Explanation for Machine Translation Errors: Generation and Evaluation

Anonymous ACL submission

Abstract

The fine-grained annotations of translation errors have been widely applied in machine translation researches such as translation quality estimation, designing automatic evaluation metrics, but these annotations only contain information such as error type, location, and severity, the reasons of the errors are not annotated. Since explaining why an annotated text span is erroneous is important for building the trustworthy machine translation models, we manually build the first resource for evaluating the quality of the explanation for the errors. We tested large language models (LLMs) on this evaluation resource, and found that LLMs failed to deliver trustworthy explanations for the machine translation errors. So, we propose a hard chain-of-thought (H-CoT) approach that induces the explanation for the errors step-by-step via hard chains. Experiments on the evaluation resource show that H-CoT greatly improves the explanation quality over LLMs without H-CoT.

1 Introduction

With the recent development of neural networks and large language models (LLMs), machine translation (MT) systems achieve steady progress in the translation quality. Although they perform well in certain circumstances, there still exist various type of errors that need further study. Multidimensional Quality Metrics (MQM)[1](Lommel et al., 2014a,b) is the fine-grained schema fit for translation error analysis. It contains explicit error annotation and has been successfully applied in researches of metrics task (Freitag et al., 2021a,b) and quality estimation (Zerva et al., 2022).

Despite its success, MQM annotation only includes information such as error type, location, and severity. There is no explanation for the translation errors, that is, the reason of why an annotated text span is an erroneous translation is not explained. This impedes the interpretability of current researches and the building of trustworthy MT models.

Source	为什么抄手有6 个?
Translation	Why are there 6 **<v>copiers</v>?**
Type	Accuracy/Mistranslation
Severity	Major
Reference	Why are there six **wontons**?
Explanation	There is a translation error in the target, **"**抄手**"** should be translated as **"wontons"**; so, change **"copiers"** to **"wontons"**.

Table 1: An example of the manually annotated explanation for the translation error type of mistranslation. Given the source and its translation, MQM schema annotates the erroneous text span tagged between <v> and </v> in the translation with its error type and severity. We annotate the explanation for this error. Reference is optional in the explanation generation approach.

In this paper, we manually build the first resource for evaluating the quality of the explanation for the translation errors, and propose an automatic explanation generation approach given the MQM annotations.

In the evaluation resource building, we define different explanation guidelines for different types of the translation errors annotated in MQM. Table 1 lists an example of the error type of mistranslation with its explanation. The explanation includes the source text span which is mistranslated, and the target translation into which the mistranslation should be corrected. For other error types, the explanation template is reformulated accordingly.

In the automatic explanation generation, we tested LLMs ability in explaining the translation errors on this evaluation resource, and found that LLMs failed in generating the trustworthy explanations. To solve this problem, we propose a hard chain-of-thought (H-CoT) approach that induces the explanation for LLMs through a hard chain of reasoning steps. Experiments on the built evaluation resource consisting of Chinese-to-English and

[1] https://themqm.org/error-types-2/typology/

English-to-German sets show that our H-CoT can effectively improves LLMs ability in explaining the translation errors with significant improvement. In summary, the contributions of our work are as follows:

- We manually build the first evaluation resource for evaluating approaches of the explanation generation for the translation errors.
- We tested LLMs on the evaluation resource and found that LLMs failed in generating trustworthy explanations for the translation errors.
- We propose H-CoT to effectively induce the explanation ability of LLMs, and the experimental results show the significant improvements achieved by H-CoT.

2 Related Work

Explanation for MT errors is based on MQM annotations. So we introduce the MQM schema at first, then we introduce the automatic explanation generation approaches.

2.1 MQM Schema

MQM schema was first introduced in Lommel et al. (2014a,b) as a measurement and analysis framework for MT errors. It is adopted in Freitag et al. (2021a,b) for the metrics task which examines how well an automatic evaluation metric for MT correlates with human judgements. They annotated the fine-grained errors according to the MQM schema, and found that these annotations are more trustworthy for the task. These annotations are subsequently used in the quality estimation task which estimates the quality of MT output without relying on reference translations (Zerva et al., 2022). Due to the success of MQM annotations, they are widely adopted in series of WMT evaluation campaigns, and the annotations are enriched to incorporate more translation results of WMT 2020-2023 submissions[2]. Despite the success of MQM annotations, they do not contain explanation for the translation errors, which hampers the building of trustworthy MT models or LLMs. We create the manual explanation resource based on MQM annotations in this paper.

[2] https://github.com/google/wmt-mqm-human-evaluation

2.2 Explanation Generation

Nowadays, explainable natural language processing (NLP) gains more and more interests in trustworthy NLP models (Danilevsky et al., 2020; Lyu et al., 2024). A valid explanation can raise trust in the many NLP systems humans interact with daily (e.g., chatbots, machine translation engines, recommendation systems, and many others). It usually considers two themes in explainable NLP: building explanation resource and explanation generation approaches.

