

Evaluating Coreference Consistency in Chinese-to-English Dialogue Translation

Xu Yang¹, Zhengxian Gong¹, Junhui Li^{1,2*}, Jinlong Yang³,
Min Zhang³, Shimin Tao³, Daimeng Wei³,

¹School of Computer Science and Technology, Soochow University, Suzhou, China

²Jiangsu Key Lab of Language Computing, China

³Huawei Translation Services Center, Beijing, China

{yangxiaoxu778}@gmail.com, {zhxgong, jhli}@suda.edu.cn

{yangjinlong7, zhangmin186, taoshimin, weidaimeng}@huawei.com

Abstract

Coreference consistency is essential for preserving discourse coherence in dialogue translation, yet it remains underexplored in the context of large language models (LLMs). In this paper, we present a comprehensive study of coreference preservation in Chinese-to-English dialogue translation. We construct a new dataset based on the RiSAWOZ corpus, annotated with source-side zero pronouns and coreference chains. Translations are obtained from a diverse set of LLMs of varying sizes. To evaluate whether coreference relations are maintained, we propose the *Coreference Consistency Score* (CCS), which quantifies the extent to which translated mentions preserve original coreferential chains. We further introduce an LLM-as-Judge protocol for assessing the faithfulness of translated mentions. Our results reveal that, despite producing fluent translations, many LLMs, including both small and large models, struggle to maintain cross-lingual coreference, highlighting the need for more discourse-aware dialogue translation models.

1 Introduction

Cross-lingual understanding has attracted increasing attention in NLP, driven by the need for machine translation (MT) systems that preserve not only sentence-level meaning but also discourse-level coherence (Lyu et al., 2021; Lei et al., 2022; Zhang et al., 2023b; Vu et al., 2024). A central element of discourse coherence is coreference, where multiple expressions refer to the same entity (Soon et al., 2001; Zhao and Ng, 2007), enabling coherent reference tracking and conversational continuity. Coreference has been extensively studied in monolingual settings (Yu et al., 2019; Khosla et al., 2021; Zheng et al., 2023), and related phenomena such as pronoun translation, zero pronouns, and discourse consistency have also been widely explored in machine translation. However, the broader question of

*Corresponding author: Junhui Li

[Context] usr: 坐 高铁/gao_tie 的话有哪个车? [Which high-speed train is available?] sys: 有 G90/G90, 成都东-北京西。 [There is the G90 train, from Chengdu East to Beijing West.] [Source] usr: 那一等座票价是多少? 它/ta 几点能到北京? [Auto Translation] sys: How much is the first-class ticket, and what time will we arrive in Beijing? [Reference] sys: What is the price for a first-class ticket? When will it arrive in Beijing?
[Context] usr: 姑苏区有没有价位便宜的 酒店/jiu_dian 推荐一下。 [Could you recommend any affordable hotels in Gusu District?] usr: 这附近有什么热门的学校校园吗? [Are there any popular schools or campuses nearby?] sys: 苏州大学/su_zhou_da_xue 就很热门。 [Soochow University is quite popular.] usr: 能地铁直达吗? [Can I reach there directly by subway?] [Source] sys: 可以地铁直达 (苏州大学/su_zhou_da_xue), 直接去。 [Auto Translation] sys: Yes, you can take the subway directly to the hotel. [Reference] sys: You can take the subway directly there.

Figure 1: Examples of Chinese-to-English dialogue translation in which source-side coreferent mentions are not preserved in the generated translations.

how coreference chains as a whole are preserved in dialogue translation systems, particularly modern LLM-based MT systems, less well understood.

Dialogue translation further complicates this problem due to informal expressions, frequent speaker turns, and strong reliance on context. As shown in Figure 1, current systems, including strong LLM-based models such as LLaMA3-8B-Instruct (Meta, 2024), often fail to preserve coreference consistency. In the first example, the model correctly translates *G90* but misrenders its coreferential pronoun 它/ta (it) as *we*, leading to a change in entity interpretation. The

challenge is further pronounced in Chinese dialogue due to the pervasive use of zero pronouns. As in the second example, a dropped pronoun referring to 苏州大学/*su_zhou_da_xue* (*Soochow University*) is incorrectly resolved as *the hotel*, introducing a wrong entity. These errors highlight a broader issue: even when translations are fluent, current MT systems struggle to maintain consistent entity references across dialogue turns, resulting in degraded discourse coherence and even distorted meaning.

In this paper, we systematically investigate coreference preservation in Chinese-to-English dialogue translation. We conduct both quantitative and qualitative analyses to evaluate how well current MT systems maintain source-side coreference relations, highlighting limitations of even advanced LLMs.¹ Although our findings reveal these limitations, we also employ LLMs in constrained roles for data annotation and evaluation. These components are supported by manual verification, inter-annotator agreement analysis, and classifier evaluation. All test-set annotations are manually constructed, and reference translations are produced by professional translators.

Our main contributions are as follows:

- We construct a Chinese-to-English dialogue translation dataset with 10,000/600/600 dialogues for train/dev/test, annotated with zero pronouns and coreference chains, with full manual verification on the test set.
- We propose a novel Coreference Consistency Score (CCS) and an LLM-as-judge framework for evaluating coreference preservation in dialogue translation. Through comprehensive experiments across diverse LLMs, we show that even strong LLMs struggle to maintain discourse-level coreference consistency.

2 Chinese-to-English Coreference Consistency Evaluation Dataset

In this section, we describe the construction of our Chinese-to-English dialogue translation dataset designed to evaluate coreference consistency. We detail the process of obtaining English translations (Section 2.1) and the annotation of source-side zero pronouns (Section 2.2) and coreference chains

¹Some acceptable translations may restructure discourse or resolve references differently while remaining natural to human readers. We focus on preservation of source-side coreference relations rather than treating all deviations as errors.

(Section 2.3). The source dialogues are derived from RiSAWOZ² (Quan et al., 2020), a large-scale Chinese corpus containing 11.2k multi-turn dialogues between clients and customer service agents, totaling over 150k utterances across 12 domains. We select RiSAWOZ because its multi-turn, task-oriented dialogues involve rich contextual dependencies and frequent zero-pronoun usage, making it well suited for studying discourse-level coreference preservation in Chinese-to-English dialogue translation. All utterances in RiSAWOZ are segmented using the HanLP toolkit³ (He and Choi, 2021) to facilitate annotation and analysis.

Following the original split, the dataset is divided into a training set of 10k dialogues and development and test sets of 600 dialogues each. Annotations for the training and development sets are initially generated automatically using LLMs. In contrast, annotations for the test set are created entirely manually from scratch to ensure the highest quality and to avoid any potential bias introduced by LLM-generated outputs.

2.1 English Translation Annotation

We use GPT-4o-mini (OpenAI, 2024) to translate the Chinese dialogues into English. Each utterance is tagged with a speaker label (*usr* for clients or *sys* for customer service agents), and an entire dialogue is translated in a single pass. To ensure utterance-level alignment, we verify that the speaker sequence in the translation matches the source. Misaligned translations are discarded, and the LLM is prompted again until correct alignment is achieved. The prompt used for translation is shown in Figure 5 (Appendix D).

Translation of Named Entities. Since named entities are critical for maintaining coreference chains, two trained master’s students manually verify all entity translations appearing in these chains to ensure their accuracy.

Test Set Translation. The test set is translated by a professional human translator.⁴ Given the conversational style of the source dialogues, the translator is instructed to produce English translations that are both natural and orally fluent, while faithfully preserving the original meaning.

²<https://github.com/terryqj0107/RiSAWOZ>

³<https://github.com/hankcs/HanLP>

⁴The translator is a native Chinese speaker, certified at the highest level in both interpretation and translation by CATI, and has passed TEM-8. She received compensation for her professional translation services.

