Can Multi-agent Help Disambiguation in Multi-domain Translation?

Zhibo Man[1,2], **Shaoyang Xu**[3], **Yujie Zhang**[1,2†], **Yi Feng**[1,2], **Yuanmeng Chen**[1,2]
Yufeng Chen[1,2], **Jinan Xu**[1,2], **Wenxuan Zhang**[3†]
[1]Key Laboratory of Big Data & Artificial Intelligence in Transportation
[2]School of Computer Science and Technology,
Beijing Jiaotong University, Beijing, China
[3]Singapore University of Technology and Design, Singapore
{zhiboman, yjzhang}@bjtu.edu.cn, wxzhang@sutd.edu.sg

Abstract

Large language models (LLMs)-based multi-agent systems have recently shown strong potential for machine translation (MT). However, their application to multi-domain translation (MDT) remains under-explored, particularly in addressing cross-domain word ambiguity. To investigate whether multi-agent approaches can help disambiguation in MDT, we propose a **m**ulti-**a**gent **c**ollaborative **d**isambiguation framework for MDT (**MACD**), which leverages the collaborative capabilities of LLMs for disambiguation. MACD consists of four cooperating agents responsible for *domain allocation*, *general translation*, *domain disambiguation*, and *translation fusion*. Experimental results show that MACD significantly improves translation performance across multiple domains and enhances disambiguation accuracy. Our approach reveals several findings on multi-agent collaboration in resolving word ambiguities.[1]

1 Introduction

In recent years, large language models (LLMs) have achieved promising results in machine translation (MT), demonstrating strong potential for practical applications (Qian et al., 2024; Lu et al., 2025; Liang et al., 2025; Feng et al., 2025; Qi et al., 2025, 2026). However, their performance in multi-domain translation (MDT) remains unsatisfactory, primarily due to severe cross-domain word ambiguity, where the same word requires different translations depending on the domain context (Man et al., 2025). For example, the English word "*cell*" should be translated as the Chinese word "监狱 (*a small room in a prison for holding prisoners*)" in the *Law* domain, while in the *Science* domain it corresponds to "细胞 (*the basic structural and functional unit of living organisms*)". Therefore, disambiguation in MDT has become a critical challenge.

Regarding the above-mentioned challenge, previous work mainly focuses on two aspects: (1) Conventional Multi-domain Neural Machine Translation (NMT) (Jiang et al., 2020; Lee et al., 2022; Man et al., 2023, 2024), which incorporates sentence-level and word-level domain labels as domain features into NMT framework; and (2) LLM-based Multi-domain Translation (Hu et al., 2024; Ye et al., 2025; Man et al., 2025), which explores how to improve MDT performance under the LLMs. Compared with (1), LLM-based MDT relies on prompting strategies to incorporate domain information and steer LLMs toward domain-specific translations. Despite their promise, these methods are primarily prompting-based and do not fully exploit the diverse capabilities of LLMs.

To resolve this limitation, researchers explore multi-agent translation (MAT) frameworks (Wu et al., 2024; Kim, 2025), in which LLMs assume different roles throughout the translation process. For example, *Transagents* (Wu et al., 2025) assigns specific translation roles to multiple LLM-based agents to simulate real-world translation workflows. MAT leverages the complementary strengths of LLMs across these roles, yet this naturally raises the question: **Can multi-agent approaches help disambiguation in multi-domain translation?**

To address the above research question, we propose a **m**ulti-**a**gent **c**ollaborative **d**isambiguation framework for multi-domain translation (**MACD**). Specifically, MACD consists of four collaborative agents. Domain allocation agent identifies domain labels for input sentences, enabling accurate assignment to subsequent agents and facilitating effective collaboration. General translation agent captures cross-domain general knowledge, while domain disambiguation agent resolves ambiguous words by generating translations with clear domain-specific features. Translation fusion agent collaborates with other agents to integrate candidate translations into a coherent final output. Together, these agents en-

†Corresponding author.
[1]GitHub Repository: https://github.com/Manowen/MACD

Findings of the Association for Computational Linguistics: ACL 2026, pages 18234–18244
July 2-7, 2026 ©2026 Association for Computational Linguistics

