AgentDiscoTrans: Agentic LLMs for Discouse-level Machine Translation

Anonymous ACL submission

Abstract

In this paper, we propose AgentDiscoTrans, a novel agentic framework for document-level machine translation that leverages specialized LLM agents to process long documents at the discourse level. Our system segments an input document into coherent discourse units—drawing inspiration from the theories presented in *Attention, Intentions, and the Structure of Discourse*—and then translates each unit using a Translation Agent that incorporates contextual information from a dynamic Memory. The Memory Agent updates and maintains critical translation cues such as discourse markers, entity mappings, noun-pronoun mappings, and phrase translations, ensuring inter-discourse consistency. Experiments on multiple datasets, including the TED test sets from IWSLT2017, the mZPRT corpus, and the WMT2022 dataset, demonstrate that our system outperforms competitive baselines (such as NLLB, Google Translate, and DELTA) in terms of automatic metrics (d-BLEU, d-COMET, TER) and human evaluations focusing on both General Quality and Discourse Awareness. Our ablation studies further validate the importance of both discourse segmentation and Memory updating for achieving high-quality translations.

1 Introduction

Document-level machine translation (DocMT) faces challenges far beyond those of sentence-level translation, as it must capture long-range dependencies, inter-sentential relationships, and evolving discourse phenomena to preserve narrative flow, consistent terminology, and cultural nuances (Kim et al., 2019; Maruf et al., 2021). Early neural approaches attempted to model entire documents or incorporate neighboring context, but by treating documents as sequences of isolated sentences, they often produced incoherent outputs. The advent of large language models such as GPT-3, GPT-4, and LLaMA (Brown et al., 2020; Chowdhery et al., 2022; Touvron et al., 2023) has opened new avenues by generating long, contextually coherent text. Recent studies leveraging LLMs for DocMT (Wang et al., 2023a; Wu and Hu, 2023; Wu et al., 2024) demonstrate improved global discourse handling; however, issues such as input length constraints and noisy long-context representations can still lead to content omissions and terminology inconsistencies (Karpinska and Iyyer, 2023).

Prior work has explored various neural methods for DocMT. Approaches that incorporate discourse phenomena—demonstrated by Bawden et al. (2018) and Maruf and Haffari (2018)—improve coherence by modeling referring expressions and cohesive ties, while attention mechanisms and hierarchical structures (Wang et al., 2017; Zhang et al., 2018; Tan et al., 2019) aim to better capture inter-sentential relationships. Yet, many systems either translate one sentence at a time, losing essential cross-sentence cues, or process entire documents as a single block, leading to practical limitations.

In parallel, the agentic paradigm in LLMs offers a promising alternative by decomposing translation into specialized subtasks. Recent systems, such as DELTA (Wang et al., 2025), show that multi-level memory and agent-based frameworks can enhance translation consistency by explicitly modeling inter-discourse relationships. Building on this idea, our work models DocMT as a series of interrelated discourse-level tasks. Drawing on theoretical foundations from *Attention, Intentions, and the Structure of Discourse* (Grosz and Sidner, 1986) and Rhetorical Structure Theory (Mann and Thompson, 1988), as well as definitions of discourse as a coherent group of contiguous sentences (Jurafsky, 2000), we segment documents into self-contained discourse units that are then translated independently. Our agentic framework, which comprises a Discourse Agent, a Memory Agent, and a Translation Agent, overcomes the pitfalls of both

sentence-level and full-document translation methods, providing a robust solution for ultra-long texts.

Our contributions are:

1. **AgentDiscoTrans** (Section 3): We propose a novel agentic translation system for document-level machine translation that leverages large language models as core agents to achieve coherent and consistent translations across entire documents. Our system shows an average improvement of **3.1** d-BLEU points over sentence-level baselines, demonstrating the efficacy of our agentic approach for the DocMT task.

2. **Military-Domain Parallel Corpus** (Section 4.1): We present a curated, document-level parallel corpus in the military domain for the English–Chinese language pair, addressing a critical gap in resources for specialized domain-specific machine translation. The dataset consists of twelve documents with an average document length of **1,500** words in English and **2,000** words in Chinese.

3. **Comprehensive Evaluation** (Section 5.2): We conduct extensive evaluations on standard benchmark datasets as well as the newly introduced military-domain corpus. Our study includes a detailed correlation analysis between human translators and LLM-based agents. Compared with sentence-level baselines, our system achieves an average improvement of **3.1** d-BLEU points, and shows an average improvement of **4.5** d-BLEU points over commercial systems across multiple language pairs.

4. **Ablation Study** (Section 5.4): A thorough ablation study quantifies the contributions of each agent in our system. For the Chinese-to-English language direction, incorporating the Discourse Agent leads to an average improvement of **2.5** d-BLEU points, while the addition of the Memory Agent contributes an average improvement of **2.4** d-BLEU points, over a baseline sentence-level translation system.

