Evaluating Coreference Consistency in Chinese-to-English Dialogue Translation

Anonymous ACL submission

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. All code and data will be released.

1 Introduction

In recent years, cross-lingual understanding has gained increasing attention in natural language processing (NLP), particularly with the growing demand for machine translation systems that preserve not only surface meaning but also deeper discourse-level phenomena (Lyu et al., 2021; Lei et al., 2022; Zhang et al., 2023b; Vu et al., 2024). A key component of discourse coherence is coreference, where multiple expressions refer to the same entity (Soon et al., 2001). Coreference allows speakers to avoid repetition and maintain continuity, which is essential for the flow and interpretability of conversations. While coreference resolution has been extensively studied in monolingual settings (Yu et al., 2019; Khosla et al., 2021; Zheng et al., 2023), little is known about how coreference relations are preserved or disrupted when translating dialogues across languages.[1] Compared with other translation tasks, such as news translation, dialogue translation presents distinct challenges due to its informal style, frequent speaker turns, and reliance on contextual coherence.

[Context]
usr: 坐 高铁/gao_tie 的话有哪个车？
sys: 有 G90/G90，成都东-北京西。
[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。
……
usr: 这附近有什么热门的学校校园吗？
sys: 苏州大学/su_zhou_da_xue 就很热门。
usr: 能地铁直达吗？
[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.

Figure 1 illustrates two examples of Chinese-

[1] In this paper, we define dialogue translation as translating an entire dialogue from the source language into the target language. This differs from cross-lingual chat translation (Farajian et al., 2020), where speakers use different languages within the same conversation.

to-English dialogue translation generated by `LLaMA3-8B-Instruct` (Meta, 2024), highlighting typical coreference inconsistencies. In the first example, the source expressions *G90* and 它*/ta* refer to the same *train* entity. However, the model correctly translates *G90* while incorrectly rendering 它*/ta* as the first-person plural pronoun *we*, altering the intended meaning and misleadingly suggesting that the speaker is planning a trip rather than inquiring about the train. The challenge is further pronounced in Chinese dialogue due to the pervasive use of zero pronouns. In the second example, a dropped pronoun refers to 苏州大学*/su_zhou_da_xue*. The model fails to recover this implicit reference and instead translates it as *the hotel*, introducing a contextually incorrect entity. These examples highlight a broader issue: despite producing fluent translations, current MT systems, including advanced LLM-based models, often struggle to preserve discourse-level coreference, resulting in degraded coherence and distorted meaning in translated dialogues.

In this paper, we systematically investigate coreference preservation in Chinese-to-English dialogue translation. We conduct both quantitative and qualitative analyses to assess how well current machine translation systems maintain source-side coreference relations, highlighting the limitations of even advanced LLMs.[2]

Our main contributions are as follows:

- We construct a high-quality Chinese-to-English dialogue translation dataset with 10,000, 600, and 600 dialogues for the training, development, and test sets, respectively. Source dialogues are annotated with zero pronouns and coreference chains, with all test annotations manually verified.
- We propose a novel automatic metric CCS for evaluating coreference consistency, along with an LLM-as-judge framework for assessing the faithfulness of translated mentions.
- We provide a comprehensive evaluation of diverse LLMs, revealing persistent challenges in maintaining coreference consistency in dialogue translation.

[2]We note that some acceptable translations may legitimately restructure discourse or resolve references differently while remaining natural to human readers. Our evaluation focuses on the preservation of source-side coreference relations rather than treating all such deviations as errors.

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 (Section 2.3). The source dialogues are derived from RiSAWOZ [3] (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. All utterances are segmented using HanLP toolkit [4] 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, 2024a) 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 3 (Appendix C).

Translation of Named Entities. Named entities are crucial for preserving coreference chains. To ensure the correctness of entity translations, two trained master's students manually verify all entity translations.

Test Set Translation. The test set is translated by a professional human translator.[5] Given the conversational style of the source dialogues, the translator is instructed to produce English translations that

[3]https://github.com/terryqj0107/RiSAWOZ

[4]https://github.com/hankcs/HanLP

[5]The translator is a native Chinese speaker, certified at the highest level in both interpretation and translation by CATTI, and has passed TEM-8.

are both natural and orally fluent, while faithfully preserving the original meaning.

