11). M-ETA measures the accuracy of entity translation by comparing the predicted entity IDs in the translation to the corresponding gold entity IDs. It reflects the proportion of correctly translated entities within the output. COMET, on the other hand, is a neural-based evaluation metric that assesses overall translation quality by comparing system outputs against human references. It generates a score representing the adequacy and fluency of the translation. Together, these metrics offer complementary insights into both entity-level precision and general translation performance.

4.1 Dataset

In this study, we used the Mintaka dataset (12), a complex and multilingual resource designed for end-to-end question-answering models. The dataset consists of 20,000 question-answer pairs originally collected in English, each annotated with Wikidata entities. These pairs have been translated into eight additional languages: Arabic, French, German, Hindi, Italian, Japanese, Portuguese, and Spanish, resulting in a total of 180,000 samples. For evaluation purposes, we selected a subset of 50 sentences from each language, totaling 500 sentences. This subset allowed us to perform focused and controlled experiments. The dataset enabled us to assess question-answering models in a multilingual context, maintaining entity-related information across translations.

An example of a data entry from the dataset is shown in the snippet below:


```
{
  "id": "Q2461698_0",
  "wikidata_id": "Q2461698",
  "entity_types": [
    "Fictional entity"
  ],
  "source": "Who are the main
      antagonistic forces in the World
      of Ice and Fire?",
  "targets": [
    {
      "translation": "Chi sono le
          principali forze antagoniste
          nel mondo delle Cronache del
          ghiaccio e del fuoco?",
      "mention": "mondo delle Cronache
          del ghiaccio e del fuoco"
    }
  ],
  "source_locale": "en",
  "target_locale": "it"
}
```


Each entry in the dataset includes an ‘id‘, which is a unique identifier combining the Wikidata entity ID and the question ID , and a ‘wikidata_id‘, which refers to the specific entity being mentioned . The ‘entity_types‘ field lists the types of entities involved, such as "Fictional entity." The ‘source‘ field contains the original English question, while the ‘targets‘ field includes the translated question in the target language (in this case, Italian), along with the ‘mention‘ of the entity in the translation. The ‘source_locale‘ and ‘target_locale‘ fields specify the language codes for the source and target languages, respectively.

Model	ar_AE	de_DE	es_ES	fr_FR	it_IT	ja_JP	ko_KR	th_TH	tr_TR	zh_TW	Average
m2m100_418M	0.00	12.00	0.00	12.00	0.00	0.00	0.00	0.00	0.00	0.00	2.40
m2m100_1.2B	0.00	12.00	0.00	24.00	0.00	0.00	4.00	0.00	6.00	0.00	4.60
llama3.1_8B-Instruct	0.00	2.00	12.00	4.00	6.00	0.00	6.00	0.00	0.00	0.00	3.00
qwen2.5_3B-Instruct	0.00	10.00	30.00	38.00	16.00	4.00	6.00	6.00	16.00	0.00	12.60
qwen2.5_7B-Instruct	10.00	16.00	34.00	48.00	26.00	10.00	8.00	6.00	16.00	2.00	17.60
gemma2-9b-it	14.00	36.00	48.00	54.00	34.00	12.00	18.00	12.00	32.00	0.00	26.00
gemma2+EL (with given ID)	**96.00**	**88.00**	**98.00**	**80.00**	**92.00**	**88.00**	**80.00**	**94.00**	**92.00**	**40.00**	**84.80**
NER+gemma2+EL	70.00	48.00	88.00	66.00	70.00	50.00	66.00	64.00	72.00	34.00	62.80

Table 1: M-ETA scores for each translation model across ten languages.

Model	ar_AE	de_DE	es_ES	fr_FR	it_IT	ja_JP	ko_KR	th_TH	tr_TR	zh_TW	Average
m2m100_418M	59.02	71.24	73.53	75.37	72.15	61.87	63.27	55.93	66.44	62.02	66.08
m2m100_1.2B	61.22	73.44	76.10	79.78	75.87	63.81	66.98	58.83	68.78	63.70	68.85
llama3.1_8B-Instruct	72.38	79.70	78.22	72.84	86.56	80.72	85.14	45.75	85.76	71.89	75.89
qwen2.5_3B-Instruct	74.39	81.16	85.75	84.22	83.17	79.99	83.39	67.81	74.64	77.92	79.24
qwen2.5_7B-Instruct	81.31	82.96	86.86	86.50	83.02	84.93	84.39	70.92	82.25	82.41	82.55
gemma2-9b-it	86.75	86.31	89.09	88.69	85.99	86.39	89.46	80.42	89.96	84.64	86.77
gemma2+EL (with given ID)	**90.53**	**89.59**	**93.08**	**92.28**	**93.47**	**92.89**	**92.25**	**90.49**	**93.04**	88.89	**91.65**
NER+gemma2+EL	90.26	86.86	90.52	91.06	90.92	89.13	91.31	88.16	92.88	**89.42**	90.05

