Translation of Multifaceted Data without Re-Training of Machine Translation Systems

Hyeonseok Moon[1], **Seungyoon Lee**[1], **Seongtae Hong**[1]
Seungjun Lee[1], **Chanjun Park**[2], **Heuiseok Lim**[1†]
[1]Department of Computer Science and Engineering, Korea University
[2]Upstage
[1]{glee889,dltmddbs100,ghdchlwls123,dzzy6505,limhseok}@korea.ac.kr
[2]chanjun.park@upstage.ai

Abstract

Translating major language resources to build minor language resources becomes a widely-used approach. Particularly in translating complex data points composed of multiple components, it is common to translate each component separately. However, we argue that this practice often overlooks the interrelation between components within the same data point. To address this limitation, we propose a novel MT pipeline that considers the intra-data relation[1] in implementing MT for training data. In our MT pipeline, all the components in a data point are concatenated to form a single translation sequence and subsequently reconstructed to the data components after translation. We introduce a Catalyst Statement (CS) to enhance the intra-data relation, and Indicator Token (IT) to assist the decomposition of a translated sequence into its respective data components. Through our approach, we have achieved a considerable improvement in translation quality itself, along with its effectiveness as training data. Compared with the conventional approach that translates each data component separately, our method yields better training data that enhances the performance of the trained model by 2.690 points for the web page ranking (WPR) task, and 0.845 for the question generation (QG) task in the XGLUE benchmark.

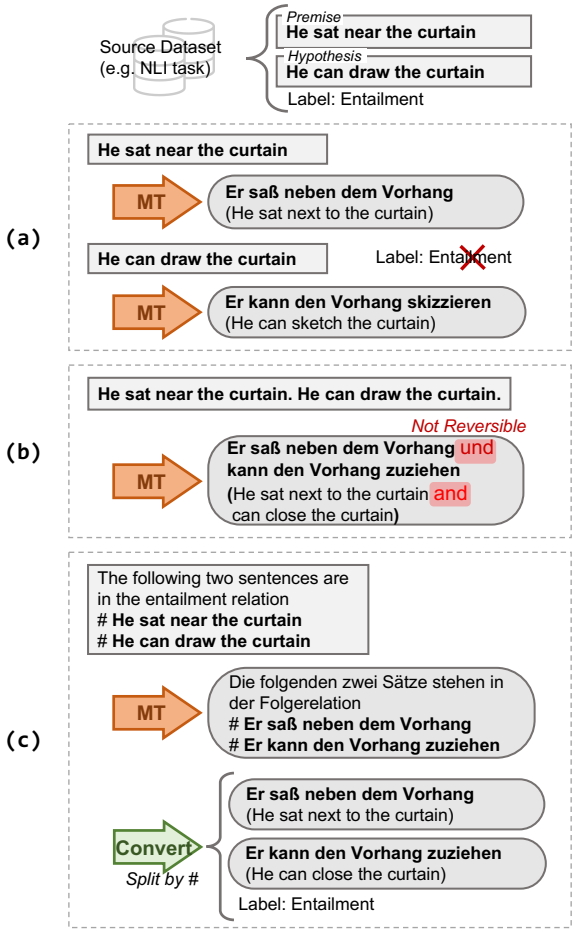

Figure 1: Example of challenges in data translation

1 Introduction

Machine translation (MT) has been developed to aid human-level utilization, with its primary focus on the accurate translation of any given sequence (*i.e.*, ensuring semantic preservation and syntactic fluency) (Specia et al., 2020; Guzmán et al., 2019; Martindale et al., 2019; Rei et al., 2020). As previous MT systems have demonstrated relatively low performance (Daems et al., 2017; Vilar et al.), their translation outputs are hardly utilized as another data source. With the ongoing advancement of MT research, the translation performance of MT systems becomes comparable to the expert human-level (Costa-jussà et al., 2022; Peng et al., 2023b), and subsequently several attempts have recently emerged to utilize MT system for the data translation process (Cui et al., 2023; Li et al., 2023; Liang et al., 2020; Peng et al., 2023a; Kakwani et al., 2020; Chen et al., 2022; Bassignana et al., 2023). Particularly, several non-English datasets are vig-

[1]Note that a single data point is composed of multiple components. In this sense, we use the term "interrelation among data components" as the same meaning as the "intra-data relation within a data point"

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 2088–2108

orously being constructed by translating English datasets (Bassignana et al., 2023; Adelani et al., 2023; Abulkhanov et al., 2023).

In applying MT to data translation, one concern we raise is the conservation of the intra-data relation during MT process. Depending on the task composition, a single data point may comprise multiple components. For example, each data point of natural language inference (NLI) task comprises three components; namely, the hypothesis and the premise along with one label. In translating such multifaceted components, we encounter dilemmas in determining the input unit, considering MT systems generally take a singular sequence.

In dealing with this situation, current research predominantly translates each individual data component separately (Turc et al., 2021; Bigoulaeva et al., 2023). However, we argue that such data translation approaches may not yield optimal results, as the interrelation among components in the same data can easily be disregarded. As shown in Figure 1-(a), the translated pair may not accurately maintain the original label, despite the absence of any error in their respective translations. This can further derive performance degradation of the model trained with these translated datasets, as the purpose of the task is generally represented within the interrelation among data components.

Theoretically, this issue can partially be alleviated by simply concatenating all the components in a single sequence for translation. Then in translating each component, MT system can refer the semantics of other components in the same sequence. However, in this case, MT system often merges all the components and generates an inseparable result to form a natural context. As shown in Figure 1-(b), this presents challenges in distinguishing data components from the translated sequence.

Upon these considerations, we propose a simple yet effective MT pipeline for the data translation that can be applied to any MT systems without further re-training. In particular, we propose a relation-aware translation that strategically concatenates multifaceted components into a singular sequence, as in Figure 2. Especially in concatenating data components, we discern the following two aspects: **(i)** the inter-relation between components should be considered in a concatenated sequence. **(ii)** translated sequence should be reversible (*i.e.* can explicitly be converted to the translated data components). To attain these objectives, we introduce **Indicator Token (IT)** and **Catalyst Statement (CS)**. **IT** is basically designed to distinguish the location of each data component and help conversion of the translated sequence into the translated components. **CS** is devised to specify the definite relation between each component in the concatenated sequence for enhancing the interrelation between components. Constructed sample is shown in Figure 1-(c).

For validation, we select multilingual benchmark tasks in which the maintenance of the interrelation among data components plays a critical role. Specifically, we adopt the XNLI dataset (Conneau et al., 2018) and select two tasks in an XGLUE benchmark (Liang et al., 2020): Web Page Ranking (WPR) and Question Generation (QG). We construct training data for up to five languages (German, French, Chinese, Hindi, and Vietnamese) by translating the English dataset existing within each dataset. Subsequently, by evaluating the performance of the models trained on each translated data, we estimate the validity of each data translation strategy. Notably, our proposed data translation pipeline demonstrates a more effective strategy to attain high-quality training data, compared to the individual translation of each data component.

2 Related Works

Attempts to construct training data with MT systems can broadly be divided into two major approaches. The first approach aims to construct a task-specific MT system by training with any corpus specially constructed for reaching intended goal (Phang et al., 2020; Ramponi and Plank, 2020; Carrino et al., 2020; Lewis et al., 2020; Duan et al., 2019; Shen et al., 2018). For instance, Sowański and Janicki (2023) trained a new translation model with a manually curated domain-specific dataset, then made a Polish training corpus for virtual assistant by translating English dataset. However, these attempts encounter difficulties in utilizing newly released assets.

In contrast, the second approach covers attempts to use publicly released NMT models without any modification, in constructing datasets via translation (Mozannar et al., 2019; Croce et al., 2019; Bassignana et al., 2023; Adelani et al., 2023; Abulkhanov et al., 2023; Sorokin et al., 2022). Representatively, commercialized NMT systems such as DeepL [2] (Croce et al., 2019; Bassignana et al.,

[2]https://www.deepl.com/translator

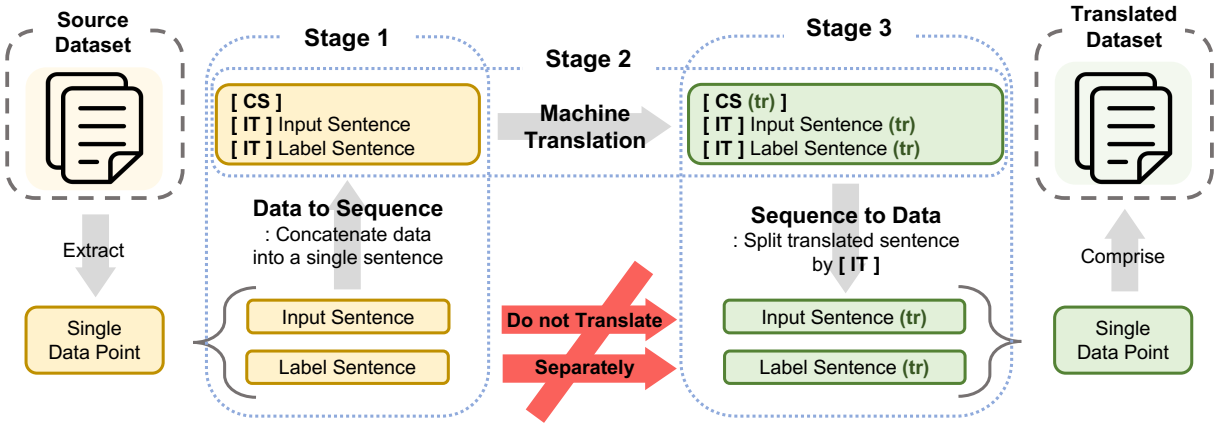

Figure 2: Relation-Aware translation pipeline. To explain the overall process, we assume data comprises two components: input sentence and label sentence. In this figure, **(tr)** represents the corresponding translated unit.

