DaCoM: Strategies to Construct Domain-specific Low-resource Language Machine Translation Dataset

Junghoon Kang[1], Keunjoo Tak[1], Joung Su Choi[1], Myunghyun Kim[2], Junyoung Jang[3], Youjin Kang[1],

[1]AI Center, HD Korea Shipbuilding & Offshore Engineering,
[2]DT Innovation Deparment, HD Hyundai Samho,
[3]School of Computing, Korea Advanced Institute of Science & Technology,
Correspondence: youjinkang@hd.com

Abstract

Translation of low-resource languages in industrial domains is essential for improving market productivity and ensuring foreign workers have better access to information. However, existing translators struggle with domain-specific terms, and there is a lack of expert annotators for dataset creation. In this work, we propose DaCoM, a methodology for collecting low-resource language pairs from industrial domains to address these challenges. DaCoM is a hybrid translation framework enabling effective data collection. The framework consists of a large language model and neural machine translation. Evaluation verifies existing models perform inadequately on DaCoM-created datasets, with up to 53.7 BLEURT points difference depending on domain inclusion. DaCoM is expected to address the lack of datasets for domain-specific low-resource languages by being easily pluggable into future state-of-the-art models and maintaining an industrial domain-agnostic approach.

1 Introduction

Foreign workers play an essential role in many industries. The emergence of neural networks and Large Language Models (LLMs) has accelerated the development of Machine Translation (MT), improving the quality of translation between different languages (Bahdanau et al., 2015; Wu et al., 2016; Vaswani et al., 2017; Zhang et al., 2023) and enabling workers of various nationalities to work together. However, despite the improvement, MT still struggles in certain domains of Low-Resource Languages (LRLs) (Kudugunta et al., 2023; Zhu et al., 2023) due to insufficient training data and technical terminology (Hayakawa and Arase, 2020).

Several studies have proposed to create datasets for translation of domain-specific LRLs, but most of them are focused on specific domains such as medicine, law, or religion (Anastasopoulos et al., 2020; Jaworski et al., 2023; Goyal et al., 2022). These datasets are often built by crawling or automatically generating data from websites like Wikipedia (Schuster et al., 2022; Schwenk et al., 2021). However, this general method is ineffective in constructing industrial domain data in LRLs due to the poor quality (Her and Kruschwitz, 2024; Haque et al., 2021).

The following are the reasons why collecting pair data of the industrial domain in LRLs is challenging:

The difficulty of collecting terminology and colloquial data. Terminology and colloquialisms are often used in industrial domains. For example, the South Korean construction site term "뺑끼"[1](Ppaengkki), which means paint, is derived from the Japanese "ペンキ(Penki)", which is also derived from the Dutch "Pek". However, these terms are usually not included in general-purpose language databases and require empirical knowledge of the field.

Lack of terminology due to industry culture differences. Due to different developed industries in different countries, some countries may not have a specific industry. In this case, domain concepts may not exist in other regions (Xiao, 2010). For example, in the shipbuilding industry, the term "pre-outfitting" means "the process of installing electrical, plumbing, etc. before a ship is assembled," but there is no term for this concept in landlocked countries like Mongolia or Kazakhstan.

In this paper, we propose a data collection system, DaCoM (**Da**ta **Co**nstruction through **M**essenger), to overcome the problem of low-resource data in industrial domains. DaCoM includes a translation framework consisting of a domain-specific glossary, a large language model

[1] https://en.wiktionary.org/w/index.php?title=%EB%BA%91%EB%81%BC&oldid=62233079

Proceedings of the 31st International Conference on Computational Linguistics: Industry Track, pages 612–624
January 19–24, 2025. ©2025 Association for Computational Linguistics

(LLM), and a neural machine translation model (NMT). It is applied to a messenger for tasks to help translate domain terms into appropriate LRLs. Finally, we build the automatically collected data into an industrial domain-specific low-resource language dataset through a validation procedure.

We construct a dataset leveraging DaCoM in the shipbuilding domain to verify the effectiveness of the system. By evaluating various models on the constructed dataset, it is confirmed that we have built a challenging dataset that is difficult for existing models to translate. In particular, the sub-dataset containing domain-specific terms shows a difference of up to approximately 53 BLEURT points compared to the sub-dataset without domain-specific terms. In addition, human evaluation certifies that the dataset constructed by DaCoM has high quality while the highest-scored model in the dataset still has room to improve.