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.

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 to ensure that recovered sentences are natural and grammatically correct, and (3) strategies for handlling 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 4 (Appendix C).

Test Set Annotation. Two native Chinese-speaking master's students 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. Inter-annotation agreement statistics are reported in Appendix 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. We first generate coreference chians automatically using the HanLP coreference resolution module (He and Choi, 2021),[6] To ensure quality, To ensure quality, 100 dialogues are randomly sampled and manually reviewed by two graduate-level annotators, with incorrect links corrected by removing erroneous mentions.

Test Set Annotation. For the test set, coreference chains are annotated entirely manually using a web-based annotation tool[7] that allows annotators to visualize dialogues and link coreferent mentions. The tool outputs structured coreference chains, reducing both annotation time and cognitive load.

Two native Chinese-speaking master's students perform the 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. Inter-annotator agreement statistics are reported in 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.

The data statistics also show that the distributions of zero pronouns, coreference chains, and annotated mentions in the test set closely match those of the training and development sets. This alignment ensures that evaluation results are reliable, even though the test set is fully manually annotated, whereas the training and development sets are automatically annotated with subsequent manual refinement.

[6] https://www.hanlp.com/semantics/functionapi/participle

[7] https://anonymized/

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 $C = \{c_1, \cdots, c_m\}$ 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 $C = \{c_1, \ldots, c_m\}$ in a source dialogue X, we first obtain the corresponding translations $t = \{t_1, \ldots, t_m\}$ in the target dialogue Y via word alignment between X and Y.[8] Each t_i may consist of zero, one, or multiple words. CCS for a single chain C is defined as:

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

where the denominator $\binom{m}{2}$ is the total number of unique mention pairs in C, and $\text{R}\left(t_i, t_j\right)$ 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`.

[8]We 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.

3.1.1 Coreferenre Consistency Classifier

As in Eq. 1, computing CCS requires identifying whether the translations of two source-side mentions are coreferential in the target dialgoue. To this end, we construct a classifier that predicts coreference relations based on contextual representations of translated mentions.

Training Instances. For each coreference chain $C = \{c_1, \cdots, c_m\}$ in a parallel dialogue pair (X, Y) from the training set, we collect the translations of all mentions, denoted as $T = \{t_1, \cdots, t_m\}$. 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`. 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:

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

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 (Wang et al., 2023; 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[9], 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 assess the quality of each metion translation using a five-level scale: *Perfect, Good, Acceptable, Poor, and Unacceptable*. These qualitative labels are mapped to numerical scores as follows: 75-100 for *Perfect*, 50-74 for *Good*, 25-49 for *Acceptable*, 1-24 for *Poor*, and 0 for *Unacceptable*, respectively. The final LLM-as-Judge score is computed as the average of the numerical ratings across all evaluated mentions. The full prompt used in this procedure is provided in Appendix D.

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-base`[10] (Meta, 2024). Fine-tuning details are provide in Appendix B.

Translation Models. We evaluate two categories of translation models:

- **Specialized Translation Models:** These models are specifically designed for multilingual translation. We use `NLLB-200-3.3B`[11] (NLLB Team et al., 2024) and `MADLAD-400-3B`[12] (Kudugunta et al., 2023).

- **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-source models include `Qwen2.5-7B/14B/72B-Instruct`[13] (Yang et al., 2024), and `Llama-3-8B-Instruct`[14]. The closed-source models include `GPT-3.5-Turbo` (Meta, 2022) and `GPT-4o-mini` (OpenAI, 2024b). Some of these open-source models are fine-tuned on our constructed training set; fine-tuning details are provided in Appendix B.

For all models, we perform document-to-document translation, i.e., translating the entire dialogue in a single pass.[15] The prompt used for the LLMs is presented in Figure 3 of Appendix C.

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-da`[16] (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 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).

[9] https://help.aliyun.com/zh/model-studio/developer-reference/what-is-qwen-llm

[10] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

[11] https://huggingface.co/facebook/nllb-200-3.3B

[12] https://huggingface.co/google/madlad400-3b-mt

[13] https://huggingface.co/Qwen/Qwen2.5-7B-Instruct

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

[15] Since 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 classification performance on the test set, obtained by repeating negative instance sampling three times.