Table 2: COMET scores for each translation model across ten languages

4.2 Benchmarking Base Models

In this section, we present a comparative analysis of several translation models, including both general-purpose large language models (LLMs) and models fine-tuned for translation tasks. The models evaluated in our experiments include `m2m100_418M`, `m2m100_1.2B`, `llama3.1_8B-Instruct`, `qwen2.5_3B-Instruct`, `qwen2.5_7B-Instruct`, and `gemma2-9b-it`.

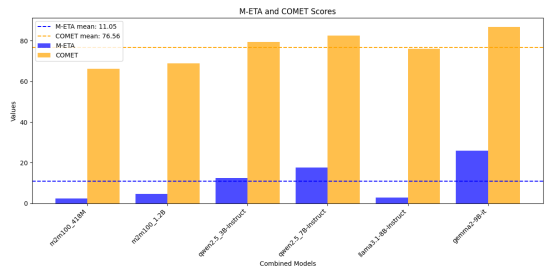

Figure 1: Comparison of different base models

The results in Table 1 and in Table 2 show that `gemma2-9b-it` consistently outperforms all other models across different evaluation metrics. It achieves the highest scores in both the M-ETA and COMET benchmarks, demonstrating its ability to generate high-quality translations across a wide range of languages. In contrast, the performance of the other models, including both the translation-specific `m2m100` variants and the general-purpose LLMs, tends to be more variable depending on the language pair.

Notably, the `gemma2-9b-it` model's robust performance highlights its strengths in handling multilingual translation tasks, setting it apart from the other models in this evaluation.

4.3 NER Models Comparison

In the context of integrating a Named Entity Recognition (NER) model into the translation pipeline, selecting a model that optimizes both precision and entity recognition capability is crucial. The results obtained, shown in Figure 3, reveal a clear difference in the performance of the tested NER models. The `dslim/bert-base-NER` and `dslim/bert-large-NER` models achieve similar results, with a precision of 0.75, a recall of 0.62, and an F1-score of 0.68. These models offer good precision, but they do not excel in recovering entities.

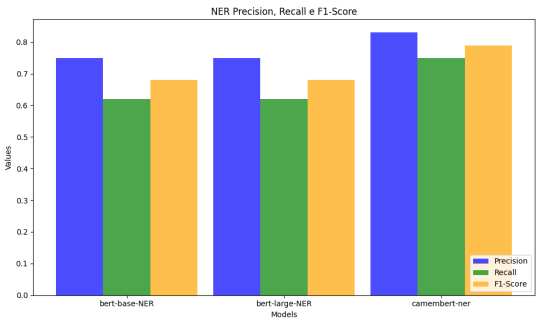

Figure 2: Comparison of different NER models

In contrast, the Jean-Baptiste/camembert-ner model demonstrates significantly better performance, with a precision score of 0.83, a recall of 0.75, and an F1-score of 0.79, making it the preferred choice for applications requiring strong entity recognition capabilities. These results suggest that integrating the `Jean-Baptiste/camembert-ner` model into the translation pipeline would provide more effective entity handling in translated texts, improving both the quality and overall accuracy of the process.

4.4 Entity Linking Implementation and Comparison

Entity linking (EL) was implemented in two distinct ways: one utilizing gold data with predefined entity IDs, and the other using entity IDs extracted from a NER model. While the gold data approach guarantees the correct linking of entities, the NER-based approach simulates a more realistic setting where entity recognition may be less precise.