Overall, our contribution is as follows

- We propose DaCoM, a methodology for collecting LRLs translation pair data in industrial domains. To the best of our knowledge, this is the first work to address data construction system for domain-specific and LRLs pair datasets.
- The translation system used in DaCom is a hybrid system consisting of a basic domain-specific dictionary, LLM, and NMT to construct translation pair data, which can be easily plugged into various models.
- Through extensive experiments and analysis, we demonstrate that domain-specific datasets collected using DaCoM reveal limitations in the performance of existing translators.

2 Related Work

To overcome the shortcomings of MT methods that find relationships between patterns by using massive amounts of data, research in the field of translation has begun to utilize NMT and LLMs. Accordingly, methodology and research on applying LRLs and domain-specific languages, which remained limitations in the traditional MT field, have also been conducted (Hedderich et al., 2020).

Low-resource languages LRLs hinder the effective training of MT models due to a lack of data, and translation quality is lower than in high-resource languages. To mitigate these issues, Johnson et al. (2017) improved LRL translation quality by training an NMT model for multiple languages simultaneously, sharing parameters across languages. Artetxe et al. (2017) used monolingual data to learn translation mapping through iterative back-translation and Denoising-Autoencoder. Goyal et al. (2021, 2022) created the Flores-101 and Flores-200 benchmarks for LRLs and multilingual MT, covering 101 and 200 languages, verified by professional translators. Recently, NLLB Team et al. (2022) and Kudugunta et al. (2023) proposed multilingual NMT models for more than 200 and 450 languages each by training extensive data for LRLs.

Domain-specific language Domain-specific MT requires a higher level of accuracy and context awareness than general domain translation because general language models do not sufficiently cover domain-specific terms and expressions. Müller et al. (2019) proposed a method of maintaining robust translation performance across various domains through inter-domain transfer learning. Khiu et al. (2024) investigated domain similarity and size of a corpus in LRLs MT and revealed the affection of domain similarity.

Data Collection For translation systems that deal with LRLs and specific domains, there are many difficulties in collecting appropriate data. To solve this problem, Mubarak (2018) used crowdsourcing to build speech and language resources for various annotation tasks. They proposed recommendations for task design and data quality management for high-quality data. Bañón et al. (2020) proposed a technology to automatically collect and refine large-scale parallel corpora from various web pages.

3 Pilot Experiments

To collect high-quality pair data automatically, we designed pilot experiments comparing NMT model and LLM. The setting is in Appendix A in detail. There is a limitation in capturing the nuances of domain-specific terms due to out-of-vocabulary (Alves et al., 2023). Therefore, we leverage LLM's powerful text generation ability and NMT's robustness to low-resource language translation to overcome the limitations. To this end, we experiment with a system that allows LLM to correct input text using definitions of domain terms and NMT to translate the corrected text. We have made glossaries for 30 terms from the construction domain

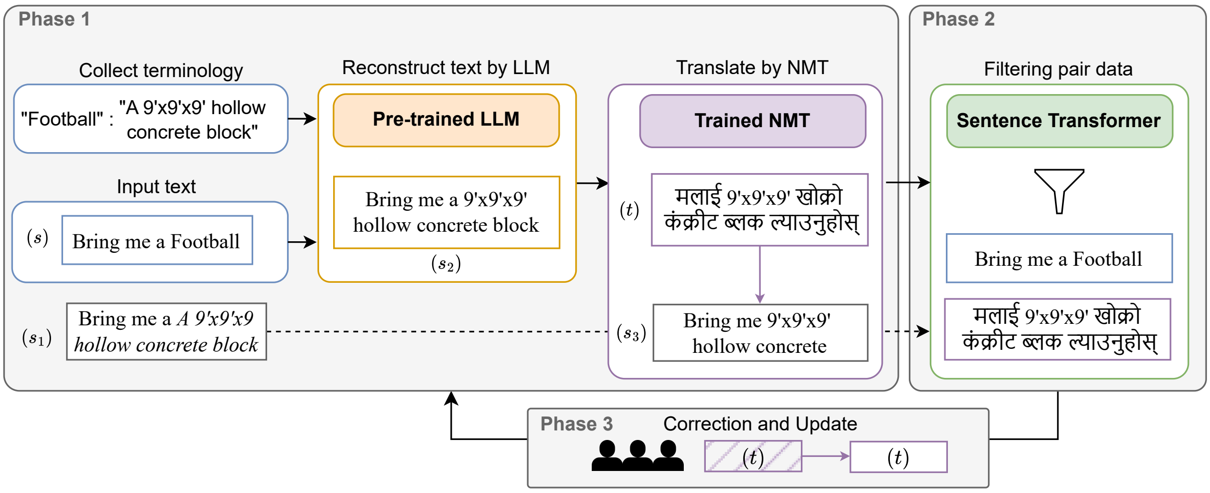
Phase 1
Collect terminology
"Football" : "A 9'x9'x9' hollow concrete block"
Input text
(s) Bring me a Football
(s1) Bring me a A 9'x9'x9 hollow concrete block
Reconstruct text by LLM
Pre-trained LLM
Bring me a 9'x9'x9' hollow concrete block
(s2)
Translate by NMT
Trained NMT
(t) मलाई 9'x9'x9' खोक्रो कंक्रीट ब्लक ल्याउनुहोस्
(s3) Bring me 9'x9'x9' hollow concrete
Phase 2
Filtering pair data
Sentence Transformer
Bring me a Football
मलाई 9'x9'x9' खोक्रो कंक्रीट ब्लक ल्याउनुहोस्
Phase 3
Correction and Update
(t)
(t)