2 Related Work

Research on document-level machine translation (DocMT) has evolved along two main lines. One stream, the document-to-sentence (Doc2Sent) approach, integrates contextual signals from neighboring sentences into the translation process. Early work in this direction (Wang et al., 2017; Tan et al., 2021; Lyu et al., 2021) employs architectures that encode preceding or surrounding sentences to improve the translation of the current sentence. While these methods provide useful context, they tend to treat each sentence as an isolated unit during generation and do not fully capture cross-sentence dependencies. Subsequent studies have highlighted that this separated encoding can lead to fragmented discourse representations and missed target-side cues (Sun et al., 2022; Bao et al., 2021; Li et al., 2023).

In contrast, document-to-document (Doc2Doc) approaches aim to translate multiple sentences jointly, thereby modelling long-range dependencies directly. Systems proposed by Zhang et al. (2020); Liu et al. (2020a) and further refined by Lupo et al. (2022); Bao et al. (2021) incorporate hierarchical attention mechanisms to capture inter-sentential relationships. Although Doc2Doc models better preserve coherence, they face challenges in scaling to ultra-long documents and are prone to errors such as content omissions, as reported by Wang et al. (2023a) and Karpinska and Iyyer (2023).

Recent advances have leveraged large language models (LLMs) for DocMT, capitalizing on their capability to handle long contexts. Studies such as Wang et al. (2023a); Wu et al. (2024) have demonstrated that LLMs can process document-level inputs and generate more context-aware translations. Yet, these systems are not without shortcomings; when operating over extensive texts, they may still produce inconsistencies in terminology and omit critical information.

A separate but related line of research focuses on agentic LLMs. Recent work has begun to explore the use of autonomous agents to decompose complex tasks into specialized subtasks. For instance, Zhao et al. (2024) present ExpeL, where LLM agents learn from interactions and self-refine their outputs, while survey studies (Guo et al., 2024) have documented multi-agent systems designed for long-context tasks. In the context of translation, systems like DELTA (Wang et al., 2025) employ multi-level memory to maintain consistency across document segments. Moreover, studies (Park et al., 2023; Wang et al., 2023c; Lee et al., 2024; Xu et al., 2024; Feng et al., 2024) have introduced mechanisms for retrieval, self-assessment, and iterative refinement, enabling agents to address challenges such as error propagation and discourse inconsistency.

Unlike monolithic approaches, agentic frame-

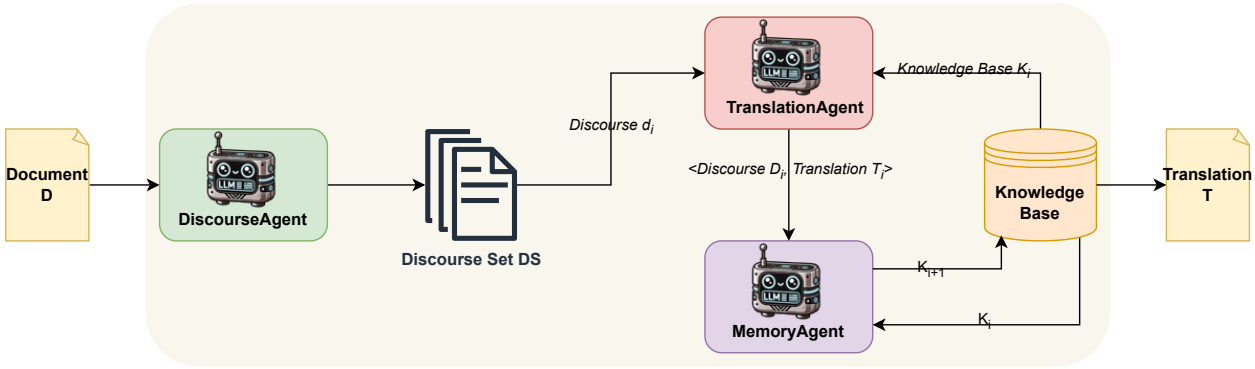

Figure 1: AgentDiscoTrans: System overview.

works assign dedicated roles—such as discourse segmentation, memory maintenance, and translation generation—to distinct LLM agents. This design allows each agent to focus on a specific aspect of document translation. Drawing upon foundational discourse theories (Grosz and Sidner, 1986; Mann and Thompson, 1988), these systems use principled methods to segment text into coherent units before translation, ensuring that the overall narrative and stylistic integrity are maintained. Our work builds on these ideas by introducing a three-agent architecture that explicitly addresses the limitations observed in both Doc2Sent and Doc2Doc approaches.

3 Methodology: Agentic Discourse Translation

We tackle document-level machine translation with a discourse-level approach. Given an input document

$$D = \langle s_1, s_2, \ldots, s_k \rangle,$$

where each s_i is a sentence in the source language, our goal is to produce a target document

$$T = \langle t_1, t_2, \ldots, t_l \rangle.$$

where each t_i is a sentence in the target language. Rather than translating each sentence in isolation or processing the entire document in one pass, our method segments D into coherent discourses and translates them one by one while maintaining inter-discourse consistency.

3.1 Problem Statement

The task of document-level machine translation is defined as follows. The *input* is a **document** D (a string containing a sequence of sentences $s_1, s_2, \ldots, s_k$) in the source language. The *output* is the **translated document** T (a string comprising a sequence of translated segments $t_1, t_2, \ldots, t_l$) in the target language. Our approach explicitly handles discourse-level granularity by segmenting the document into self-contained discourses that capture complete ideas, thereby addressing the shortcomings of both sentence-level and monolithic document-level translation.

