Deerlu at SemEval-2025 Task 2: Wikidata-Driven Entity-Aware Translation: Boosting LLMs with External Knowledge

Lu Xu
Sapienza NLP Group, Sapienza University of Rome
xu@diag.uniroma1.it

Abstract

This paper presents an entity-aware machine translation system that significantly improves named entity translation by integrating external knowledge from Wikidata with Large Language Models (LLMs). While LLMs demonstrate strong general translation capabilities, they struggle with named entities that require specific cultural or domain knowledge. We address this challenge through two approaches: retrieving multilingual entity representations using gold Wikidata IDs, and employing Relik, an information extraction tool, to automatically detect and link entities without gold annotations. Experiments across multiple language pairs show our system outperforms baselines by up to 63 percentage points in entity translation accuracy (m-ETA) while maintaining high overall translation quality. Our approach ranked 3rd overall and 1st among non-finetuned systems on the SemEval-2025 Task 2 leaderboard. Additionally, we introduced language-specific post-processing further enhances performance, particularly for Traditional Chinese translations.

1 Introduction

Machine translation (MT) has witnessed remarkable advancements in recent years, largely driven by neural approaches and, more recently, large language models (LLMs). Despite these improvements, the accurate translation of named entities remains a significant challenge.

Named entities—proper names referring to people, famous landmarks, and cultural artifacts, which often require specialized handling that goes beyond standard translation procedures. The challenge of named entity translation is multifaceted. Many entities demand specific localized forms in the target language (e.g., country names, famous landmarks). Some entities, particularly those relating to cultural artifacts like books, movies, and products, may have official translations or established conventions in target languages that must be adhered to for accurate communication.

Traditional MT systems typically struggle with named entities for several reasons. First, named entities are often rare in training data, leading to poor representation in the model's parameters. Second, ambiguity in entity references frequently requires contextual or world knowledge to resolve correctly. Third, domain-specific or culturally-specific entities demand specialized knowledge that general MT systems may lack.

These challenges become even more pronounced when translating between languages with different writing systems or culturally distant contexts. For instance, translating English entity names into languages like Chinese, Japanese, or Arabic involves not just semantic transfer but also phonetic adaptation and cultural localization.

To address these challenges, we explore how LLMs can be enhanced with external knowledge to better handle these challenging cases. Our approaches leverage Wikidata as an external knowledge source and demonstrate significant improvements in both entity translation accuracy and overall translation quality. Our main contributions are as follows:

- We establish baseline performance for entity-aware translation using state-of-the-art LLMs (Qwen-Plus, Qwen-Max, and GPT-4o-mini) with simple prompting strategies
- We propose a novel approach using gold Wikidata ID to retrieve multilingual entity information before translation, and leverage these information to guide the translation
- We develop a more practical approach using Relik (Orlando et al., 2024), an information extraction tool, to automatically identify entities and retrieve their multilingual represen-

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 1802–1809
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

tations without relying on gold entity annotations

- We implement language-specific post-processing techniques to address issues such as simplified/traditional Chinese character conversion

2 Related Work

The challenge of translating named entities has been a longstanding issue in machine translation research. Our work addresses the task introduced by (Conia et al., 2025) in SemEval-2025 Task 2, which highlights the limitations of existing translation models in handling named entities that require more than literal translation.

Previous work by (Conia et al., 2024) established the foundation for this task by demonstrating the effectiveness of retrieval-augmented generation (RAG) from multilingual knowledge graphs. While (Zhao et al., 2020) showed promising results with knowledge-enhanced translation in the biomedical domain, such approaches often struggle to generalize beyond specific domains to broader translation scenarios.

The recent emergence of LLM offers a potential solution to these generalization challenges. (Zhu et al., 2024) demonstrated impressive zero-shot translation abilities with LLMs, while (Zhang et al., 2023) and (Guo et al., 2024) explored various prompting strategies to enhance their translation quality. These models show particular promise for low-resource language pairs (Dai et al., 2025) and challenging linguistic phenomena (Nicholas and Bhatia, 2023). Despite these advances, (Guerreiro et al., 2023) identified that LLMs remain prone to hallucinations when translating entities they have limited knowledge of, highlighting the need for augmentation with external knowledge sources.

