SheffieldGATE at SemEval-2025 Task 2: Multi-Stage Reasoning with Knowledge Fusion for Entity Translation

Xinye Yang , Kalina Bontcheva , Xingyi Song
School of Computer Science
The University of Sheffield
{xyang138, k.bontcheva, x.song}@sheffield.ac.uk

Abstract

This paper describes the machine translation system submitted to the SemEval-2025 Entity-Aware Machine Translation Task by the SheffieldGATE Team. We proposed a multi-agent entity-aware machine translation system that operates through three distinct reasoning stages: entity recognition, knowledge enhancement, and translation decision-making. The innovation in our approach lies in leveraging large language models to generate contextually relevant queries during the knowledge enhancement stage, extracting candidate entities and their translations from external knowledge bases. In the final translation decision-making stage, we employ fine-tuned large language models to denoise the retrieved knowledge, selecting the most relevant entity information to ensure accurate translation of the original text. Experimental results demonstrate our system's effectiveness. In SemEval-2025 Task 2, our system ranks first among all systems in Spanish entity translation metrics and third in Italian. For systems that do not use gold standard entity IDs during test set inference, ours achieves the highest overall scores across four language pairs: German, French, Italian, and Spanish.

1 Introduction

Machine translation has made significant progress in recent years, but it still faces substantial challenges when processing texts containing named entities. Existing research focuses mainly on improving overall translation quality using metrics such as BLEU (Papineni et al., 2002), which treat all words equally. However, from a human comprehension perspective, different components within a sentence do not contribute equally to the quality of translation. Named entities often carry information crucial for accurate communication, cultural transmission, and domain-specific translation (Li et al., 2018). Particularly in dynamic contexts such as social media, even state-of-the-art translation models encounter significant difficulties when handling named entities (Rikters and Miwa, 2024).

The challenges in named entity translation primarily stem from several key factors. First, the continuous emergence of new entities makes it difficult for translation systems based on fixed vocabularies to adapt. Second, the correct translation of entities often depends on context. Additionally, in informal settings such as social media, users often employ entities in creative ways, further complicating the translation task. These challenges underscore the necessity of developing specialized entity-aware machine translation approaches.

To effectively address the challenges in named entity translation, this paper makes the following key contributions: First, we design a three-stage reasoning framework based on Large Language Models (LLMs) (Wei et al., 2023), specifically optimised for named entity translation. Second, we propose innovative entity query generation mechanisms that effectively integrate information from external knowledge bases. Experimental results show that our approach achieves significant improvements across multiple language pairs in named entity translation tasks. These findings provide new insights for the future development of entity-aware machine translation systems.

The remainder of this paper is organised as follows. Section 2 reviews related work, Section 3 describes the system design, Section 4 presents experimental results and analysis, and Section 5 concludes the paper and explores directions for future work.

2 Related Work

Named entity translation has emerged as a key challenge in machine translation. SemEval-2025 Task 2 (Conia et al., 2025) marks the first entity-aware machine translation evaluation task, providing a standardised assessment framework for re-

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 1495–1503

search in this field. The task uses the XC-Translate dataset (Conia et al., 2024), which represents the first large-scale manual annotated dataset focused on cross-cultural entity translation. Based on these developments, we present a review of major research progress in this field.

2.1 Entity-Aware Machine Translation

Entity-aware machine translation focuses on improving the translation of texts containing named entities. Recent research has developed several neural network architectures to enhance entity translation accuracy. These include entity classifiers embedded within encoder-decoder frameworks (Xie et al., 2022) and multi-task learning strategies that combine Named Entity Recognition (NER) with translation tasks (Rikters and Miwa, 2024). However, these methods face challenges in handling dynamic entities and informal text. Social media presents particular difficulties, where users often employ entities creatively, adding complexity to the translation task.