Regarding building explanation resource, Rajani et al. (2019); Aggarwal et al. (2021) manually build the explanation resource for commonsense QA. They annotated the explanations for commonsense reasoning, with the exception that Aggarwal et al. (2021) include positive and negative factors in the explanation. Chen et al. (2023) build the dataset of XplainLLM for understanding LLM decision making in QA. They integrated knowledge graph into the annotation process to make decision making more transparent and reliable. There is no explanation work for MT errors, while MT errors exhibit distinct properties such as bilingual scenario, different error types, and fine-grained error locations in a sentence, deserving careful explanation for them with a manually created resource.

Regarding explanation generation approaches, recent studies generally adopt the sequence-to-sequence or language model generation methods. Given the training set of explanation generation, the sequence-to-sequence models are trained/fine-tuned to use the commonsense statement/QA pair as the source to generate the reason as the target (Rajani et al., 2019; Jon et al., 2020; Wan and Huang, 2020; Aggarwal et al., 2021). In the language model generation approaches, they prompt the auto-regressive language models to generate explanations (Konar et al., 2020; Chen et al., 2023; Cheng et al., 2023; Xu et al., 2023). One study that is closely related to our work is NEON (Cheng et al., 2023), which finds conflict points in an error statement of commonsense, but it is still distinctive to the task of generating explanations for the translation errors in this paper. InstructScore is a tool to produce both an evaluation score for a generated text and a human readable diagnostic report (Xu et al., 2023). Explanation for the translation errors is a byproduct of InstructScore, but we found that their diagnostic report does not correlate well with the manually built MQM annotations, result-

	Chinese-to-English		English-to-German	
Error Type	Quantity	Proportion(%)	Quantity	Proportion(%)
Accuracy/Addition	21	1.17	5	0.52
Accuracy/Mistranslation	946	52.56	302	31.66
Accuracy/Omission	98	5.44	19	1.99
Accuracy/Source language fragment	15	0.83	42	4.40
Fluency/Grammar	207	11.50	132	13.84
Fluency/Inconsistency	25	1.39	17	1.78
Fluency/Punctuation	118	6.56	112	11.74
Fluency/Register	1	0.06	17	1.78
Fluency/Spelling	83	4.61	43	4.51
Fluency/Character encoding	0	0	2	0.21
Locale convention	10	0.56	3	0.31
Style/Awkward	252	14.00	236	24.74
Terminology	15	8.33	23	2.41
Non-translation	5	0.28	0	0
Source error	4	0.22	1	0.10
Severity	Quantity	Proportion(%)	Quantity	Proportion(%)
Major	953	52.94	206	21.59
Minor	847	47.06	748	78.41

Table 2: Statistics of the explanation resource for the translation errors. The explanations are manually annotated for all error types.

ing in unsatisfied explanation quality. We present the performance of InstructScore in section 6.2.

3 The Resource for Evaluating the Explanation for the MT Errors

We manually explain the MQM annotated errors to build the evaluation resource. MQM schema designs a hierarchy of translation error types and severity classes. Freitag et al. (2021a) refine the severity classes into two tags: Major and Minor, with deleting the original two tags: Neutral and Critical, due to their subjective nature.

We select MQM annotations on translation results submitted in WMT2022 Chinese-to-English and English-to-German general translation task to explain the errors. This annotation set contains translation results of 15 participated teams in Chinese-to-English task and 15 participated teams in English-to-German task. To avoid repetitive explanations on the same error, we do not annotate explanations for all MQM errors, and just uniformly select equal number of sentences for each participating team to annotate. Note that the sentences of each team do not overlap in the source side. In the end, we annotate 1.8K explanations for Chinese-to-English translation errors and 1.0K explanations for English-to-German translation errors, covering all source side sentences in the WMT2022 general translation test sets. Table 2 lists the statistics of the explanations for the error types and severity classes.

In particular, for each translation error type, we design an explanation template to formulate the reason of why a specific text span is a translation error and the manner of how to correct it. Table 3 lists the various templates for all error types. We have two professional annotators for each translation direction to fill in the slots in the templates. The annotation process starts by letting the annotators learn the guideline of the MQM errors (Freitag et al., 2021a), then the annotators annotate a small portion and send to each other to check the quality. In the last, the annotators begin the formal annotation, and check each other the quality through sampling. The process iterates until no problem can be found in the sampling.

4 Explanation Generation

We test the explanation generation ability of LLMs by directly using prompt, which consists of the task instruction, few-shot examples, and the input of an MQM error. LLMs include LLaMA2-7B and LLaMA2-13B (Touvron et al., 2023). The few-shot examples are selected in the full set of the released MQM errors. We manually annotate the explanations for the errors in these few-shot examples, and we check them not to overlap with our evaluation resource. The test is conducted in both the general case and the error type specific cases, respectively. Both cases show that the tested LLMs can not deliver the trustworthy explanations for the translation errors. The test results are presented in the experimental section 5.2.