To effectively integrate external knowledge, robust entity recognition and linking capabilities are essential. Existing systems like BLINK (Wu et al., 2020), GENRE (De Cao et al., 2020), and cross-lingual approaches (Botha et al., 2020) provide valuable capabilities but often require significant computational resources. Our work overcomes these constraints by leveraging ReLiK (Orlando et al., 2024), which provide en efficient method for identifying entities in text and connecting them to knowledge graph entries.

Our work builds upon these foundations, combining the strengths of LLMs with external knowledge retrieval to address the specific challenges of entity-aware machine translation. By leveraging tools like ReLiK for entity detection and Wikidata for multilingual entity information, we create a system that significantly improves named entity translation.

3 System Description

This section describes our entity-aware machine translation approach, which enhances LLMs with external knowledge to improve translation of sentences containing named entities.

3.1 Baseline Systems

We conducted experiments using three pre-trained LLMs: Qwen-plus, Qwen-max, and GPT-4o-mini. Our initial baseline used minimal prompting, instructing the model to translate from source to target language with a simple system prompt without any inspiration of entity information.

In particular, we noticed that the dataset we are worked with is composed of questions, so we implemented a second baseline with a more explicit prompt to address cases where models attempted to answer questions rather than translate them. This explicit identification of the sentence improved task adherence but did not resolve the fundamental challenge of entity translation.

3.2 Entity-Enhanced Translation

To overcome the limitations in entity translation, we developed two approaches that incorporate external knowledge:

Gold Entity Knowledge Integration. Our first approach leverages gold Wikidata IDs provided in the dataset. First we query the Wikidata API to retrieve entity names in both source and target languages, and enhance the translation prompt by including these entity mappings. Then let LLMs translate sentences with explicit knowledge of the correct entity representations. This approach significantly improves translation accuracy and overall translation quality.

Automatic Entity Detection and Knowledge Retrieval. However, gold entity annotations are impractical in real-world scenarios, so we further developed an approach free the system from gold entity annotations. We employ *Relik* (Orlando et al., 2024), an information extraction tool, to identify potential named entities in the source text. For each detected entity, we query Wikidata to retrieve corresponding entity names in both source and target languages. These automatically retrieved entity mappings are integrated into the translation prompt to guide LLMs during the translation process. While this approach does not achieve the same level of accuracy as using gold entity data, it significantly outperforms baselines and represents a practical solution for real-world applications.

3.3 Post-Processing for Language-Specific Challenges

For Traditional Chinese (zh-TW) translations, we implemented a targeted post-processing step using the *zhconv* tool to convert any Simplified Chinese characters in the output to Traditional Chinese. This addresses the common issue where LLMs produce mixed character sets despite instructions to use Traditional Chinese.

3.4 System Components

Our entity-aware translation system integrates the following key components:

- Pre-trained LLMs: Qwen-plus, Qwen-max, and GPT-4o-mini
- *Relik* for entity extraction and linking to Wikidata
- Wikidata API for cross-lingual entity name retrieval
- *zhconv* for Traditional Chinese character conversion

The prompt templates and detailed examples of entity-enhanced prompts are provided in the Appendix. We proposed a novel approach to integrating external knowledge into LLM-based translation, specifically targeting the challenging problem of entity translation across languages.

4 Results and Analysis

In this section, we present our experimental results on the entity-aware machine translation task, comparing our system against baselines and analyzing factors affecting translation quality and entity handling across different settings.

4.1 Main Results

Table 1 summarizes the performance of our systems evaluated on m-ETA (entity translation accuracy), COMET (Rei et al., 2020) (overall translation quality), and the overall score (harmonic mean of m-ETA and COMET).

	Average across all languages		
System	**M-ETA**	**COMET**	**Overall**
GPT-4o-mini	0.317	0.903	0.469
GPT-4o-mini-relik	0.724	0.931	0.815
GPT-4o-mini-gold	0.851	0.943	0.895
Qwen-plus	0.257	0.879	0.398
Qwen-plus-relik	0.728	0.922	0.814
Qwen-plus-gold	0.880	0.940	0.909
Qwen-max-relik	0.713	0.928	0.806
Qwen-max-gold	0.883	0.947	0.914
Our system-relik	0.714	0.928	0.807
Our system-gold	**0.890**	**0.948**	**0.917**

Table 1: Performance comparison of different system configurations on the entity-aware translation task. Systems with "-relik" use automatic entity detection, while "-gold" systems use gold Wikidata entity information.

Our experiments reveal a clear performance gap between baseline LLMs without entity knowledge and our enhanced approaches. Without external entity information, models like GPT-4o-mini and Qwen-plus achieve m-ETA scores of only 0.317 and 0.257 respectively, confirming the difficulty LLMs face in correctly translating named entities using only their parametric knowledge.