2023) or Google Translator [3] (Mozannar et al., 2019; Lee et al., 2018), along with publicly released NMT models (Costa-jussà et al., 2022; Fan et al., 2021) are adopted to construct multilingual training datasets (Adelani et al., 2023). However, prior approaches using existing assets with modification have encountered limitations in performing accurate data translation considering the interrelation among components comprising each data. Considering these attempts and their challenges, we focus on establishing an easy-to-implement pipeline for data translation that utilizes MT systems without model modification and takes into account intra-data relation.

3 Machine Translation for Machine

3.1 Problem Statement

In this study, we focus on potential issues of translating data constituting multiple components using conventional MT systems. Take QG task as an example, in which data constitutes passage x and question y as its components. We should note that there exists definite relation between these components: x is a passage that can derive a question y, and y is a question that can be retrieved from the passage x.

Ideally, in translating (x, y) to obtain a translated pair (x', y'), the semantic relation between (x, y) should be preserved after translation. To ensure relation-considering translation, the MT system should consider both components together even in translating respective components. This can be represented as an inference objective of maximizing probabilities displayed in Equation (1).

$$p(x'_i|x'_{<i}, x, y) \quad , \quad p(y'_i|y'_{<i}, x, y) \tag{1}$$

However, since MT system takes only a singular sequence, it can be challenging to impose additional constraints beyond the translating sequence. Consequently, in the conventional scenario, each component composing the same data point is individually translated instead, with an inference objective shown in Equation (2).

$$p(x'_i|x'_{<i}, x) \quad , \quad p(y'_i|y'_{<i}, y) \tag{2}$$

In this scenario, we argue that the efficacy of translated components as training data is inevitably diminished due to the lack of consideration for the intra-data relation. Theoretically, this issue can partially be alleviated by simply concatenating two components before translation, as the MT system can simultaneously refer to the context of all components. This entails translation with an inference objective similar to Equation (3), where ";" denotes any form of sequence concatenation.

$$p(z'_i|z'_{<i}, z) \quad \text{where} \quad z = [x; y] \tag{3}$$

Following the above equation, x in z can be translated by referring the semantics of y in z and vice versa. Subsequently, x' and y' can be yielded within the consideration of inter-relation between x and y.

However, in this case, the translated sequence z' might not be separated into x' and y'. As the major objective of the MT system is gaining fluent context, MT systems frequently insert conjunctions between two components and merge them into inseparable semantic unit, if necessary. Then

[3] https://translate.google.com/

Task	CS Type	Catalyst Statement
NLI	**Concat**	The following is a pair of sentences that are related to each other
	Relation	The following two sentences are in the **[LABEL]** relation
WPR	**Concat**	The following is a group of sentences that are related to each other
	Relation	Using the first sentence as a query, we obtained the following search results. We evaluate these results as **[LABEL]**
QG	**Concat**	The following is a pair of sentences that are related to each other
	Relation	The second sentence is a **question** that can be generated after reading the first **passage**

Table 1: Catalyst Statements (CSs) adopted in our experiments. Samples of constructed translation sequences are shown in Table 10.

the translated sequence cannot be converted to the translated data component whether the translation is perfect or not.

In essence, the primary challenges in data translation can be encapsulated as follows:

- Translating individual components hardly considers the intra-data relation within the same data point.
- When these are concatenated into a single sequence without any consideration, the translated sequence might not be restored back into the respective data components.

3.2 Our Solution: Relation-Aware Translation

To address these issues, we propose a viable strategy for performing data translation via any conventional MT framework without any modification. Our strategy involves a simple three-stage pipeline as shown in Figure 2.

First, we concatenate multifaceted components into a singular sequence to enable data translation through any form of MT systems (Data to Sequence). In concatenating instances, we integrate catalyst statement (**CS**) and indicator token (**IT**) to enhance the interrelation between data components and better distinguish the location of each data component after translation. **CS** is inserted in the head of the sequence, defining the relation between data components. **IT** is attached directly in front of each data component. Samples of constructed translation sequences are shown in Table 10, and we elaborate each role of **CS** and **IT** is elaborated in our subsequent sections.

Then we translate the concatenated sequence through the MT system. In implementing MT, we expect IT to be preserved intact after translation. If **IT** is not preserved after translation, we inevitably discard that data as we can hardly discriminate translated units for each data component. This may incur a degree of data loss; however, by conducting extensive experiments, we demonstrate that this process enables us to obtain high-quality training data from the remaining dataset.

After translation, we extract data components from the translated sequence (Sequence to Data). Specifically, we distinguish each translated component by splitting the translated sequence by the **IT**. Throughout this process, we can obtain the translated dataset, where each data point is translated with the consideration of intra-data relation.

3.3 Indicator Token (IT)

In cases where two or more components constituting the data are concatenated, the most intuitive way of ensuring the sequence can be re-segmented after translation, is to accurately specify each boundary. This can be performed based on a simple punctuation ('.'). Yet a more definitive criterion is necessary as a single component can comprise multiple sentences, and punctuation can frequently be substituted to conjunctions after translation, as depicted in Figure 1. In this regard, we prepend **IT** to each data component in concatenating data into a single sequence to distinguish the location of each component after translation. Notably, we expect **IT** to remain intact after translation, thereby we can obtain a translated data point by segmenting the translated sequence by **IT**.

Representatively, we experiment with the following simple instances: @, #, *. We take a single character form concerning any harms of semantics derived by the **IT**. We recognize that there may exist more effective instances of **IT** beyond the three examples we experimented with; we remain a room for improvement. In this paper, we focus on analyzing the impact of **IT** itself in data translation.

3.4 Catalyst Statement (CS)

By translating concatenated sequences, we can theoretically consider relation between the components within the data point. However, in such cases,

it might be challenging to discern how these components are directly related to each other, as naive concatenation can retains semantic separation between components within the same sequence, and thereby MT systems can hardly catch their semantic relation.

To enhance the interrelation between components within the same sequence, we propose to add an additional sentence that represents the definite relation. The purpose of its introduction is to signify the interrelated ties among the data components within the sequence to be translated, and to provide assistance by making these relations even explicit during the translation process. In essence, the aim is to substitute the task of translating seemingly semantically-separated statements with the attempt to translate a semantically-related single unit.

We denote this additional sentence as a **CS**. Particular examples we adopt in this study are shown in Table 1. We define the following two types of **CS**: directly defining the relation between components (**Relation CS**) and merely serving to connect components into a single sequence (**Concat CS**).

We use only simplified samples where other elements are excluded to objectively analyze the impact of considering intra-data relation during data translation. Specifically, these two sentences can be distinctly differentiated depending on the method of defining the relation of components. While there are potentially more possible CSs than the two we selected, we conduct experiments solely with these two representative samples to clarify our objective.

4 Experimental Settings

4.1 Dataset Details

We validate the effectiveness of our approach with the XNLI dataset (Conneau et al., 2018) and selected two tasks in the XGLUE benchmark (Liang et al., 2020) (WPR and QG). To acquire more general results, we conduct experiments in two to five languages for each dataset. Detailed statistics and composition of each dataset are described in Appendix B

4.2 Evaluation Details

We evaluate the validity of translation based on two primary criteria. The first is the **data reversibility**. As we have pointed out, if we translate concatenated sequence, respective components can be merged into a non-reversible element. We regard it as a translation failure, as it can hardly be utilized as training data. In estimating reversibility, we measure the percentage of the reversible data among translated sequences.

The second criterion pertains to the **quality** of the translated data. The main objective of our MT pipeline is enhancing the value of translated data as training instances by considering intra-data relation during the translation process. To validate our goal, we evaluate the performance of the model trained on the translated data. We estimate the label accuracy for evaluating performance of NLI and WPR tasks and measure ROUGE-L (Lin, 2004) for QG task. To deepen our evaluation, we compare this with the quality of the translation quality (estimated with BLEU score (Post, 2018)) and the results from the LLM evaluation (Liu et al., 2023; Chen et al., 2023).