Figure 1: Pipeline of DaCoM. In phase 1, PaLM2-unicorn and GNMT are used as LLM and NMT model. In phase 2, LaBSE is used as a sentence-transformer.

	Acc (%)	COMET	METEOR	BERTScore
NMT	13.3	75.8	66.1	68.5
LLM	57.3	76.6	64.6	67.1
LLM + NMT	**76.0**	**82.6**	**69.5**	**76.1**

Table 1: Results on pilot experiments. Google Translate and Gemini 1.0 pro are used for NMT and LLM, respectively, and Acc(%) denotes accuracy from human evaluation.

and asked Gemini 1.5 pro (Reid et al., 2024) to generate five appropriate English sentences each using the term. This method is inspired by research on automatic dataset construction through LLMs (Schick and Schütze, 2021; Wu et al., 2022).

In the pilot experiments, we used Google Translate (GNMT) (Wu et al., 2016) as an NMT system and Gemini 1.0 pro (Anil et al., 2023a) as an LLM[2]. Translation results were evaluated by COMET (Rei et al., 2020), METEOR (Banerjee and Lavie, 2005), BERTScore (Zhang et al., 2020), and human evaluation. We back-translated the Korean translation into English to measure the performance of automatic indicators and treated it as the target text. At this time, for more accurate semantic encoding, terms in the reference text were heuristically replaced with definitions. For human evaluation, 3 experts were asked to judge if the translation was correct in a blind setting for models, and then each instance was majority voted.

As a result, Table 1 shows that the method to utilize both LLM and NMT model achieves the highest translation quality. This result proves that the implicit knowledge of LLM and the multilingual token-matching ability of the NMT model can improve the quality of domain-specific and LRLs translation. Therefore, we introduce a data construction system in Section 4 leveraging the LLM and NMT model which primarily collects robust and high-quality pair data for translation.

4 Dataset Construction

We propose DaCoM, a system for constructing low-resource language translation datasets in industrial domains. DaCoM consists of a three-phase pipeline: (1) a translation service for efficient data collection, (2) automatic collection of data pairs, and (3) validation and calibration of the collected data.

4.1 Phase 1: Translation service for data collection

Build translation features into the communication tools used by the company or industry to include natural language usage patterns. The process pipeline for the translation service consists of a glossary, LLM, and NMT translator. First, we customize a glossary of commonly used terms in the domain by conducting on-site interviews, technical resources, web scraping, etc. We aim to collect around 2,000 terms or less depending on the size of the industry at this stage.

Next, when a user sends a message, the input source is divided into individual words, and specialized terms are extracted by referencing a constructed terminology dictionary. In this context, the users exchanging messages speak different

[2] GNMT and LLMs were used on June 2, 2024, at GMT+9

languages and communicate about work instructions or related topics. Subsequently, through an appropriate prompt to the LLM (refer to Table 9), the input source is refined into a text that can be accurately translated according to the context by consulting the terminology glossary. For example, as shown in Figure 1, "Bring me a Football" is segmented into [“Bring”, “me”, “a”, “Football”] and the term used in the construction industry, {Football: 9’x9’x9’ hollow concrete block}[3] is included as a candidate. It then reconstructs the phrase into an easy-to-understand sentence, such as “Bring me a 9’x9’x9’ hollow concrete block."