In Comparison to the above one-step prompt

Error Type	Explanation Template
Accuracy/Addition	There is no information about [err] in the source, but it is included in the translation; so, delete [err].
Accuracy/Mistranslation	There is an error in the translation, [src] should be translated as [answer]; so, change [err] to [answer].
Accuracy/Omission	There is no translation for [src]; so, it should be translated as [answer] and added between [position] and [position].
Accuracy/Source language fragment	The translation of [src] in the source is wrong; so, change [err] to [answer].
Fluency/Grammar	There is a grammatical error in the translation; so, change [err] to [answer].
Fluency/Inconsistency	There is an inconsistency in the translation, [src] is translated as [err] in the missing context; so, change [err] to [answer].
Fluency/Punctuation	There is a punctuation error in the translation, [src] should be translated as [answer]; so, change [err] to [answer].
Fluency/Register	There is a grammatical error in the translation that does not fit the context; so, change [err] to [answer].
Fluency/Spelling	There is a spelling error in the translation, [err] should be spelled as [answer]; so, change [err] to [answer].
Fluency/Character encoding	There is a garbled character in the translation; so, change [err] to [answer].
Locale convention	There is a format error in the translation, [src] should be translated as [answer]; so, change [err] to [answer].
Style/Awkward	The style of the translation does not conform to language conventions; so, change [err] to [answer].
Terminology	There is a terminology in the translation that is inappropriate for context; so, change [err] to [answer].
Non-translation	It is impossible to reliably characterize distinct errors in the target; so, change [err] to [answer].
Source error	There is an error in the source.

Table 3: Explanation templates for the translation errors. Slots specified in [ ] are to be filled in per error. [src] denotes the source span that is erroneously translated, [err] denotes the erroneous span in the translation, [answer] denotes the correction of the error, and [position] denotes the precise position in the translation to insert the correction.

approach, we propose H-CoT based on the observation that the explanation for the MT errors has specific natures that can be decomposed into subtasks. Take the error type of mistranslation for example as shown in Table 1, the explanation can be decomposed into three subtasks: finding the source text span that is mistranslated, generating the correction for the mistranslation, and composing the final explanation according to the template shown in Table 3. H-CoT organizes these subtasks into a hard chain of thoughts with prefixed steps. It has the advantage that the explanation ability of LLMs can be induced step by step, better than generating explanation in one step directly. Table 4 lists the H-CoT steps for the example. Since different error types have different subtasks, we adapt the H-CoT steps accordingly. Note that the second step in Table 4 is based on reference, while in some applications reference is not always available. So we design a reference-free prompt for the second step shown in Table 5.

We also test running word alignment on parallel sentences of {source, reference} pair and {source, translation} pair to find text spans aligned to the errors, then compose the final explanation based on the found text spans. The results show that the word alignment method is significantly inferior to H-CoT. The comparison result is presented in section 6.1.

5 Experiment

We conduct the explanation generation experiment on our manually built Chinese-to-English and English-to-German evaluation resource.

5.1 Experimental Setting

We use LLMs of LLaMA2-7B and LLaMA2-13B for the explanation generation. We use greedy search with a length limit of 256. Four-shot examples are used in composing the prompt for all LLMs in Chinese-to-English task, and six-shot examples are used in English-to-German task. One step of direct prompting for the LLMs are set as our baselines. For our H-CoT approach, we use LLaMA2-13B as the basis model. We use A100 graphics cards to run the LLaMA2 models.

Evaluation Metrics. We evaluate the quality of the explanation through two categories of metrics: sentence-level metrics and span-level metrics. The sentence-level metrics include BLEU (Papineni

	Prompts/Actions
First step	You are a language assistant to find the phrase in the source that is aligned to the phase marked between <v> and </v> in the translation. Here are four examples: Source: 实时分享金融/财会/建筑最新考试资讯。 Translation: Share the latest <v>financial/financial</v>/building exam information in real time. The phrase 'financial/financial' in the translation is aligned to the phrase in the source: '金融/财会' ... Source: 为什么抄手有6 个? Translation: Why are there 6 <v>copiers</v>? The phrase 'copiers' in the translation is aligned to the phrase in the source:
Second step	You are a language assistant to find the phrase in the reference that is aligned to the phase marked between <v> and </v> in the translation. Here are four examples: Translation: Share the latest <v>financial/financial</v>/building exam information in real time. Reference: Real time sharing of the latest examination information of Finance/ Accounting/Architecture. The phrase 'financial/financial' in the translation is aligned to the phrase in the reference: 'Finance/ Accounting' ... Translation: Why are there 6 <v>copiers</v>? Reference: Why are there six wontons? The phrase 'copiers' in the translation is aligned to the phrase in the reference:
Third step	Compose the final explanation: There is an error in the translation, **"**抄手**"** should be translated as **"wontons"**; so, change **"copiers"** to **"wontons"**.

Table 4: H-CoT steps for the example of Table 1. The first step is to find the source span that is mistranslated. The second step is to find the correction of the mistranslation. The third step is to compose the final explanation based on the outputs of the previous two steps and the template.