4.3 Model Details

For implementing MT, we employed the multilingual MT systems capable of processing multiple languages, NLLB (Costa-jussà et al., 2022) and M2M100 (Conneau et al., 2020). Considering the verification scale and our resource constraints, we select distilled version of the original large-scale MT models: NLLB-600M, NLLB-1.3B, and M2M100-418M. After data translation, the translated data are fine-tuned with multilingual pre-trained language models to evaluate their value as training data. For NLI and WPR tasks, we adopt the XLM-R base model (Conneau et al., 2020), and for the QG task, we implement with the mT5 base model (Xue et al., 2021). Implementation details are included in Appendix A.

5 Results and Discussion

5.1 Simple concatenation does not guarantee the reversibility

In our preliminary discussions, we highlighted the issue that translating a concatenated sequence of data components may results in inseparable translated results, that cannot be converted to data components. This section provides experimental evidence supporting this claim. For each data point, we create a single sequence by concatenating data components with a '#' symbol and examine the preservation rate of '#' in the translated sequence. As Figure 3 demonstrates, the majority of cases fail to properly maintain the integrity of the '#'. For German training dataset in NLI task, the NLLB-

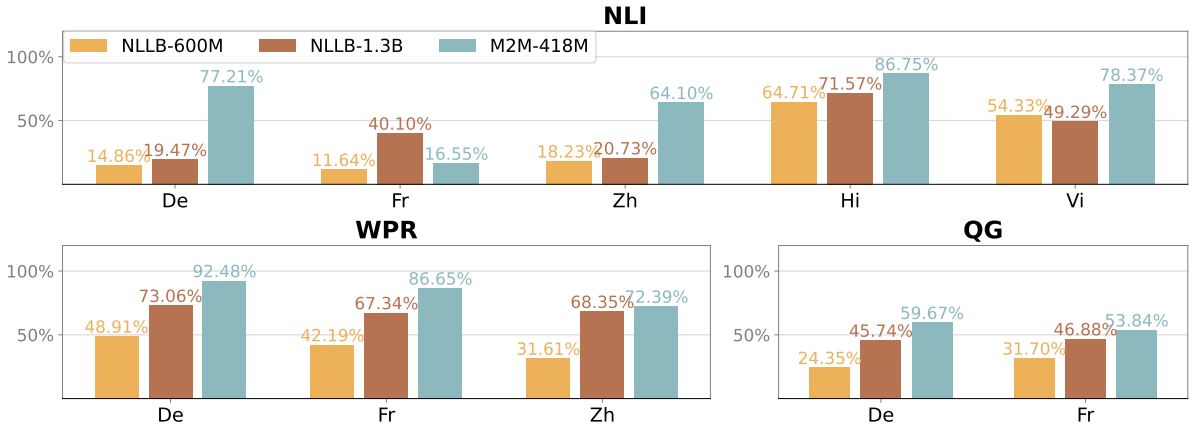

Figure 3: Data reversibility per NMT model and target dataset. For each data point, we create a single sequence by concatenating data components with a '#' symbol and examine the preservation rate of '#' in the translated sequence.

1.3B model preserved only 19.47% data points, indicating that approximately 80% of translated sequence can not be utilized as a data component. This underscores that mere concatenation is not a viable strategy for data translation and emphasizes the need for a thoughtful approach that considers relational aspects to ensure effective data translation. The drop in reversibility caused by superficial concatenation is a common tendency in our experiments, and the results for all combinations of IT and CS are presented in Appendix D.

5.2 Adding CS and considerate IT selection can be a solution

In this section, we verify that adding CS and prudent selection of IT can relieve the above challenge. We empirically assess the impact of incorporating IT and CS we designed, on the degree of reversibility. Figure 4 illustrates the average of reversibility over the five-language NLI data translated with NLLB-1.3B, for each case. As evidenced by our experimental findings, altering IT significantly influences reversibility. Particularly, utilizing '@' as IT can yield over a 25% increase in reversibility compared to using '#'. Additionally, the inclusion of CS contributes to enhanced reversibility. Notably, the performance of the Relation CS, which defines clearer relations among components, surpasses that of the Concat CS, which assigns weaker relationships among them. This underscores the effectiveness of our proposed IT+CS methodology in aiding data translation strategies. Further analysis of these impacts is presented in the subsequent sections.

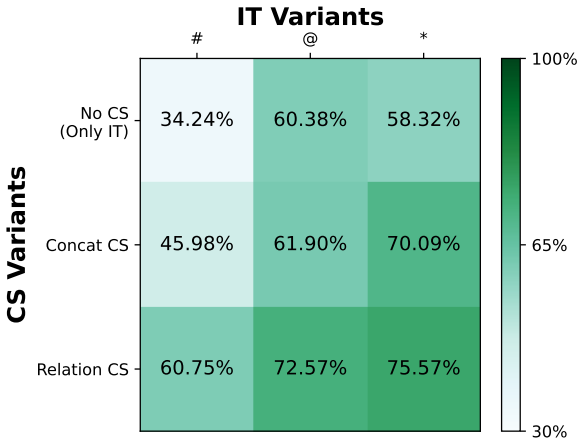

Figure 4: Reversibility after translation for IT and CS variants.

5.3 IT+CS enhances effectiveness as training data

We can considerably enhance reversibility through IT+CS, but our MT process inevitably incurs data loss while individual translation of each component would preserve a whole dataset. However, we contend that even though individually translated datasets may exhibit a larger quantity, their quality is likely to be compromised. Note that the primary focus of this study lies in enhancing the value of translated data as training instances. Considering this, we verify the substantial effectiveness of our approach against the individual translation of each data component, by comparing the performance of the model trained with each translated dataset. We report performances of CS variants utilizing '#' as IT. Experimental results presented in Table 2 demonstrate the following implications.

Task	WPR				QG		
Language	**De**	**Fr**	**Zh**	**Avg**	**De**	**Fr**	**Avg**
Separate	48.630	47.491	47.620	47.913 (-)	24.181	25.424	24.802 (-)
No CS	48.420	50.146	47.462	48.676 (+0.763)	24.733	25.715	25.224 (+0.422)
Concat CS	48.576	50.132	47.707	48.805 (+0.892)	24.781	25.657	25.219 (+0.417)
Relation CS	**50.066**	**50.593**	**48.908**	**49.855 (+1.942)**	**24.996**	**25.837**	**25.416 (+0.614)**
Performance of the trained model (vs model trained with individually translated data)							
Separate	**100%**						
No CS	48.91%	42.19%	31.61%	40.90% (-59.10%)	24.35%	31.70%	28.03% (-71.97%)
Concat CS	59.72%	61.48%	44.81%	55.34% (-44.66%)	33.10%	31.37%	32.24% (-67.76%)
Relation CS	82.41%	87.61%	69.70%	79.91% (-20.09%)	35.94%	41.82%	38.88% (-61.12%)
Quantity of the training data (vs individually translated data)							

Table 2: Performance of the model trained with each translated dataset. **Separate** refers to the individual translation of each data component.

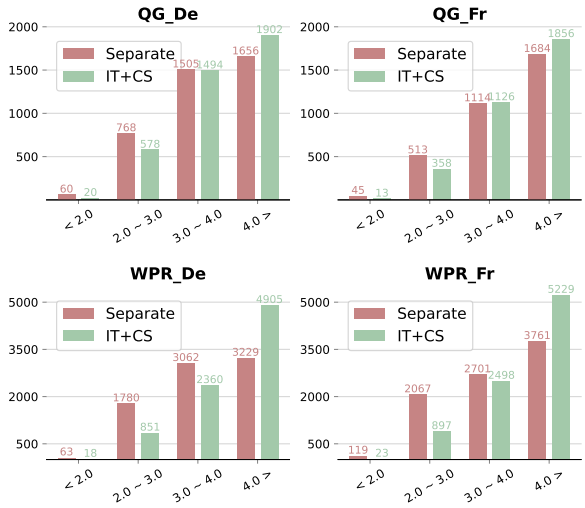

Figure 5: LLM evaluation results. We prompts ChatGPT to provide 0 - 5 scale quality score for each data point. Y-axis represents the quantity of instances which score is in the score range in X-axis.

Even small quantity, we can obtain high-quality data While fewer in quantity compared to those translated individually for each component, relation-aware translations (*i.e.*, **No CS**, **Concat CS** and **Relation CS**) demonstrated superior performance. Even in cases where only 28% of the QG data was preserved, the relation-aware translation exhibited greater effectiveness than the 100% training data generated by translating each component separately. These results validate our framework as an effective pipeline for acquiring high-quality training data.

Relation-aware translation makes better data The experimental results demonstrate that all methods concatenating data components for data translation outperform separate translation. Specifically, enhancing the interrelation between data components defined in CS led to improved performance. This underscores the significance of considering inter-component relationships in data translation, as highlighted by our motivation. Particularly we can obtain considerable performance improvement both for QG and WPR, compared to translating each component individually.