Adding gold Wikidata entity information dramatically improves performance, with m-ETA scores increasing by approximately 60 percentage points across all models. More importantly, our Relik-based approach, which automatically identifies entities without relying on gold annotations, achieves m-ETA scores of 0.714-0.728, representing a practical solution for real-world scenarios.

Table 2 compares our systems with top performers on the SemEval leaderboard. Our gold-enhanced system ranked 3rd overall and achieved the highest COMET score (94.76%) among all non-finetuned systems. Notably, while not officially submitted, our Relik-based system (71.35% m-ETA, 92.82% COMET) would have outperformed the best non-gold and non-finetuned system on the leaderboard, demonstrating the effectiveness of our approach in practical scenarios.

4.2 Impact of Entity Knowledge Integration

To understand how different levels of entity information affect translation quality, we conducted a controlled experiment using GPT-4o-mini across ten language pairs. We compared three conditions: (1) no external entity information, (2) source language entity information only, and (3) complete entity information for both source and target lan-

System	Uses Gold	Finetuned	LLM Name	m-ETA	COMET	Overall score	Rank
Top System	Yes	Yes	Qwen2.5	89.10%	94.74%	91.79%	1
Top Non-gold System	No	Yes	GPT-4o-mini	77.13%	91.81%	83.63%	17
Top Non-gold & Non-finetuned System	No	No	Qwen2.5	68.24%	91.64%	78.17%	19
Ours-Gold(Top Non-finetuned System)	Yes	No	Qwen2.5-max	88.95%	**94.76%**	91.74%	3
Ours-Relik	No	No	Qwen2.5-max	71.35%	92.82%	80.68%	-

Table 2: Comparison with top systems on the SemEval-2025 Task 2 leaderboard. Our Relik-based system was not submitted to the official leaderboard.

guages.

Language	w/o Info	Source Info	Full Info
EN-ZH	32.44%	33.30%	64.61%
EN-AR	25.53%	25.45%	91.18%
EN-DE	35.11%	35.57%	84.97%
EN-IT	36.19%	38.21%	91.72%
EN-JA	31.42%	33.48%	87.43%
EN-KO	30.76%	29.59%	86.32%
EN-ES	42.19%	45.05%	89.68%
EN-TH	13.05%	12.97%	89.79%
EN-TR	35.03%	35.14%	75.45%
EN-FR	34.86%	37.44%	89.53%
Average	31.66%	32.62%	85.07%

Table 3: Entity translation accuracy (m-ETA) with varying levels of external entity information across language pairs. Language codes: Chinese (ZH), Arabic (AR), German (DE), Italian (IT), Japanese (JA), Korean (KO), Spanish (ES), Thai (TH), Turkish (TR), and French (FR).

As shown in Table 3, providing only source language entity information yields minimal improvement (31.66% to 32.62% on average), suggesting that recognizing the entity alone is insufficient. The model requires the target language entity information to perform accurate translation. When complete entity information is provided, performance improves dramatically across all language pairs, with an average increase of over 53 percentage points.

This pattern is particularly striking for language pairs with significant linguistic distance from English. For Thai, m-ETA increases from 13.05% to 89.79% when complete entity information is provided, highlighting how critical external knowledge is for translating entities into languages with different writing systems or cultural contexts.

4.3 Handling Traditional Chinese

For Chinese translations, we observed that models frequently produced a mixture of Simplified and Traditional Chinese characters despite explicit instructions to generate Traditional Chinese. This inconsistency significantly affected evaluation metrics when comparing against gold Traditional Chinese references.

Table 4 demonstrates the effectiveness of our post-processing approach using the *zhconv* tool. This simple yet effective step improved m-ETA scores by approximately 6-16 percentage points across all models, with the most significant improvement seen in GPT-4o-mini (from 64.61% to 80.64%).

Interestingly, we also observed that without additional entity information but with the same post-processing step, LLMs generally performed better when asked to translate into Simplified Chinese rather than Traditional Chinese, suggesting a potential bias in model training toward more widely used character sets.

Systems	w/o zhconv	w zhconv
GPT-4o-mini	64.61%	80.64%
Qwen2.5-plus	73.93%	80.51%
Qwen2.5-max	74.78%	80.76%

Table 4: Impact of post-processing to convert Simplified to Traditional Chinese on m-ETA scores.