2.2 Knowledge-Enhanced Neural Machine Translation

Researchers are exploring ways to incorporate external knowledge into translation systems to improve the handling of named entities. Zeng et al.(Zeng et al., 2023) explored dictionary-based methods for entity translation, although this approach struggled with ambiguous entities. Conia et al.(Conia et al., 2024) proposed using multilingual knowledge graphs for retrieval-augmented generation(RAG), offering new perspectives for cross-cultural machine translation. These studies demonstrate that the effective use of external knowledge significantly improves the quality of entity translation.

2.3 Large Language Models for Translation

Large language models have transformed translation through their extensive pre-training data and capabilities. These models demonstrate superior entity disambiguation and translation performance compared to traditional neural machine translation models, primarily due to their exposure to vast amounts of multilingual data. The introduction of Chain-of-Thought reasoning (Wei et al., 2023) provides a new paradigm for complex language understanding tasks. However, these models may struggle with new entities or domain-specific knowledge, where their pre-trained knowledge can be outdated or imprecise.

3 System Description

We present **SHEF Machine Translation System** a multi-agent entity-aware machine translation system that employs a multi-stage reasoning mechanism. Through three key steps, entity extraction, knowledge enhancement, and translation decision, our system achieves high-quality entity translation.

The overall architecture of the system is illustrated in Figure 1.

3.1 System Architecture Overview

Our system implements a three-stage reasoning mechanism based on multiagent collaboration, breaking down translation tasks into subtasks. We incorporate a knowledge enhancement module that leverages external knowledge bases to improve the translation accuracy of entities and implement a verification and optimisation mechanism through a second agent to detect reasoning failures and maintain quality control.

3.2 Model Selections

Due to training resource constraints, we choose Llama-3.3-70B-Instruct (Aaron Grattafiori, 2024) as our base model. To optimise the performance of the model while reducing computational overhead, we employ QLoRA(Dettmers et al., 2023) for parameter-efficient fine-tuning. The model is specifically optimised for three tasks: entity recognition, knowledge fusion, and translation decision-making. We also integrate DeepSeek-R1(DeepSeek-AI et al., 2025) as a verification model.

3.3 Multi-stage Reasoning Mechanism

3.3.1 Entity Recognition Stage

In the first stage the LLM (Large Language Model) acts as a named entity recognizer, precisely extracting key entities from the source text. We design specific prompt templates to guide the model in identifying entities that have the most significant semantic impact on the original text. This precise entity identification lays the foundation for subsequent knowledge enhancement and translation.

3.3.2 Knowledge Enhancement Stage

The second stage implements LLM (Large Language Model)-based query enhancement with effi-

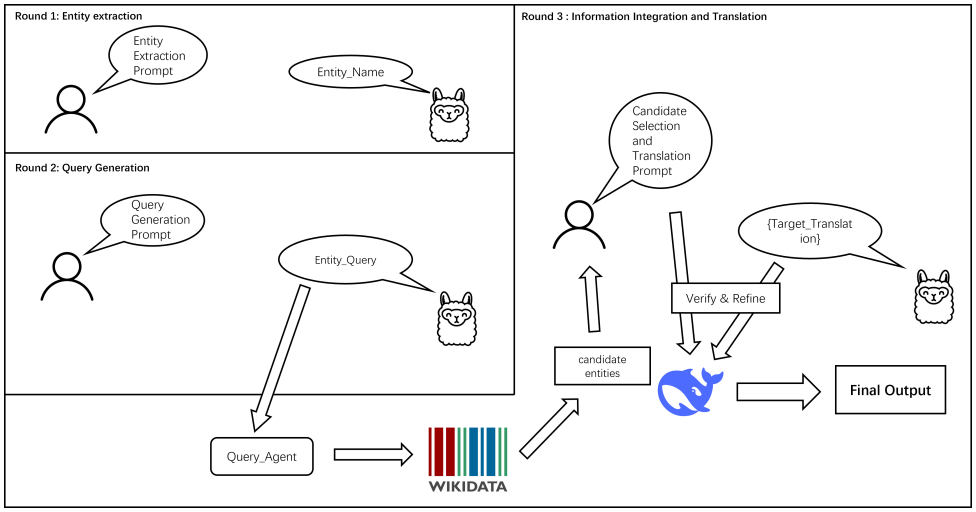

Figure 1: The architecture of our multi-agent cross-lingual entity translation system. The system consists of three main stages: (1) entity recognition, (2) knowledge enhancement with retrieval, and (3) translation decision-making, followed by a verification and optimisation module. The prompt used for this stage is provided in the Section 6 : Appendix