5.4 LLM Evaluation

To elaborate a more meticulous analysis of the impact of the IT+CS strategy on training data translation, we perform a LLM evaluation on the translated data (Chen et al., 2023). Chen et al. (2023) proposed an evaluation measure utilizing ChatGPT (OpenAI, 2022), that estimates the effectiveness of each data point as a training instance. Drawing inspiration from the previous study, we estimate the utility of each translated dataset as a training source. We adopt GPT3.5-turbo for evaluating each data point. Experimental results are illustrated in Figure 5, with additional details provided in Appendix C.

As observed from the experimental results, the IT+CS approach significantly increases the proportion of data scoring in the higher range (4.0 >) compared to the method of translating each data individually, while notably reducing the proportion of data scoring in the lower range (2.0 $\sim$ 3.0). This demonstrates the efficacy of our proposed framework in data translation and highlights the vulnerability of strategies translating each data component

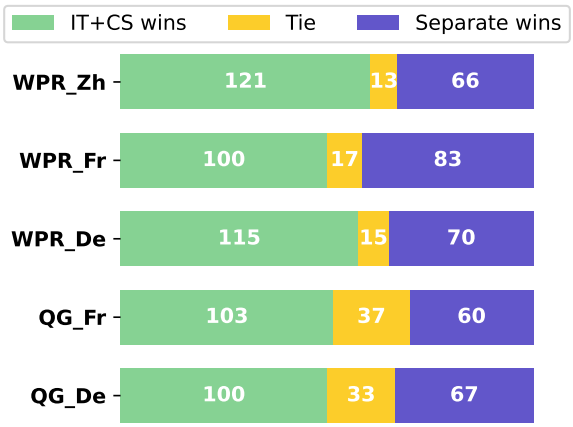

Figure 6: LLM evaluation results. We prompts GPT-4o to provide pairwise evaluation between Separate and IT+CS for each data point.

Task	NLI				
Language	**De**	**Zh**	**Fr**	**Hi**	**Vi**
Separate	38.887	26.634	56.281	14.004	41.796
No CS	39.456	26.455	56.664	14.147	42.437
Concat CS	38.398	26.474	56.598	**14.437**	42.966
Relation CS	**39.774**	**26.714**	**57.006**	14.425	**43.094**
BLEU score of the common training data					
Separate	66.607	**64.830**	68.862	**67.605**	66.846
No CS	65.589	64.088	**69.840**	65.329	69.062
Concat CS	67.505	62.495	69.561	65.449	69.741
Relation CS	**68.204**	64.092	68.543	65.230	**70.220**
Performance of the trained model					

Table 3: Performance of the model trained with each translated dataset. We derive a common index set for each language that intersects all the translated datasets. Then we extract a subset for each translated dataset, which indices are all included in the common index set.

separately.

Additionally, to facilitate a more intuitive analysis and comparison of the translation results of Separate and IT+CS, we employed GPT-4o for pairwise evaluation. We randomly sampled 200 data points from each dataset and qualitatively compared the translation quality of both approaches. Detailed prompts used for the evaluation are provided in Appendix C and illustrated in Figure 9. To minimize the effects of the LLM's positional bias (Wang et al., 2023), we randomized the input order of Separate and IT+CS for each evaluation. The experimental results are presented in Figure 6. As shown in the figure, the IT+CS strategy yields qualitatively superior translations compared to translating each component separately.

5.5 IT+CS enhances explicit translation quality

As XNLI provides human-crafted references for each language, the quality of translated data can explicitly be measured with conventional translation methods. In line with this, we analyze the impact of the relation-considering translation on translation quality. We extract overlapping portions from reversible datasets to form common training data and compare the translation quality and model performance trained on these data.

As shown in Table 3, IT+CS significantly improves translation quality. This improvement is particularly pronounced in alphabetic languages, with Vietnamese showing a 1.298-point enhancement in translation quality and a corresponding 3.374-point improvement in model performance compared to the Separate translation approach. However, performance degradation is observed in non-alphabetic languages, likely attributable to MT performance itself. We plan further analysis on this phenomenon.

5.6 IT+CS on the MT model variants

To verify the general applicability of the framework we propose, we evaluate the performance across three different models. The experimental results are presented in Table 4.

	NLLB-600M		NLLB-1.3B		M2M-418M	
	Separate	**IT+CS**	**Separate**	**IT+CS**	**Separate**	**IT+CS**
			WPR			
De	48.630	**49.519**	47.560	**50.710**	48.950	**49.845**
Fr	47.491	**48.729**	46.721	**49.208**	49.060	**50.151**
Zh	47.620	**49.056**	46.891	**49.326**	47.301	**49.191**
Avg	47.913	**49.101**	47.057	**49.748**	48.437	**49.729**
			QG			
De	24.181	**25.420**	25.535	**25.607**	23.960	**25.135**
Fr	25.424	**25.876**	25.833	**26.053**	24.676	**25.970**
Avg	24.802	**25.648**	25.684	**25.830**	24.318	**25.552**

Table 4: Experiment on model variants. We report the performance of the model trained with each translated dataset.

Here, we report on the performance of '*' IT, which consistently exhibits high performance across all three models. Specifically, we can obtain 2.690 and 0.845 point performance improvement, for each WPR and QG. As can be seen from our experimental results, our method outperforms the separate translation approach in terms of performance across all MT systems, all datasets, and languages. This validates the broad applicability of

our method as a data translation framework.

5.7 Qualitative Analysis

To delve deeper into the effectiveness of IT+CS in data translation and for the further analyses of translation results, we examine actual translation outcomes along with cross attention map analysis. The results are described in Appendix E. Through our experimental results, we affirm the practical superiority of the IT+CS strategy in its application to data translation, especially for the multifaceted data structure.

6 Conclusion

This study explored challenges encountered when implementing data translation through MT frameworks. We highlighted that individual translation of each data component neglects their interrelations, leading to a compromise in data quality. While composing a singular sequence by concatenating all the components theoretically can alleviate this, it also introduces the limitation of the inability to restore data components from the translated sequence. As a solution, we introduced a relation-considering translation pipeline that integrates IT and CS. This approach led to a substantial enhancement in the quality of training data as opposed to separate component translation. Our empirical findings underscored the paramount importance of inter-component relation in data translation, emphasizing that considering this relation can facilitate high-level data translations. This progression lays a foundation for future data translation research.

Limitation

We identify three potential constraints of our experimental setting. Firstly, variants of IT and RP were only tested under three specific cases. We were unable to validate every possible case, and there may exist other optimal types of IT or RP. While it is challenging to claim our results as optimal, our experiment conclusively affirmed that even subtle changes in IT can lead to evident performance improvement, and reinforcing the interrelation within each data component by concatenating RP can result in superior quality training data. Our experimental design encapsulates sufficient discussion to reach this conclusion.

The second limitation pertains to the variants of NMT models. We employed only three types of NMT models. Testing against a wider array of translation models could significantly enhance the general applicability of our study, but this was hindered by our resource constraints. Nevertheless, our experiments cover the difference in the model size (NLLB-600M and NLLB-1.3B) and the difference in the NMT training data or training strategy (NLLB and M2M) to induce more generalizable results.

Lastly, we confined language variants in our experiments. Due to resource constraints, we could not experiment with all languages provided by XGLUE and XNLI. However, we set up more than two languages for each task, to ensure our results not be biased towards any specific language. We deemed the varied performance and tendencies across different languages within NLI as a significant discovery. We did not perform further analyses as such discovery may fall beyond the scope of this paper, but we present an interesting scope for future research.

Ethics Statement

We utilized the publicly available XGLUE benchmark and XNLI datasets. We adhere strictly to the copyright of the original research in relation to the language resources and translated data used. Given that the utility and validity of XLGUE and XNLI have been established in numerous prior studies, we confirm that there were no distinct ethical issues encountered in our usage of these datasets.

Acknowledgements

This work was partly supported by ICT Creative Consilience Program through the Institute of Information & Communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (IITP-2024-RS-2020-II201819, 20%), Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIT) (RS-2024-00398115, Research on the reliability and coherence of outcomes produced by Generative AI, 40%) and Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(NRF-2021R1A6A1A03045425, 40%).

References

Dmitry Abulkhanov, Nikita Sorokin, Sergey Nikolenko, and Valentin Malykh. 2023. Lapca: Language-agnostic pretraining with cross-lingual alignment. In

Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 2098–2102.

David Ifeoluwa Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba O Alabi, Yanke Mao, Haonan Gao, and Annie En-Shiun Lee. 2023. Sib-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects. *arXiv preprint arXiv:2309.07445*.

Elisa Bassignana, Filip Ginter, Sampo Pyysalo, Rob Goot, and Barbara Plank. 2023. Multi-crossre a multi-lingual multi-domain dataset for relation extraction. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 80–85.

Irina Bigoulaeva, Viktor Hangya, Iryna Gurevych, and Alexander Fraser. 2023. Label modification and bootstrapping for zero-shot cross-lingual hate speech detection. *Language Resources and Evaluation*, pages 1–32.