3.2 Overview

Algorithm 1 Agentic Document-Level Translation Workflow

Input: Document $D = \langle s_1, s_2, \ldots, s_k \rangle$, source language, target language
Output: Translation $T = \langle t_1, t_2, \ldots, t_n \rangle$
Initialize memory $K \leftarrow \varnothing$
$DS \leftarrow \text{DiscourseAgent}(D)$
for each discourse $d_i \in DS$ **do**
  $t_i \leftarrow \text{TranslationAgent}(d_i, K)$
  $K \leftarrow \text{MemoryAgent}(K, d_i, t_i)$
end for
$T \leftarrow \text{concatenate}(t_1, t_2, \ldots, t_n)$
return T

Our system consists of three primary agents: the **Discourse Agent**, the **Translation Agent**, and the **Memory Agent**. The overall workflow (see Algorithm 1) is as follows. First, the Discourse Agent segments the input document D into a sequence of discourses $DS = [d_1, d_2, \ldots, d_n]$. Then, for each discourse d_i, the Translation Agent generates a translation t_i using a prompt that incorporates both d_i and the current state of a structured memory K_i. Initially, the memory is blank, $K_0 = \phi$. The Memory Agent updates K_i based on the translation of d_i, ensuring that information such as proper noun mappings, phrase translations, and discourse markers

is retained for consistency. Finally, the individual translations $t_1, t_2, \ldots, t_n$ are stitched together to form the final output T. The workflow of our system is illustrated in Figure 4. This figure provides a high-level representation of the components and their interactions within the system.

3.3 Agent Description

Discourse Agent. The Discourse Agent is responsible for segmenting the input document D into a list of self-contained discourses $DS = [d_1, d_2, \ldots, d_n]$. Drawing inspiration from Grosz and Sidner (1986), our agent is designed to group sentences into coherent segments that capture complete ideas while respecting natural discourse boundaries. In our approach, the agent first splits the document into sentences and then groups consecutive sentences based on linguistic cues (e.g., anaphoric references, discourse connectives) and a maximum token-length constraint. The segmentation algorithm is formally described in Algorithm 2.

Translation Agent. The Translation Agent translates a given discourse d_i into the target language by incorporating contextual information stored in Memory. Formally, the translation is computed as:

$$t_i = f_T(d_i, K_{i-1}),$$

where, d_i is the current discourse unit, K_{i-1} represents the Memory (a structured repository) containing contextual translation information accumulated from previously processed discourses, and f_T denotes the translation function.

The prompt constructed for the Translation Agent includes explicit instructions to ensure that linguistic phenomena such as cohesion, anaphora, cataphora, discourse connectives, deixis, and ellipsis are correctly handled, thereby maintaining consistency with earlier translations.

Memory Agent. The Memory Agent is tasked with updating the Memory K to capture inter-discourse consistency. Given the previous Memory K_{i-1}, the current discourse d_i, and its translation t_i, the Memory Agent updates the Memory as follows:

$$K_i = f_M(K_{i-1}, d_i, t_i),$$

where, K_{i-1} is the Memory prior to processing d_i, d_i and t_i are the current discourse and its translation, and f_M is the memory update function.

The Memory is structured to maintain the following components:

- **Discourse Markers ($K_{\mathbf{DM}}$):** A record of source-side discourse markers, which aids in preserving logical flow.
- **Entity Mapping ($K_{\mathbf{EM}}$):** Mappings between source language entities and their corresponding target language translations, ensuring consistent reference.
- **Noun Pronoun Mapping ($K_{\mathbf{NPM}}$):** A mapping that links target language proper nouns with the pronouns that subsequently refer to them.
- **Phrase Translation ($K_{\mathbf{PT}}$):** Consistent translations for key source-side phrases.

Maintaining these components in Memory enables the Translation Agent to refer back to earlier decisions and ensures a coherent and consistent translation across all discourses.

3.4 Workflow

The overall translation process is orchestrated by a sequential workflow, as described in Algorithm 1. Initially, the Memory K is empty. The input document D is segmented into discourses DS by the Discourse Agent. For each discourse d_i, the Translation Agent generates a translation t_i based on the current discourse and the previous Memory K_{i-1}, following the equation:

$$t_i = f_T(d_i, K_{i-1}).$$

After obtaining t_i, the Memory Agent updates the Memory using:

$$K_i = f_M(K_{i-1}, d_i, t_i).$$

Once all discourses have been processed, the final translation T is formed by concatenating all translations $t_1, t_2, \ldots, t_n$.