cient entity retrieval. Our approach consists of two main components:

Query Construction: We prompt the entity name (extracted in the first stage) along with the original text to the LLM (Large Language Model). This leverages the information from the original text and the model's pre-trained knowledge to generate enhanced query representations containing two key elements: the standardised entity name and a contextual entity description.

Entity Retrieval: As illustrated in Figure 2, our retrieval agent leverages the Wikidata API to initially retrieve entity names. The agent significantly narrows the search space by pruning retrieved named entities, retaining only the top 10 most relevant candidates. After obtaining these candidate entities, we employ SentenceTransformers (Reimers and Gurevych, 2019) to encode both the LLM (Large Language Model)-generated query representations and the names and descriptions of all candidates. The agent then determines the final similarity ranking using a weighted cosine similarity approach, wherein similarity scores for names and descriptions are computed separately and combined using predefined weights α and β.

This hybrid approach combines LLM (Large Language Model) knowledge enhancement with efficient retrieval techniques, enhancing semantic understanding while maintaining computational efficiency.

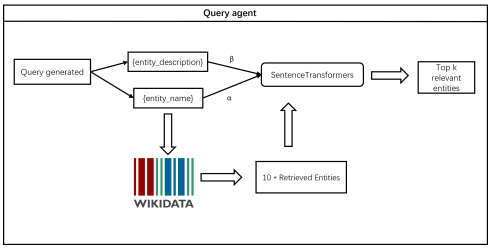

Figure 2: Overview of the entity retrieval process. The query agent retrieves candidate entities from Wikidata using the entity name and prunes them to the top 10 most relevant candidates. A weighted cosine similarity ranking is then applied to determine the top k entities.

3.3.3 Translation Stage

The third stage leverages the LLM (Large Language Model)'s semantic understanding and denoising capabilities for translation decisions. We prompt the large language model with the original text and information from retrieved candidate entities. This includes names and descriptions in the target language. Drawing on semantic understanding capabilities from pre-training, the LLM (Large Language Model) identifies which candidate entity best matches the original context. This effectively filters out candidates that are superficially similar but semantically distinct. Such filtering eliminates interference at the semantic level. During training, we implement negative sample learning. We provide potential entities and randomly replace one candidate with an irrelevant entity to introduce noise. Through this fine-tuning approach, we acti-

System	German		Spanish		French		Italian		Average	
	M	C	M	C	M	C	M	C	M	C
Llama-3.3-70B	36.10	88.50	45.34	91.37	38.98	87.78	40.74	89.54	40.29	89.30
Llama-3.3-70B_full_system	**85.57**	**92.82**	**90.5**	**93.91**	**90.05**	**91.93**	**93.02**	**94.68**	**89.8**	**93.3**
Ablation Study (Llama-3-8B)										
llama8b_baseline	24.15	85.71	30.25	89.21	21.39	83.92	25.77	86.27	25.39	86.28
llama8b_full_system	64.58	90.30	72.29	91.88	57.73	89.55	70.05	91.63	66.16	90.84
llama8b_deepseek_verify	70.43	89.05	75.39	87.77	63.83	89.07	74.34	91.78	71.00	89.42
llama_without_wiki	28.60	86.00	32.90	88.82	24.87	85.17	26.60	87.08	28.24	86.77

Table 1: Results across selected languages with M-ETA (M) and Comet (C) scores. The upper section shows results for our main experiments with Llama-3.3-70B, while the lower section presents our ablation study using Llama-3-8B.

vate the LLM (Large Language Model)'s inherent denoising capability. This enables accurate translation decisions in complex contexts.