You are a language assistant to correct the error in the translation that is marked between <v> and </v>. Here are four examples: Source: 实时分享金融/财会/建筑最新考试资讯。 Translation: Share the latest <v>financial/financial</v>/ building exam information in real time. The phrase 'financial/financial' should be corrected as: 'Finance/Accounting' ... Source: 为什么抄手有6个? Translation: Why are there 6 <v>copiers</v>? The phrase 'copiers' should be corrected as:

Table 5: Reference-free prompt of the second step to generate the correction of the mistranslation.

et al., 2002) and COMET (Rei et al., 2020) that are widely adopted in evaluating language generation tasks. They are computed against the manually annotated explanations. The span-level metrics include Source Accuracy (Src Acc) and Target Accuracy (Tgt Acc), where Src Acc computes the ratio of correctly finding the source span that are erroneously translated, and Tgt Acc computes the ratio of correctly generating the rectification of the error. Take the manual explanation in Table 1 for example, if "抄手" appears in the corresponding slot of a generated explanation, we deem it correct when counting Src Acc. If "wontons" appears in the corresponding slot, we deem it correct when counting Tgt Acc. All counting is based on matching of the whole span.

Besides the two span-level metrics, there is one special metric for the error type of omission whose explanation includes the inserting position of the omitted translation. We denote the inserting position accuracy as Insert Acc. Note that not all span-level metrics are suitable for all error types. For example, there is no Src Acc for grammar errors. We report them error-specifically.

5.2 Main Results

Results of the experiments with references. Table 6 reports the evaluation results when reference translations are available. LLaMA2-7B/13B$_{\text{OneStep}}$ are baseline methods using the one step prompt. It shows that H-CoT drastically surpasses the baselines across all error types and all language pairs. When we look at the error type of mistranslation, which takes up the largest portion of the translation errors as shown in Table 2, Src Acc and Tgt Acc are significantly improved by H-CoT, indicating that step-wise prompting LLMs with clear instruct to find text spans aligned to the errors is more effective than directly prompting LLMs to generate the explanation. The improved Src Acc and Tgt Acc also bring improvements in sentence level BLEU and COMET. This improvement trend generalizes to other error types across the language pairs. Regarding the two baselines, LLaMA2-13B performs better than LLaMA2-7B

	Accuracy/Mistranslation							
With Reference	Chinese-to-English				English-to-German			
	BLEU	COMET	Src Acc	Tgt Acc	BLEU	COMET	Src Acc	Tgt Acc
LLaMA2-7B$_{\text{OneStep}}$	62.30	72.91	23.47	14.69	74.47	76.74	20.86	37.08
LLaMA2-13B$_{\text{OneStep}}$	66.38	76.34	27.21	15.64	75.49	76.60	44.70	18.21
H-CoT	**76.09**	**79.23**	**43.13**	**42.49**	**88.14**	**85.32**	**63.91**	**58.29**
	Accuracy/Omission							
With Reference	Chinese-to-English				English-to-German			
	BLEU	COMET	Tgt Acc	Insert Acc	BLEU	COMET	Tgt Acc	Insert Acc
LLaMA2-7B$_{\text{OneStep}}$	58.83	74.29	20.41	**19.39**	62.79	70.82	21.05	5.26
LLaMA2-13B$_{\text{OneStep}}$	60.11	74.73	29.59	17.35	74.66	73.85	**36.84**	**31.58**
H-CoT	**77.46**	**79.95**	**39.80**	18.37	**76.00**	**75.70**	**36.84**	**31.58**
	Terminology							
With Reference	Chinese-to-English				English-to-German			
	BLEU	COMET	Src Acc	Tgt Acc	BLEU	COMET	Src Acc	Tgt Acc
LLaMA2-7B$_{\text{OneStep}}$	75.45	80.51	20.00	6.67	85.34	83.62	69.57	30.43
LLaMA2-13B$_{\text{OneStep}}$	79.99	82.42	26.67	26.67	86.58	82.04	**73.91**	47.83
H-CoT	**93.77**	**87.93**	**66.67**	**80.00**	**92.55**	**89.32**	**73.91**	**86.96**
	Style/Awkward							
With Reference	Chinese-to-English				English-to-German			
	BLEU	COMET	Tgt Acc	-	BLEU	COMET	Tgt Acc	-
LLaMA2-7B$_{\text{OneStep}}$	50.47	67.71	12.30		41.89	66.70	22.46	
LLaMA2-13B$_{\text{OneStep}}$	50.21	68.50	12.30		44.69	68.32	43.22	
H-CoT	**60.46**	**73.82**	**36.11**		**62.40**	**73.95**	**56.78**	
	Grammar							
With Reference	Chinese-to-English				English-to-German			
	BLEU	COMET	Tgt Acc	-	BLEU	COMET	Tgt Acc	-
LLaMA2-7B$_{\text{OneStep}}$	46.53	68.19	6.28		49.69	65.76	15.91	
LLaMA2-13B$_{\text{OneStep}}$	47.47	68.19	8.21		51.24	67.59	21.97	
H-CoT	**55.83**	**70.80**	**22.71**		**56.32**	**69.88**	**41.67**	
	Others							
With Reference	Chinese-to-English				English-to-German			
	BLEU	COMET	Src Acc	Tgt Acc	BLEU	COMET	Src Acc	Tgt Acc
LLaMA2-7B$_{\text{OneStep}}$	60.96	78.25	32.73	11.28	64.38	74.28	30.86	22.46
LLaMA2-13B$_{\text{OneStep}}$	64.00	79.54	41.82	14.79	60.13	74.15	48.15	21.61
H-CoT	**69.76**	**83.03**	**58.18**	**34.24**	**73.14**	**81.02**	**81.48**	**39.83**