Casimiro Pio Carrino, Marta R Costa-jussà, and José AR Fonollosa. 2020. Automatic spanish translation of squad dataset for multi-lingual question answering. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5515–5523.

Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srinivasan, Tianyi Zhou, Heng Huang, et al. 2023. Alpagasus: Training a better alpaca with fewer data. *arXiv preprint arXiv:2307.08701*.

Yang Chen, Chao Jiang, Alan Ritter, and Wei Xu. 2022. Frustratingly easy label projection for cross-lingual transfer. *arXiv preprint arXiv:2211.15613*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Danilo Croce, Alexandra Zelenanska, and Roberto Basili. 2019. Enabling deep learning for large scale question answering in italian. *Intelligenza Artificiale*, 13(1):49–61.

Yiming Cui, Ziqing Yang, and Xin Yao. 2023. Efficient and effective text encoding for chinese llama and alpaca. *arXiv preprint arXiv:2304.08177*.

Joke Daems, Sonia Vandepitte, Robert J Hartsuiker, and Lieve Macken. 2017. Identifying the machine translation error types with the greatest impact on post-editing effort. *Frontiers in psychology*, 8:1282.

Xiangyu Duan, Mingming Yin, Min Zhang, Boxing Chen, and Weihua Luo. 2019. Zero-shot cross-lingual abstractive sentence summarization through teaching generation and attention. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3162–3172.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. The flores evaluation datasets for low-resource machine translation: Nepali–english and sinhala–english. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6098–6111.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, NC Gokul, Avik Bhattacharyya, Mitesh M Khapra, and Pratyush Kumar. 2020. Indicnlpsuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961.

Kyungjae Lee, Kyoungho Yoon, Sunghyun Park, and Seung-won Hwang. 2018. Semi-supervised training data generation for multilingual question answering. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. Mlqa: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7315–7330.

Haonan Li, Fajri Koto, Minghao Wu, Alham Fikri Aji, and Timothy Baldwin. 2023. Bactrian-x: A multilingual replicable instruction-following model with low-rank adaptation. *arXiv preprint arXiv:2305.15011*.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, et al. 2020. Xglue: A new benchmark datasetfor cross-lingual pre-training, understanding and generation. In *Proceedings of the*

2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 6008–6018.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. Gpteval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

Marianna Martindale, Marine Carpuat, Kevin Duh, and Paul McNamee. 2019. Identifying fluently inadequate output in neural and statistical machine translation. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 233–243.

Hussein Mozannar, Elie Maamary, Karl El Hajal, and Hazem Hajj. 2019. Neural arabic question answering. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 108–118.

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

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023a. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*.

Keqin Peng, Liang Ding, Qihuang Zhong, Li Shen, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. 2023b. Towards making the most of chatgpt for machine translation. *arXiv preprint arXiv:2303.13780*.

Jason Phang, Iacer Calixto, Phu Mon Htut, Yada Pruksachatkun, Haokun Liu, Clara Vania, Katharina Kann, and Samuel Bowman. 2020. English intermediate-task training improves zero-shot cross-lingual transfer too. 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 557–575.

Matt Post. 2018. A call for clarity in reporting bleu scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191.

Alan Ramponi and Barbara Plank. 2020. Neural unsupervised domain adaptation in nlp—a survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6838–6855.

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.

Shi-qi Shen, Yun Chen, Cheng Yang, Zhi-yuan Liu, Mao-song Sun, et al. 2018. Zero-shot cross-lingual neural headline generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(12):2319–2327.

Nikita Sorokin, Dmitry Abulkhanov, Irina Piontkovskaya, and Valentin Malykh. 2022. Ask me anything in your native language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 395–406.

Marcin Sowański and Artur Janicki. 2023. Slot lost in translation? not anymore: A machine translation model for virtual assistants with type-independent slot transfer. In *2023 30th International Conference on Systems, Signals and Image Processing (IWSSIP)*, pages 1–5. IEEE.

Lucia Specia, Zhenhao Li, Juan Pino, Vishrav Chaudhary, Francisco Guzmán, Graham Neubig, Nadir Durrani, Yonatan Belinkov, Philipp Koehn, Hassan Sajjad, et al. 2020. Findings of the wmt 2020 shared task on machine translation robustness. In *Proceedings of the Fifth Conference on Machine Translation*, pages 76–91.

Iulia Turc, Kenton Lee, Jacob Eisenstein, Ming-Wei Chang, and Kristina Toutanova. 2021. Revisiting the primacy of english in zero-shot cross-lingual transfer. *arXiv preprint arXiv:2106.16171*.

David Vilar, Jia Xu, and Hermann Ney. Error analysis of statistical machine translation output.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023. Large language models are not fair evaluators. *arXiv preprint arXiv:2305.17926*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.

A Implementation Details

All experiments were performed using the RTX A6000. A single GPU was utilized for the training of an individual model, with early stopping criteria applied within 20 epochs. The learning rates selected for all tests were chosen among 1e-04, 3e-05, 5e-05, or 1e-04. The XLM-r base model was

fine-tuned using a lr of 2e-05, while the mT5-base model employed a lr of 1e-05. HuggingFace (Wolf et al., 2020) provided the foundation for all model configurations and training pipeline.

B Dataset Details

We validate the effectiveness of our approach with the XNLI dataset (Conneau et al., 2018) and selected two tasks in the XGLUE benchmark (Liang et al., 2020). The **NLI** dataset comprises pairs of sentences with a label categorizing the semantic relationship between the two sentences into one of three classifications: entailment, contradiction, or neutral. This task aims to illustrate the effectiveness of considering semantic relationships during translations.

In the **WPR** task, the goal is to predict the relevance of a web page to a given query. Each instance is a 4-part tuple: query, web page title, web page snippet, and label. The relevance label includes ratings from Perfect (4) to Bad (0). This task is included to verify the effectiveness of our approach in dealing with more than two components.

QG is a generation task, comprising a passage and a question that could originate from the given passage. In this case, we investigate the generalizability of our approach for lengthier translation units. To acquire more general results, we conduct experiments in five languages: English (En), Chinese (Zh), French (Fr), Vietnamese (Vi), and Hindi (Hi). These were chosen based on generally conceived resource quantity differences and the shared alphabetic character system. We validate all five languages for NLI. We select three existing languages (De, Fr, Zh) among the five for WPR, and two languages (De, Fr) for QG. Detailed statistics of each dataset are described in Table 5.

As we take the simplest form of IT, our experimented IT types can be included in the training dataset. To address this, we elaborate the count of data points that include each IT in Table 5. In comparison to the total data volume, the counts of data containing IT are deemed negligible. Given the data reversibility is not 100%, we postulate that the bias resulting from omitted data will likely be minimal.

Task	XNLI	WPR	QG
Train			
Num of data	392,702	99,997	100,000
Containing #	55	2,101	709
Containing @	80	1,558	79
Containing *	66	998	374
Validation			
Num of data	2,490	10,008	10,000
Containing #	0	240	25
Containing @	0	144	13
Containing *	0	109	34
Test			
Num of data	5,010	10,004	10,000
Containing #	3	234	30
Containing @	0	167	11
Containing *	0	107	46

Table 5: Data statistics.

C LLM Evaluation Details

In our study, we leverage LLMs to assess the quality of datasets translated through various methodologies. Zheng et al. (2023) indicate that LLMs' ability to align with preferences identified through both controlled experiments and crowdsourced methodologies exhibits a remarkable concordance rate exceeding 80%. This evidence underscores the potential of advanced language models in reflecting human judgments. Furthermore, following the methodology described by Chen et al. (2023) in utilizing LLMs as ChatGPT for data quality assessment, applying filtering criteria based on LLM evaluations resulted in a significant reduction of low-quality datasets, leading to improved performance of the trained model. This outcome serves as evidence of the effectiveness of employing LLMs for data quality assessment purposes.

To tailor the evaluation process to our specific needs, we adapted the prompts from (Chen et al., 2023) to assess the quality of our translated datasets and employed GPT-3.5-turbo as our evaluator. We conducted quality assessments on sentences translated from English source sentences in the XGLUE dataset's test set to De, Zh, and Fr using our method. The prompts used for each task were customized to reflect our evaluation criteria, illustrating the adaptability and precision of our methodology in assessing translation quality across diverse data contexts.