3.4 Verification and Optimisation Mechanism

The system integrates an independent verification and optimisation module using the DeepSeek-R1(DeepSeek-AI et al., 2025) model for the detection of reasoning failures. Our failure detection rules focus on three aspects: completeness check of entity recognition, relevance verification of generated query, and semantic consistency assessment between translation results and the original text. Based on these rules, the verification model performs comprehensive reviews of the three-stage reasoning process, identifying potential errors and improving the results. The prompt used for this stage is provided in the Appendix.

3.5 Data Augmentation

We built a specialised three-stage dialogue dataset based on the XC-Translate dataset (Conia et al., 2024) to train our main model, focusing specifically on four language pairs: English-German (en-de), English-French (en-fr), English-Spanish (en-es), and English-Italian (en-it). Starting with gold-standard entity IDs provided in the XC-Translate dataset, we retrieved entity names and descriptions from Wikidata. These retrievals formed the basis for our first stage, which focused on entity recognition tasks and our second stage, which was dedicated to query generation. To augment the data, our entity retrieval module was employed to obtain entities most analogous for precise alignment, while semantically unrelated entities were intentionally introduced as negative examples during the third stage. This dual approach prevents the model from overgeneralising and forming incorrect associations, enabling the LLM (Large Language Model) to make accurate translation decisions not only when presented with similar entity information, but also across diverse inputs, rather than relying solely on similarity features. By combining these three stages of dialogue reasoning, the LLM (Large Language Model) better utilises information from the source text. It also leverages both its pre-training capabilities and its ability to interact with the external knowledge base.

3.6 Training Setup

Leveraging the excellent cross-lingual generalisation capabilities of large language models, where training on a language pair improves translation performance across other pairs (Yang et al., 2024), we implement a multilingual joint training strategy rather than developing separate models for each language pair. This approach maximises the model's inherent cross-lingual abilities, while substantially reducing computational resource requirements. All our experiments were conducted on 4 NVIDIA A100 GPUs (80GB each). Detailed information on model training hyperparameters, data preprocessing procedures, and experimental environment configurations is provided in Appendix Table 2.

4 Experiments and Results

4.1 Baseline Setup

We utilise a fine-tuned Llama-3.3-70B-Instruct (Aaron Grattafiori, 2024) as our baseline model, which undergoes the same multilingual joint training strategy on the XC-Translate dataset (Conia et al., 2024) as our system. The baseline was trained using data from four language pairs: English-German, English-French, English-Italian, and English-Spanish, without any data augmentation preprocessing. The detailed statistics of our

dataset are presented in Appendix Table 3. For each language pair, we reserved 10% of the samples as a validation set to monitor the training process. The training parameters remain consistent for both systems. The prompts used for baseline experiments are available in Section 6: Appendix.

4.2 Hyperparameters

In our system, we introduced three key hyperparameters to optimise the retrieval and translation process:

- α: The weight assigned to entity names generated by the LLM (Large Language Model) when retrieving from Wikidata. This parameter controls how much importance is given to the entity's name during the retrieval process.
- β: The weight assigned to entity descriptions generated by the LLM (Large Language Model) when retrieving from Wikidata. This parameter determines how much the system should consider the entity's description during retrieval.
- k: The number of candidate entities provided to the LLM (Large Language Model) in the final stage for translation decision-making. This parameter controls how many potential entity matches the model considers before making its final translation decision.

Together, these hyperparameters allow us to balance the relative importance of entity names versus descriptions (α and β) and control the breadth of candidates considered (k) during the entity translation process. In our submitted system, we assign $\alpha = 0.5$ and $\beta = 0.5$, ensuring equal contribution from both components. To optimise the balance between computational efficiency and retrieval effectiveness, we select (k = 3), which prevents excessive input length while providing sufficient candidate entities for downstream processing.