Table 6: Performances of the explanation generation approaches under the condition that the references are available.

in most cases. Only in occasional cases such as Tgt Acc in English-to-German mistranslation, LLaMA2-7B leads LLaMA2-13B by a significant margin.

Regarding the error type of omission, Insert Acc is quite hard to improve. The omitted translation has difficulty in finding the right position to insert. There is no significant performance difference between the one step prompt baseline methods and H-CoT.

Results of the experiments without references. Table 7 reports the evaluation results when reference translations are not available. It shows that the condition of having no reference poses a great challenge in explanation generation compared to Table 6. For example, there are significant drops in performances of the explanation for the error type of mistranslation. BLEU drops from 76.09 to 65.35, COMET drops from 79.23 to 77.61, and Tgt Acc drops from 42.49 to 10.57 in Chinese-to-English task. Since the first step in H-CoT for mistranslation is the same for both with reference and without reference conditions, Src Acc remains unchanged. Despite the challenge, H-CoT still drastically outperforms baselines for all error types and language pairs, demonstrating the advantage of H-CoT in explanation generation.

Results summarized over all error types. Different to Table 6 and 7 that report error-specific performances, Table 8 reports the summarized performance. The summarization is in two ways: universal and concatenation. In the universal way, we use the same examples for all types of error[3]. In the concatenation way, we concatenate all error-specific explanations reported in Table 6 and 7. H-CoT in Table 8 is in the concatenation way. It shows that the concatenation way performs bet-

[3]The examples include two mistranslation examples, one/two omission examples, and one/two terminology examples for Chinese-to-English/English-to-German tasks.

	Accuracy/Mistranslation							
Without Reference	Chinese-to-English				English-to-German			
	BLEU	COMET	Src Acc	Tgt Acc	BLEU	COMET	Src Acc	Tgt Acc
LLaMA2-7B$_{\text{OneStep}}$	58.58	70.32	21.88	3.07	69.20	72.56	23.84	3.31
LLaMA2-13B$_{\text{OneStep}}$	61.99	74.19	37.42	3.59	71.00	73.52	44.04	2.98
H-CoT	**65.35**	**77.61**	**43.13**	**10.57**	**76.24**	**77.50**	**63.91**	**5.96**
	Accuracy/Omission							
Without Reference	Chinese-to-English				English-to-German			
	BLEU	COMET	Tgt Acc	Insert Acc	BLEU	COMET	Tgt Acc	Insert Acc
LLaMA2-7B$_{\text{OneStep}}$	55.96	73.23	12.24	15.31	65.06	72.17	10.53	10.53
LLaMA2-13B$_{\text{OneStep}}$	57.56	72.79	**14.29**	22.45	63.57	70.20	5.26	**26.32**
H-CoT	**59.02**	**73.88**	**14.29**	**25.51**	**68.13**	**74.11**	**10.53**	**26.32**
	Terminology							
Without Reference	Chinese-to-English				English-to-German			
	BLEU	COMET	Src Acc	Tgt Acc	BLEU	COMET	Src Acc	Tgt Acc
LLaMA2-7B$_{\text{OneStep}}$	72.68	78.77	26.67	0.00	84.78	80.16	69.57	13.04
LLaMA2-13B$_{\text{OneStep}}$	74.87	78.91	33.33	0.00	83.71	79.76	78.26	17.39
H-CoT	**77.51**	**81.86**	**66.67**	0.00	**83.96**	**82.43**	**78.26**	**26.09**
	Style/Awkward							
Without Reference	Chinese-to-English				English-to-German			
	BLEU	COMET	Tgt Acc	-	BLEU	COMET	Tgt Acc	-
LLaMA2-7B$_{\text{OneStep}}$	48.00	66.02	1.59		37.50	62.47	3.39	
LLaMA2-13B$_{\text{OneStep}}$	47.02	66.24	1.19		44.07	65.48	5.51	
H-CoT	**51.00**	**70.94**	**6.75**		**55.04**	**68.12**	**6.78**	
	Grammar							
Without Reference	Chinese-to-English				English-to-German			
	BLEU	COMET	Tgt Acc	-	BLEU	COMET	Tgt Acc	-
LLaMA2-7B$_{\text{OneStep}}$	46.33	68.00	4.35		47.75	64.12	3.03	
LLaMA2-13B$_{\text{OneStep}}$	46.54	68.13	4.35		49.39	66.25	**11.36**	
H-CoT	**53.40**	**70.07**	4.35		**52.05**	**67.52**	**11.36**	
	Others							
Without Reference	Chinese-to-English				English-to-German			
	BLEU	COMET	Src Acc	Tgt Acc	BLEU	COMET	Src Acc	Tgt Acc
LLaMA2-7B$_{\text{OneStep}}$	61.32	78.13	27.27	4.67	62.32	72.32	33.33	10.17
LLaMA2-13B$_{\text{OneStep}}$	62.15	78.90	41.81	12.84	59.98	72.96	56.79	8.05
H-CoT	**65.28**	**80.91**	**58.18**	**15.95**	**69.19**	**80.27**	**81.48**	**13.56**