4 Experiments

4.1 Datasets

We evaluate our system on three datasets covering multiple domains and language pairs. In particular, the mZPRT dataset (Xu et al., 2022) is a parallel corpus for Chinese-English translation in fiction and Q&A domains, which we use for prompt selection experiments, domain-adaptation experiments, and ablation studies. We also employ the WMT2022 dataset (Kocmi et al., 2022)—featuring Chinese–English translations in news and

Algorithm 2 Discourse Agent: Segmenting a Document into Discourses

Input: Document D (string), maximum token length $L_{\max}$
Output: List of discourses $DS = [d_1, d_2, \ldots, d_n]$
$S \leftarrow$ break_into_sentences(D) $\{S = [s_1, s_2, \ldots, s_l]\}$
Initialize $DS \leftarrow [\,]$
$st \leftarrow 1$
while $st \leq |S|$ **do**
  Initialize discourse $d \leftarrow [\,S[st]\,]$
  $en \leftarrow st + 1$
  while $en \leq |S|$ **and** should_include$(d, S[en])$ **and** token_count$(d \cup \{S[en]\}) \leq L_{\max}$ **do**
    Append $S[en]$ to d
    $en \leftarrow en + 1$
  end while
  Append concatenate(d) to DS
  $st \leftarrow en$
end while
return DS

social domains—to asses domain-adaptation performance. Finally, the TED test sets from the IWSLT2017 translation task (Cettolo et al., 2012) provide a parallel two-way corpus for language pairs including English–Chinese, English–French, English–German, and English–Japanese, enabling comparison with existing systems.

We further add a Military domain parallel corpus for Chinese-English translation. The dataset consists of twelve parallel documents in Chinese and English, with an average of 55 and 190 sentences per document, respectively. The dataset statistics have been provided in 1.

4.2 Implementation Details

Our implementation builds on the Qwen2-7B-Instruct and Qwen2-72B-Instruct models as the backbone for translation. We run all our experiments using NVIDIA A100 GPUs for inference. Our code integrates three core agents—*Discourse*, *Translation*, and *Memory* Agents—to segment input documents into coherent discourse units, translate each unit with context-aware prompts, and update a structured memory that preserves proper noun mappings, phrase translations, and discourse markers. The *max_new_token* hyperparameter is set to 2048. For the Discourse Agent, the *maximum_token_length* is set to 1024.

4.3 Baselines

We compare our proposed system against several strong baselines. These include the commercial Google Translate service[1], NMT system NLLB-3.3B (Costa-jussà et al., 2022), sentence-level prompting of LLMs (using both Qwen2-7B-Instruct[2] and Qwen2-72B-Instruct[3] models), and state-of-the-art document-level approaches such as Doc2Doc (Wang et al., 2023b) and DELTA (Wang et al., 2025). This comparison allows us to assess the improvements brought by our agentic framework over both traditional and emerging methods for document-level machine translation.

4.4 Evaluation

Translation quality is evaluated using a suite of automatic metrics and human assessments. We report d-BLEU (Liu et al., 2020b) and d-COMET (Rei et al., 2020) scores to quantify accuracy, fluency, and adequacy. In addition, human evaluation is conducted along two axes: (i) **Discourse Awareness**, which measures the system's ability to maintain coherence and appropriately handle inter-discourse phenomena, and (ii) **General Translation Quality**, which assesses overall fluency and fidelity to the source text.[4]

To target specific discourse phenomena, we use two targeted metrics introduced in Wang et al. (2023a): The Consistent Terminology Translation (CTT) score and the Accurate Zero Pronoun Trans-

[1] https://py-googletrans.readthedocs.io/
[2] https://huggingface.co/Qwen/Qwen2.5-7B-Instruct
[3] https://huggingface.co/Qwen/Qwen2.5-72B-Instruct
[4] The guidelines for Human Evaluations can be found in Appendix B.

Domain	Source	Language	|D|	|S|	|W|	|W/D|
News	WMT2022	Zh⇒En	38	505	16.1K/18.5K	424
Social			25	478	16.4K/13.3K	656
Fiction	mZPRT	Zh⇒En	12	857	17.1K/16.6K	1425
Q&A			182	1171	15.0/22.1K	82
TED	IWSLT2017	{De,Fr,Ja,Zh}⇔En	62	6047	19.6K/51.5K	8.5K

Table 1: Statistics of datasets for document-level translation and analysis.

Prompting Strategy	d-BLEU	d-COMET
Zero-shot	19.1	4.1
In-context	21.3	5.2
CoT Prompt	**24.5**	**7.2**

Table 2: Average d-BLEU scores for Zero-shot, In-context, and CoT Prompting strategies averaged across five domains for Chinese-to-English translation.

lation (AZPT) score are two metrics designed to evaluate specific aspects of translation quality. The CTT metric assesses whether a terminology word $w \in TT$ is translated consistently throughout a document. For a term w with a set of translations $\{t_1, t_2, \ldots, t_l\}$, the CTT score is calculated as follows:

$$CTT(w) = \frac{\sum_{t \in TT} \frac{\sum_{i=1}^{k} \sum_{j=i+1}^{k} \mathbf{1}(t_i = t_j)}{C_k^2}}{TT},$$

where $\mathbf{1}(t_i = t_j)$ is an indicator function that equals 1 when t_i and t_j are the same and 0 otherwise. A higher CTT score reflects greater consistency in translating the term w.