2.2 Zero Pronoun Annotation

Chinese is a pro-drop language, where subjects and objects are often omitted when they can be inferred from context. These omitted elements are referred to as *zero pronouns*, as they function like pronouns but lack explicit surface forms. In contrast, English is a non-pro-drop language, where overt pronouns explicitly serve the function of maintaining discourse continuity (Rao et al., 2015). Since zero pronouns in Chinese are typically translated into explicit pronouns in English, we annotate zero pronouns in the Chinese dialogues to enable a more accurate evaluation of coreference consistency in dialogue translation. The source text with annotated zero pronouns will be used for coreference chain annotation, facilitating the subsequent computation of coreference consistency scores.

We use GPT-4o-mini to annotate zero pronouns in Chinese dialogues. The model is guided by carefully designed annotation guidelines, which include: (1) rules for identifying the scope of zero pronouns (Yang et al., 2022), (2) principles for encouraging the naturalness of recovered sentences, (3) principles for encouraging their grammatical correctness, and (4) strategies for handling special cases, such as colloquial expressions. To further improve performance, we include two representative dialogues with manually annotated zero pronouns as in-context learning examples. These examples help the model better recover omitted subjects or objects by leveraging broader context. The prompt used for annotation is shown in Figure 6 (Appendix D).

Test Set Annotation of Chinese Dialogues. Two native Chinese-speaking master’s students, serving as volunteer annotators, manually annotate the positions of zero pronouns and their corresponding explicit pronouns in the test set. Before starting, annotators receive comprehensive training, which includes discussion of 10 sample dialogues to ensure a consistent understanding of the task. In each annotation round, 55 dialogues are assigned to each annotator, with an overlap of 10 dialogues used to evaluate inter-annotator agreement. Discrepancies in the overlapping annotations are resolved through discussion. This process is repeated iteratively until the entire test set is annotated. The inter-annotator agreement, measured by Cohen’s Kappa, is 0.638 for zero pronoun annotation (see Appendix A.1 and A.2).

2.3 Coreference Chain Annotation

Coreference chains, which represent groups of mentions referring to the same real-world entity, are a central concept in coreference resolution (Ng, 2010; Pradhan et al., 2011). After recovering zero pronouns, we annotate coreference chains over all mentions, including the recovered zero pronouns. We first generate coreference chains automatically using the HanLP coreference resolution module (He and Choi, 2021).⁵ To improve the quality of the automatically generated annotations, we randomly sample 100 dialogues for manual inspection. The two annotators review the predicted links and correct clear errors by removing erroneous mentions.

Test Set Annotation of Chinese Dialogues. For the test set, coreference chains are annotated entirely manually using a web-based annotation tool that allows annotators to visualize dialogues and link coreferent mentions. The tool outputs structured coreference chains, reducing both annotation time and cognitive load. A screenshot of the annotation interface is provided in Appendix A.3.

The annotators perform the subsequent annotation. Prior to the main phase, they undergo training that covers the annotation guidelines, example dialogues, and proper use of the tool. Each annotation round assigns approximately 55 dialogues to each annotator, with 10 overlapping dialogues used to assess inter-annotator agreement. Disagreements are resolved through discussion, and this process is repeated iteratively until all dialogues are fully annotated. The inter-annotator agreement is 0.694 for coreference chain annotation (see Appendix A.3).

2.4 Data Statistics

Table 1 summarizes key statistics of our dataset. On average, each dialogue contains 13.57 utterances, with 150.11 words on the Chinese (source) side and 179.22 words on the English (target) side. Each dialogue also includes approximately 7.07 zero pronouns and 4.69 coreference chains, totaling 18.23 annotated mentions. In the test set, zero pronouns account for 25.78% of all mentions, highlighting their prevalence in Chinese dialogues. Figure 4 in Appendix B shows a dialogue example with annotated zero pronouns and coreference chains.

⁵<https://www.hanlp.com/semantics/functionapi/participle>

Set	#Dialogue	#Utterance	#Avg. Word _{src}	#Avg. Word _{tgt}	#Avg. ZPs	#Avg. CoChain	#Avg. Mention
Train	10,000	134,580	147.66	176.36	7.12	4.87	17.81
Dev	600	8,116	146.66	179.11	6.34	4.90	17.72
Test	600	9,286	194.39	227.10	6.66	4.38	25.83
All	11,200	151,982	150.11	179.22	7.07	4.69	18.23

Table 1: Data statistics of our coreference consistency evaluation dataset. The reported averages are calculated per dialogue. *CoChain* refers to coreference chains. As zero pronouns (ZPs) are included within coreference chains, the average number of mentions also accounts for them.

3 Coreference Consistency Evaluation

Given a parallel dialogue pair $(X, Y) = \{(x_i, y_i) |_{i=1}^N\}$ with N utterances, and a coreference chain $C = \{c_1, \dots, c_m\}$ containing m mentions in X , we expect that the translations of these mentions should also maintain coreference relations in the target dialogue Y . In other words, for any two mentions c_i and c_j in C , an inconsistency occurs if their translations are not coreferential in Y .

To evaluate this, we first introduce an automatic metric, which quantifies how well translations preserve source-side coreference relations (Section 3.1). Then we propose an LLM-as-judge evaluation which evaluates translation quality from a coreference consistency perspective (Section 3.2).

3.1 Coreference Consistency Score

To properly evaluate coreference consistency in translation, we propose the *Coreference Consistency Score* (CCS). CCS measures the proportion of mention pairs with a coreference chain whose coreference relation is correctly preserved after translation.

Formally, given a coreference chain $C = \{c_1, \dots, c_m\}$ in a source dialogue X , we first obtain the corresponding translations $t = \{t_1, \dots, t_m\}$ in the target dialogue Y via word alignment between X and Y .⁶ Each t_i may consist of zero, one, or multiple words. CCS for a single chain C is defined as:

$$\text{CCS}(C) = \frac{\sum_{1 \leq i < j \leq m} \mathbb{1}(\text{R}(t_i, t_j))}{\binom{m}{2}}, \quad (1)$$

where the denominator $\binom{m}{2}$ is the total number of unique mention pairs in C , and $\text{R}(t_i, t_j)$ is a predicate (details in Section 3.1.1) that returns True if t_i and t_j are coreferential in Y , and False otherwise. The indicator function $\mathbb{1}(\cdot)$ returns 1 when the input is True and 0 when it is False.

⁶We employ awesome-align (Dou and Neubig, 2021) to generate word alignments for each utterance pair, explicitly including zero pronouns on the source side. The tool is available at <https://github.com/neulab/awesome-align>.

While the CCS definition above applies to a single coreference chain, it can be naturally extended to an entire dialogue or dataset by aggregating the numerators and denominators across all chains and computing the overall score. Singleton chains are excluded because they contain no coreferential mention pairs and therefore do not contribute to CCS.

3.1.1 Coreference Consistency Classifier

As $\mathbb{1}(\text{R}(t_i, t_j))$ in Eq. 1, computing CCS requires identifying *whether the translations of two source-side mentions are coreferential in the target dialogue*. To this end, we construct a classifier that predicts coreference relations based on contextual representations of translated mentions.

Training Instances. For each coreference chain $C = \{c_1, \dots, c_m\}$ in a parallel dialogue pair (X, Y) from the training set, we collect the translations of all mentions, denoted as $T = \{t_1, \dots, t_m\}$. Each pair (t_i, t_j) within the same chain is labeled as a positive instance if both t_i and t_j contain more than one word. In contrast, translations of two mentions from different chains are treated as negative instances. To maintain a balanced training set, we randomly sample negative instances such that the number of positive and negative instances is equal.

Classification. We first encode the entire target-side dialogue Y using Llama-3.1-8B-base (Meta, 2024), obtaining contextualized hidden states for all tokens. For each translated mention t_i , we compute its representation h_i by averaging the hidden states of its inside tokens. For a translation pair (t_i, t_j) , we concatenate their representations $[h_i; h_j]$ and feed the result into a two-layer feed-forward network:

$$p = \text{softmax}(W_2 \tanh(W_1[h_i; h_j])), \quad (2)$$

where $W_1 \in \mathbb{R}^{d \times 2d}$ and $W_2 \in \mathbb{R}^{2 \times d}$ are learnable parameters, and d is the hidden size of the LLM.