Table 7: Performances of the explanation generation approaches under the condition that no references are available.

ter than the universal way in all cases. This manifests that the explanation generation should be conducted error-specifically. Finally, H-CoT significantly improves the overall performances for all error types no matter the reference translations are available or not.

6 Analysis

6.1 Comparison to Word Alignment Based Explanation Generation

Since most explanations contain information about the text spans that are aligned to the translation errors, we test using word alignment based approach to extract those information for composing the explanation. We use fastalign[4] to align the MQM data. To maintain the corpus statistically sound, we also include large parallel corpus from WMT2022 Chinese-English shared task into the data for running the alignment. Mistranslation, which is the major source of the errors in Chinese-to-English task, is experimented under the condition of having references.

Table 9 lists the results in this study. H-CoT shows the clear advantage over the word alignment approach. Even in the alignment related span level performances, H-CoT achieves much better accuracy, especially in Tgt Acc that measures the accuracy of the correction of the errors. This indicates that LLMs have the better capability in locating the spans aligned to the errors than the traditional word alignment tool fastalign.

6.2 Comparison to InstructScore

InstructScore (Xu et al., 2023) is capable of generating the explanations for the translation errors, but their errors are inconsistent with the MQM annotated translation errors due to their framework of using GPT4 (Kocmi and Federmann, 2023) and

[4] https://github.com/clab/fast_align

All Errors	Chinese-to-English		English-to-German	
With Reference	BLEU	COMET	BLEU	COMET
LLaMA2-7B$_{\text{OneStep}}$(Universal)	55.21	69.33	58.95	70.37
LLaMA2-7B$_{\text{OneStep}}$(Concatenate)	58.63	72.61	61.54	71.26
LLaMA2-13B$_{\text{OneStep}}$(Universal)	58.33	71.73	61.35	71.28
LLaMA2-13B$_{\text{OneStep}}$(Concatenate)	61.75	74.77	61.80	72.76
H-CoT	**70.87**	**78.21**	**74.78**	**79.19**
Without Reference	BLEU	COMET	BLEU	COMET
LLaMA2-7B$_{\text{OneStep}}$(Universal)	53.24	68.20	57.04	68.01
LLaMA2-7B$_{\text{OneStep}}$(Concatenate)	56.11	70.90	57.73	69.01
LLaMA2-13B$_{\text{OneStep}}$(Universal)	55.37	70.09	58.12	69.54
LLaMA2-13B$_{\text{OneStep}}$(Concatenate)	58.23	73.08	60.05	70.46
H-CoT	**61.62**	**76.16**	**65.65**	**74.55**

Table 8: The summarized performance of the explanation generation approaches for all error types.

	BLEU	COMET	Src Acc	Tgt Acc
Word Alignment	64.51	75.99	31.83	9.41
H-CoT	**76.09**	**79.23**	**43.13**	**42.49**

Table 9: The comparison between using word alignments and using H-CoT for the explanation generation. The comparison result is for the error type of mistranslation on the Chinese-to-English task.

	BLEU	COMET
InstructScore$_{\text{TP}}$	6.48	60.52
LLaMA2-13B$_{\text{TP}}$	62.36	74.86
H-CoT$_{\text{TP}}$	71.15	78.00
InstructScore$_{\text{TP+FP}}$	4.91	47.67
InstructScore$_{\text{TP+FN}}$	3.46	49.92

Table 10: Evaluation results for the explanations generated by InstructScore.

human instruction to build the synthetic data for finetuning LLaMA. The finetuned LLaMA evaluates the translation result automatically, resulting in the number and span of errors different to the manually labeled MQM errors.