Finally, the reconstructed text is translated into the target language. Since language-specific token size has a significant impact on translation performance, we select the optimal translator by considering the performance of each translator for the source and target languages. In this study, we utilize GNMT (Wu et al., 2016), following the results of previous pilot experiments as NMT systems still often outperform LLM translation for LRLs.(Son and Kim, 2023).

4.2 Phase 2: Automatic pair data collection using LLM, NMT

In this phase, data pairs are collected after filtering out inappropriate content such as hate speech, personal information, and incorrect pairs resulting from automatic pair generation. First, to remove hate speech and personal information, we identified high-frequency words using a Bag-of-Words approach and heuristically filtered them as stopwords (Akuma et al., 2022; Pandey et al., 2022). This method was empirically chosen over profanity detection models and entity detection models for Korean source texts.

Next, we identified potential errors in DaCoM that could arise from (1) input text refinement by the LLM and (2) target text generation by the NMT. We performed similarity-based filtering at these two stages. For similarity calculations, we utilized the BERT-based LaBSE model(Feng et al., 2022), which is beneficial for LRLs and demonstrates consistent performance across multiple languages.

First, to filter out sentences incorrectly refined by the LLM, we compare text (s_1), which replaces domain-specific terms in the input text with their meanings, and text (s_2), refined by the LLM, using $\cos(s_1, s_2) \geq \theta_1$, with threshold, θ. Next, we apply a final filter using $[\cos(s_2, t) \geq \theta_2] \cup [\cos(s_2, s_3) \geq \theta_3]$ for text (t), translated into the target language, and text (s_3), back-translated into the source language. Each θ is chosen empirically.

Source Text	Counts
domain-specific words	1,714
unique domain-specific words	531
total # of tokens	14,518
average token length per sentence	7
domain-specific sentences	1,414
everyday life sentences	660
total # of sentences	2,074

Table 2: Statistics for DaCoM-created dataset

4.3 Phase 3: Correction and System update

In phase 2, the filtered text is verified by experts (interpreters or multilingual proficient individuals). Due to the scarcity of domain experts fluent in multiple languages, we requested at least one expert per target language to validate the target text and correct them. Using the corrected target text and its back-translation into the source language, we applied the same filter as in phase 2 to minimize bias. To improve data collection capabilities, we analyzed the data pairs extracted from the validation process and added domain-specific terms to the glossary used in phase 1.

5 Experiments

We apply DaCoM to the shipbuilding industry and build a dataset with Korean sources with English, Thai, Nepali, Uzbek, and Vietnamese targets to evaluate the performance of different translators.

5.1 Environment

Dataset We built a glossary of terms in the shipbuilding domain[4] and configured a prompt for the LLM to reconstruct the input sentence in general terms by referring to the collected terms. We selected model PaLM2-unicorn[5] (Anil et al., 2023b) as the LLM. The source language used in the experiment was Korean, and the LLM was leveraged to refine the domain terms as well as correct grammar and typos. The LLM reorganized the sentences to consider syllable block and spacing according to

[3] https://www.designingbuildings.co.uk/wiki/Glossary_of_construction_slang_and_other_terms

[4] https://standard.go.kr/KSCI/portalindex.do, https://parl.ns.ca/woodenships/terms.htm