The output $p \in \mathbb{R}^2$ represents the predicted probability distribution over the two classes: *True* for coreferential and *False* for non-coreferential.

3.2 LLM-as-Judge Coreference Consistency Evaluation

Recent studies (Zhang et al., 2023a; Liu et al., 2024) have shown that strong LLMs correlate well with human judgments. Motivated by these findings, we adopt two high-performing models, GPT-4 (gpt-4-turbo-2024-04-09) (OpenAI, 2023) and Qwen-Plus (Yang et al., 2024)⁷, referred to as GPT-4-Judge and Qwen-Plus-Judge, as automatic evaluators for coreference consistency.

In this evaluation, all source-side mentions are explicitly marked, and an LLM is prompted to directly assign a score between 0 and 100 for each translated mention, based on whether the coreference relation is correctly preserved, the referent is clearly identifiable in the target, and the translation is natural and unambiguous. To facilitate interpretation, we group the continuous scores into five categories: *Perfect* (75–100), *Good* (50–74), *Acceptable* (25–49), *Poor* (1–24), and *Unacceptable* (0). The final LLM-as-Judge score is computed as the average of all mention-level scores. The full prompt is provided in Appendix E.

4 Experimentation

4.1 Experimental Settings

Classifier Training. We construct a training set of 31,304 instances, balanced between positive and negative examples. Our classifier is based on Llama-3.1-8B-base⁸ (Meta, 2024). Fine-tuning details are provide in Appendix C.

Translation Models. We evaluate two categories of translation models:

- **Specialized Translation Models:** These models are specifically designed for multilingual translation, including NLLB-200-3.3B⁹ (NLLB Team et al., 2022), MADLAD-400-3B¹⁰ (Kudugunta et al.,

2023), Qwen-mt-lite¹¹ (Team, 2025), and Hunyuan-translation¹² (Zheng et al., 2025).

- **LLMs:** This category includes both general-purpose pretrained LLMs and LLMs fine-tuned on our translation dataset. All models are based on decoder-only transformer architectures and trained with a causal language modeling objective. The open-weights models include Qwen2.5-7B/14B/72B-Instruct¹³ (Yang et al., 2024), and Llama-3-8B-Instruct¹⁴. The closed-source models include GPT-3.5-Turbo (OpenAI, 2023) and GPT-4o-mini (OpenAI, 2024). Some of these open-weights models are fine-tuned on our constructed training set; fine-tuning details are provided in Appendix C.

For all models, we perform document-to-document translation, i.e., translating the entire dialogue in a single pass.¹⁵ During translation, source-side zero pronouns are not explicitly provided to the models. The prompt used for the LLMs is presented in Figure 5 of Appendix D.

Evaluation metrics. Besides CCS for coreference consistency evaluation, we also report regular translation metrics, including document-level BLEU (d-BLEU) (Papineni et al., 2002) and document-level COMET (d-COMET) (Vernikos et al., 2022). Specifically, we apply reference-based metric wmt22-comet-da¹⁶ (Rei et al., 2022) for computing d-COMET.

Additionally, we introduce *Mention Translation Rate* (MTR), a new metric that measures the proportion of source-side coreferential mentions that are translated. Importantly, MTR does not consider whether the translation is correct or faithful; it only checks whether a mention is translated at all, i.e., aligned via word alignment to one or more

⁷<https://help.aliyun.com/zh/model-studio/developer-reference/what-is-qwen-llm>

⁸<https://huggingface.co/meta-llama/Llama-3.1-8B>

⁹<https://huggingface.co/facebook/nllb-200-3.3B>

¹⁰<https://huggingface.co/google/madlad400-3b-mt>

¹¹<https://help.aliyun.com/zh/model-studio/qwen-mt-api>

¹²<https://huggingface.co/tencent/Hunyuan-MT-7B>

¹³<https://huggingface.co/Qwen/Qwen2.5-7B-Instruct>

¹⁴<https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct>

¹⁵Since NLLB does not support translating entire dialogues in a single pass, we divide each dialogue into segments of three consecutive utterances and translate each segment separately.

¹⁶<https://huggingface.co/Unbabel/wmt22-comet-da>

Accuracy	Recall	Precision	F1
96.75	98.41	95.25	96.86

Table 2: Average coreference classification performance on the test set, obtained by repeating negative instance sampling three times.

tokens in the target dialogue. MTR therefore complements CCS by capturing whether coreferential mentions, including zero pronouns, appear in the output, independently of their correctness.

4.2 Experimental Results

Classification Performance. We first evaluate our coreference classifier on its ability to determine whether the translations of two source-side mentions are coreferential. Following the procedure described in Section 3.1.1, we construct a test set consisting of positive instances (i.e., mention translation pairs from the same coreference chain) and an equal number of randomly sampled negative instances (i.e., pairs from different chains). To mitigate variance introduced by negative sampling, we repeat the sampling process three times and report the averaged results. Evaluation metrics include overall accuracy, as well as precision, recall, and F1 score for the positive (*coreferential*) class. As shown in Table 2, the coreference classifier achieves 96.86% in F1 score, indicating strong performance and suggesting that the classifier is reliable for computing CCS.

Translation Performance. Table 3 presents the translation performance of different models on the test set. High MTR scores indicate that most source-side mentions in chains are translated in the target dialogue, regardless of whether they are translated correctly or faithfully. Based on these results, we make the following observations:

- **Human translations preserve coreference much better than machine translations.** The evaluation results across coreference consistency metrics, including CCS, GPT-4-Judge, and Qwen-Plus-Judge, reveal a significant gap between human translations (i.e., reference) and those generated by LLMs. For instance, the reference translations achieve a high CCS score of 92.74%, indicating that human translators are highly effective at preserving coreference relationships across languages. In contrast, all evaluated LLMs yield

CCS scores below 50.00%, suggesting that nearly half of the coreferential mention pairs in the source language are not maintained in the translations. This discrepancy underscores the challenges translation systems still face in handling cross-lingual discourse phenomena such as coreference.

- **Traditional MT metrics do not fully capture coreference consistency.** Although d-BLEU and d-COMET generally correlate with translation quality, they are less sensitive to discourse-level coreference preservation. For example, GPT-4o-mini and GPT-3.5-turbo differ by only 1.84/2.30 COMET/BLUE points, but their CCS/GPT-4-Judge/Qwen-Plus-Judge scores differ by 10.80/20.24/7.93 points. These results suggest that dedicated discourse-aware metrics such as CCS and LLM-as-judge evaluation are necessary for assessing cross-lingual coreference consistency.
- **Fine-tuning LLMs on dialogue translation dataset improves coreference preservation.** Models fine-tuned on our dialogue dataset consistently achieve higher scores across evaluation metrics. For instance, fine-tuning Qwen2.5-14B-Instruct increases the CCS score from 35.35% to 37.13% and the d-COMET score from 0.8811 to 0.8927, showing that task-specific adaptation enhances coreference consistency.

4.3 Discussion

Impact of including zero pronouns on CCS. Zero pronouns account for approximately one-quarter of all mentions in the source-side coreference chains. To better understand how well coreference consistency is preserved in translation, we compare CCS scores across three categories: all mention pairs, pairs excluding zero pronouns, and pairs that include at least one zero pronoun. As shown in Table 4, the CCS score for pairs involving zero pronouns is significantly lower than that for pairs excluding them. This result indicates that maintaining coreference consistency for zero pronouns is substantially more challenging than for overt mentions. We attribute this performance gap to the implicit nature of zero pronouns, which are not explicitly marked in the source text and thus are more likely to be overlooked or misinterpreted during translation.