So, we manually align the errors between InstructScore and MQM. We define the errors appearing in both InstructScore detections and MQM annotations as TP, the errors appearing in InstructScore detections but not in MQM annotations as FP, and the errors appearing in MQM annotations but not in InstructScore detections as FN. Take the MQM annotated errors as gold annotations, then the precision of InstructScore detections is: #TP / (#TP + #FP), and the recall is: #TP / (#TP + #FN). We compute the precision and recall of InstructScore in Chinese-to-English task under the condition of having references. The precision is 49.26%, and the recall is 66.39%, which demonstrate that InstructScore behaves distinctively to MQM annotations with few overlaps.

To check the quality of the explanations generated by InstructScore, we evaluate its explanation for TP. For FP, we set gold explanation as "This is not an error.". For FN, we set InstructScore explanation also as "This is not an error". Table 10 lists the evaluation results. In the evaluation on TP, we include LLaMA2-13B and H-CoT for comparison since gold explanation for TP data is consistent. It shows that InstructScore performs much worse than LLaMA2-13B and H-CoT on TP data. For the performances on TP+FP and TP+FN data, InstructScore performs unsatisfactorily since large portion of the errors are not consistent between InstructScore and MQM.

7 Conclusion

Fine-grained error annotations from MQM schema contain information such as error type, location, and severity, but they do not explain why an annotated text span is an erroneous translation. It is important to know the reasons of the translation errors in building the trustworthy MT or large language models. So, we study the problem of explaining the translation errors by building the evaluation resource and designing the explanation generation approach. In building the evaluation resource, we manually explain the reasons error-specifically. In designing the explanation generation approach, we test typical LLMs on the evaluation resource at first, and find that the LLMs failed in delivering the trustworthy explanations. To address this issue, we propose H-CoT to induce the explanation ability of the LLMs step-by-step. Experiments show that H-CoT can effectively enhance the explanation ability of the LLMs on the evaluation resource.

Limitations

The explanation generation experiments have not included more LLMs such as GPT4 for study. The explanation resource building and the corresponding experiments for other language pairs were not carried out. There is limitation on the coverage of both the types of LLMs and the language pairs. Besides, this paper does not involve error detection and rectification based on the explanation, we leave them as future research.

Ethics Statement

The data used in our work are freely downloadable from MQM annotations github (`https://github.com/google/wmt-mqm-human-evaluation`). The codes and models of LLaMA2 are freely downloadable from web.

References

Shourya Aggarwal, Divyanshu Mandowara, Vishwajeet Agrawal, Dinesh Khandelwal, Parag Singla, and Dinesh Garg. 2021. Explanations for CommonsenseQA: New Dataset and Models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3050–3065, Online. Association for Computational Linguistics.

Zichen Chen, Jianda Chen, Mitali Gaidhani, Ambuj Singh, and Misha Sra. 2023. Xplainllm: A qa explanation dataset for understanding llm decision-making.

Sijie Cheng, Zhiyong Wu, Jiangjie Chen, Zhixing Li, Yang Liu, and Lingpeng Kong. 2023. Unsupervised explanation generation via correct instantiations. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence*.

Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. 2020. A survey of the state of explainable AI for natural language processing. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 447–459, Suzhou, China. Association for Computational Linguistics.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021a. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. 2021b. Results of the WMT21 metrics shared task: Evaluating metrics with expert-based human evaluations on TED and news domain. In *Proceedings of the Sixth Conference on Machine Translation*, pages 733–774, Online. Association for Computational Linguistics.

Josef Jon, Martin Fajcik, Martin Docekal, and Pavel Smrz. 2020. BUT-FIT at SemEval-2020 task 4: Multilingual commonsense. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 374–390, Barcelona (online). International Committee for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.

Anandh Konar, Chenyang Huang, Amine Trabelsi, and Osmar Zaiane. 2020. ANA at SemEval-2020 task 4: MUlti-task learNIng for cOmmonsense reasoNing (UNION). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 367–373, Barcelona (online). International Committee for Computational Linguistics.

Arle Lommel, Aljoscha Burchardt, and Hans Uszkoreit. 2014a. Multidimensional quality metrics (mqm): A framework for declaring and describing translation quality metrics. *Tradumàtica: tecnologies de la traducció*, 0:455–463.

Arle Richard Lommel, Aljoscha Burchardt, Maja Popovic, Kim Harris, and Hans Uszkoreit. 2014b. Using a new analytic measure for the annotation and analysis of mt errors on real data. In *Proceedings of the 17th Annual Conference of the European Association for Machine Translation. Annual Conference of the European Association for Machine Translation (EAMT-14)*.

Qing Lyu, Marianna Apidianaki, and Chris Callison-Burch. 2024. Towards faithful model explanation in nlp: A survey.

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

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4932–4942, Florence, Italy. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Jiajing Wan and Xinting Huang. 2020. KaLM at SemEval-2020 task 4: Knowledge-aware language models for comprehension and generation. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 543–550, Barcelona (online). International Committee for Computational Linguistics.

Wenda Xu, Danqing Wang, Liangming Pan, Zhenqiao Song, Markus Freitag, William Wang, and Lei Li. 2023. INSTRUCTSCORE: Towards explainable text generation evaluation with automatic feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5967–5994, Singapore. Association for Computational Linguistics.