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.

- **CCS correlates more strongly with LLM-as-judge metrics than with traditional translation metrics.** While d-BLEU and d-COMET provide general translation quality measures, they are less sensitive to coreference consistency. In contrast, the CCS metric exhibits stronger agreement with the GPT-4-Judge and Qwen-Plus-Judge scores, suggesting that LLM-as-judge approaches, guided by carefully designed prompts, are more suitable for evaluating discourse-level phenomena like coreference.

- **Fine-tuning LLMs 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.

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 translates each utterance with access to preceding source-side context; and document-to-document (D2D) translation, which translates the entire dialogue in a single pass, and is adopted as the default setting in this work. We provide the prompts for the three paradigms in Appendix E.

Table 5 compares model performance across these paradigms. We observe consistent improvements as we move from S2S to CA and D2D translation across all models. While BLEU and COMET scores benefit from additional context, the most significant gains are observed in CCS, underscoring the importance of both source-side and target-side context for maintaining coherence in dialogue translation.

Model	d-BLEU	d-COMET	MTR	CCS	GPT-4-Judge	Qwen-Plus-Judge
- (Reference)	100.00	100.00	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
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	46.03	89.12	**93.29**	43.88	56.62	**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**	92.27	**48.13**	**57.60**	78.94

Table 3: Performance comparison on the test set. The best and second best results are highlighted in **bold** and underlined, respectively.

Model	All Mentions	Overt Only	ZPs Only
- (Reference)	92.74	95.35	89.01
NLLB-200-3.3B	29.33	31.98	25.54
GPT-4o-mini	**48.13**	**51.11**	**43.87**
Qwen2.5-72B-Inst	43.88	47.19	39.15
Llama-3-8B-base$_{\text{tuned}}$	37.83	40.39	34.17
Llama-3-8B-Inst$_{\text{tuned}}$	37.72	40.21	34.16
Qwen2.5-7B-Inst$_{\text{tuned}}$	37.54	40.18	33.77
Qwen2.5-14B-Inst$_{\text{tuned}}$	37.13	41.40	31.03

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
GPT-4o-mini	S2S	34.64	87.30	32.57
	CA	37.43	88.17	34.39
	D2D	**46.31**	**90.87**	**48.13**
Qwen2.5-72B-Inst	S2S	40.41	87.92	37.00
	CA	44.90	88.45	39.25
	D2D	46.03	89.12	43.88

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

Case study. Figure 2 presents three illustrative cases in which the MT system (Llama-3-8B-Instruct) fails to correctly resolve source-side coreference. As a result, the generated translations misrepresent the underlying entity relations and may introduce confusion for readers.

In the first example, the system translates 车*/che* as *bus*, failing to establish its coreferential relation with *G160*次列车*/G160_ci_lie_che* in the subsequent sentence, which clearly denotes a specific train service. This error breaks the coreference chain and results in an inconsistent interpretation of the entity. In the second example, the system similarly fails to recognize that 他家*/ta_jia* corefers to 酒店*/jiu_dian* in preceding context. Lacking this contextual understanding, the model renders 他家*/ta_jia* literally as *his home*, rather than the intended reference to the hotel, theyby distorting the discourse context. The third example illustrates an error involving zero pronoun. In the source dialogue, the omitted pronoun 它*/ta* refers to the previously mentioned entity 这个航班*/zhe_ge_hang_ban*. However, the system fials to recover this referential link and incorrectly translates it as *I*. Such errors not only result in incorrect entity references but also disrupt discourse coherence, ultimately reducing the readability and reliability of the translated dialogue.

5 Related Work

5.1 Dialogue Translation

OpenSubtitles (Lison and Tiedemann, 2016)[17] is one of the most widely used datasets for dialogue translation. However, many previous studies (Voita et al., 2018; Miculicich et al., 2018) primarily treat it as a document-level translation resource, often overlooking valuable dialogue-specific information such as utterance boundaries and speaker annotations. BConTrasT, introduced in the WMT2020 Chat Translation Task (Farajian et al., 2020), is an English-to-German dataset specifically designed for dialogue translation. However, it focuses

[17] https://www.opensubtitles.org/

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？
[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？
sys: 标准间当然有了,那是他们/ta_men的标配。
[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点五十五分从高崎国际机场起飞。
[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.

on translating bilingual conversations from one language to another, providing a more dialogue-oriented benchmark. Building on this dataset, subsequent works have explored a variety of dialogue-specific features, including speaker role preferences, discourse coherence, translation consistency, and speaker identification (Liang et al., 2021a,b, 2022; Zhou et al., 2023).