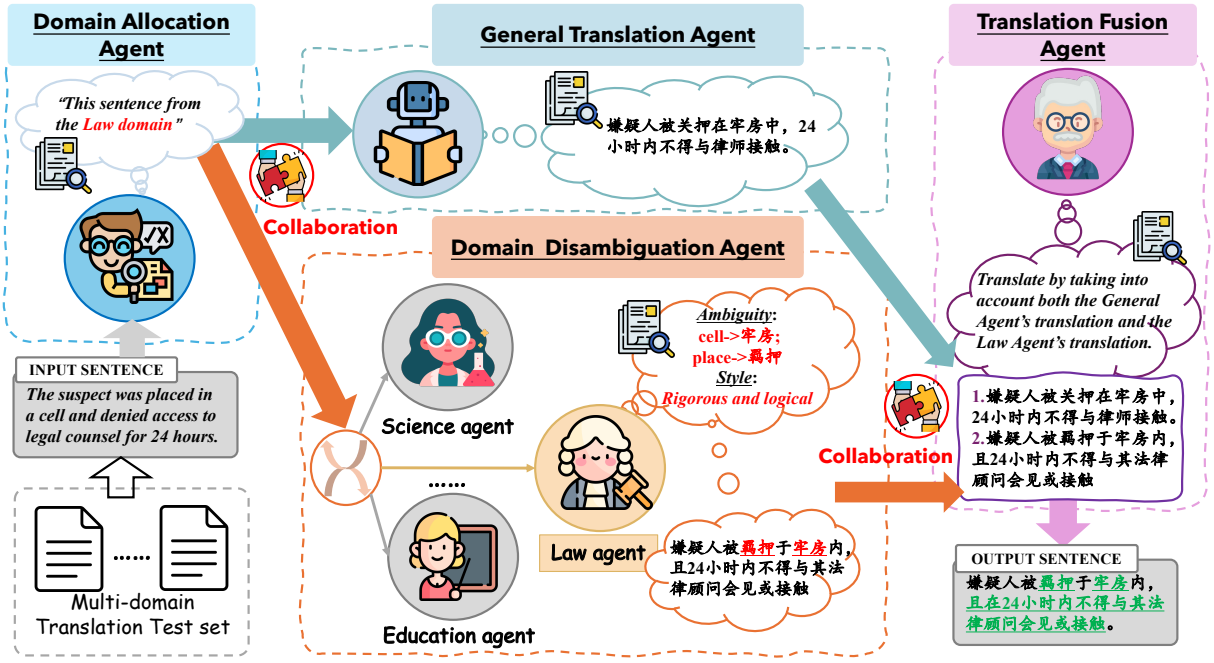

Figure 1: The MACD framework: an English-to-Chinese translation case to illustrate the collaborative process.

able MACD to effectively address cross-domain word ambiguities.

Our contributions are threefold: (1) We propose a simple yet effective autonomous collaborative multi-agent framework for MDT to resolve cross-domain ambiguity. **(2)** We design specialized agents with distinct roles to perform cross-domain disambiguation in a coordinated manner. **(3)** Experiments on public MDT benchmarks show that our method significantly improves both translation quality and disambiguation accuracy, yielding several meaningful findings and analyses.

2 Approach

2.1 Multi-agent Translation Background

A multi-agent translation framework views translation as a collaborative process among specialized agents with distinct roles (*e.g.,* analysis, drafting, post-editing), reflecting real-world workflows and enabling LLMs to operate within a unified system (Wu et al., 2024, 2025). Formally, given a source text $x \in \mathcal{L}_s$, MAT produces a target translation $y \in \mathcal{L}_t$ via a sequence of agent transformations:

$$y = \mathcal{A}_K \circ \mathcal{A}_{K-1} \circ \cdots \circ \mathcal{A}_1(x) \quad (1)$$

where each agent $\mathcal{A}_k$ represents a role-specific operation and exchanges intermediate representations with other agents.

2.2 MACD for Multi-domain Translation

Our goal is to build a multi-agent collaborative disambiguation framework for MDT, we leverage the prompt-driven capabilities of LLMs to instantiate different agents. As shown in Figure 1, MACD designs prompting strategies to simulate expert roles for LLMs, thereby guiding LLMs to act as different agents. Compared with prior work that primarily relies on prompt engineering, our approach explicitly decomposes MDT into four sub-tasks, which reduces the burden of long prompting contexts and enables collaboration among multiple agents, thereby resolving cross-domain word ambiguities.

Domain Allocation Agent. As shown in Figure 1, to facilitate collaboration in the domain allocation process, we prompt LLMs to discriminate the domain label of the input sentence from a predefined set, *e.g.*, *Education, Laws, News, Science, Spoken*. Specifically, given an input sentence x, we use a prompt to let the $\text{LLM}_{\text{allocation}}(\cdot)$ infer its domain label $d \in \mathcal{D}$, where $\mathcal{D}$ denotes a predefined set of domains. The inferred domain label then guides subsequent translation steps.

$$d = \text{LLM}_{\text{allocation}}(x) \quad (2)$$

General Translation Agent. To provide a stable semantic anchor for translation, we first introduce a general translation agent LLM_{a_g} that focuses on preserving the overall meaning and fluency of the

source sentence. Specifically, the source sentence x produces a translation candidate from a general perspective, without explicitly emphasizing domain-specific distinctions.

$$y_g = \text{LLM}_{a_g}(x) \tag{3}$$

Domain Disambiguation Agent. Resolving cross-domain word ambiguities is a key objective of our approach. To this end, we design a domain disambiguation agent conditioned on the inferred domain label d. As illustrated in Figure 1, accurate translation requires capturing the domain-specific meanings of expressions such as "牢房" and "羁押", since their correct renderings depend on the *Law* domain. Moreover, this agent generates translations that reflect domain-specific stylistic preferences. Conditioned on the source sentence x and the inferred domain label d, then LLM_{a_d} produces a domain-specific translation candidate.

$$y_d = \text{LLM}_{a_d}(x, d) \tag{4}$$

By explicitly incorporating domain information, this agent is better able to resolve ambiguous expressions and reflect domain-specific lexical and stylistic preferences. As a result, the two agents play complementary roles: y_g provides a semantically faithful reference, while y_d contributes domain-specific disambiguation knowledge.