For pairwise comparison, we use the prompt shown in Figure 9. To minimize the effects of LLM's positional bias (Wang et al., 2023), we randomly set the input order of Separate and IT+CS for each evaluation. Additionally, we designed the evaluation setup with a more refined prompt to en-

	Task	NLI					QG		WPR			Avg
	Language	De	Zh	Fr	Hi	Vi	De	Fr	De	Fr	Zh	
@	No CS	70.36%	32.63%	**65.89%**	80.63%	73.49%	39.88%	**43.82%**	85.20%	72.00%	39.96%	60.39%
	Concat CS	81.04%	**56.18%**	48.27%	**96.52%**	63.76%	43.33%	30.73%	79.43%	69.16%	50.72%	61.91%
	Relation CS	**85.63%**	52.38%	65.33%	95.74%	**74.20%**	**61.16%**	42.29%	**87.68%**	**88.51%**	**73.00%**	**72.59%**
#	No CS	14.86%	11.64%	18.23%	64.71%	54.33%	24.35%	31.70%	48.91%	42.19%	31.61%	34.25%
	Concat CS	25.81%	25.62%	27.72%	79.34%	**70.94%**	33.10%	31.37%	59.72%	61.48%	44.81%	45.99%
	Relation CS	**46.44%**	**36.12%**	**54.07%**	**82.69%**	70.86%	**35.94%**	**41.82%**	**82.41%**	**87.61%**	**69.70%**	**60.77%**
*	No CS	40.27%	25.09%	41.90%	83.60%	71.20%	51.39%	53.18%	79.33%	76.03%	61.35%	58.33%
	Concat CS	68.63%	51.70%	57.50%	**91.12%**	**78.08%**	61.65%	58.42%	80.73%	80.59%	72.61%	70.10%
	Relation CS	**75.08%**	**55.52%**	**66.78%**	89.16%	75.94%	**69.01%**	**70.80%**	**84.56%**	**90.83%**	**78.13%**	**75.58%**

Table 6: Percentage of data reversibility after translation under NLLB-600M.

sure a more objective assessment.

System Prompt:

```
We would like to request your feedback on the
performance of AI assistant in response to the
passage and the given question displayed
following.

passage: [passage]
question: [question]
```

User Prompt:

```
Please rate according to the [dimension] of the
response to the passage and the question. Each
assistant receives a score on a scale of 0 to 5,
where a higher score indicates a higher level of
the [dimension].
Please first output a single line containing
the value indicating the scores. In the
subsequent line, please provide a comprehensive
explanation of your evaluation, avoiding any
potential bias.
```

Figure 7: Prompt template for ChatGPT evaluation of QG task. We evaluated each data point with the 0-5 scale quality score.

System Prompt:

```
We would like to request your feedback on the
performance of AI assistant in response to
the query and the given title and snippet
displayed following

query: [query]
title: [title]
snippet: [snippet]
```

User Prompt:

```
Please rate according to the {dimension} of the
response to the passage and the question. Each
assistant receives a score on a scale of 0 to 5,
where a higher score indicates higher level of
the {dimension}.
Please first output a single line containing
the value indicating the scores. In the
subsequent line, please provide a comprehensive
explanation of your evaluation, avoiding any
potential bias.
```

Figure 8: Prompt template for ChatGPT evaluation of WPR task. We evaluated each data point with the 0-5 scale quality score.

D Data Reversibility

When employing a translation model to translate data, reversibility is a crucial factor. High reversibility directly impacts the number of translated data instances and contributes to increasing the variability of data during model training. We experimented with all combinations of relation prompts and indicator tokens.

As indicated in Table 6, regardless of the type of indicator token used, leveraging relation prompt results in significantly high average reversibility. Notably, when the '#' indicator token was used, reversibility improved by approximately 77% when considering relations. However, the No RP scenario shows the lowest level of preservation during the translation process. This trend is also observed in other translation models we adopted. This suggests that leveraging a relation-considered relation prompt in translation can be an appropriate means to secure more data amount, regardless of the language or task.

E Qualitive Analysis

Table 7, Table 8 and Table 9 show the sample of translated components of each dataset using NLLB-600M. It compares the "Separate" and "IT+CS" methods, with CS using '#' symbol. It indicates that the original English text was translated into German. Additionally, LLM Eval Score represents the results obtained using the approach detailed in Section 5.4. Table 8 presents exemplars of results for each translation method applied to the

System Prompt:

```
Please act as an impartial judge and evaluate the quality of the two statements.
You will be given a English source statement and two {language} translated statements.
Each statement is generated by an AI translation assistant.
You should choose the statement that correctly evaluated each response and provided better quality
explanation to their assessment.
Your evaluation should consider factors such as the correctness, fluency, relevance,
accuracy and coherence.
Begin your evaluation by comparing the two statements and provide a short explanation.
Avoid any position biases and ensure that the order in which the responses were presented does not
influence your decision.
Do not allow the length of the responses to influence your evaluation.
Do not favor certain names of the assistants.
Be as objective as possible.
After providing your explanation, output your final verdict by strictly following this format:
"[[A]]" if Statement A is better, "[[B]]" if Statement B is better, and "[[C]]" for a tie.""",
```

User Prompt:

```
[Source]
{source}
[The Start of Statement A]
{translation 1}
[The End of Statement A]
[The Start of Statement B]
{translation 2}
[The End of Statement B]
```

Figure 9: Prompt template for pairwise comparison by GPT4o. We compare translation results derived from **Separate** and **IT+CS** strategies. `{source}` denotes source statement and `{translation 1}` and `{translation 2}` denote its transltaion resutls. `{source}` is set to be English, and each translation is in `{language}` language.

WPR dataset. In terms of contextual similarity and terminological consistency, "IT+CS" outperforms "Separate" offering a more faithful interpretation of the original text.

Table 9 illustrates examples of results obtained from various translation methods applied to the QG dataset. As evident from the examples, translations that consider relation exhibit greater fidelity in preserving the content, sentence structure, and similarity of the passage and question to the reference.

Additionally, the results of the attention map analysis are depicted in Figure 10, Figure 11, Figure 12. As can be observed from the results, when translation is conducted using the No CS approach, it is possible to refer to the context of each other, however, it frequently leads to the loss of the given IT. Conversely, when translating multifaceted data via the IT+CS method, semantic interreference between data components is observed, leading to superior translation quality and effective preservation of IT.

Sample 1 - QG	
Before Translation	- **passage**: Markice Moore as Andrew on The Walking Dead (AMC) Markice Moore, an actor best known for his role as Andrew on season three of The Walking Dead, is recovering at home after being hospitalized for a suicide attempt, according to The Hollywood Reporter. - **question**: walking dead star suicide attempt
Separate	- **passage**: Markice Moore als Andrew in The Walking Dead (AMC) Markice Moore, ein Schauspieler, der für seine Rolle als Andrew in der dritten Staffel von The Walking Dead bekannt ist, erholt sich nach einem Selbstmordversuch nach dem Krankenhaus nach dem Hollywood Reporter. - **question**: Selbstmordversuch mit einem starken Sternen
IT only	Markice Moore als Andrew in The Walking Dead (AMC) Markice Moore, ein Schauspieler, der am besten für seine Rolle als Andrew in der dritten Staffel von The Walking Dead bekannt ist, erholt sich nach einem Selbstmordversuch im Krankenhaus.
IT+CS	- **passage**: Markice Moore als Andrew in The Walking Dead (AMC) Markice Moore, ein Schauspieler, der am besten für seine Rolle als Andrew in der dritten Staffel von The Walking Dead bekannt ist, erholt sich nach einem Selbstmordversuch nach dem Krankenhaus nach dem Hollywood Reporter. - **question**: Walking Dead Star Selbstmordversuch
Sample 2 - WPR	
Before Translation	- **query**: twitw - **title**: Twitter Developer Platform — Twitter Developers - **snippet**: Twitter is the best place in the world for businesses and people to connect. Since the early days of Twitter people have used the public, live, and conversational nature of the platform to engage with businesses.
Separate	- **query**: Schlagwort - **title**: Twitter-Entwickler-Plattform Twitter-Entwickler - **snippet**: Seit den frühen Tagen von Twitter nutzen Menschen die öffentliche, live und konversationsartige Natur der Plattform, um mit Unternehmen zu interagieren.
IT only	- Twitter ist der beste Ort der Welt für Unternehmen und Menschen, um sich zu verbinden. Seit den frühen Tagen von Twitter haben die Menschen die öffentliche, live und konversative Natur der Plattform genutzt, um mit Unternehmen zu interagieren.
IT+CS	- **query**: twitw - **title**: Twitter Developer Platform Twitter Developers - **snippet**: Twitter ist der beste Ort der Welt für Unternehmen und Menschen, um sich zu verbinden. Seit den frühen Tagen von Twitter haben die Menschen die öffentliche, live und konversative Natur der Plattform verwendet, um mit Unternehmen zu interagieren.

Table 7: System-level qualitative analysis.