Model	d-BLEU	d-COMET	MTR	CCS	GPT-4-Judge	Qwen-Plus-Judge
- (Reference)	100.00	96.55	94.22	92.74	87.03	89.84
Specialized Translation Models						
NLLB-200-3.3B	21.00	79.51	89.37	29.33	44.43	53.76
MADLAD-400-3B	21.78	81.67	81.96	30.91	50.07	55.58
Qwen-mt-lite	34.42	87.92	84.27	35.28	48.46	74.62
Hunyuan-translation	33.13	87.59	82.61	34.97	44.74	72.89
LLMs						
Llama-3-8B-Instruct	33.92	86.96	83.45	35.42	50.26	72.32
+ fine-tuned	44.79	89.14	88.01	37.72	56.36	77.37
Qwen2.5-7B-Instruct	38.66	87.72	82.68	35.05	46.79	73.79
+ fine-tuned	44.51	89.11	86.19	37.54	51.49	75.49
Qwen2.5-14B-Instruct	39.54	88.11	83.10	35.35	48.51	76.87
+ fine-tuned	45.62	89.27	83.49	37.13	52.48	77.35
Qwen2.5-72B-Instruct	<u>46.03</u>	<u>89.12</u>	93.29	<u>43.88</u>	<u>56.62</u>	78.98
GPT-3.5-turbo	44.01	89.03	88.27	37.33	37.36	71.01
GPT-4o-mini	46.31	90.87	<u>92.27</u>	48.13	57.60	<u>78.94</u>

Table 3: Performance comparison on the test set. The best and second best results are highlighted in **bold** and underlined, respectively. GPT-4-Judge and Qwen-Plus-Judge denote LLM-as-judge evaluation scores.

Model	All Mentions	Overt Only	ZPs Only
- (Reference)	92.74	95.35	89.01
NLLB-200-3.3B	29.33	31.98	25.54
Qwen-mt-lite	35.28	38.74	30.25
Hunyuan-translation	34.97	38.83	29.27
Llama-3-8B-Instruct	37.72	40.21	34.16
Qwen2.5-7B-Instruct	37.54	40.18	33.77
Qwen2.5-14B-Instruct	37.13	41.40	31.03
Qwen2.5-72B-Instruct	<u>43.88</u>	<u>47.19</u>	<u>39.15</u>
GPT-4o-mini	48.13	51.11	43.87

Table 4: CCS comparison across different mention pair categories: *All Mentions* considers all mention pairs; *Overt Only* includes only mention pairs excluding zero pronouns (ZP); *ZPs Only* includes only mention pairs where at least one mention is a zero pronoun.

Model	Paradigm	BLEU	COMET	CCS
MADLAD-400-3B	S2S	21.13	79.87	29.67
	CA	-	-	-
	D2D	21.78	81.67	30.91
Qwen2.5-72B-Inst	S2S	40.41	87.92	37.00
	CA	44.90	88.45	39.25
	D2D	<u>46.03</u>	<u>89.12</u>	<u>43.88</u>
GPT-4o-mini	S2S	34.64	87.30	32.57
	CA	37.43	88.17	34.39
	D2D	46.31	90.87	48.13

Table 5: Performance comparison across various translation paradigms including Sent2Sent (S2S), Context-Aware (CA) and Doc2Doc (D2D).

Comparison of different translation paradigms. Document-level translation can be achieved through various paradigms, including sentence-to-sentence (S2S) translation, which translates each utterance independently; context-aware (CA) translation, which incorporates preceding source-side

context; and document-to-document (D2D) translation, which translates the entire dialogue in a single pass and serves as the default setting in this work. The prompts for all paradigms are provided in Appendix F.

Table 5 compares model performance across these paradigms. Performance consistently improves from S2S to CA and D2D across all models. Although BLEU and COMET benefit from additional context, the largest gains are observed in CCS, highlighting the importance of discourse context for preserving coreference consistency in dialogue translation.

Case study. Figure 2 presents three examples where the MT system (Llama-3-8B-Instruct) fails to preserve source-side coreference, resulting in incorrect or misleading entity references.

In the first example, the system translates 车/*che* (*vehicle*) as *bus*, failing to link it with G160次列车/*G160_ci_lie_che* (*G160 Train*) mentioned later. This breaks the coreference chain and leads to an inconsistent interpretation. In the second example, the system fails to recognize that 他家/*ta_jia* (*his home*) refers to the previously mentioned 酒店/*jie_dian* (*hotel*), and thus translates it literally as *his home*, distorting the discourse context. The third example involves a zero pronoun: the omitted pronoun 它/*ta* (*it*) refers to 这个航班/*zhe_ge_hang_ban* (*this flight*), but the system incorrectly translates it as *I*. These errors disrupt discourse coherence and reduce translation reliability.

Case 1 [Source] usr: 你在嘛? 下周一深圳去杭州的车/che有吗? [Context] sys: 有的, 下周一上午有一辆G1604次列车/G1604_ci_lie_che, 你是要上午走吗? [Yes, there's a G1604 train in the morning. Are you planning to leave in the morning?] [Auto Translation] usr: Are you there? Is there any bus from Shenzhen to Hangzhou next Monday? [Reference] usr: Are you there? Is there a train going from Shenzhen to Hangzhou next Monday?
Case 2 [Context] usr: 酒店/jiu_dian里面有标准间吧? [Do they have standard rooms?] sys: 标准间当然有了,那是他们/ta_men的标配。[Of course, they have standard rooms, which are their standard offerings.] [Source] usr: 那就好, 他家/ta_jia地址在哪里? [Auto Translation] usr: That's good. Where is his home address? [Reference] usr: That's good. Where is their address?
Case 3 [Context] sys: 这个航班/zhe_ge_hang_ban起飞时间是中午十一时五十五分从高崎国际机场起飞。[The departure time for this flight is 11:55 AM from Gaoqi International Airport.] [Source] usr: 好的, (它/ta)不能延误吧? [Auto Translation] usr: Okay, won't I be late? [Reference] usr: Alright, it won't be delayed, will it?

Figure 2: Case studies illustrating the MT system's failure to resolve source-side coreference, resulting in incorrect and misleading translations.

5 Related Work

5.1 Discourse-Driven Machine Translation

Discourse phenomena remain a major challenge in machine translation (Hardmeier, 2012; Voita et al., 2018). A substantial body of research focuses on pronoun translation, particularly in morphologically rich and pro-drop languages. Resources such as the PROTEST test suite (Guillou and Hardmeier, 2016) and the WMT cross-lingual pronoun prediction shared tasks (Guillou et al., 2016) systematically evaluate pronoun handling in morphologically

rich languages. In parallel, zero-pronoun translation and recovery are also extensively studied for pro-drop languages (Wang et al., 2023). Another line of research focuses on improving lexical translation consistency, ensuring that source terms are translated consistently across a document (Guillou, 2013; Lyu et al., 2021; Lei et al., 2024).

More recently, several studies have explored the use of coreference information in machine translation. Ohtani et al. (2019) and Vu et al. (2024) incorporate source-side coreference information into context-aware neural MT to better model discourse relations. Closely related to our work, Yehudai et al. (2023) investigate the coreference capabilities of MT systems and propose methods for improving coreference-aware translation. Unlike Yehudai et al. (2023), which mainly focuses on sentence-level coreference, our work provides a systematic evaluation framework for measuring how well complete source-side coreference chains are preserved in dialogue translation.

5.2 Discourse-related Evaluation Metrics

Traditional MT evaluation metrics such as BLEU (Papineni et al., 2002) and COMET (Rei et al., 2022) mainly assess sentence-level translation quality and are less sensitive to discourse-level phenomena. To address this limitation, several discourse-aware metrics have been proposed. APT (Werlen and Popescu-Belis, 2017) evaluates pronoun translation accuracy, while LTCR (Lyu et al., 2021) measures lexical translation consistency across documents. DiscoScore (Zhao et al., 2023) evaluates discourse coherence using contextual representations, whereas BLONDE (Jiang et al., 2022) and TCP (Sun et al., 2022) provide broader discourse-level evaluation covering phenomena such as tense consistency, conjunction usage, and pronoun translation. In addition, several studies develop discourse-oriented test suites for evaluating MT systems on specific discourse phenomena (Voita et al., 2018, 2019; Lei et al., 2024). Unlike previous work, we introduce CCS, a metric specifically designed to measure whether source-side coreference relations are consistently preserved in dialogue translation.