Translation Fusion Agent. To obtain the final translation, we introduce a translation fusion agent that integrates the complementary strengths of the two candidates instead of relying on heuristic reranking. Specifically, based on the source sentence x, the inferred domain d, and the candidate translations y_g, y_d, the fusion agent $\text{LLM}_{\text{fusion}}$ synthesizes the final translation as follows.

$$y = \text{LLM}_{\text{fusion}}(x, d, y_g, y_d) \tag{5}$$

By jointly considering the semantic adequacy of y_g and the domain-specific accuracy of y_d, the fusion agent produces a translation that is both faithful to the source meaning and consistent with the target domain. For all agents, we provide five human-annotated exemplars per domain in the prompt to guide the model toward domain-consistent translation behavior. The detailed prompting design for each agent is provided in Appendix B.

3 Experimental Results and Analysis

Datasets. We evaluate our approach on two publicly available MDT datasets: German-to-English (Aharoni and Goldberg, 2020) and UM-Corpus (Tian et al., 2014), and detailed statistics are provided in Appendix A. In addition, test sets for evaluating cross-domain word ambiguities are derived from DMDTEval (Man et al., 2025).

Experimental Settings. In this work, we adopt three metrics: (1) SacreBLEU (Post, 2018);[2] (2) COMET scores (Rei et al., 2020);[3] (3) Disambiguation accuracy (Man et al., 2025). The backbone of the language agent is *Qwen3-30B-A3B-Instruct-2507*.[4] We compare our approach with the following representative methods. (1) Self-consistency (SC) generates multiple translation candidates and selects the final output via majority voting . In our experiments, we generate five candidates based on the previous work (He et al., 2024). (2) Multi-agent Translation (MAT) decomposes MDT into three agents, where we refer to prior work in assigning the roles of *Editor*, *Translator*, and *Proofreader* (Wu et al., 2024, 2025). (3) Single Prompting with Multiple Steps (SPMS) performs all four steps—*domain allocation*, *general translation*, *domain disambiguation*, and *translation fusion*—within a single prompt (Hu et al., 2024).

3.1 Key Findings for Main Results

Table 1 shows that the comparison of previous methods and our approach. We further provide the following findings and analysis of the main experimental results.

Finding 1: MACD achieves more stable and consistent improvements across multiple domains. As shown in Table 1, we observe that although SPMS injects domain information through a single prompt, MACD achieves further gains by decomposing the task across multiple collaborative agents. Compared with SPMS, MACD improves performance by +0.88 BLEU, showing that collaboration among multi-agent can effectively enhance translation quality.

Finding 2: MACD improves overall translation quality by effectively resolving ambiguities. We observe that both SPMS and MACD explicitly include a disambiguation step, whereas SC and MAT do not. Notably, SPMS and MACD achieve

[2] `Signature: nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.1.0`

[3] We use the `wmt22-comet-da` model to generate COMET scores.

[4] https://huggingface.co/Qwen/Qwen3-30B-A3B-Instruct-2507

EN-ZH	**Education**	**Laws**	**News**	**Science**	**Spoken**	**AVG**
SC	39.02/86.61	64.59/88.23	34.82/85.06	34.07/81.76	23.27/80.16	39.15/84.36
MAT	39.24/86.62	65.06/89.15	35.23/85.73	34.13/82.05	24.29/81.53	39.59/85.02
SPMS	41.20/87.03	65.64/89.40	35.63/85.62	35.53/83.72	25.18/82.36	40.64/85.63
Ours	**41.78/87.53**	**66.24/89.88**	**37.06/86.74**	**36.39/84.51**	**26.14/82.45**	**41.52/86.22**
DE-EN	**IT**	**Koran**	**Law**	**Medical**	**Subtitles**	**AVG**
SC	36.85/80.80	18.35/73.05	37.48/83.95	39.78/83.18	29.08/78.22	32.31/79.84
MAT	37.44/81.25	18.93/73.56	37.99/84.45	40.22/83.68	29.71/78.67	32.86/80.32
SPMS	37.65/81.35	19.10/73.70	38.25/84.55	40.65/83.95	29.85/78.85	33.10/80.48
Ours	**38.54/81.79**	**20.36/75.40**	**38.26/85.12**	**42.28/85.23**	**30.39/79.75**	**33.97/81.46**

Table 1: BLEU/COMET scores on the English-to-Chinese (*EN-ZH*) and German-to-English (*DE-EN*) translation tasks with Qwen-3-30B. The best results are highlighted in **bold**.

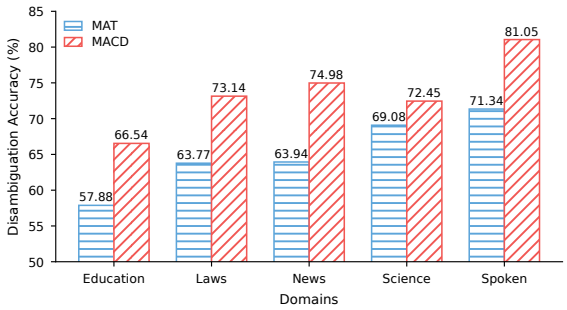

Figure 2: The comparison of disambiguation accuracy (%) on EN-ZH translation task

the best and second-best average performance, respectively, suggesting that explicit disambiguation benefits translation quality in both single-prompt and multi-agent settings.