	Reference(En)	
Query	100% cotton racerback camisoles 1XL	
Title	Sofra Women's 100% Cotton Racerback Tank Top - amazon.com	
Snippet	Sofra Women's 100% Cotton Racerback Tank Top ... A bit worried what will happen when I wash them as they are 100% cotton. I think they will be ok to wear under other things, the fabric is slightly shear so probably not good with nothing over it. You get what you pay for is the lesson here. Read more. Helpful.	
	Separate(De)	**IT+CS(De)**
Query	100% Baumwoll-Rennstreifenhemden 1XL	100% Baumwoll-Racerback-Shemisoles 1XL
Title	Sofra Frauen 100% Baumwoll-Racerback Tank Top - amazon.com	Sofra Frauen 100% Baumwoll-Racerback Tank Top - amazon.com
Snippet	Sofra Frauen 100% Cotton Racerback Tank Top... ein wenig besorgt, was passiert, wenn ich sie wasche, da sie 100% Baumwolle sind. Ich denke, sie werden in Ordnung sein, um unter andere Dinge zu tragen, der Stoff ist leicht scheren, so wahrscheinlich nicht gut mit nichts über. Sie bekommen, was Sie bezahlen ist die Lektion hier. Lesen Sie mehr. hilfreich.	Sofra Women's 100% Cotton Racerback Tank Top... Ein bisschen besorgt, was passiert, wenn ich sie wasche, da sie 100% Baumwoll sind. Ich denke, sie werden in Ordnung sein, um unter anderen Dingen zu tragen, der Stoff ist leicht scheren, so dass wahrscheinlich nicht gut mit nichts darüber. Sie bekommen, was Sie bezahlen ist die Lektion hier. Lesen Sie mehr. Hilfreich.
LLM Score	**2**	**4**

	Reference(En)	
Query	llc online application for florida	
Title	Corporations - Division of Corporations - Florida ...	
Snippet	Make all checks payable to the Florida Department of State. Check and money orders must be payable in U.S. currency drawn from a U.S. bank. Credit cards accepted for filing online are MasterCard, Visa, Discover and American Express. Prepaid Sunbiz E-File Account. Processing. File online: 2-3 business days.	
	Separate(De)	**IT+CS(De)**
Query	llc Online-Bewerbung für Florida	llc Online-Anwendung für Florida
Title	Unternehmen - Unternehmensbereich - Florida...	Corporations - Division of Corporations - Florida...
Snippet	Alle Schecks müssen an das Florida-Außenministerium gezahlt werden. Schecks und Geldbestellungen müssen in US-Währung aus einer US-Bank gezahlt werden. Kreditkarten, die für die Online-Aufgabe akzeptiert werden, sind MasterCard, Visa, Discover und American Express. Prepaid Sunbiz E-File Account. Verarbeitung. Online-Aufgabe: 2-3 Werktage.	Alle Schecks an das Florida State Department zu zahlen machen. Schecks und Geldbefehle müssen in US-Währung aus einer US-Bank gezogen werden. Kreditkarten, die für die Online-Aufgabe akzeptiert werden, sind MasterCard, Visa, Discover und American Express. Prepaid Sunbiz E-File Account. Verarbeitung. Online-Datei: 2-3 Werktage.
LLM Score	**2**	**4**

Table 8: Samples of translations result using the "Separate" and "IT+RP" method for the WPR dataset.

	Reference(En)	
Passage	There is Frost on the Coils. The coils you see on the back of your dehumidifier are called the evaporator. When the unit is running, these coils turn ice cold. As the fan draws the moist air over the coils, the humidity in the air condenses on these coils.	
Question	why do dehumidifiers stop working	
	Separate(De)	**IT+CS(De)**
Passage	Die Spulen, die Sie auf der Rückseite Ihres Entfeuchters sehen, werden als Verdampfer bezeichnet. Wenn die Einheit läuft, werden diese Spulen zum Eis kalt.	Es gibt Frost auf den Coils. Die Coils, die Sie auf der Rückseite Ihres Entfeuchters sehen, werden als Verdampfer bezeichnet. Wenn die Einheit läuft, werden diese Coils zum Eis kalt. Wenn der Lüfter die feuchte Luft über die Coils zieht, kondensiert sich die Luftfeuchtigkeit auf diesen Coils.
Question	Warum die Entfeuchter nicht mehr funktionieren	Warum hören Entfeuchter auf zu arbeiten
LLM Score	**2**	**4**
	Reference(En)	
Passage	falls on a Saturday and as such, the due date for New Hampshire Interest & ... and Business Tax returns will be due on Tuesday April 18, 2017. Return due dates for all other tax types with a due date of April 15th are not impacted by ...	
Question	nh business tax due date	
	Separate(De)	**IT+CS(De)**
Passage	Die Frist für die Erstattung von Zinsen und... und Unternehmenssteuer wird am Dienstag, 18. April 2017 verfallen. Die Frist für die Erstattung aller anderen Steuertypen mit Ablaufdatum vom 15. April wird nicht von...	fällt am Samstag und als solches wird das Fälligkeitsdatum für die New Hampshire Interest &... und Business Tax-Returns am Dienstag, 18. April 2017 fällig sein. Die Fälligkeitsdaten für alle anderen Steuerarten mit einem Fälligkeitsdatum vom 15. April sind nicht von...
Question	n Geschäftssteuer fällig	nh Geschäftssteuer Fälligkeitsdatum beeinflusst.
LLM Score	**2**	**3**

Table 9: Samples of translation result using the "Separate" and "IT+RP" method for the QG dataset.

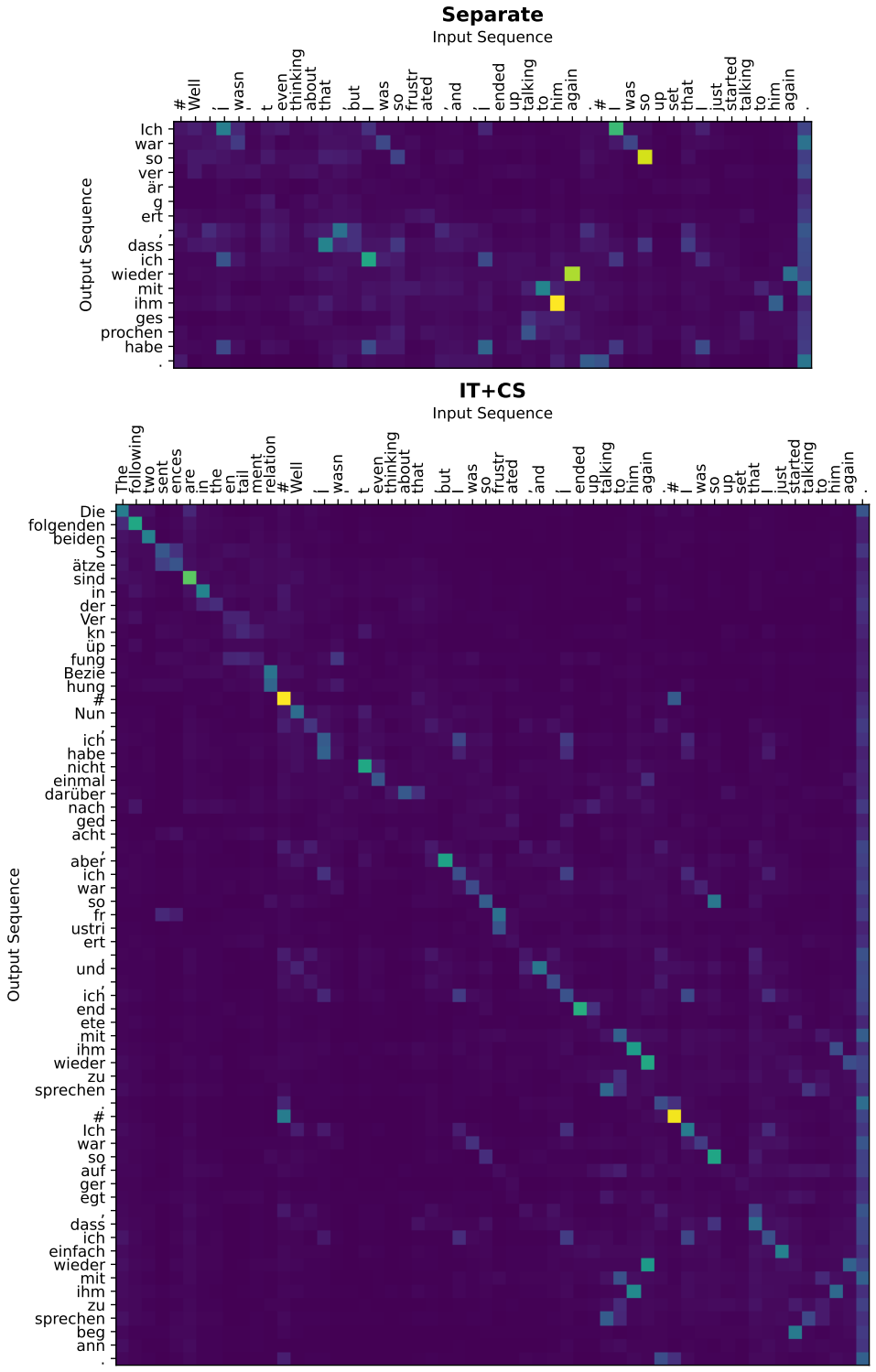

Figure 10: Cross-attention map in translating NLI data via NLLB-600M.