[5] PaLM2-unicorn was used on May 2024, at GMT+9

		NLLB-54b	MADLAD-10b	GNMT	Gemini 1.0 pro	GPT-4	Llama 3.1-70b
ko ↓ en	BLEURT	45.17	51.96	**73.05**	58.57	62.23	60.60
	COMET	64.93	70.30	**83.82**	74.82	78.06	76.68
	METEOR	28.40	37.17	**72.57**	48.48	52.04	51.13
	BERTScore	88.71	88.58	**95.44**	90.92	92.51	92.26
ko ↓ th	BLEURT	30.78	33.38	**67.87**	35.35	51.56	47.43
	COMET	61.10	67.31	**84.12**	65.65	76.20	74.33
	METEOR	19.91	31.68	**71.02**	32.20	45.81	40.45
	BERTScore	66.06	75.57	**90.08**	58.94	81.37	79.83
ko ↓ ne	BLEURT	36.07	44.16	**76.61**	56.19	60.86	57.61
	COMET	51.72	55.86	**78.67**	61.21	65.65	63.58
	METEOR	16.30	21.19	**69.48**	33.31	35.61	26.67
	BERTScore	57.17	72.03	**90.39**	75.19	80.62	78.46
ko ↓ uz	BLEURT	39.21	29.84	**76.23**	44.12	51.72	54.59
	COMET	63.82	54.84	**84.85**	66.61	70.27	73.75
	METEOR	19.40	10.73	**69.12**	25.20	29.96	32.59
	BERTScore	66.79	65.32	**88.46**	66.02	75.82	76.21
ko ↓ vi	BLEURT	33.46	39.89	**71.05**	42.46	55.44	51.84
	COMET	62.46	67.17	**84.23**	68.74	77.27	76.55
	METEOR	24.42	28.99	**70.55**	35.79	46.13	41.41
	BERTScore	64.93	75.70	**90.26**	73.99	82.31	80.79

Table 3: Evaluation results on DaCoM-created. Bold and underlined indicate the highest and the next scores, respectively.

the postpositional particle (Park et al., 2020) in consideration of Korean characteristics. The dataset, named DaCoM-created, consists of about 2,074 pairs in Korean, English, Thai, Nepali, Uzbek, and Vietnamese. Table 2 presents the statistics.

Models We evaluate translation models: NLLB-54b (NLLB Team et al., 2022), MADLAD-400-10b (Kudugunta et al., 2023), GNMT (Wu et al., 2016), Gemini 1.0 pro (Anil et al., 2023a), GPT-4 (Achiam et al., 2023)[6], and Llama 3.1-70b-Instruct (Dubey et al., 2024). Information on the prompts and hyperparameters of the models is in Appendix C.

Metric We compute the BLEURT (Sellam et al., 2020), METEOR (Banerjee and Lavie, 2005), and COMET (Rei et al., 2020) scores reported on a typical translation task. For further semantic comparison, we use BERTScores (Zhang et al., 2020) leveraging the multilingual-BERT model (Devlin et al., 2019).

5.2 Results

DaCoM helps to build industrial domain datasets in low-resource languages Table 3 shows the translation inference performance of the translators on the Korean input in the DaCoM-created dataset. GNMT (Wu et al., 2016) performs the best. However, it performs up to 9 points lower than the average COMET score reported in Zhu et al. (2023) (about 87 points). Through qualitative analysis, we infer that this result originated from domain terminology (In Table 10).

In addition, we show that other translators achieve significantly low performance on the DaCoM-created dataset, especially when English is not the source or target language. These results reveal that existing models suffer low performance on domain-specific data in LRLs. DaCoM can improve the model by providing datasets of industrial domains in LRLs.

The model's performance challenges are related to domain-specific data. To analyze the cause of the translation performance degradation in DaCoM-created, we additionally experimented with the NLLB-54b model, which had the lowest performance in DaCoM-created, on subsets. The subsets

[6] GNMT and LLMs were used on July 9, 2024, at GMT+9

		B	C	M	B.S.
en	Domain-Y	22.4	45.7	10.7	84.8
	Domain-N	69.3	83.4	58.8	92.9
th	Domain-Y	5.3	42.5	5.1	59.3
	Domain-N	59.0	80.2	42.4	75.8
ne	Domain-Y	15.6	36.6	6.1	54.2
	Domain-N	59.8	69.1	37.9	72.5
uz	Domain-Y	13.1	48.1	6.3	58.1
	Domain-N	62.0	77.3	43.3	75.9
vi	Domain-Y	8.6	43.8	7.9	58.7
	Domain-N	62.3	81.8	50.6	79.7

Table 4: Evaluation comparison of NLLB-54b on domain-specific data (Domain-Y) and general data (Domain-N) in DaCoM-created. B, C, M, and B.S. denote BLEURT, COMET, METEOR, and BERTScore, respectively

	Flu.	Term app.	Rel.	Acc.(%)
GNMT	2.52	1.62	1.80	13
DaCoM	2.84	2.86	2.71	79

Table 5: Human evaluation on a subset from the SOTA model and DaCoM. Each metric denotes accuracy, Fluent, Term Appropriate, and Reliable, respectively.

consist of randomly extracted 200 data points each from data with and without domain-specific terms.