Finding 3: MACD improves translation even in less distinctive domains. Although *Spoken* and *News* domains exhibit weaker domain-specific features (Jiang et al., 2020; Man et al., 2024, 2025), MACD still consistently outperforms MAT in these domains, achieving BLEU gains of +1.83 and +1.85 for *News* and *Spoken* in EN-ZH, respectively. This indicates that MACD can still leverage cross-domain knowledge to improve translation quality. Moreover, in Appendix C and Appendix D, we report the token cost of our experiments and present comparative results for using different LLMs for different agents.

3.2 Analysis of Disambiguation Performance

In this section, we focus on the EN–ZH translation direction to further analyze disambiguation. Figure 2 shows that MACD consistently outperforms MAT in disambiguation accuracy across all five domains, and we observe a significant improvement of +11.04 points in the *News* domain. We further find that the improvement in disambiguation accuracy also leads to better translation quality. To further illustrate the effectiveness of our approach, we present the following analyses.

Translation Case. Table 2 illustrates how MACD resolves cross-domain ambiguity through agent collaboration. In Translation Case 1, the sentence is assigned to the *Science* domain. The General Translation Agent (GTA) produces a semantically adequate translation, but selects the more generic expression “强度”, whereas the Domain Disambiguation Agent (DDA) captures the more domain-appropriate term “猛烈性”. In Translation Case 2, the sentence is assigned to the *Education* domain, which misleads the DDA into producing the incorrect translation “培养”. By contrast, the GTA leverages cross-domain shared knowledge to generate the correct translation “孵化”. In both cases, the Translation Fusion Agent (TFA) integrates the complementary strengths of the two agents, yielding translations that are both semantically accurate and domain-appropriate. This demonstrates that effective disambiguation in MDT depends on both domain-specific knowledge and cross-domain shared knowledge.

Effect of the Number of Collaborative Agents on Disambiguation. We further investigate how the number of collaborative agents affects disambiguation performance. The goal of this experiment is to examine whether ambiguity resolution benefits more from relying only on the agent matched to the sentence’s assigned domain or from involving

	Translation Case 1	Translation Case 2
SRC	When moving over land, a typhoon is gradually decreased in severity.	Establish productivity promotion centers and technology enterprises incubation bases.
REF	当台风在陆地上移动时，其猛烈性逐渐减弱。	建立生产力促进中心和科学技术企业孵化基地。
DAA	Science	Education
GTA(y_g)	当台风在陆地上移动时，其强度逐渐减弱。*(ensures overall semantic adequacy, but lacks domain-specific nuance)*	建立生产力促进中心和科学技术企业孵化基地。*(leverages cross-domain shared knowledge and captures the correct meaning of "incubation")*
DDA(y_d)	当台风在陆地上移动时，其猛烈性逐渐减弱。*(captures domain-specific lexical choice, highlighting "猛烈性")*	建立生产力促进中心和技术企业培养基地。*(is misled by the assigned domain and renders "incubation" as a more education-oriented expression)*
TFA(y)	当台风在陆地上移动时，其猛烈性逐渐减弱。*(merges candidates, ensuring semantic and domain-specific correctness)*	建立生产力促进中心和科学技术企业孵化基地。*(integrates the two candidates and selects the semantically appropriate translation)*

Table 2: Two EN-ZH translation cases illustrating the complementary roles of the General Translation and Domain Disambiguation Agents. Red text indicates the correct translation, while blue text indicates the incorrect translation.

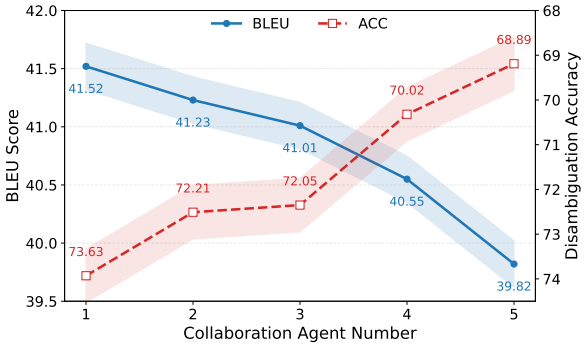

Figure 3: Average BLEU and disambiguation accuracy with different numbers of collaboration agents.

multiple agents across domains. To this end, we vary the number of participating agents {1-5}.[5] We evaluate each setting using average BLEU and disambiguation accuracy. As shown in Figure 3, we find that using only the sentence's assigned domain agent achieves the best performance on both metrics. This result indicates that domain-matched agents capture the most relevant knowledge for resolving ambiguities.

Can Longer Context Help Disambiguation? To examine whether richer context can still benefit disambiguation under a longer-input setting, we reorganize the EN–ZH sentence-level test set by grouping every 100 sentences within each domain into a single input unit. This results in 4 subsets for *Education*, 4 for *Law*, 15 for *News*, 5 for *Science*, and 4 for *Spoken*. Each subset can be regarded as a long text that provides substantially richer contextual information than isolated sentences. During evaluation, we ensure that the LLM only accesses the context within the current 100-sentence subset. We then evaluate disambiguation accuracy on these subsets.

[5] 1 = only the agent corresponding to the sentence's assigned domain; 5 = all domain agents; 2–4 = randomly selected subsets of agents.

Methods	Edu	Law	News	Sci	Spo
Sent	66.54	73.14	74.98	72.45	81.05
Long	**75.23**	**85.66**	**81.90**	**83.24**	**88.13**

Table 3: Disambiguation accuracy (%) under sentence-level and long-text settings on the EN–ZH test set.