6 Conclusion

This work presents a systematic evaluation of coreference consistency in Chinese-to-English dialogue translation. We construct a dialogue dataset anno-

tated with zero pronouns and coreference chains, and propose the *Coreference Consistency Score* (CCS) to automatically measure whether source-side coreferential relations are preserved in translation. We further complement CCS with an LLM-as-judge evaluation to assess discourse-level translation quality. Experimental results show a substantial gap between human translations and both specialized MT systems and state-of-the-art LLMs. Although most systems translate coreferential mentions on the surface, they frequently fail to maintain consistent entity references across dialogue turns, particularly in cases involving zero pronouns. These findings highlight that coreference consistency remains a challenging discourse-level problem for current translation models.

Limitations

This study has several limitations.

First, our analysis is restricted to Chinese-to-English dialogue translation, with a particular focus on task-oriented conversations. While this setting is well suited for studying discourse-level phenomena such as coreference and zero pronouns, the findings may not directly generalize to other language pairs, dialogue genres, or non-dialogue text.

Second, our annotation and evaluation pipeline relies on automatic word alignment, specifically awesome-align. The multilingual aligner is used both to construct training data for the mention-pair classifier and to compute the Mention Translation Rate (MTR), while alignment errors may also affect the proposed Coreference Consistency Score (CCS). Thus, the results may depend on the choice and configuration of the aligner, and evaluating alternative alignment methods or settings is an important direction for future work. Although we observe strong classifier performance, CCS should be interpreted as an approximate, rather than perfect, measure of coreference preservation.

Third, CCS assumes that source-side coreferential relations should be preserved in the target translation whenever possible. In practice, some translations may legitimately alter discourse structure or resolve references differently while remaining acceptable to human readers, which CCS may penalize.

Fourth, our LLM-as-judge evaluation, while motivated by prior work showing strong correlation with human judgments, is still an automatic assessment and may reflect model-specific biases. Incorporating human evaluation remains an important

direction for future work.

Finally, although the reference translations were produced by professional bilingual translators, they were not translated by native English speakers, which may affect the naturalness of some target-side expressions.

Acknowledgements

The authors would like to thank the anonymous reviewers for their constructive feedback. We use AI assistants solely to polish writing and check syntax errors.

References

- Zi-Yi Dou and Graham Neubig. 2021. [Word alignment by fine-tuning embeddings on parallel corpora](#). In *Proceedings of EACL*, pages 2112–2128.
- Liane Guillou. 2013. [Analysing lexical consistency in translation](#). In *Proceedings of DiscoMT*, pages 10–18.
- Liane Guillou and Christian Hardmeier. 2016. [PROTEST: A test suite for evaluating pronouns in machine translation](#). In *Proceedings of LREC*, pages 636–643.
- Liane Guillou, Christian Hardmeier, Preslav Nakov, Sara Stymne, Jörg Tiedemann, Yannick Versley, Mauro Cettolo, Bonnie Webber, and Andrei Popescu-Belis. 2016. [Findings of the 2016 WMT shared task on cross-lingual pronoun prediction](#). In *Proceedings of WMT: Shared Task Papers*, pages 525–542.
- Christian Hardmeier. 2012. [Discourse in statistical machine translation: A survey and a case study](#). *Discours: Revue de linguistique, psycholinguistique et informatique*, pages 525–542.
- Han He and Jinho D. Choi. 2021. [The stem cell hypothesis: Dilemma behind multi-task learning with transformer encoders](#). In *Proceedings of EMNLP*, pages 5555–5577.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [Lora: Low-rank adaptation of large language models](#). In *Proceedings of ICLR*.
- Yuchen Jiang, Tianyu Liu, Shuming Ma, Dongdong Zhang, Jian Yang, Haoyang Huang, Rico Sennrich, Ryan Cotterell, Mrinmaya Sachan, and Ming Zhou. 2022. [BlonDe: An automatic evaluation metric for document-level machine translation](#). In *Proceedings of NAACL: HLT*, pages 1550–1565.
- Sopan Khosla, Juntao Yu, Ramesh Manuvinaurike, Vincent Ng, Massimo Poesio, Michael Strube, and Carolyn Rosé. 2021. [The CODI-CRAC 2021 shared](#)

- task on anaphora, bridging, and discourse deixis in dialogue. In *Proceedings of the CODI-CRAC 2021 Shared Task on Anaphora, Bridging, and Discourse Deixis in Dialogue*, pages 1–15.
- Sneha Kudugunta, Isaac Rayburn Caswell, Biao Zhang, Xavier Garcia, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. *MADLAD-400: A multilingual and document-level large audited dataset*. In *Proceedings of NeurIPS*, pages 67284–67296.
- J. Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.
- Xiangyu Lei, Junhui Li, Shimin Tao, and Hao Yang. 2024. *Evaluation dataset for lexical translation consistency in Chinese-to-English document-level translation*. In *Proceedings of LREC-COLING*, pages 6575–6581.
- Yikun Lei, Yuqi Ren, and Deyi Xiong. 2022. *CoDoNMT: Modeling cohesion devices for document-level neural machine translation*. In *Proceedings of COLING*, pages 5205–5216.
- Yuxuan Liu, Tianchi Yang, Shaohan Huang, Zihan Zhang, Haizhen Huang, Furu Wei, Weiwei Deng, Feng Sun, and Qi Zhang. 2024. *Calibrating llm-based evaluator*. In *Proceedings of LREC-COLING*, pages 2638–2656.
- Xinglin Lyu, Junhui Li, Zhengxian Gong, and Min Zhang. 2021. *Encouraging lexical translation consistency for document-level neural machine translation*. In *Proceedings of EMNLP*, pages 3265–3277.
- Meta. 2024. Introducing meta llama 3: The most capable openly available llm to date. <https://ai.meta.com/blog/meta-llama-3/>.
- Vincent Ng. 2010. *Supervised noun phrase coreference research: The first fifteen years*. In *Proceedings of ACL*, pages 1396–1411.
- NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Hefernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barraud, Gabriel Mejia-Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. *No language left behind: Scaling human-centered machine translation*. *CoRR*, abs/2207.04672.
- Takumi Ohtani, Hidetaka Kamigaito, Masaaki Nagata, and Manabu Okumura. 2019. *Context-aware neural machine translation with coreference information*. In *Proceedings of DiscoMT*, pages 45–50.
- OpenAI. 2023. *Gpt-3.5 turbo*.
- OpenAI. 2023. *Gpt-4 technical report*. *CoRR*, abs/2303.08774.
- OpenAI. 2024. *Gpt-4o mini: Advancing cost-efficient intelligence*.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. *Bleu: a method for automatic evaluation of machine translation*. In *Proceedings of ACL*, pages 311–318.
- Sameer Pradhan, Lance Ramshaw, Mitch Marcus, Martha Palmer, Ralph Weischedel, and Nianwen Xue. 2011. *CoNLL-2011 shared task: Modeling unrestricted coreference in ontonotes*. In *Proceedings of CoNLL: shared task*, pages 1–27.
- Jun Quan, Shian Zhang, Qian Cao, Zizhong Li, and Deyi Xiong. 2020. *RiSAWOZ: A large-scale multi-domain Wizard-of-Oz dataset with rich semantic annotations for task-oriented dialogue modeling*. In *Proceedings of EMNLP*, pages 930–940.
- Sudha Rao, Allyson Ettinger, Hal Daumé III, and Philip Resnik. 2015. *Dialogue focus tracking for zero pronoun resolution*. In *Proceedings of NAACL*, pages 494–503.
- Ricardo Rei, José GC De Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André FT Martins. 2022. *Comet-22: Unbabel-ist 2022 submission for the metrics shared task*. In *Proceedings of WMT*, pages 578–585.
- Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. 2001. *A machine learning approach to coreference resolution of noun phrases*. *Computational Linguistics*, 27:521–544.
- Zewei Sun, Mingxuan Wang, Hao Zhou, Chengqi Zhao, Shujian Huang, Jiajun Chen, and Lei Li. 2022. *Re-thinking document-level neural machine translation*. In *Findings of ACL*, pages 3537–3548.
- Qwen Team. 2025. *Qwen-mt: Where speed meets smart translation*. <https://qwenlm.github.io/blog/qwen-mt/>.
- Giorgos Vernikos, Brian Thompson, Prashant Mathur, and Marcello Federico. 2022. *Embarrassingly easy document-level MT metrics: How to convert any pretrained metric into a document-level metric*. In *Proceedings of WMT*, pages 118–128.
- Elena Voita, Rico Sennrich, and Ivan Titov. 2019. *When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion*. In *Proceedings of ACL*, pages 1198–1212.
- Elena Voita, Pavel Serdyukov, Rico Sennrich, and Ivan Titov. 2018. *Context-aware neural machine translation learns anaphora resolution*. In *Proceedings of ACL*, pages 1264–1274.