4.4 Additional Findings

Our experiments revealed several additional insights about entity-aware translation:

Impact on Low-Resource Languages The benefits of entity knowledge integration are particularly pronounced for low-resource languages. For instance, Thai showed one of the most dramatic improvements (13.05% to 89.79%) when complete entity information was provided. Similar trend can be find in Korean. Due to the scarcity of data, LLMs struggled more with low-resource languages when no entity information was provided. However, when we retrieved entity information using Relik or gold annotations and provided entity information to the model, the performance improved significantly. In some cases, the performance for low-resource languages surpassed that of rich-resource languages like German, as shown in Table 3.This suggests that external knowledge can effectively compensate for limited training data in the model's parameters.

Task Comprehension Given that all sentences in the dataset are questions, we found that explicitly indicating the sentence to be translated in the prompt improved model performance. Without this specification, models occasionally attempted to answer the question rather than translate it, particularly in baseline configurations.

Prompt Language We explored whether prompting in the target language rather than English would improve performance. Results showed minimal and inconsistent effects across language pairs, suggesting that the availability of external entity knowledge is a much stronger determinant of performance than the language of instruction.

These findings collectively underscore the importance of external knowledge integration for entity-aware translation and highlight the effectiveness of our proposed approaches in addressing this challenging aspect of machine translation.

5 Conclusion

We introduced an entity-aware machine translation system that improves the translation of named entities by integrating external knowledge from Wikidata. Using gold Wikidata IDs or the Relik tool for automatic entity extraction, our approach outperforms baseline models and ranks highly on the official leaderboard. This demonstrates the effectiveness of incorporating external knowledge to address the limitations of LLMs in handling named entities during translation tasks.

This paper presented an effective approach to entity-aware machine translation by enhancing LLMs with external knowledge retrieval. We demonstrated that while state-of-the-art LLMs possess impressive general translation capabilities, they struggle significantly with named entity translation, particularly for culturally-specific entities or those requiring specialized knowledge.

Our experimental results across multiple language pairs confirm that integrating entity knowledge from Wikidata substantially improves both entity translation accuracy and overall translation quality. The gold entity knowledge integration approach achieved near-optimal performance (m-ETA of 89.0%, COMET of 94.8%), ranking among the top systems in the SemEval-2025 Task 2 competition. More importantly, our practical Relik-based approach, which automatically identifies and links entities without requiring gold annotations, achieved competitive results (m-ETA of 71.4%, COMET of 92.8%) while being applicable to real-world translation scenarios.

Analysis of our approach revealed several key insights: (1) providing complete entity information in both source and target languages is crucial for accurate entity translation, (2) automatic entity detection with knowledge retrieval is highly effective for practical applications, and (3) targeted post-processing for specific language challenges, such as Traditional Chinese character conversion, can yield substantial gains.

Our work contributes to the ongoing efforts to make machine translation more reliable for real-world scenarios where accurate handling of named entities is essential for effective cross-cultural communication.

Limitations

While our approach significantly improves entity-aware machine translation, several limitations should be acknowledged:

First, our system's effectiveness is contingent upon the quality and coverage of Wikidata. For low-resource languages or specialized domains, Wikidata may lack comprehensive entity information or accurate translations. Furthermore, as a dynamic knowledge source that undergoes frequent updates, Wikidata's evolving nature may lead to inconsistent translations of certain entities over time, potentially affecting reproducibility.

Second, the Relik-based approach, though effective in practical scenarios, introduces potential error propagation. Inaccuracies in entity detection or linking to incorrect Wikidata entries directly impact translation quality. Our analysis shows that approximately 15-20% of translation errors with the Relik approach stem from entity linking failures rather than translation model limitations, detailed analysis can be find in Appendix.

Third, our post-processing solutions, such as Chinese script conversion, introduce language-specific complexity that doesn't generalize well across all language pairs. This approach requires maintaining separate conversion pipelines for different writing systems, increasing implementation complexity.

Finally, despite strong performance, our approach still requires multiple API calls to external services for each sentence containing entities, introducing latency that may be problematic for real-time applications or high-volume translation services.

Acknowledgments

This work has been carried out while Lu Xu was enrolled in the Italian National Doctorate on Artificial Intelligence at Sapienza University of Rome.

References

Jan A Botha, Zifei Shan, and Dan Gillick. 2020. Entity linking in 100 languages. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7833–7845.

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.