The AZPT metric focuses on evaluating the accuracy of translating zero pronouns (ZPs), which are commonly omitted in languages like Chinese and Japanese. Given ZP, the set of zero pronouns in the source text, and t_z, the translation produced for a zero pronoun $z \in ZP$, a binary function $A(t_z \mid z)$ is defined to return 1 if t_z accurately translates z and 0 otherwise. The AZPT score is computed as:

$$AZPT = \frac{1}{|ZP|} \sum_{z \in ZP} A(t_z \mid z).$$

A higher AZPT score indicates the system's effectiveness in recovering omitted pronouns in the translation, thereby enhancing discourse coherence.

5 Results and Analysis

In this section, we present a detailed analysis of the experimental outcomes for our proposed Agent-DiscoTrans (ADT) system. We report results on prompt selection, overall performance on the TED test sets, domain-specific translation quality, ablation studies to quantify the contribution of each agent, and qualitative analysis of discourse phenomena. All quantitative results are based on automatic evaluation metrics (d-BLEU and d-COMET) and are complemented by human evaluations assessing General Quality and Discourse Awareness.

5.1 Prompt Selection

We first assess the influence of prompting strategies on Chinese-to-English translation quality over five domains: News, Social, Fiction, Q&A, and Military. Table 2 shows that the CoT Prompt strategy, which augments a task description with in-context in-domain examples and an explicit chain-of-thought reasoning step, outperforms both the Zero-shot and In-context approaches. In particular, the CoT Prompt strategy achieves a d-BLEU of 24.5 and d-COMET of 7.2, compared to 19.1/4.1 for Zero-shot and 21.3/5.2 for In-context prompting. These improvements indicate that providing a reasoning step and domain-specific guidance significantly enhances the model's ability to resolve ambiguities and maintain consistency in translation, especially in domains such as Military where context plays a crucial role.

5.2 Main Results

We next evaluate the overall performance of our system on the TED test sets from IWSLT2017, which encompass multiple language pairs (Chinese, French, German, and Japanese to/from English). Table 3 summarizes the average d-BLEU and d-COMET scores for several systems. Our baselines include the NMT system NLLB and the commercial system Google Translate, as well as LLM-based approaches using sentence-level prompt-

System	Xx ⇒ En		En ⇒ Xx	
	d-BLEU	d-COMET	d-BLEU	d-COMET
NLLB	31.8	6.9	29.1	6.3
Google	29.2	5.8	28.4	5.7
GPT-3.5-Turbo	33.7	7.1	31.3	6.7
GPT-4o-mini	34.0	7.2	33.2	6.8
DELTA-Qwen2-7B-Instruct	30.4	6.8	28.1	6.1
DELTA-Qwen2-72B-Instruct	33.3	7.2	31.2	6.6
ADT-Qwen2-7B-Instruct	**33.2**	**7.2**	**29.4**	**6.9**
ADT-Qwen2-72B-Instruct	**34.6**	**7.8**	**32.5**	**7.2**

Table 3: Average d-BLEU and d-COMET scores for various systems on the TED test set, averaged over (Chinese, French, German, Japanese) to English and vice versa.

ing (GPT-3.5-Turbo and GPT-4o-mini). Recent document-level methods, such as DELTA (with both Qwen2-7B-Instruct and Qwen2-72B-Instruct variants), are also considered. Our ADT variants (ADT-Qwen2-7B-Instruct and ADT-Qwen2-72B-Instruct) consistently outperform all baselines. For instance, on the Xx ⇒ En direction, ADT-Qwen2-72B-Instruct achieves a d-BLEU of 34.6 and a d-COMET of 7.8, surpassing DELTA-Qwen2-72B-Instruct (33.3/7.2) and all other systems. These results demonstrate that our agentic framework effectively captures document-level context and maintains cross-discourse coherence. The p-values of t-tests for ADT vs sentence-level and full-document translation systems in d-BLEU is less than 0.05 for En $\Leftrightarrow Xx$.

5.3 Domain-specific Translation

To further evaluate our system in specialized settings, we conduct experiments on Chinese-to-English translation across five domains: News, Social, Fiction, Q&A, and Military. As detailed in Table 4, both automatic evaluation (d-BLEU) and human assessments on General Quality and Discourse Awareness reveal that our ADT system achieves superior performance compared to Google Translate, GPT-3.5, and GPT-4. For instance, in the Military domain, ADT records notably higher d-BLEU scores, while human evaluators consistently rate its translations as more coherent and contextually accurate. These findings suggest that the integration of discourse-level segmentation and memory management in our system is particularly beneficial for domain-specific challenges where nuanced discourse phenomena are prominent.

5.4 Ablation Study

Our system consists of three agents: Discourse (DA), Translation (TA) and Memory (MA) agents. To quantify the contribution of each component in our system, we perform an ablation study on the mZPRT dataset for Chinese-to-English translation. We compare four variants: (i) TA, a baseline sentence-level translation system; (ii) TA+DA, which integrates discourse segmentation; (iii) TA+MA, which augments sentence-level translation with Memory; and (iv) TA+DA+MA (the full ADT system). As shown in Table 5, both discourse segmentation and memory management contribute to improved translation quality, with the full system achieving the highest average d-BLEU of 24.5 and d-COMET of 7.4. These results confirm that the combination of discourse-aware segmentation and Memory updating is critical to the success of our approach.