4.3 Evaluation Metrics

To evaluate our system, we employ two complementary metrics:

- **M-ETA** (Conia et al., 2024): Measures the proportion of correctly translated named entities by comparing predicted entity translations against gold standard references.
- **COMET** (Rei et al., 2020): A neural-based metric that evaluates overall translation quality by comparing machine translations to human references, providing scores for translation fluency and adequacy.

These metrics enable us to assess both entity translation accuracy and general translation quality.

4.4 Ablation Study

To evaluate component contributions in our framework, we conducted a comprehensive ablation study across all language pairs. Given the computational intensity of our full framework and due to significant time and hardware constraints, we performed these ablation experiments on the smaller Llama-3-8B model rather than the larger variants. This smaller-scale experimentation, shown in the lower section of Table 1, still provided valuable insights into component effectiveness. We compared our system against several variants: (1) a baseline system (llama8b_baseline) trained without our three-stage reasoning framework, (2) our standard system with all components (llama8b_full_system), (3) our system with an enhanced verification component (llama8b_deepseek_verify), and (4) a system without Wikidata retrieval (llama_without_wiki) that uses Chain-of-Thought (CoT) to emphasise entities.

4.4.1 Impact of Three-Stage Reasoning Framework

Our results demonstrate the substantial impact of our proposed framework on named entity translation. Comparing the baseline system with our standard implementation reveals an average improvement of 40.77 in M-ETA scores (from 25.39 to 66.16). COMET scores also improved by 4.56 (from 86.28 to 90.84). This significant performance gap underscores the importance of our structured approach. The three-stage framework effectively decomposes the complex task into manageable subproblems.

4.4.2 Importance of the Verification Component

Our error analysis revealed that most translation errors originated from the entity recognition stage. The verification component effectively addresses this issue. The enhanced verification system (llama8b_deepseek_verify) further improves M-ETA scores by 4.84 points (from 66.16 to 71.00)

compared to our standard system. This improvement is consistent across all language pairs, with the largest gains observed in German (5.85 points) and Italian (4.29 points).

While the verification component slightly reduces COMET scores in some languages, particularly Spanish (from 91.88 to 87.77), it maintains or improves scores in others. This suggests a trade-off between entity translation accuracy and overall fluency in certain contexts. Nevertheless, the substantial M-ETA improvements justify the inclusion of this component in mission-critical entity translation scenarios.

4.4.3 Knowledge Retrieval Importance and CoT Limitations

We observe that using CoT to emphasise entities without external knowledge base (Wikidata) retrieval shows minimal improvement. The average M-ETA score increased slightly from the baseline's 25.39 to 28.24. However, this improvement is negligible compared to our knowledge-enhanced systems. The full system outperforms the CoT-only approach by 37.92 M-ETA. COMET scores show a similar pattern, with the full system scoring 4.07 points higher than the CoT-only variant.

These findings indicate that while CoT is effective for complex reasoning tasks in general, its benefits are surprisingly modest for entity-aware translation, which requires broader and more comprehensive knowledge. External knowledge retrieval proves to be the critical component in our framework. The Wikidata integration provides authoritative entity information that may not be fully captured in the model's parametric knowledge. In our complete system, CoT serves not as a standalone solution but as an essential mechanism for integrating and reasoning with knowledge retrieved from external sources, enabling more effective use of this information during the translation process.

4.4.4 Language-Specific Patterns

Our ablation study reveals consistent patterns across languages. The performance gains are most pronounced for Spanish and Italian, with German and French showing relatively lower improvements. This pattern aligns with our main results in Table 1 and supports our observation about linguistic distance affecting entity translation performance. Even the lowest-performing pair (English-German) shows a substantial improvement compared to the baseline.