The subsets with domain-specific terms were labeled 'Domain-Y' and those without domain-specific terms were labeled 'Domain-N', which are shown in Table 4. The experimental results show that translation performance on the dataset with domain-specific terms degrades by **up to 53.7 points** on the BLEURT metric compared to the dataset without terms. As a result, we found that the presence of domain-specific terms affects the translator's performance.

DaCoM is a high-performance translator according to human evaluation. Table 5 shows the human evaluation scores for the translation results of the SOTA model (GNMT) in Table 3 and DaCoM system. We asked three shipbuilding experts, fluent in Korean and English, to evaluate 100 random samples containing pairs of Korean and English text with domain-specific terms. Annotators were instructed to evaluate the target text based on three criteria: 'Fluent' for assessing the fluency of the text, 'Term Appropriate' for verifying the correct use of domain-specific terms, and 'Reliable' for ensuring the target text conveys the same meaning as the source text. Each score is out of 3, and the average score per instance was used. The criteria for each metric is described in Appendix D. Finally, we measure accuracy (Acc.) by identifying cases where at least two out of three evaluators assign a score of 2 or higher for the 'Reliable' metric, assigning an accuracy score of 1 to such cases and 0 otherwise. The experimental results show that the Korean-English datapair built with DaCoM scores well on all three metrics, while the SOTA model scores poorly. This result ensures the dataset's quality while largely excluding the possibility that the dataset caused the performance degradation of the translators.

Combination		Similarity
PaLM2-unicorn + MADLAD-10b	en	74.2
	th	63.6
	ne	68.6
	uz	37.8
	vi	66.7
Gemini 1.5 pro + MADLAD-10b	en	66.3
	th	59.3
	ne	63.9
	uz	35.1
	vi	62.0
Llama 3.1-70b + MADLAD-10b	en	70.3
	th	62.9
	ne	66.9
	uz	37.3
	vi	64.6
Gemini 1.5 pro + GNMT	en	78.0
	th	78.6
	ne	76.4
	uz	74.4
	vi	76.5
Llama 3.1-70b + GNMT	en	82.8
	th	87.3
	ne	87.4
	uz	85.1
	vi	86.9

Table 6: Similarity between DaCoM-created dataset and translation results from the collaboration of various LLMs and NMT models

	Text
Source	엠티하게 자분캔 가져와 (Please bring the magnetic powder can for MT*.)
DaCoM	Please bring the empty magnetic powder can.
Source	영국아, 가서 용접해 (Yongguk, go and weld.)
DaCoM	England, go do some welding.

Table 7: Error analysis of translation results from DaCoM. *MT=Magnetic Test

DaCoM can be integrated as a plugin into various models. In the DaCoM system, we generated target sentences using various models (PaLM2-unicorn, Gemini 1.5 pro, Llama 3.1-70b-Instruct, MADLAD-400-10b, GNMT). [7] These models were different from those employed in DaCoM-created, and their similarity to the references of DaCoM-created was compared. Table 6 shows that English target sentences generated by different LLMs and NMT systems are generally similar to the reference. We used sentence-transformers (Reimers and Gurevych, 2019) with LaBSE model (Feng et al., 2022) for calculating the similarity. Notably, the combination of the Llama 3.1-70b-Instruct and the GNMT showed the highest similarity to DaCoM-created.These results demonstrate the pluggability of DaCoM.

6 Error Anlaysis

In Table 7, error cases of pair data from DaCoM generation are shown. The spelling of "영국"(Yongguk), which represents a person's name used in the Table, is written the same in Korean as "England"(pronounced as 'Yongguk'). DaCoM still has a limitation in handling homonyms and name translations, like other NMT or LLM translation systems. We plan to address this issue in depth in future work.

7 Conclusion

In this study, we propose DaCoM, a system for collecting low-resource translation datasets specialized for industrial domains. Extensive experiments and analysis demonstrate that the datasets constructed by DaCoM have high translation reliability. The experiments also indicate that existing translators show suboptimal translation performance due to the lack of domain-specific data pairs. In conclusion, we expect DaCoM to accelerate the improvement of translators by providing high-quality datasets that meet the unique translation requirements of LRLs and industrial domains.