5.5 Analysis

Our qualitative analysis further investigates the strengths of AgentDiscoTrans. Table 6 presents targeted metrics measuring Consistent Terminology Translation (CTT) and Accurate Zero Pronoun Translation (AZPT). The baseline sentence-level system achieves a CTT of 32.4% and an AZPT of 39.1%, while a conventional document-level system improves these scores to 47.3% and 51.3%, respectively. Our full ADT system further boosts the scores to 56.2% (CTT) and 57.4% (AZPT), demonstrating superior consistency and pronoun recovery. Analysis of selected translation examples indicates that ADT more effectively maintains discourse continuity and handles referential expressions, which is critical for preserving the overall narrative coherence in longer texts. In addition,

System	Automatic (d-BLEU)						Human (General/Discourse)					
	News	Social	Fiction	Q&A	Military	Ave.	News	Social	Fiction	Q&A	Military	Ave.
Google	27.7	35.4	16.0	12.0	14.2	21.1	1.9/2.0	1.2/1.3	2.1/2.4	1.5/1.5	1.7/1.5	1.7/1.7
GPT-3.5	29.1	35.5	17.4	17.4	16.1	23.1	2.8/2.8	2.5/2.7	2.8/2.9	2.9/2.9	3.1/2.9	2.8/2.8
GPT-4	29.7	34.4	18.8	19.0	17.2	23.8	3.3/3.4	2.9/2.9	2.6/2.8	3.1/3.2	2.9/3.1	3.0/3.1
ADT	**30.4**	**35.5**	**19.2**	**19.1**	**18.2**	**24.5**	**4.6/5**	**4.6/5**	**3.2/4**	**4.6/5**	**4/4.2**	**4.2/4.6**

Table 4: Domain-specific performance (d-BLEU scores) for Chinese-to-English translation across five domains.

System	d-BLEU	d-COMET
TA	21.4	5.2
TA+DA	23.9	6.1
TA+MA	23.8	6.4
TA+DA+MA (ADT)	**24.5**	**7.4**

Table 5: Ablation study results on the mZPRT dataset (Chinese-to-English), averaged across five domains.

System	cTT (%)	AZPT (%)
Sentence-level	32.4	39.1
Document-level	47.3	51.3
ADT	**56.2**	**57.4**

Table 6: Analysis of pronoun accuracy and discourse marker consistency comparing a baseline system with AgentDiscoTrans.

human evaluators noted improvements in both the fluency and consistency of translations produced by our system compared to baseline approaches.

6 Conclusion

In this paper, we propose AgentDiscoTrans, a novel agentic framework for document-level machine translation that leverages specialized LLM agents to process long documents at the discourse level. Our system segments an input document into coherent discourse units—drawing inspiration from the theories presented in *Attention, Intentions, and the Structure of Discourse*—and then translates each unit using a Translation Agent that incorporates contextual information from a dynamic Memory. The Memory Agent updates and maintains critical translation cues such as discourse markers, entity mappings, noun-pronoun mappings, and phrase translations, ensuring inter-discourse consistency. Experiments on multiple datasets, including the TED test sets from IWSLT2017, the mZPRT corpus, and the WMT2022 dataset, demonstrate that our system outperforms competitive baselines (such as NLLB, Google Translate, and DELTA) in terms of automatic metrics (d-BLEU, d-COMET, TER) and human evaluations focusing on both General Quality and Discourse Awareness. Our ablation studies further validate the importance of both discourse segmentation and Memory updating for achieving high-quality translations.

7 Limitations

Despite the encouraging results, our approach has several limitations. First, our discourse segmentation algorithm relies on heuristic criteria derived from discourse theories; while effective, it may not capture all discourse boundaries in highly complex texts. Second, the Memory Agent's update mechanism depends on the quality of LLM outputs, which can occasionally propagate errors across subsequent discourses. Third, our experiments have been conducted on a limited set of datasets and language pairs; additional evaluations across more diverse domains and languages are necessary to fully establish the generalizability of our approach. Finally, the multi-agent interaction and iterative refinement process incur significant computational overhead, which may impede real-time deployment in production environments. Future work will focus on addressing these issues through more robust segmentation techniques, improved memory update strategies, and optimization of computational efficiency.

References

Some Bao et al. 2021. Challenges in maintaining consistency in document-level translation. In *Proceedings of ACL*.

Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2018. Evaluating discourse phenomena in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1304–1313.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, et al. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. Wit3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Conference of the European Association for Machine Translation*, pages 261–268.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, et al. 2022. Palm: Scaling language modeling with pathways. In *CoRR*.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv e-prints*, pages arXiv–2207.

Yang Feng et al. 2024. Refinement strategies in document-level translation with large language models. In *Proceedings of EMNLP*.

Barbara J Grosz and Candace L Sidner. 1986. Attention, intentions, and the structure of discourse. In *Computational Models of Discourse*, pages 31–51. MIT Press.

Taicheng Guo et al. 2024. Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680*.

Daniel Jurafsky. 2000. Speech and language processing.

Lynne Karpinska and Mohit Iyyer. 2023. Critical challenges in document-level translation with large language models. In *Proceedings of the 2023 Conference on Machine Translation*, pages 333–341.