As shown in Table 3, the long-text setting consistently achieves higher disambiguation accuracy than the sentence-level setting across all five domains. We observe clear improvements in every domain, with especially large gains in *Education*, *Legal*, and *Science*, indicating that longer context substantially helps disambiguation. This suggests that richer contextual information within the same domain provides stronger semantic and discourse cues, enabling the LLMs to infer the intended meaning of ambiguous words more accurately.

4 Conclusion

In this work, we introduce MACD to explore and address the research question: *Can multi-agent systems help disambiguation in multi-domain translation?* MACD consistently improves both translation quality and disambiguation accuracy across all domains. Our findings demonstrate that domain-specific agent collaboration effectively resolves word ambiguities. In future work, we plan to extend our approach to document-level translation, as broader context provides richer semantic and pragmatic information beyond the sentence level.

Limitations

This work still has some limitations. Our current framework remains a preliminary exploration of multi-agent collaboration, as it still relies on manually designed prompts and relatively simple interaction among agents. In addition, our study focuses on sentence-level translation, while document-level context could naturally provide richer cues for ambiguity resolution. In future work, we will explore more advanced collaboration mechanisms and extend the framework to document-level translation.

Acknowledgements

This research is supported by the National Natural Science Foundation of China (No. 62376019, 62476023, 61976015, 61976016, 61876198, and 61370130) and the National Research Foundation, Singapore under its National Large Language Models Funding Initiative (AISG Award No. AISG-NMLP-2024-005). The authors would like to thank the anonymous reviewers for their valuable comments and suggestions to improve this paper.

References

Roee Aharoni and Yoav Goldberg. 2020. Unsupervised domain clusters in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7747–7763, Online. Association for Computational Linguistics.

Zhaopeng Feng, Jiayuan Su, Jiamei Zheng, Jiahan Ren, Yan Zhang, Jian Wu, Hongwei Wang, and Zuozhu Liu. 2025. M-MAD: Multidimensional multi-agent debate for advanced machine translation evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7084–7107, Vienna, Austria. Association for Computational Linguistics.

Zhiwei He, Tian Liang, Wenxiang Jiao, Zhuosheng Zhang, Yujiu Yang, Rui Wang, Zhaopeng Tu, Shuming Shi, and Xing Wang. 2024. Exploring human-like translation strategy with large language models. *Transactions of the Association for Computational Linguistics*, 12:229–246.

Tianxiang Hu, Pei Zhang, Baosong Yang, Jun Xie, Derek F. Wong, and Rui Wang. 2024. Large language model for multi-domain translation: Benchmarking and domain CoT fine-tuning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5726–5746, Miami, Florida, USA. Association for Computational Linguistics.

Haoming Jiang, Chen Liang, Chong Wang, and Tuo Zhao. 2020. Multi-domain neural machine translation with word-level adaptive layer-wise domain mixing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1823–1834, Online. Association for Computational Linguistics.

Ahrii Kim. 2025. A preliminary study of AI agent model in machine translation. In *Proceedings of the Tenth Conference on Machine Translation*, pages 583–586, Suzhou, China. Association for Computational Linguistics.

Jiyoung Lee, Hantae Kim, Hyunchang Cho, Edward Choi, and Cheonbok Park. 2022. Specializing multi-domain NMT via penalizing low mutual information. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10015–10026, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yupu Liang, Yaping Zhang, Zhiyang Zhang, Yang Zhao, Lu Xiang, Chengqing Zong, and Yu Zhou. 2025. Single-to-mix modality alignment with multimodal large language model for document image machine translation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12391–12408, Vienna, Austria. Association for Computational Linguistics.

Hongyuan Lu, Zixuan Li, Zefan Zhang, and Wai Lam. 2025. SLoW: Select low-frequency words! automatic dictionary selection for translation on large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 898–913, Suzhou, China. Association for Computational Linguistics.

Zhibo Man, Yuanmeng Chen, Yujie Zhang, and Jinan Xu. 2025. DMDTEval: An evaluation and analysis of LLMs on disambiguation in multi-domain translation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 6065–6082, Suzhou, China. Association for Computational Linguistics.

Zhibo Man, Kaiyu Huang, Yujie Zhang, Yuanmeng Chen, Yufeng Chen, and Jinan Xu. 2024. ICL: Iterative continual learning for multi-domain neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7732–7743, Miami, Florida, USA. Association for Computational Linguistics.

Zhibo Man, Yujie Zhang, Yuanmeng Chen, Yufeng Chen, and Jinan Xu. 2023. Exploring domain-shared and domain-specific knowledge in multi-domain neural machine translation. In *Proceedings of Machine Translation Summit XIX, Vol. 1: Research Track*, pages 99–110, Macau SAR, China. Asia-Pacific Association for Machine Translation.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on*

Machine Translation: Research Papers, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Rui Qi, Zhibo Man, Yufeng Chen, Fengran Mo, Jinan Xu, and Kaiyu Huang. 2025. SoT: Structured-of-thought prompting guides multilingual reasoning in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 11024–11039, Suzhou, China. Association for Computational Linguistics.

Rui Qi, Fengran Mo, Yufeng Chen, Xue Zhang, Shuo Wang, Hongliang Li, Jinan Xu, Meng Jiang, Jian-Yun Nie, and Kaiyu Huang. 2026. Language-coupled reinforcement learning for multilingual retrieval-augmented generation. *Preprint*, arXiv:2601.14896.