- Huy Hien Vu, Hidetaka Kamigaito, and Taro Watanabe. 2024. [Context-aware machine translation with source coreference explanation](#). *Transactions of the Association for Computational Linguistics*, 12:856–874.
- Longyue Wang, Siyou Liu, Mingzhou Xu, Linfeng Song, Shuming Shi, and Zhaopeng Tu. 2023. [A survey on zero pronoun translation](#). In *Proceedings of ACL*, pages 3325–3339.
- Lucie M. Werlen and Andrei Popescu-Belis. 2017. [Validation of an automatic metric for the accuracy of pronoun translation \(apt\)](#). In *Proceedings of DiscoMT*, pages 17–25.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, et al. 2024. [Qwen2.5 technical report](#). *CoRR*, abs/2412.15115.
- Jing Yang, Sujian Li, Shan Gao, Zihan Li, and Wen Zhang. 2022. [Corefdpr: A joint model for coreference resolution and dropped pronoun recovery in chinese conversations](#). *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:571–581.
- Asaf Yehudai, Arie Cattan, Omri Abend, and Gabriel Stanovsky. 2023. [Evaluating and improving the coreference capabilities of machine translation models](#). In *Proceedings of EACL*, pages 980–992.
- Xintong Yu, Hongming Zhang, Yangqiu Song, Yan Song, and Changshui Zhang. 2019. [What you see is what you get: Visual pronoun coreference resolution in dialogues](#). In *Proceedings of EMNLP-IJCNLP*, pages 5123–5132.
- Xinghua Zhang, Bowen Yu, Haiyang Yu, Yangyu Lv, Tingwen Liu, Fei Huang, Hongbo Xu, and Yongbin Li. 2023a. [Wider and deeper llm networks are fairer llm evaluators](#). *CoRR*, abs/2308.01862.
- Zhen Zhang, Junhui Li, Shimin Tao, and Hao Yang. 2023b. [Lexical translation inconsistency-aware document-level translation repair](#). In *Findings of ACL*, pages 12492–12505.
- Shanheng Zhao and Hwee Tou Ng. 2007. [Identification and resolution of Chinese zero pronouns: A machine learning approach](#). In *Proceedings of EMNLP-CoNLL*, pages 541–550.
- Wei Zhao, Michael Strube, and Steffen Eger. 2023. [DiscoScore: Evaluating text generation with BERT and discourse coherence](#). In *Proceedings of EACL*, pages 3865–3883.
- Boyuan Zheng, Patrick Xia, Mahsa Yarmohammadi, and Benjamin Van Durme. 2023. [Multilingual coreference resolution in multiparty dialogue](#). *Transactions of the Association for Computational Linguistics*, 11:922–940.
- Mao Zheng, Zheng Li, Bingxin Qu, Mingyang Song, Yang Du, Mingrui Sun, and Di Wang. 2025. [Hunyuan-mt technical report](#). *CoRR*, abs/2509.05209.
- Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyang Luo. 2024. [LlamaFactory: Unified efficient fine-tuning of 100+ language models](#). In *Proceedings of ACL: System Demonstrations*, pages 400–410.

A Annotation Agreement

A.1 Cohen’s Kappa Coefficient

Cohen’s Kappa (κ) is a statistical measure used to evaluate inter-annotator agreement while correcting for chance agreement. It is defined as:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \quad (3)$$

where:

- p_o is the observed agreement between annotators,
- p_e is the expected agreement by chance.

A κ value of 1 indicates perfect agreement, whereas a value of 0 indicates agreement equivalent to chance. According to Landis and Koch (1977), κ values between 0.61 and 0.80 are interpreted as substantial agreement.

A.2 Zero Pronoun Annotation Agreement

To assess the reliability of zero pronoun (ZP) annotation, we compute inter-annotator agreement using Cohen’s Kappa (κ) between two annotators. Agreement is calculated on the overlapping subset of dialogues annotated by both annotators.

The average κ score for ZP annotation is 0.638, which falls into the substantial agreement category. This result indicates that, despite the inherent ambiguity of zero anaphora in Chinese, the annotation guidelines and training procedure enable a high degree of consistency across annotators.

A.3 Coreference Chain Annotation Agreement

We further evaluate inter-annotator agreement for coreference chain annotation using Cohen’s Kappa. This task requires annotators to identify and link all mentions referring to the same entity throughout a dialogue and is inherently more challenging due to longer-range dependencies and contextual ambiguity.

On the overlapping subset, the average κ score for coreference chain annotation is 0.694, which also corresponds to substantial agreement. This level of agreement demonstrates that annotators are able to consistently identify coreference relations, even for a complex discourse-level annotation task, and supports the overall quality of the annotated test set. In addition, we use a dedicated web-based annotation tool (Figure 3), which helps reduce annotation complexity and ambiguity.

B A Dialogue Example

Figure 4 presents a Chinese dialogue example alongside its English translation, with recovered zero pronouns and four annotated coreference chains.

C Fine-Tuning Model Setting

We fine-tune the classifier on Llama-3.1-8B-base. We employ LoRA (Hu et al., 2022) using rank 8, scaling factor $\alpha = 16$, and 0.0 dropout. Training is performed on two NVIDIA 3090 GPUs with a per-device batch size of 1, and gradients accumulation over 32 steps. The learning rate is set to 1e-4, and the model is trained for 6 epochs to mitigate the risk of overfitting.

During fine-tuning for dialogue translation, we also employ LoRA (Hu et al., 2022) fine-tuning approach based on the LLaMA-Factory framework¹⁷ (Zheng et al., 2024), with supervised fine-tuning (SFT) as the training stage. The LoRA target modules are set to all tunable modules to enable comprehensive parameter adaptation, with a rank of 8, scaling factor $\alpha = 16$, and 0.0 dropout. The training data is from the training set of RiSAWOZ, augmented with target-side translation from GPT-4o. Training is conducted in a multi-NVIDIA 3090 GPUs environment with a per-device batch size of 2 and gradient accumulation over 8 steps. The initial learning rate is set to 1e-4 with a warmup ratio of 0.1. We utilize the AdamW optimizer and a cosine learning rate scheduler. The fine-tuning process lasts for 4 epochs, with mixed precision training (FP16) enabled to improve computational efficiency. Other training parameters follow the default settings of the framework. The distributed training timeout is set to a large value to ensure stability during multi-GPU training.

In inference, all non-API translation experiments were conducted using the LLaMA-Factory framework with num_beams=3, do_sample=True, max_new_tokens=1024, temperature=0.95, top_p=0.7, and top_k=50, while both repetition_penalty and length_penalty were set to 1.0. For models accessed through official APIs, we used the default or provider-supported decoding settings.