Limitations

Our system effectively collects real data by integrating a high-performance translator for domain-specific LRLs into a chat messenger. As a result, the dataset primarily consists of conversational language with limited written expression. In future work, we plan to improve our system by adding a process to collect formal sentences as well, utilizing data augmentation with LLMs.

Additionally, while DaCoM was applied only to the shipbuilding domain in this paper, we confirmed through pilot experiments that it can also be effectively applied to various industrial domains.

Ethics Statement

We removed all personal information and hate speech when collecting data through DaCoM. We also notified system users in advance of our data collection plans and only used users who agreed to provide their data.

Acknowledgments

We would like to thank YoungOk Kim, Joonyoung Park, Chunhwan Jung, InIl Kim, and Junghyun Cho for their support to this project.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report.

Stephen Akuma, Tyosar Lubem, and Isaac Terngu Adom. 2022. Comparing bag of words and tf-idf with different models for hate speech detection from live tweets. *International Journal of Information Technology*, 14(7):3629–3635.

Duarte Alves, Nuno Guerreiro, João Alves, José Pombal, Ricardo Rei, José de Souza, Pierre Colombo, and Andre Martins. 2023. Steering large language

[7] GNMT, PaLM2-unicorn, Gemini 1.5 pro were used on Nov. 22, 2024, at GMT+9.

models for machine translation with finetuning and in-context learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11127–11148, Singapore. Association for Computational Linguistics.

Antonios Anastasopoulos, Alessandro Cattelan, Zi-Yi Dou, Marcello Federico, Christian Federmann, Dmitriy Genzel, Franscisco Guzmán, Junjie Hu, Macduff Hughes, Philipp Koehn, Rosie Lazar, Will Lewis, Graham Neubig, Mengmeng Niu, Alp Öktem, Eric Paquin, Grace Tang, and Sylwia Tur. 2020. TICO-19: the translation initiative for COvid-19. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online. Association for Computational Linguistics.

Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Slav Petrov, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy P. Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul Ronald Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, and et al. 2023a. Gemini: A family of highly capable multimodal models. *CoRR*, abs/2312.11805.

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernández Ábrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan A. Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vladimir Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, and et al. 2023b. Palm 2 technical report. *CoRR*, abs/2305.10403.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2017. Unsupervised neural machine translation. *arXiv preprint arXiv:1710.11041*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. 2020. ParaCrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic bert sentence embedding. *Preprint*, arXiv:2007.01852.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzman, and Angela Fan. 2021. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Preprint*, arXiv:2106.03193.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Rejwanul Haque, Chao-Hong Liu, and Andy Way. 2021. Recent advances of low-resource neural machine translation. *Machine Translation*, 35(4):451–474.

Takeshi Hayakawa and Yuki Arase. 2020. Fine-grained error analysis on English-to-Japanese machine translation in the medical domain. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 155–164, Lisboa, Portugal. European Association for Machine Translation.

Xingwei He, Zhenghao Lin, Yeyun Gong, A-Long Jin, Hang Zhang, Chen Lin, Jian Jiao, Siu Ming Yiu, Nan Duan, and Weizhu Chen. 2024. AnnoLLM: Making large language models to be better crowdsourced annotators. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, pages 165–190, Mexico City, Mexico. Association for Computational Linguistics.

Xingwei He and Siu Ming Yiu. 2022. Controllable dictionary example generation: Generating example sentences for specific targeted audiences. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 610–627, Dublin, Ireland. Association for Computational Linguistics.

Michael A Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2020. A survey on recent approaches for natural language processing in low-resource scenarios. *arXiv preprint arXiv:2010.12309*.

Wan-hua Her and Udo Kruschwitz. 2024. Investigating neural machine translation for low-resource languages: Using Bavarian as a case study. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*, pages 155–167, Torino, Italia. ELRA and ICCL.

Rafał Jaworski, Sanja Seljan, and Ivan Dunđer. 2023. Four million segments and counting: Building an english-croatian parallel corpus through crowdsourcing using a novel gamification-based platform. *Information*, 14(4).

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Eric Khiu, Hasti Toossi, Jinyu Liu, Jiaxu Li, David Anugraha, Juan Flores, Leandro Roman, A. Seza Doğruöz, and En-Shiun Lee. 2024. Predicting machine translation performance on low-resource languages: The role of domain similarity. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1474–1486, St. Julian's, Malta. Association for Computational Linguistics.