Chrysoula Zerva, Frédéric Blain, Ricardo Rei, Piyawat Lertvittayakumjorn, José G. C. de Souza, Steffen Eger, Diptesh Kanojia, Duarte Alves, Constantin Orăsan, Marina Fomicheva, André F. T. Martins, and Lucia Specia. 2022. Findings of the WMT 2022 shared task on quality estimation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 69–99, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

A Appendix

A.1 The Prompts for The Baselines

We list the example prompts for the one step prompt baselines in Table 11 and Table 12 for the error type of mistranslation and omission, respectively.

A.2 Explanation Annotators

The annotators for the Chinese-to-English task are postgraduate students proficient in Chinese and English. The annotators for the English-to-German task are postgraduate students majoring in German and being proficient in English.

There is an error in the translation, which is marked between "<v>" and "</v>". Please give a concise explanation in one sentence about the error according to the information given below.
source: 实时分享金融/财会/建筑最新考试资讯。
translation: Share the latest <v>financial/financial</v>/building exam information in real time.
reference: Real time sharing of the latest examination information of Finance/ Accounting/Architecture.
category: Accuracy/Mistranslation
severity: major
explanation: There is an error in the translation, "金融/财会" should be translated as "Finance/ Accounting"; so, change "financial/financial" to "Finance/ Accounting".
source: 教育时评:拯救被拐儿童要靠什么?
translation: Education Commentary: <v>What is it to</v> save abducted children?
reference: Education news commentary: What should we do to rescue abducted children?
category: Accuracy/Mistranslation
severity: major
explanation: There is an error in the translation, "要靠什么" should be translated as "What should we do to"; so, change "What is it to" to "What should we do to".
source: 今天是冬至节日，送错比较让人失望
translation: Today is winter solstice and it is more disappointing to <v>make mistakes</v>
reference: Today is the Winter Solstice, and delivering the wrong thing is quite disappointing.
category: Accuracy/Mistranslation
severity: major
explanation: There is an error in the translation, "送错" should be translated as "deliver the wrong thing"; so, change "make mistakes" to "deliver the wrong thing".
source: 外卖从没送达
translation: Take away never <v>delivered</v>
reference: The takeout order never arrived.
category: Accuracy/Mistranslation
severity: minor
explanation: There is an error in the translation, "送达" should be translated as "arrived"; so, change "delivered" to "arrived".

There is an error in the translation, which is marked between "<v>" and "</v>". Please give a concise explanation in one sentence about the error according to the information given below.
source: 摄影师也会很认真教你动作。
translation: The photographer will also be very serious to teach you <v>action</v>.
reference: Photographer would patiently teach you to pose.
category: Accuracy/Mistranslation
severity: major
explanation:

Table 11: The baseline prompt for the error type of mistranslation.

There is an omission of translating a source phrase, which is marked between "<v>" and "</v>". Please give a concise explanation in one sentence about the error according to the information given below.
source: <v>油管视频</v>起底美军20年嗜血杀戮真相-新华网
translation: The truth of the bloodthirsty killing of the US military for 20 years-Xinhuanet
reference: YouTube video uncovers the truth of US army's bloodthirsty killing of 20 years - Xinhuanet
category: Accuracy/Omission
severity: major
explanation: There is no translation for "油管视频"; so, it should be translated as "YouTube video" and added to the beginning of the translation.
source: 我<v>现在</v>每次都要付运费
translation: I have to pay the freight every time
reference: I have to pay the delivery fee every time now.
category: Accuracy/Omission
severity: major
explanation: There is no translation for "现在"; so, it should be translated as "now" and added to the end of the target.
source: 跟汉奸跟左癞都不要试图讲理，都是<v>一根儿筋</v>，无理可讲
translation: Don't try to reason with the traitor and leftists. They are both unreasonable. reference: Don't try to reason with rebels and crazy Leftists; they are all one track minded and unreasonable.
category: Accuracy/Omission
severity: major
explanation: There is no translation for "一根儿筋"; so, it should be translated as "one track minded and" and added between "both" and "unreasonable".
source: 莲子的作用：味甘涩，性平，<v>归心</v>，脾，肾经。
translation: Function of lotus seeds: sweet taste, flat nature, heart, spleen, kidney channel. reference: Functions of lotus seeds: sweet and astringent in taste, flat in nature, return to heart, spleen and kidney channels.
category: Accuracy/Omission
severity: major
explanation: There is no translation for "归心"; so, it should be translated as "return to" and added between "," and "heart".

There is an omission of translating a source phrase, which is marked between "<v>" and "</v>". Please give a concise explanation in one sentence about the error according to the information given below.
source: 不要把人吓倒在起跑线-<v>新华网</v>
translation: Do not scare people at the starting line.
reference: Don't scare people off at the starting line - Xinhuanet
category: Accuracy/Omission
severity: major
explanation:

Table 12: The baseline prompt for the error type of omission.