These ablation results validate our design choices. They emphasise the necessity of both the three-stage reasoning approach and external knowledge integration. The CoT technique provides only marginal benefits by itself. The verification component offers substantial improvements in entity translation accuracy, particularly for challenging cases missed in the initial recognition stage.

When comparing these results with our main results in Table 1, we observe an interesting pattern. The performance gap between our knowledge-enhanced systems and variants without external knowledge appears more pronounced with larger models. This suggests that larger models (such as Llama-3.3-70B used in our main experiments) derive significantly greater benefits from external knowledge resources. The experimental results confirm that these larger models possess enhanced abilities to leverage structured knowledge for complex reasoning tasks, with superior prompt understanding and reasoning capabilities. This finding further emphasises the importance of our knowledge-augmented approach, particularly when applied to larger-scale foundation models.

4.5 Result and Analysis

Experimental results demonstrate that our proposed three-stage reasoning framework significantly enhances LLMs' named entity translation capabilities. As shown in Table 1, our system achieves excellent performance across four language pairs, with average M-ETA scores reaching 89.79 and COMET scores of 93.33. Notably, we observe clear performance variations between different language pairs: Italian (M-ETA: 93.02) and Spanish (M-ETA: 90.5) perform best, while German (M-ETA: 85.57) shows relatively lower scores. We attribute this disparity primarily to the greater linguistic distance between German and the source language (English), as German's compound word formation and complex morphological structures pose additional challenges for entity recognition and translation.

Our comprehensive ablation study provides further insights into these results. The substantial performance gap between knowledge-enhanced and knowledge-free variants confirms that external knowledge retrieval is the critical component in our framework. While Chain-of-Thought reasoning alone provides minimal benefits for entity translation, its integration with external knowledge substantially amplifies performance. This syn-

ergy is particularly pronounced in our Llama-3.3-70B implementation, suggesting that larger models possess enhanced abilities to leverage structured knowledge for entity translation due to their superior prompt understanding and reasoning capabilities.

The verification component in our framework addresses a key source of entity translation errors identified in our analysis, further improving entity translation accuracy. These improvements are consistent across all language pairs, with the language-specific patterns in our ablation study mirroring those in our main experiments - Spanish and Italian showing the largest gains, and German the lowest, albeit still substantial. This consistency across model scales reinforces our observation that linguistic distance significantly impacts entity translation performance, even when employing a multilingual joint training strategy designed to leverage cross-lingual generalisation capabilities.

5 Conclusion

This paper presents a multi-agent entity-aware machine translation system that addresses key challenges in named entity translation. Our main contribution lies in designing a three-stage LLM (Large Language Model) reasoning framework (entity recognition, knowledge enhancement, and translation decision-making) specifically optimised for named entity translation, utilising innovative entity query generation mechanisms that effectively integrate information from external knowledge bases. Experimental results demonstrate the effectiveness of our approach, which ranked first in Spanish entity translation metrics in SemEval-2025 Task 2, and achieved the highest overall scores across four language pairs (German, French, Italian, and Spanish) among systems that do not use gold standard entity IDs during test set inference.

Future work will focus on addressing the unique challenges of named entity translation in social media environments and developing more suitable approaches for informal texts and culture-specific expressions.

Limitations

Despite our system's excellent performance, several limitations remain, including the substantial computational resources required for deploying multiple large language models and the increased latency from sequential processing of our multi-stage reasoning framework. Additionally, our system still struggles with informal entity expressions commonly found on social media platforms.

Acknowledgments

This work has been supported by the UK's innovation agency (Innovate UK) grant 10039055 (approved under the Horizon Europe Programme as vera.ai, EU grant agreement 101070093).

References

Abhinav Jauhri et al. Aaron Grattafiori, Abhimanyu Dubey. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

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.

DeepSeek-AI, Daya Guo, Dejian Yang, and Haowei Zhang et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Preprint*, arXiv:2305.14314.