Yuqian Dai, Chun Fai Chan, Ying Ki Wong, and Tsz Ho Pun. 2025. Next-level cantonese-to-mandarin translation: Fine-tuning and post-processing with llms. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 427–436.

N De Cao, G Izacard, S Riedel, and F Petroni. 2020. Autoregressive entity retrieval. In *ICLR 2021-9th International Conference on Learning Representations*, volume 2021. ICLR.

Nuno M Guerreiro, Duarte M Alves, Jonas Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo, and André FT Martins. 2023. Hallucinations in large multilingual translation models. *Transactions of the Association for Computational Linguistics*, 11:1500–1517.

Ping Guo, Yubing Ren, Yue Hu, Yunpeng Li, Jiarui Zhang, Xingsheng Zhang, and He-Yan Huang. 2024. Teaching large language models to translate on low-resource languages with textbook prompting. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15685–15697.

Gabriel Nicholas and Aliya Bhatia. 2023. Lost in translation: large language models in non-english content analysis. *arXiv preprint arXiv:2306.07377*.

Riccardo Orlando, Pere-Lluís Huguet Cabot, Edoardo Barba, and Roberto Navigli. 2024. Relik: Retrieve and link, fast and accurate entity linking and relation extraction on an academic budget. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 14114–14132.

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.

Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2020. Scalable zero-shot entity linking with dense entity retrieval. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6397–6407.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023. Prompting large language model for machine translation: A case study. In *International Conference on Machine Learning*, pages 41092–41110. PMLR.

Yang Zhao, Jiajun Lu, and Jinfu Chen. 2020. Knowledge graph enhanced neural machine translation via multi-task learning on sub-entity granularity. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4495–4505. International Committee on Computational Linguistics.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. Multilingual machine translation with large language models: Empirical results and analysis. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781.

A Appendix A

A.1 Prompt Templates

This appendix provides the prompt templates used in our experiments and examples of how they are implemented for specific translation tasks.

A.1.1 Basic Baseline Prompt

```
Prompt-0 You are an expert translator. Translate from {SOURCE_LANGUAGE} to {target_language}.
```

Only provide the translation without explanations.

{sentence}

A.1.2 Enhanced Baseline Prompt (Explicitly Identifying the Sentence)

Prompt-1 You are an expert translator. Translate from {SOURCE_LANGUAGE} to {target_language}.

Only provide the translation without explanations.

The sentence is: {sentence}

A.1.3 Source Only Entity-Enhanced Prompts

Prompt-2 You are an expert translator. Translate from {SOURCE_LANGUAGE} to {target_language}.

Only provide the translation without explanations.

The sentence contains an entity, and its mention is provided.

If the mention is ‘Label not found’, translate the entire sentence on your own.

The sentence is: {source_text}

The mention is {source_title}

A.1.4 Full Entity-Enhanced Prompts

Prompt-3 You are an expert translator. Translate from {SOURCE_LANGUAGE} to {target_language}.

Only provide the translation without explanations.

The sentence contains an entity. The entity name is specified in both the source and target languages. When you translate the sentence, please use the specified mention in the target language.

If the mention is ‘Label not found’, translate it it by yourself.

The sentence is: {source_text}

The entity in {SOURCE_LANGUAGE} is {source_title}, in {target_language} is {target_title}.

A.2 Examples of Language-Specific Prompts

A.2.1 Traditional Chinese Prompt

Prompt-4-zh 您是翻譯專家。英譯漢（繁體）。只翻譯譯文，不做解釋。

句子中存在實體。實體名稱在源語言和目標語言中均已指定。您翻譯句子時，請使用目標語言中指定的提及。如果目標語言中的提及是’Label not found’，請自行翻譯。

句子：{source_text}

英文中的實體是{source_title}。

中文（繁體）中的實體是{target_title}。

A.2.2 Japanese Prompt

Prompt-4-JA あなたは熟した翻者です。英から日本に翻してください。明はせずに翻のみを翻してください。

文にはエンティティがあります。エンティティ名はソス言とタゲット言の方で指定されています。文を翻するときは、タゲット言で指定された言及を使用してください。タゲット言の言及が「Label not found」の合は、自分で翻してください

文: {source_text}

英のエンティティは {source_title} です。

日本では {target_title} です。

A.2.3 Italian Prompt

Prompt-4-IT Sei un traduttore esperto. Traduci dall’inglese all’italiano. Traduci solo la traduzione senza spiegare.