5.2 Discourse-related Evaluation Metrics

While traditional evaluation metrics such as BLEU (Papineni et al., 2002) and COMET (Rei et al., 2022) perform reliably at the sentence level, they fall short in capturing cross-sentential discourse phenomena, which are crucial for document-level translation. To address this limitation, several discourse-aware evaluation metrics have been proposed. The APT metric (Accuracy of Pronoun Translation) is widely used to specifically assess pronoun translation accuracy (Werlen and Popescu-Belis, 2017). Lyu et al. (2021) introduce LTCR (Lexical Translation Consistency Rate) to measure how consistently words and phrases from the source document are translated across the entire target document. Zhao et al. (2023) propose DiscoScore, a parameterized discourse metric that uses BERT to evaluate discourse coherence from multiple perspectives. Differently, Jiang et al. (2022) propose BLONDE, a comprehensive metric that evaluates various discourse phenomena such as named entity consistency, tense consistency, and pronoun ellipsis. Similarly, Sun et al. (2022) introduce TCP, another comprehensive metric that focuses on tense consistency, conjunction usage, and pronoun translation. Besides, related studies build different test suits from different discourse category test MT system. Finally, previous studies have also developed various test suites specifically designed to evaluate the ability of machine translation systems to handle different discourse phenomena (Voita et al., 2018, 2019; Lei et al., 2024). Unlike the studies mentioned above, this paper introduces CSC, a metric designed to measure coreference consistency between the source document and the target translation, an aspect that has not been explicitly explored before.

6 Conclusion

This work presents a systematic evaluation of coreference consistency in Chinese-to-English dialogue translation. We construct a dialogue dataset annotated 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, the proposed Coreference Consistency Score (CCS) relies on automatic word alignment and a learned coreference consistency classifier. Errors in alignment or classification may propagate to the final CCS values, potentially affecting evaluation accuracy. 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. Finally, 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 large-scale human evaluation remains an important direction for future work.

References

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. In *Proceedings of EACL*, pages 2112–2128.

M. Amin Farajian, António V. Lopes, André F. T. Martins, Sameen Maruf, and Gholamreza Haffari. 2020. Findings of the WMT 2020 shared task on chat translation. In *Proceedings of WMT*, pages 65–75.

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. 2021. 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 Manuvinakurike, 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.

Yunlong Liang, Fandong Meng, Yufeng Chen, Jinan Xu, and Jie Zhou. 2021a. Modeling bilingual conversational characteristics for neural chat translation. In *Proceedings of ACL-IJCNLP*.

Yunlong Liang, Fandong Meng, Jinan Xu, Yufeng Chen, and Jie Zhou. 2022. Scheduled multi-task learning for neural chat translation. In *Proceedings of ACL*.

Yunlong Liang, Chulun Zhou, Fandong Meng, Jinan Xu, Yufeng Chen, Jinsong Su, and Jie Zhou. 2021b. Towards making the most of dialogue characteristics for neural chat translation. In *Proceedings of EMNLP*.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of LREC*, pages 923–929.

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. 2022. Chatgpt: Optimizing language models for dialogue. `https://openai.com/blog/chatgpt`.

Meta. 2024. Introducing meta llama 3: The most capable openly available llm to date. `https://ai.meta.com/blog/meta-llama-3/`.

Lesly Miculicich, Dhananjay Ram, Nikolaos Pappas, and James Henderson. 2018. Document-level neural machine translation with hierarchical attention networks. In *Proceedings of EMNLP*, pages 2947–2954.

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 Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, 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. 2024. No language left behind: Scaling human-centered machine translation. *CoRR*, abs/2207.04672.

OpenAI. 2023. Gpt-4 technical report. *CoRR*, abs/2303.08774.

OpenAI. 2024a. Gpt-4o mini: advancing cost-efficient intelligence. `https://openai.com/research/gpt-4`.