Zhongwei Li, Xuancong Wang, Ai Ti Aw, Eng Siong Chng, and Haizhou Li. 2018. Named-entity tagging and domain adaptation for better customized translation. In *Proceedings of the Seventh Named Entities Workshop*, pages 41–46, Melbourne, Australia. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. *Preprint*, arXiv:2009.09025.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *Preprint*, arXiv:1908.10084.

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.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Shufang Xie, Yingce Xia, Lijun Wu, Yiqing Huang, Yang Fan, and Tao Qin. 2022. End-to-end entity-aware neural machine translation. *Machine Learning*, 111(3):1181–1203.

Xinye Yang, Yida Mu, Kalina Bontcheva, and Xingyi Song. 2024. Optimising LLM-driven machine translation with context-aware sliding windows. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1004–1010, Miami, Florida, 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.

6 Appendix

6.1 Prompt Details

6.1.1 Prompt : Entity Extraction

The system extracts the main named entity from the given text using the following prompt:

```
Extract the main named entity from the
    following text: {text}
```

6.1.2 Prompt : Query Generation

Once the entity is extracted, the system generates a structured query in JSON format to retrieve relevant knowledge from Wikidata:

```
Generate a query in JSON format (with name,
    description) for {Entity_Name} based on the
    following text: {text}
```

6.1.3 Prompt : Candidate Selection and Translation

The final step involves selecting the most appropriate candidate entity from Wikidata and translating the given text into the target language. The selection and translation process is guided by the following prompt:

```
Select the most appropriate candidate entities
    based on the following text and translate
    the following text to {target_language}
    based on its translation in the target
    language:

Candidate entities:
[
    {
        "Original name": {Entity_Name},
        "Target name":
            {Entity_Name_in_Target_Language},
        "Description":
            {Entity_Description_in_Target_Language}
    },
```

```
    ... {Other Candidate Entities} ...
]

Source Text: {Text}
```

6.1.4 Prompt for Reasoning Failure Detection and Translation Refinement

Below is the full prompt used for reasoning failure detection and translation refinement.

```
You are a verification and optimization module,
    designed to detect reasoning failures and
    refine translation outputs. Process the
    input according to the following steps and
    comply with the output format to output the
    results.

Step 1: Detection of reasoning failure based on
    the following three aspects
1. Entity Recognition Completeness
   - Identify key entities in the source text.
   - Compare with the first round response to
       find omissions, misinterpretations, or
       incorrect additions.

2. Query Relevance
   - Verify if the second-round query
       misdescribes the extracted entities.

3. Semantic Consistency
   - Compare translated text with the original
       meaning.
   - Detect shifts in meaning, tone, or
       cultural nuances.

Step 2: Improving translation based on the
    following aspects
- Correct identified reasoning failures.
- Ensure terminology consistency.
- Improve clarity, fluency, and naturalness
    while preserving intent.
- Provide a step-by-step justification for each
    correction.

Example Output Format:
[REASONING ANALYSIS]
Detailed breakdown of detected failures and
    their reasons.

[IMPROVED TRANSLATION]
<result/>
{final_translation_with_justification}
</result>

Input:
- Original Text: {original_text}
- Dialogue History:
    {three_rounds_dialogue_history}
```

6.1.5 Baseline Prompt

```
You are a helpful translation assistant.
    Translate the following text from English
    to {target_language}. Provide only the
    translation without any additional
    information: {text}
```

6.2 Key Parameters

6.2.1 Training Parameters

Parameter	
Learning rate	1.0×10^{-4}
Training epochs	5.0
Learning rate scheduler	Cosine
Warmup ratio	0.1
Precision	BF16
Random seed	42

Table 2: Key Training Parameters for SemEval 2025 Task A4

6.3 Dataset Distribution

Language Pair	Training Set	Test Set
English-German	4,087	5,876
English-French	5,531	5,465
English-Italian	3,739	5,098
English-Spanish	5,160	5,338

Table 3: Dataset Distribution by Language Pair