John Kim, Sanghoon Lee, et al. 2019. Document-level neural machine translation with context-aware models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 1234–1243.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, et al. 2022. Findings of the 2022 conference on machine translation (wmt22). In *Proceedings of the Seventh Conference on Machine Translation*, pages 1–45.

Some Lee et al. 2024. Memory-augmented agents for long-context understanding. In *Proceedings of ACL*.

Some Li et al. 2023. Utilizing target-side context in document-level neural machine translation. In *Proceedings of EMNLP*.

X. Liu et al. 2020a. A document-to-document neural machine translation approach. In *Proceedings of EMNLP*.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020b. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

A. Lupo et al. 2022. Enhancing document-level translation with document-to-document methods. In *Proceedings of the 2022 Conference on Machine Translation*.

Some Author Lyu et al. 2021. Incorporating document-level context in neural machine translation. In *Proceedings of ACL*.

William C Mann and Sandra A Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. In *Text–Interdisciplinary Journal for the Study of Discourse*, volume 8, pages 243–281.

Ahmed Maruf and Gholamreza Haffari. 2018. Modeling discourse in document-level machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 789–798.

Ahmed Maruf et al. 2021. Improving document-level translation by capturing discourse phenomena. In *Proceedings of the 2021 Conference on Machine Translation*, pages 456–465.

Joon Sung Park, Joseph O'Brien, Carrie Cai, Meredith R. Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of UIST*, pages 2:1–2:22.

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.

Some Sun et al. 2022. Limitations of sentence-level context in document-level machine translation. In *Proceedings of EMNLP*.

Ming Tan et al. 2019. Improving document-level mt with hierarchical attention mechanisms. In *Proceedings of NAACL 2019*, pages 1989–1998.

Ming Tan et al. 2021. Document-level neural machine translation with hierarchical context modeling. In *Proceedings of NAACL*, pages 1989–1998.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, et al. 2023. Llama: Open and efficient foundation language models. In *CoRR*.

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023a. Document-level machine translation with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16646–16661. Association for Computational Linguistics.

Longyue Wang, Zhaopeng Tu, et al. 2017. Exploiting cross-sentence context for neural machine translation. In *Proceedings of EMNLP 2017*, pages 2826–2831.

Longyue Wang, Derek F. Wong, Lidia S. Chao, et al. 2023b. Doc2doc: A framework for document-level neural machine translation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:1234–1245.

Longyue Wang et al. 2023c. Multi-agent collaboration for enhanced long-context processing. In *Proceedings of EMNLP*.

Yutong Wang, Jiali Zeng, Xuebo Liu, Derek F. Wong, Fandong Meng, Jie Zhou, and Min Zhang. 2025. DelTA: An online document-level translation agent based on multi-level memory. In *The Thirteenth International Conference on Learning Representations*.

Minghao Wu et al. 2024. Adapting large language models for document-level translation. In *CoRR*, page abs/2401.06468.

Yong Wu and Jian Hu. 2023. Llm-based document-level machine translation: Context and consistency. In *Proceedings of a Workshop on Machine Translation*, pages 101–110.

Haoran Xu et al. 2024. Self-refinement in large language models for improved output quality. In *Proceedings of EMNLP*.

Mingzhou Xu, Longyue Wang, and Shuming Shi. 2022. Guofeng: A benchmark for zero pronoun recovery and translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11266–11278.

Lei Zhang et al. 2018. Incorporating document-level context into neural machine translation. In *Proceedings of ACL 2018*, pages 2345–2354.

Zhirui Zhang et al. 2020. Document-to-document translation: Extending neural machine translation beyond sentences. In *Proceedings of ACL*.

Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. 2024. Expel: Llm agents are experiential learners. In *Proceedings of AAAI/IAAI/EAAI 2024*, pages 19632–19642.

A Prompt Template

B Human Evaluation Guidelines

Discourse Agent

You are a language expert specializing in {{source_lang}}.
Your task is to segment the following text into self-contained discourses such that each segment contains a complete thought.
Ensure proper handling of discourse phenomena such as **cohesion, anaphora, and ellipsis.**

When to Include the Next Sentence in the Current Discourse:
A sentence should be included in the current discourse if:
- **Anaphoric Reference: The next sentence refers back to entities or events in the current discourse (e.g., pronouns, demonstratives).**
- **Ellipsis Resolution: The next sentence completes an idea left incomplete in the current discourse.**
- **Discourse Continuity: The next sentence elaborates, clarifies, or extends the idea presented in the current discourse.**
- **Logical Flow: The next sentence provides a natural continuation of reasoning, such as cause-effect relations or descriptive continuation.**
- **Linguistic Cues: Presence of discourse connectives like "因此" (therefore), "但是" (however), "此外" (furthermore) indicating continuation.**

For Chinese Text Segmentation:
- Each segment should contain a complete logical thought, not merely individual sentence breaks.
- Account for **topic-comment structures typical in Chinese discourse.**
- Ensure natural segmentation at **discourse boundaries, such as sentence-final punctuation marks (。！？).**