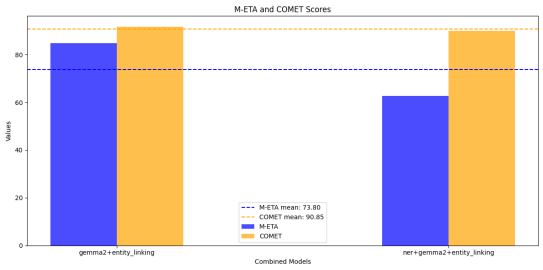

Figure 3: Comparison of Entity Linking using NER model or gold data

Table 1 presents the M-ETA scores for both configurations. The model with gold data (`gemma2+EL` with given ID) outperforms the NER-based model (`NER+gemma2+EL`) across all languages, with an average score of 84.80 compared to 62.80. This difference is expected, as the NER model is not perfect, and its output naturally introduces some errors in entity identification, affecting the overall linking performance.

However, Table 2 shows the COMET scores, where the performance gap between the two methods narrows significantly. The gold data method (`gemma2+EL` with given ID) achieves an average score of 91.65, while the NER-based method (`NER+gemma2+EL`) scores 90.05. The smaller difference in COMET scores suggests that, despite the inherent imperfections in the NER model, the system performs almost equally well when evaluating semantic quality, particularly in real-world scenarios where gold data is unavailable.

While it is clear that using gold data provides superior performance, especially for M-ETA, it is important to consider the trade-off in real-world applications. Gold data is often not available in practice, making the NER-based method a more feasible option. Moreover, despite the lower precision in entity linking, the NER-based approach still delivers competitive performance, particularly in terms of COMET scores, demonstrating its robustness and effectiveness in a real-world setting.

5 Conclusion

Our experiments revealed that employing gold entity IDs leads to significantly higher accuracy at the entity level, as demonstrated by M-ETA scores. However, when using entity IDs predicted by the NER model, a scenario that better reflects real-world conditions, the overall translation quality remains competitive, as indicated by only a modest drop in COMET scores. This highlights an inherent trade-off: while gold data sets the upper benchmark for performance, the realistic NER-based approach still ensures robust translation quality despite its inevitable imperfections.

Moreover, our results confirm that building the pipeline on a strong base model such as Gemma2-9b-it is critical to achieving high semantic fidelity, as the incorporation of entity-specific information substantially mitigates errors common in multilingual translation.

Looking ahead, future work should focus on improving the accuracy of NER, particularly in low-resource languages, to reduce the gap between ideal and real-world performance. Further, exploring more advanced, context-aware entity linking techniques could help resolve ambiguities more effectively. Finally, extending our approach to domain-specific applications (e.g., legal, medical, or technical translation) may further validate and enhance the adaptability of entity-aware machine translation systems.

Overall, our findings underscore the potential of integrating entity-aware strategies into machine translation systems to produce more accurate and context-sensitive translations across diverse languages.

References

[1] Dream AI. Linking extracted entities to wikidata: Why and how?, 2023. Accessed: 2025-02-20.

[2] Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, and Ahmed El-Kishky et al. Beyond english-centric multilingual machine translation, 2020.

[3] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, and Abhishek Kadian et al. The llama 3 herd of models, 2024.

[4] Ali Hatami, Ruslan Mitkov, and Gloria Corpas Pastor. Cross-lingual named entity recognition via FastAlign: a case study. In Ruslan Mitkov, Vilelmini Sosoni, Julie Christine Giguère, Elena Murgolo, and Elizabeth Deysel, editors, *Proceedings of the Translation and Interpreting Technology Online Conference*, pages 85–92, Held Online, July 2021. INCOMA Ltd.

[5] Junjie Hu, Hiroaki Hayashi, Kyunghyun Cho, and Graham Neubig. Deep: Denoising entity pre-training for neural machine translation, 2021.

[6] Jean-Baptiste. Camembert ner model, 2025.

[7] Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. Generalization through memorization: Nearest neighbor language models, 2020.

[8] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks, 2021.

[9] Zihan Liu, Feijun Jiang, Yuxiang Hu, Chen Shi, and Pascale Fung. Ner-bert: A pre-trained model for low-resource entity tagging, 2021.

[10] Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, and Bo Zheng et al. Qwen2.5 technical report, 2025.

[11] Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. Comet: A neural framework for mt evaluation, 2020.

[12] Amazon Science. Mintaka: A multilingual question-answering dataset, 2025.

[13] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units, 2016.

[14] Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, and Surya Bhupatiraju et al. Gemma 2: Improving open language models at a practical size, 2024.