¹⁷<https://github.com/hiyouga/LLaMA-Factory>

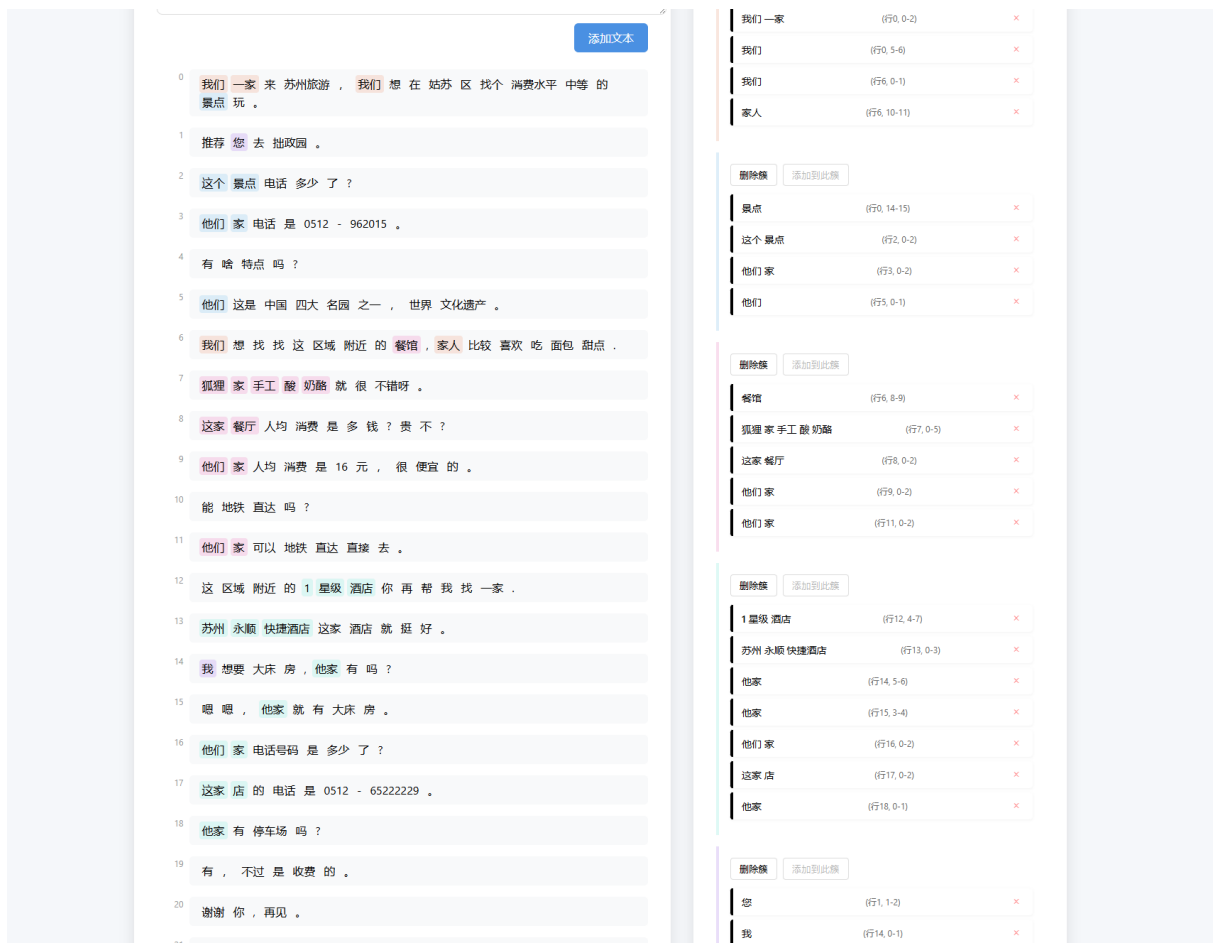


Figure 3: Web-based annotation tool used for dataset construction and coreference annotation.

D Prompts Used in Data Construction

Figure 5 presents the prompt used to obtain target-side translations under the document-to-document (D2D) translation paradigm, which is also adopted in our experimental evaluation. Figure 6 shows the prompt used for zero pronoun annotation.

E Prompt for LLM-as-Judge Evaluation

Figure 7 presents the prompt used for LLM-as-Judge evaluation.

F Prompts for Translation

Figure 8, Figure 9, and Figure 5 present the prompts used for sentence-to-sentence translation, context-aware translation, and document-to-document translation, respectively.

A Chinese Dialogue Example

usr: 你 好, [你] 可以帮 我 查下最近北京的天气情况吗? [Hello, could you help me check the recent weather conditions in Beijing?]

sys: 好的, 您 要查哪天的天气? [Certainly, which day would you like to check?]

usr: [你] 先帮 我 看看下周五的, [我] 下周五可能要出差。 [Please first check next Friday for me; I might have a business trip then.]

sys: 下周五北京天气多云转晴, 还是比较适合出行的。 [Next Friday, it will be cloudy in Beijing, turning to sunny, still good for travel.]

usr: 风沙 大不大? [Is there heavy sand and dust?]

sys: [它] 还可以, 西北风3-4级, 这个季节就是这样, 呵呵。 [The sand and dust aren't too bad, just some northwest winds at force 3-4. Typical for this season, you know.]

usr: 嗯, 是啊, 没办法。对了, [你] 能再帮 我 看看紫外线强度怎样? [Yeah, that's just how it is. By the way, could you check the UV index for me?]

sys: 紫外线强度还是有点强的, 您 出门注意做好防晒。 [The UV index is still pretty high. Make sure to protect yourself from the sun if you're heading out!]

usr: 嗯。 [你] 能再帮 我 看看 航班 吗? [Hmm. Can you help me check the flights as well?]

sys: 可以的。 您 要查哪天的? [Sure. Which day are you looking for?]

usr: 下周五的。 [Next Friday.]

sys: 下周五到北京的吗? [你] 从哪里出发? [Are you flying to Beijing next Friday? Where are you departing from?]

usr: 是的, 从成都到北京的, [我] 要直飞的, 不要中转那种的, 太麻烦。 [Yes, I'm flying from Chengdu to Beijing and would like a direct flight; no layovers, too troublesome.]

sys: 好的, [我] 给 您 查到下周五 成都直飞北京的航班 有: 四川航空3U8885。 [Got it. There's a direct flight from Chengdu to Beijing next Friday: Sichuan Airlines 3U8885.]

usr: 几点的? [What's the departure time?]

sys: 11:30从双流国际机场起飞。 [11:30 AM from Chengdu Shuangliu International Airport.]

usr: 那 [它] 到北京的时间呢? [When does it arrive in Beijing?]

sys: 不延误的情况下, [它] 到北京首都国际机场是当天下午两点二十。 [If there are no delays, it lands at Beijing Capital International Airport at 2:20 PM on the same day.]

usr: 哦, 这航班 准点率怎样? [What's this flight's on-time performance?]

sys: 准点率96.61%。 [The on-time performance rate is 96.61%.]

usr: 哦, 公务舱还有票吧? 我 要订公务舱。 [Are there still business-class seats? I would like to book business class.]

sys: 目前还有的。不过 [你] 需要早点预定, 最近机票挺紧张的, 旅游旺季。 [Yes, there are. But, you should book soon, as tickets are selling fast due to peak travel season.]

usr: 好的, 谢谢 你。 [我] 一会就定。拜拜! [Okay, thank you very much. I will book it later. Goodbye!]

sys: 不客气, 再见! [You're welcome! Goodbye!]

Figure 4: An example of a Chinese dialogue with its English translation. Recovered zero pronouns are [underwaved], and the four coreference chains are highlighted in red, blue, yellow, and green, respectively.

Prompt for document-to-document translation

Please translate the following dialogue from Chinese to English.
This dialogue is an exchange between two speakers, <usr> and <sys>.
The translated dialogue results must conform to the grammar and format of the English dialogue.
Please focus on the translation of coreference.
Please provide a translation that preserves the original dialogue style and tone.
[dialogue]:
usr: [utterance 1]
sys: [utterance 2]
.....
usr: [utterance n]
sys: [utterance n+1]

Figure 5: Prompt templates used for document-to-document translation.

Prompt for zero-pronoun annotation

usr: [.....<s>mention</s>.....]
sys: [utterance 2]
.....
usr: [utterance n]
sys: [.....<s>mention</s>.....]

Please perform zero pronoun resolution on the dialogue following the style shown in the previous examples.

Enclose the recovered mentions with <s></s> tags.