Shenbin Qian, Archchana Sindhujan, Minnie Kabra, Diptesh Kanojia, Constantin Orasan, Tharindu Ranasinghe, and Fred Blain. 2024. What do large language models need for machine translation evaluation? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3660–3674, Miami, Florida, USA. Association for Computational Linguistics.

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.

Liang Tian, Derek F. Wong, Lidia S. Chao, Paulo Quaresma, Francisco Oliveira, Yi Lu, Shuo Li, Yiming Wang, and Longyue Wang. 2014. UM-corpus: A large English-Chinese parallel corpus for statistical machine translation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1837–1842, Reykjavik, Iceland. European Language Resources Association (ELRA).

Minghao Wu, Jiahao Xu, and Longyue Wang. 2024. TransAgents: Build your translation company with language agents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 131–141, Miami, Florida, USA. Association for Computational Linguistics.

Minghao Wu, Jiahao Xu, Yulin Yuan, Gholamreza Haffari, Longyue Wan, Weihua Luo, and Kaifu Zhang. 2025. (perhaps) beyond human translation: Harnessing multi-agent collaboration for translating ultra-long literary texts. *Transactions of the Association for Computational Linguistics*, 13:901–922.

Yongshi Ye, Biao Fu, Chongxuan Huang, Yidong Chen, and Xiaodong Shi. 2025. How well do large reasoning models translate? a comprehensive evaluation for multi-domain machine translation. *arXiv preprint arXiv:2505.19987*.

English-to-Chinese					
Test set	Edu	Laws	News	Sci	Spo
	790	456	1500	503	455
German-to-English					
Test set	IT	Kor	Laws	Med	Sub
	2000	2000	2000	2000	2000

Table 4: The statistics of MDT test sets. Edu represents for the *Education* domain, Sci represents for the *Science* domain, Spo represents for the *Spoken* domain, and Sub represents for the subtitles domain.

English-to-Chinese					
Test set	Edu	Laws	News	Sci	Spo
	492	696	720	471	432
German-to-English					
Test set	IT	Kor	Laws	Med	Sub
	214	167	412	345	209

Table 5: Statistics of the number of cross-domain ambiguous words on MDT

A Datasets and Evaluation Details

Two publicly available MDT test sets are used to evaluate translation performance: UM-Corpus[6] (English-to-Chinese), which includes five domains: *Education, Law, News, Science, and Spoken* (Tian et al., 2014), and German-to-English dataset[7], which includes five domains: *IT, Koran, Laws, Medical, and Subtitles* (Aharoni and Goldberg, 2020). Detailed statistics of these datasets are presented in Table 4. To assess the model's capability in handling cross-domain word ambiguities, test sets derived from DMDTEval (Man et al., 2025) are employed, with their scale summarized in Table 5. Default hyperparameter settings in vLLM[8] are applied for inference, i.e., temperature = 0.8 and top_p = 0.95. In addition to BLEU and COMET, disambiguation accuracy (Man et al., 2025) is used as an evaluation metric, computed as follows.

$$\text{ACC} = \frac{m}{n} \tag{6}$$

where n is the total number of ambiguous words in the test set, and m is the number of words correctly translated according to their domain-specific meanings.

B Prompting Design

This section describes the prompting strategies used to instantiate different agents in MACD.

[6] http://nlp2ct.cis.umac.mo/um-corpus/
[7] http://opus.nlpl.eu/
[8] https://github.com/vllm-project/vllm

Rather than fine-tuning models, we rely on carefully designed prompts to simulate expert roles and enable LLMs to perform domain-aware reasoning and disambiguation. Specifically, MACD consists of four agents: *Domain Allocation*, *General Translation*, *Domain Disambiguation*, and *Translation Fusion*. Each agent is provided with role-specific instructions and a small set of in-context exemplars to guide its behavior.

All exemplars included in the prompts are manually annotated by the authors to ensure high quality and consistency, incurring no additional cost. We established detailed guidelines to regulate the annotation process. To avoid improvements stemming from overly uniform data distributions, all prompts in our design are entirely handcrafted and do not rely on existing multi-domain translation datasets (Tian et al., 2014; Aharoni and Goldberg, 2020). The specific English–Chinese translation tasks, along with their prompting strategies, are described as follows.

For domain-related agents, annotations follow clear criteria:

- Domain labels are assigned based on dominant semantic content and linguistic usage rather than explicit domain keywords;
- Translation exemplars prioritize semantic fidelity, fluency, and domain-consistent terminology;
- Ambiguous words are resolved according to domain-appropriate meanings rather than surface-level frequency statistics.

By grounding all prompts in human-annotated, domain-consistent examples, we ensure that the agents are guided by reliable supervision signals, which allows us to more faithfully evaluate whether multi-agent collaboration can effectively improve disambiguation in multi-domain translation.

Domain Allocation Agent The Domain Allocation Agent aims to infer the latent domain of the input sentence based purely on its linguistic characteristics. The predicted domain serves as high-level guidance for subsequent agents, enabling domain-aware translation and disambiguation without relying on explicit domain labels, as shown in Figure 4.