Ci sono entità nella frase. Il nome dell’entità è specificato sia nella lingua di origine che in quella di destinazione. Quando traduci la frase, usa la menzione specificata nella lingua di destinazione. Se la menzione nella lingua di destinazione è ’Label not found’, traducila da solo.

Frase: {source_text} L’entità in inglese è {source_title}.

In italiano è {target_title}.

B Appendix B

B.1 Error Analysis

This appendix presents a detailed error analysis of our system outputs, focusing on Qwen2.5-Max. Through careful examination of translation errors, particularly instances of entity mismatch, we identified three primary error categories:

- **Wikidata-Dataset Misalignment**: Despite using gold Wikidata IDs provided in the dataset, entity name retrieval sometimes produces translations that differ from the gold labels. This discrepancy stems from Wikidata’s continual updates since the dataset’s creation. For example, the entity "Q1024181" (Pushkin House) returns "普希金屋" in Chinese from current Wikidata, while the dataset's gold label is "普希金之家"—a subtle but evaluation-affecting difference.

As shown in Table 5, these Wikidata-dataset misalignments result in 10.8% errors on average across languages, with values ranging from 7.34% for Italian to 18.87% for Traditional Chinese. This represents an artificial evaluation penalty rather than a true translation error, as both forms may be valid translations. Further more, we observe that LLMs can fix some small misalignments, because the miss match rate dropped 1% after the LLMs translation.

- **Relik Entity Retrieval Errors**: When using Relik for automatic entity detection, the system occasionally prioritizes prominent entities over the actual target entities. For instance, in "How does Jia Jing contribute to the overall plot of Dream of the Red Chamber?", Relik identifies the famous novel "Dream of the Red Chamber" but misses the character "Jia Jing" (the actual entity of interest).

  This entity retrieval error challenge affects approximately 19.94% target labels are miss match to the dataset label, as shown in Table 5.

- **Missing Target Language Labels**: In some cases, Wikidata lacks a corresponding entity name in the target language. When this occurs, the system passes "Label not found" as the target label, leading to two typical outcomes: (1) the model produces a translation that mismatches the gold label, or (2) the model retains the source language entity name untranslated.

  As indicated in Table 5, when gold information provided, there is less than 1% instance missing target language labels, when retrive entity information uesing Relik, there is 11.93% instance missing target language labels.

Table 5 presents the miss match rate that due to these error categories across language pairs. Our analysis reveals substantial variation across languages, reflecting different challenges in entity handling for each language pair.

Additionally, we compared the performance between our gold entity system and Relik-based system. As shown in Table 6, the Relik-based system exhibits notably higher error rates across all language pairs, with an average difference of 16.93 percentage points. This gap demonstrates the significant impact of accurate entity identification on translation quality. It also illustrates the error propagation from entity linking to the final translation results.

Language	**Gold Mismatch**	**Relik Errors**	**Missing Labels**
EN-AR	8.23%	16.14%	11.08%
EN-ZH	18.87%	19.37%	11.93%
EN-DE	12.92%	20.78%	10.79%
EN-IT	7.34%	21.56%	11.59%
EN-JA	7.83%	20.16%	10.85%
EN-KO	8.85%	22.39%	16.92%
EN-ES	10.70%	18.79%	9.50%
EN-TH	8.56%	19.52%	15.09%
EN-TR	15.58%	17.77%	13.08%
EN-FR	9.09%	21.17%	8.44%
Average	10.80%	19.94%	11.93%

Table 5: Mismatch rates by error type across language pairs.Language codes: Chinese (ZH), Arabic (AR), German (DE), Italian (IT), Japanese (JA), Korean (KO), Spanish (ES), Thai (TH), Turkish (TR), and French (FR).

Language	**Gold System**	**Relik System**
EN-AR	8.47%	23.91%
EN-ZH	25.22%	35.62%
EN-DE	14.06%	30.82%
EN-IT	7.16%	25.60%
EN-JA	8.50%	26.10%
EN-KO	9.31%	28.39%
EN-ES	9.74%	26.40%
EN-TH	9.02%	30.40%
EN-TR	16.12%	30.90%
EN-FR	8.93%	28.64%
Average	11.75%	28.68%

Table 6: Error rates comparison between systems.Language codes: Chinese (ZH), Arabic (AR), German (DE), Italian (IT), Japanese (JA), Korean (KO), Spanish (ES), Thai (TH), Turkish (TR), and French (FR).