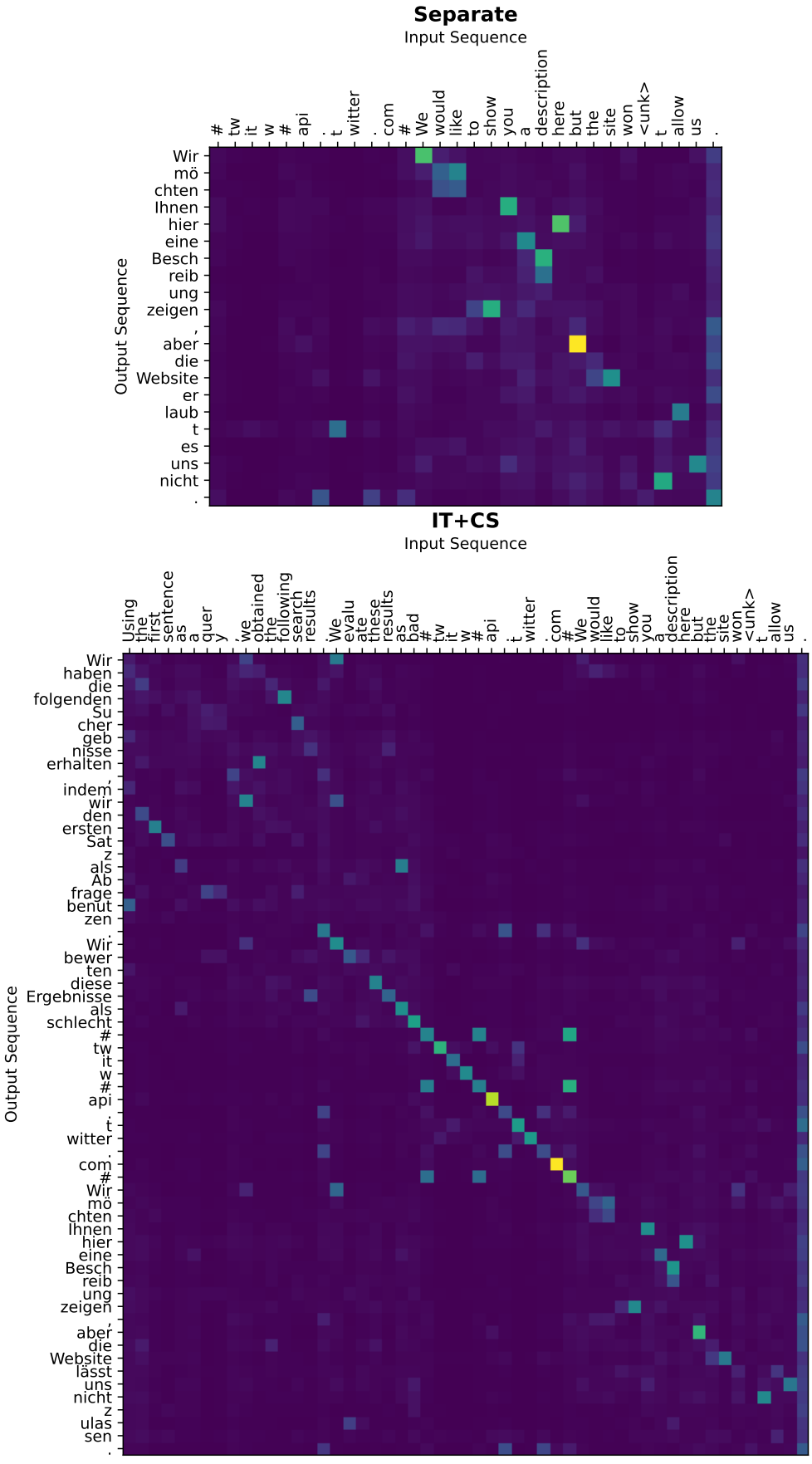

Figure 11: Cross-attention map in translating WPR data via NLLB-600M.

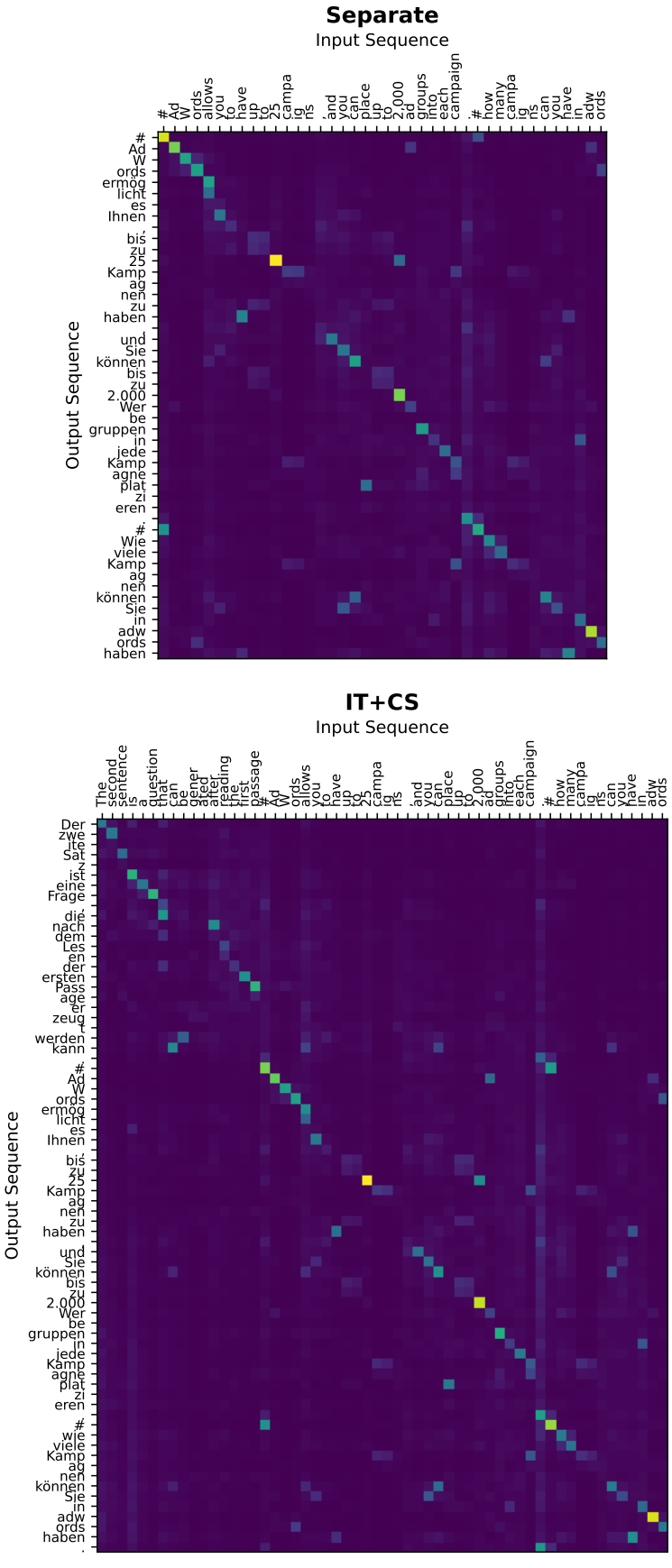

Figure 12: Cross-attention map in translating QG data via NLLB-600M.

	NLI	WPR	QG
Components	- **Premise**: One of our number will carry out your instructions minutely - **Hypothesis**: A member of my team will execute your orders with immense precision . - **Label**: Entailment	- **Snippet**: you have chosen this item to be automatically replenished at the above selected frequency - **query**: philosophy skin care - **title**: philosophy.com - skin care, fragrance, perfume, bath and ... - **label**: 4 (quality score)	- **Passage**: born on august 1 , 1779 , in frederick county , maryland , francis scott key became a lawyer who witnessed the british attack on fort mchenry during the war of 1812 . - **Question**: when was francis scott key born
No CS (Only IT)	# One of our number will carry out your instructions minutely # A member of my team will execute your orders with immense precision .	# you have chosen this item to be automatically replenished at the above selected frequency # philosophy skin care # philosophy.com - skin care, fragrance, perfume, bath and ...	# born on august 1 , 1779 , in frederick county , maryland , francis scott key became a lawyer who witnessed the british attack on fort mchenry during the war of 1812 . # when was francis scott key born
Relation CS	The following two sentences are in the **entailment** relation # One of our number will carry out your instructions minutely # A member of my team will execute your orders with immense precision .	Using the first sentence as a query, we obtained the following search results. We evaluate these results as **perfect** # you have chosen this item to be automatically replenished at the above selected frequency # philosophy skin care # philosophy.com - skin care, fragrance, perfume, bath and ...	The second sentence is a **question** that can be generated after reading the first **passage** # born on august 1 , 1779 , in frederick county , maryland , francis scott key became a lawyer who witnessed the british attack on fort mchenry during the war of 1812 . # when was francis scott key born

Table 10: Sample translation sequences. We show examples of translation sequences utilizing "#" as IT. For each task, we manually created label mapping to fit the original task objective. Specifically, we utilize [0: "entailment", 1: "neutral", 2: "contradiction"] for NER, and [4: "perfect", 3: "excellent", 2: "good", 1: "fair", 0: "bad"] for WPR.