General Translation Agent The General Translation Agent is responsible for producing a fluent and semantically faithful translation of the input

Instruction. You are a domain allocation agent. Given an English sentence, infer its most likely domain from {Education, Laws, News, Science, Spoken}. Base your decision on linguistic cues, terminology, and usage context. Output only the domain label.

Exemplars.

- English: "Students are required to submit the final report by Friday, and late submissions will incur a penalty unless prior permission is obtained."
- ☐ Output: Education
- English: "The defendant, accused of embezzling company funds over a period of two years, pleaded not guilty in court yesterday."
- ☐ Output: Law
- English: "During an early morning press conference, the prime minister announced a series of economic measures aimed at stimulating small businesses and job creation."
- ☐ Output: News
- English: "The recent experiment involving a novel neural network architecture demonstrates a significant improvement in image classification accuracy compared to traditional methods."
- ☐ Output: Science
- English: "Hey, I'm running a bit late for dinner, but I'll catch up with you as soon as I get there; don't start without me!"
- ☐ Output: Spoken

Figure 4: An example prompt for Domain Allocation Agent

Instruction. You are a general translation agent. Translate the given English sentence into Chinese, focusing on overall semantic adequacy and fluency.

Exemplars.

- English: “The system processes large amounts of structured and unstructured data daily, providing critical insights for decision-making in real-time applications.”
- Chinese: “该系统每天处理大量结构化和非结构化数据，为实时决策提供关键洞察。”
- English: “The experiment was repeated several times under varying environmental conditions to ensure the robustness and reproducibility of the observed results.”
- Chinese: “该实验在不同的环境条件下重复进行了多次，以确保观察结果的稳健性和可重复性。”
- English: “The committee will review all applications carefully before making a decision.”
- Chinese: “委员会将在做出决定之前仔细审查所有申请。”
- English: “Advances in technology have significantly changed the way people communicate and access information.”
- Chinese: “科技的进步已经显著改变了人们交流和获取信息的方式。”
- English: "Despite facing numerous challenges during the project, the team managed to complete it on time, demonstrating remarkable dedication and collaboration.”
- Chinese: “尽管在项目过程中遇到了诸多挑战，团队仍然按时完成了任务，展现出了非凡的敬业精神和协作能力。”

Figure 5: An example prompt for General Translation Agent

sentence without relying on explicit domain information. It provides a domain-agnostic baseline translation that preserves the overall meaning, serving as a complementary candidate for subsequent disambiguation and fusion, the specific prompting is shown in the Figure 5.

Domain Disambiguation Agent The Domain Disambiguation Agent focuses on resolving domain-specific word ambiguities through explicit incorporation of domain knowledge. Conditioned on the inferred domain, this agent prioritizes domain-appropriate meanings, terminology, and stylistic conventions, producing translations consistent with domain-specific usage. Each translation resolves potentially ambiguous words or phrases, adapting them to the target domain, as shown in Figure 6.

Translation Fusion Agent The Translation Fusion Agent performs reasoning-based integration of multiple candidate translations from the general and domain-specific agents. It synthesizes a

Instruction. You are a domain disambiguation translation agent. Given an English sentence and its domain, identify any word or phrase that could have multiple meanings (ambiguities), provide a brief explanation of the chosen meaning in context, and produce a domain-consistent Chinese translation that reflects the terminology and style.

Exemplars.

- Domain: Laws
- English: After reviewing all the evidence, the court decided to drop the charge against the defendant, emphasizing that there was insufficient proof to support a conviction.” Ambiguous word explanation: charge” refers to a formal legal accusation rather than an electrical or financial meaning.
- Chinese: “在审查所有证据后，法院决定撤销对被告的指控，并强调缺乏足够证据支持定罪。”
- Domain: Science
- English: “The research team conducted a series of experiments at the cellular level to understand how certain proteins interact within the context of gene expression, hoping to uncover mechanisms relevant to disease progression.” Ambiguous word explanation: “cellular” refers to biological cells, not mobile networks or other technical structures.
- Chinese: “研究团队在细胞层面进行了系列实验，以理解特定蛋白质在基因表达中的相互作用，希望揭示与疾病进展相关的机制。”
- Domain: Education
- English: “Students are required to submit their research papers by the end of the semester, and failure to do so may result in a lower grade, as the assignment constitutes a major component of the course assessment.” Ambiguous word explanation: “papers” refers to academic essays or assignments, not newspapers or legal documents.
- Chinese: “学生需在学期末提交他们的研究论文，未按时提交可能导致成绩下降，因为该作业是课程评估的重要组成部分。”
- Domain: News
- English: “During the press conference yesterday, the minister addressed several urgent issues concerning the national economy and outlined a series of policy measures aimed at stabilizing growth and improving public welfare.” Ambiguous word explanation: “addressed” here means made a formal statement or speech rather than giving a location or direction.
- Chinese: “在昨日的新闻发布会上，部长就国家经济的若干紧迫问题发表讲话，并概述了一系列旨在稳定增长和改善民生的政策措施。”
- Domain: Spoken
- English: “I’ll drop by your office later this afternoon to discuss the project details, but if you’re busy, we can postpone to tomorrow morning instead.” Ambiguous word explanation: “drop by” means visit casually or briefly, not physically dropping something.
- Chinese: “我今天下午会顺便去你办公室讨论项目细节，但如果你忙的话，我们可以改到明天上午。”

Figure 6: An example prompt for Domain Disambiguation Agent

	AVG_{BLEU}	AVG_{ACC}
MACD	41.52	73.63
$MACD_{GPT}$	**43.17**	**74.54**
$MACD_4$	**43.54**	**75.26**

Table 6: Performance of MACD with different LLM configurations. MACD: all agents use *Qwen-3-30B*; $MACD_{GPT}$: all agents use *GPT-5-mini*; $MACD_4$: four agents instantiated with different LLMs.

final translation that combines the complementary strengths of all agents and resolves any remaining ambiguities. Using longer, context-rich examples, the agent can demonstrate nuanced domain-specific decisions, as shown in Figure 7.