OpenAI. 2024b. Gpt-4o mini: advancing cost-efficient intelligence. `https://openai.com/research/gpt-4`.

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. Rethinking document-level neural machine translation. In *Findings of ACL*, pages 3537–3548.

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.

Cunxiang Wang, Sirui Cheng, Qipeng Guo, Yuanhao Yue, Bowen Ding, Zhikun Xu, Yidong Wang, Xiangkun Hu, Zheng Zhang, and Yue Zhang. 2023. Evaluating open-qa evaluation. In *Proceedings of NeurIPs*, pages 77013–77042.

Lucie M. Werlen and Andrei Popescu-Belis. 2017. Validation of an automatic metric for the accuracy of pronoun translation (apt). In *Proceedings of the Third Workshop on Discourse in Machine Translation*, 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.

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.

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.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyan Luo. 2024. LlamaFactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of ACL: System Demonstrations*, pages 400–410.

Chulun Zhou, Yunlong Liang, Fandong Meng, Jie Zhou, Jinan Xu, Hongji Wang, Min Zhang, and Jinsong Su. 2023. A multi-task multi-stage transitional training framework for neural chat translation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45:7970–7985.

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:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{3}$$

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. In addition, the use of a dedicated web-based annotation tool helps reduce annotation complexity and ambiguity, further contributing to reliable annotations.

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.

B Fine-Tuning Model Setting

We fine-tune the classifier on `Llama-3.1-8B-base`. We employ LoRA (Hu et al., 2021) 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., 2021) fine-tuning approach based on the LLaMA-Factory framework[18] (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.

C Prompts Used in Data Construction

Figure 3 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 4 shows the prompt used for zero pronoun annotation.

D Prompt for LLM-as-Judge Evaluation

Figure 5 presents the prompt used for LLM-as-Judge evaluation.

E Prompts for Translation

Figure 6, Figure 7, and Figure 3 present the prompts used for sentence-to-sentence translation, context-aware translation, and document-to-document translation, respectively.

[18] https://github.com/hiyouga/LLaMA-Factory

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.
Translation results 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 3: Prompt templates used for document-to-document translatoin.

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 4: Prompt templates used for zero-pronoun annotation.

Prompt for LLM-as-Judge evaluation

[Source]:
<Chinese dialogue>
usr: [...<id1>mention<id1>...]
sys: [utterance 2]
.....
usr: [utterance n]
sys: [...<idx>mention<idx>...]

[MT Target]:
<English dialogue>
usr: [utterance 1]
sys: [utterance 2]
.....
usr: [utterance n]
sys: [utterance n+1]

Based on the provided [Source] context, evaluate the mentions marked with <id*><id*> in the source text by assigning a score from 0 to 100, where 100 represents a perfect and unambiguous coreference translation, and 0 represents a complete failure to preserve the coreference relationship.
Please note that a score of 100 should be reserved for only the most flawless cases, where the coreference is translated with complete clarity, accuracy, and naturalness, leaving no room for ambiguity or improvement. Use the full scoring range to reflect nuanced differences in translation quality.
The evaluation results are categorized into five levels, each corresponding to a specific score range: "Perfect" (75–100), "Good" (50–74), "Acceptable" (25–49), "Poor" (1–24), and "Unacceptable" (0).
Your evaluation should consider:
1. Whether the coreference relationship is preserved correctly.
2. Whether the referents are clearly identifiable in the target.
3. Whether the translation sounds natural and unambiguous.

Return your response in the following format:
[Mention #id]:
[Score]:

Figure 5: Prompt template used for LLM-as-Judge evaluation.

Prompt for sentence-to-sentence translation

Please translate the following sentences from Chinese to English:
[utterance]

Figure 6: Prompt template for sentence-to-sentence translation.

Prompt for context-ware translation

[Chinese]:
[utterance 1]
[English]:
[utterance 1]

[Chinese]:
[utterance 2]
[English]:
[utterance 2]

[Chinese]:
[utterance 3]
[English]:
[utterance 3]

Please translate the following sentences from Chinese to English based on the above translation examples from Chinese to English.
[Chinese]:
[utterance 4]

Figure 7: Prompt template used for context-aware translation. Specifically, we use three previous source-side utterances as context.