The following resolution rules must be strictly followed:

1. Only the following ten specific pronouns are restored: 我 (I), 我们 (we), 你 (you), 你们 (you plural), 他 (he), 他们 (they, masculine/mixed), 她 (she), 她们 (they, feminine), 它 (it), 它们 (they, inanimate). Abstract pronouns (e.g., those referring to events, expletive subjects, or generic entities) are ignored.
2. Restoration is applied to omitted subjects or objects. If an explicit subject appears later in the sentence, restoration may be omitted. Omitted possessive forms are not restored.
3. For colloquial utterances, the decision to restore is based on sentence fluency: fully fluent utterances are restored, partially fluent ones are partially restored, and non-fluent ones are not restored.
4. Regarding the restoration position, insertion at the beginning of the sentence is preferred. In combinations of colloquial and standard sentences, zero anaphora in the standard sentence may be restored, while the colloquial sentence shares the restored result.

Please return the dialogue with the annotations.

[dialogue]:

usr: [utterance 1]
sys: [utterance 2]
.....
usr: [utterance n]
sys: [utterance n+1]

Figure 6: Prompt templates used for zero-pronoun annotation.

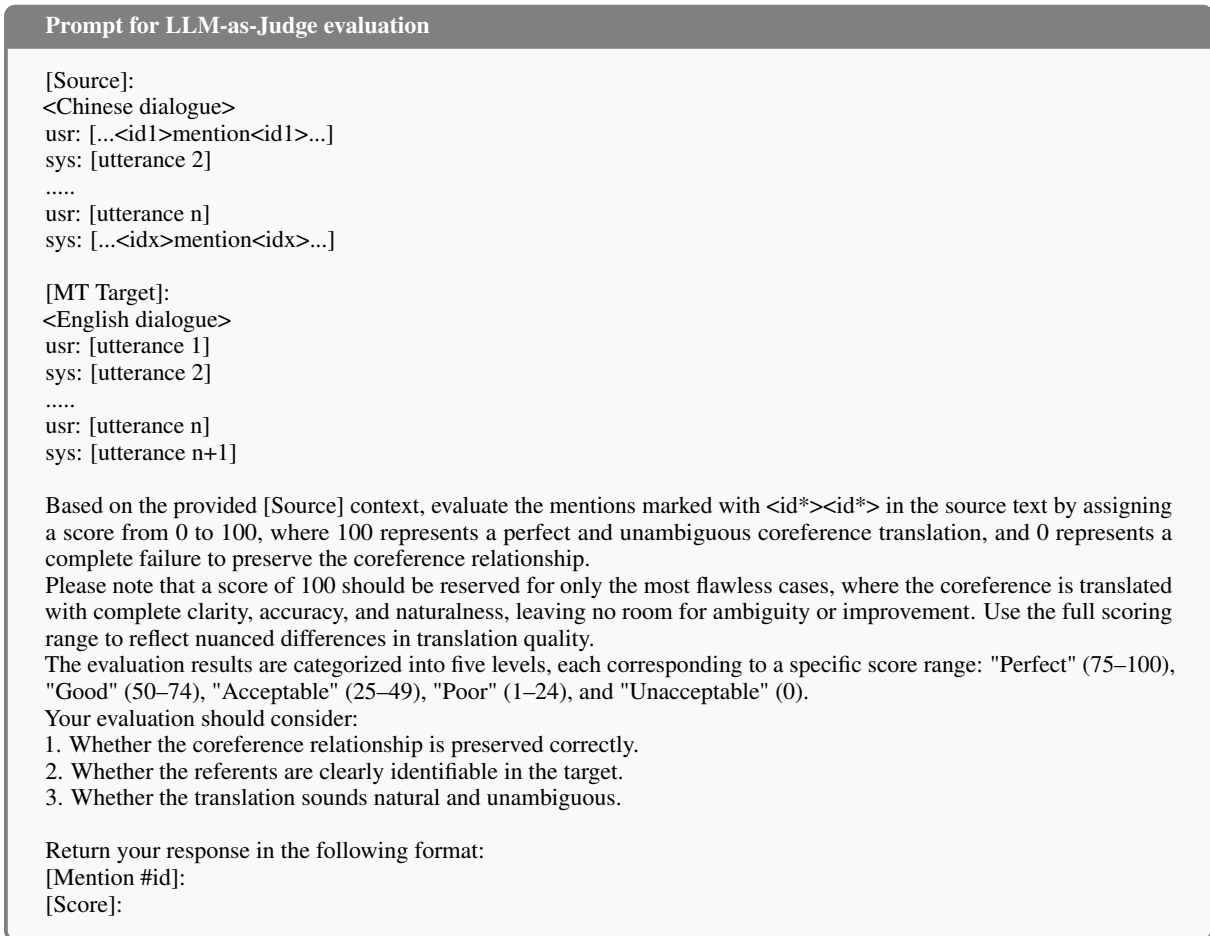


Figure 7: Prompt template used for LLM-as-Judge evaluation.

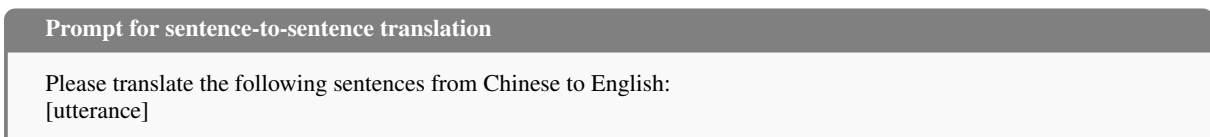


Figure 8: Prompt template for sentence-to-sentence translation.

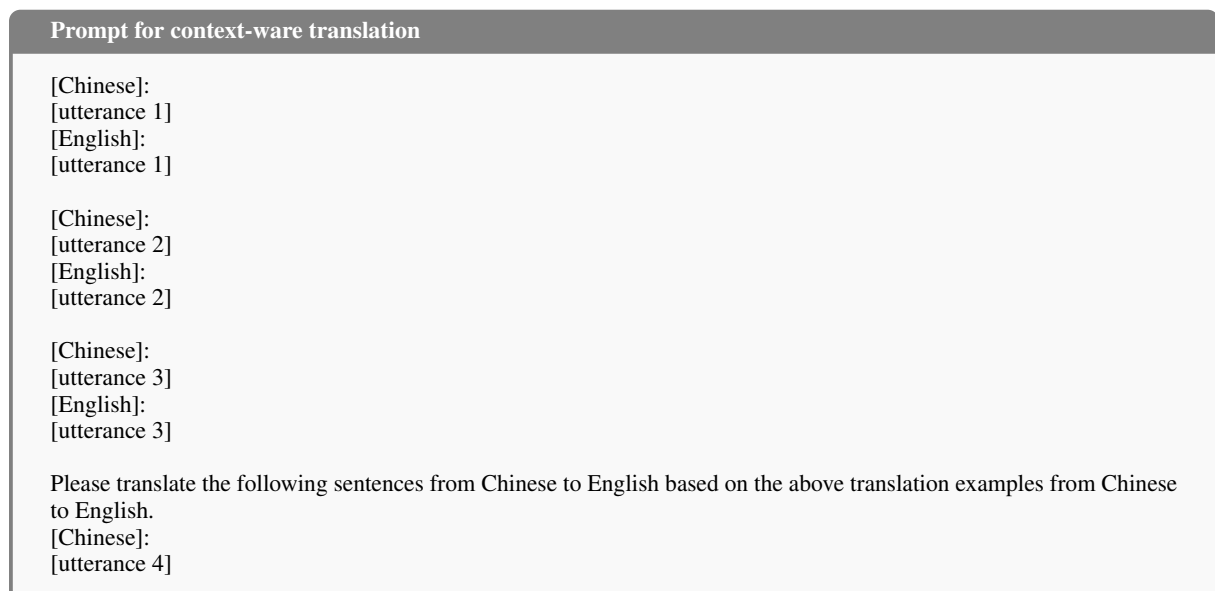


Figure 9: Prompt template used for context-aware translation. Specifically, we use three previous source-side utterances as context.