C Cost Analysis

We further analyze the computational cost of different methods by counting the total number of tokens, including the input sentences, prompting tokens, and output translations. Averaged over the five EN–ZH domains, the baseline model MAT consumes approximately 2.83×10^4 tokens per domain, indicating a relatively low cost. In comparison, MACD consumes about 4.52×10^4 tokens per domain on average, mainly due to its richer prompts that incorporate both domain-specific and general contextual information. Although MACD requires more tokens than MAT, this additional cost is accompanied by consistent improvements in disambiguation performance across all domains.

D Does MACD benefit from multiple LLMs?

In our main experiments, all agents in MACD are instantiated with the same LLM, *Qwen-3-30B*. While this setting verifies the effectiveness of the multi-agent framework, it does not show whether MACD can further benefit when different agents use different LLMs. To examine this, we compare three configurations: MACD, where all agents use *Qwen-3-30B*; $MACD_{GPT}$, where all agents use *GPT-5-mini*; and $MACD_4$. The agent–LLM pairing is based on i、preliminary experiments that selected the best-performing model for each capability. As shown in Table 6, replacing *Qwen-3-30B* with *GPT-5-mini* for all agents improves both BLEU and ACC, indicating that stronger LLMs can enhance the overall framework. More importantly, $MACD_4$ achieves the best performance, showing that using different LLMs for different agents is more effective than using the same LLM for all

Instruction. You are a translation fusion agent. Given the source sentence, its domain, and multiple candidate translations, compare them carefully and produce a final Chinese translation that best resolves domain-specific ambiguities and preserves semantic and stylistic consistency.

Exemplars.

- Domain: Laws, English: After a lengthy deliberation, the appellate court rejected the plaintiff's appeal, noting that the evidence presented did not sufficiently support the claims."
- ☐ Candidate 1: 经过长时间的审议，上诉法院驳回了原告的上诉，指出所提供的证据不足以支持其主张。" Candidate 2: 在经过长时间的审理后，法院拒绝了原告的申诉，并说明证据不足以证明其诉求。"
- Final Answer: 经过长时间的审议，上诉法院驳回了原告的上诉，指出所提供的证据不足以支 持其主张。
- Domain: Science, English: "The research team developed a computational model that demonstrates robust generalization across different datasets, suggesting its potential for broader applications in related scientific fields."
- ☐ Candidate 1: "研究团队开发了一个计算模型，在不同数据集上表现出强大的泛化能力，表明其在相关科学领域具有广泛应用的潜力。" Candidate 2: "研究小组设计了一个模型，能够很好地推广到多个数据集，显示出其在科学研究中的潜在应用价值。"
- Final Answer: "研究团队开发的计算模型在不同数据集上表现出强大的泛化能力，显示其在相关科学领域具有广泛应用的潜力。"
- Domain: Education, English: "This semester, the course emphasizes hands-on projects and collaborative exercises to help students develop practical skills and critical thinking necessary for their future careers."
- ☐ Candidate 1: "本学期的课程强调动手项目和协作练习，以帮助学生培养未来职业所需的实践技能和批判性思维。" Candidate 2: "本课程在本学期重点安排实践项目和团队活动，旨在提升学生的实际能力和思维能力，为未来职业做好准备。"
- Final Answer: "本学期课程重点安排实践项目和协作练习，以帮助学生培养未来职业所需的实践技能和批判性思维。"
- Domain: News, English: "During a late-night signing ceremony, the two countries finalized an agreement that outlines new trade regulations and cooperation measures designed to strengthen bilateral economic ties."
- ☐ Candidate 1: "在深夜的签署仪式上，两国完成了一项协议，明确了新的贸易规则和旨在加强双边经济联系的合作措施。" Candidate 2: "在昨夜举行的签字仪式中，双方敲定了一项协议，规定了新的贸易规定及合作方案，以促进两国经济关系。"
- Final Answer: "在昨夜举行的签署仪式上，两国完成了一项协议，明确了新的贸易规则和旨在加强双边经济联系的合作措施。"
- Domain: Spoken, English: "I think your suggestion is quite reasonable, but we might need to adjust a few details before implementing it to ensure everything runs smoothly."
- ☐ Candidate 1: "我认为你的建议相当合理，但在实施之前我们可能需要调整一些细节，以确保一切顺利进行。" Candidate 2: "你的提议听起来挺合理的，不过在执行之前我们可能得修改部分细节，保证整个流程顺畅。"
- Final Answer: "我觉得你的建议挺合理的，不过在执行之前我们可能需要调整一些细节，以确保一切顺利进行。

Figure 7: An example prompt for Translation Fusion Agent

agents. This indicates that the improvement of MACD comes not only from agent collaboration, but also from using different LLMs across agents.