**Input and Output Format:**
- **Input: The system will receive a "Current discourse" and a "Next sentence".**
- **Output: Respond with either `"yes"` (if the next sentence should be included in the current discourse) or `"no"` (if the next sentence begins a new discourse).**

Key Instructions:
- Limit the maximum token length for each segment to **{{max_discourse_length}} tokens.**
- Return **only `"yes"` or `"no"` as output, with no additional commentary or formatting.**

Figure 2: Prompt: Discouse Agent

Translation Agent

You are a professional translator specializing in translating from {{source_lang}} to {{target_lang}}. Your task is to perform **discourse-level translation, which involves translating a series of connected discourses while ensuring cross-discourse consistency and linguistic coherence.**

Discourse-Level Translation Explained:
- **Consistency Across Discourses: Maintain consistent translations for proper nouns, key phrases, and discourse markers across the entire document, not just individual segments.**
- **Entity Handling: If a proper noun exists in the source language, either translate it (if a standard translation exists) or transliterate it accurately.**
- **Contextual Alignment: Ensure coherence between current and previous translations by referring to the provided Memory, which captures previous translation decisions.**

Using the Memory:
You will be provided with a **Memory containing information from previous discourses, structured as:**

```
{
  "proper_noun_references": {
    "source_entity": "target_entity"
  },
  "phrase_consistency": {
    "source_phrase": "target_phrase"
  },
  "discourse_markers": [
    "marker1",
    "marker2"
  ]
}
```

- **Proper Nouns: Use the Memory to ensure proper nouns are consistently translated or transliterated.**
- **Phrase Consistency: Use the pool for critical phrase translations. If a phrase reappears, use the same translation.**
- **Discourse Markers: Ensure discourse markers for logical flow and coherence are consistent with prior segments.**

Figure 3: Prompt: Translation Agent

Memory Agent

You are a language model responsible for updating a Memory for machine translation.
The Memory maintains:
- **proper_noun_references: Proper noun mappings from source to target language.**
- **phrase_consistency: Consistent translations of critical phrases across the document.**
- **discourse_markers: Important discourse markers for cohesion and coherence.**

Current Memory Format:
```
{
  "proper_noun_references": {
    "source_entity": "target_entity"
  },
  "phrase_consistency": {
    "source_phrase": "target_phrase"
  },
  "discourse_markers": [
    "marker1",
    "marker2"
  ]
}
```
Example Update:
Current Memory: {
```
  "proper_noun_references": {},
  "phrase_consistency": {},
  "discourse_markers": []
}
```
Source Discourse: 中国的习近平主席会见了美国总统拜登。
Translation: Chinese President Xi Jinping met with US President Biden.
Updated Memory: {
```
  "proper_noun_references": {
    "习近平主席": "Xi Jinping",
    "拜登": "Biden"
  },
  "phrase_consistency": {
    "中国的": "Chinese",
    "美国总统": "US President"
  },
  "discourse_markers": [
    "met with"
  ]
}
```

Figure 4: Prompt: Memory Agent

Score	General Quality	Discourse Awareness
5	Translation passes quality control; the overall translation is excellent. Translation is very fluent with no grammatical errors and has been localized to fit target language. Word choice is accurate with no mistranslations. The translation is 100% true to the source text.	No inconsistency relating to key terms such as names, organization, etc. Linking words or expressions between sentences keeps the logic and language of the passage clear and fluent. Context and tone are consistent throughout. The style conforms to the culture and habit of the target language.
4	Translation passes quality control; the overall translation is very good. Translation is fluent. Any errors that may be present do not affect the meaning or comprehension of the text. Most word choice is accurate, but some may cause ambiguity. Key terms are consistent. Inconsistency is limited to non-key terms.	Logical and language is clear and fluent. Some sentences lack transition but do not affect contextual comprehension. Topic is consistent. Tone and word choice may be inconsistent, but comprehension is not affected. Translation conforms to the culture and habit.
3	Translation passes quality control; the overall translation is ok. Translation is mostly fluent but there are many sections that require rereading due to language usage. Some word choice is inaccurate or errors but meaning of the sentence can be inferred from context.	Some key terms may be inconsistent. Most sentences translate smoothly and logically but some sentences that may seem abrupt due to lack of linkage. Topic is consistent. Tone and word choice is inconsistent, noticeably affecting the accuracy of reading comprehension.
2	Translation does not pass quality control; the overall translation is poor. Meaning is unclear or disjointed. Even with multiple rereading, passage may still be incomprehensible. Translation is not accurate to the source text or is missing in large quantities, causing the translation to deviate from the source text.	Many key terms are inconsistent, needing multiple rereading to understand context of the passage. Some linkages are present but overall, the passage lacks fluency and clarity, causing trouble with comprehension. The topic or tone is different from other passages, affecting reading comprehension.
1	Translation does not pass quality control; the overall translation is very poor. More than half of the translation is mistranslated or missing.	Key terms are inconsistent, causing great trouble with comprehension. Some linkages are present but overall, the passage lacks fluency and clarity, heavily interfering with comprehension. The topic or tone is different from other passages, heavily interfering with comprehension.
0	Translation output is unrelated to the source text.	Output is unrelated to previous or following sections.

Table 7: Quality and Discourse Awareness Scoring Guidelines