Sneha Kudugunta, Isaac Rayburn Caswell, Biao Zhang, Xavier Garcia, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. MADLAD-400: A multilingual and document-level large audited dataset. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

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

Hamdy Mubarak. 2018. Crowdsourcing speech and language data for resource-poor languages. In *Proceedings of the International Conference on Advanced Intelligent Systems and Informatics 2017*, pages 440–447. Springer.

Mathias Müller, Annette Rios, and Rico Sennrich. 2019. Domain robustness in neural machine translation. *arXiv preprint arXiv:1911.03109*.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia-Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *CoRR*, abs/2207.04672.

Yogesh Pandey, Monika Sharma, Mohammad Kashaf Siddiqui, and Sudeept Singh Yadav. 2022. Hate speech detection model using bag of words and naïve bayes. In *Advances in Data and Information Sciences: Proceedings of ICDIS 2021*, pages 457–470. Springer.

Kyubyong Park, Joohong Lee, Seongbo Jang, and Dawoon Jung. 2020. An empirical study of tokenization strategies for various Korean NLP tasks. 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 133–142, Suzhou, China. 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.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021. Generating datasets with pretrained language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6943–6951, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tal Schuster, Sihao Chen, Senaka Buthpitiya, Alex Fabrikant, and Donald Metzler. 2022. Stretching sentence-pair nli models to reason over long documents and clusters. In *Findings of the Association for Computational Linguistics: EMNLP*.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021. WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361, Online. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Jungha Son and Boyoung Kim. 2023. Translation performance from the user's perspective of large language models and neural machine translation systems. *Information*, 14(10).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Yuxiang Wu, Matt Gardner, Pontus Stenetorp, and Pradeep Dasigi. 2022. Generating data to mitigate spurious correlations in natural language inference datasets. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2660–2676, Dublin, Ireland. Association for Computational Linguistics.

Geng Xiao. 2010. Cultural differences influence on language. *Review of European Studies*, 2.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Xuan Zhang, Navid Rajabi, Kevin Duh, and Philipp Koehn. 2023. Machine translation with large language models: Prompting, few-shot learning, and fine-tuning with QLoRA. In *Proceedings of the Eighth Conference on Machine Translation*, pages 468–481, Singapore. Association for Computational Linguistics.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2023. Multilingual machine translation with large language models: Empirical results and analysis. *Preprint*, arXiv:2304.04675.

A Pilot Experimental Setting

In the pilot experiment, given the difficulty in collecting pair data for sentences containing domain-specific terms, we relied on the rich sentence-generation capabilities of LLMs to input terms and definitions and generate sentences using those terms, as shown in Figure 2. Table 8 shows the prompts used for sentence generation. As the target language, we selected Korean, which does not use Latin script. The accuracy evaluation aimed to determine whether the meaning of the source English

Prompt
The following are the terminology used at construction sites and their definitions. Using this term according to the explanation, make 5 sentences that could be used at a construction site. Term: {TERM} - {DEFINITION}

Table 8: Example of prompt to generate sentences using domain-specific terms for pilot experiments

Terminology

Banker : A mason, typically involved in cutting and smoothing building stone

Generated Sentence

The cathedral's construction required a team of skilled *bankers* to shape the intricate stone carvings.

Terminology

Tupper : A worker who carries the hod for a bricklayer

Generated Sentence

The foreman yelled at the *tupper* to bring more mortar, as they were running low.

Figure 2: Examples of construction data for pilot experiments

sentences was accurately reflected in the predicted Korean sentences. For automatic evaluation, we used the WMT22-COMET-DA model for COMET (Rei et al., 2020) and the mBART-large (Liu et al., 2020) model for METEOR (Banerjee and Lavie, 2005) and BERTScore (Zhang et al., 2020).

B DaCoM-created

To construct DaCoM-created, we chose thresholds(θs) introduced in Section 4.2 as follows, θ_1 = 0.9, θ_2 = 0.8, and θ_3 = 0.9. These values are chosen empirically. Table 9 presents an example of prompts used for PaLM2-unicorn, the LLM employed in DaCoM. Through the prompt, LLM was requested to refine input text with terminology, typos, and grammatical errors. Empirically, we selected N=8 shots for DaCoM-created.

C Baseline Details

NLLB-54b and MADLAD-400-10b used greedy decoding, and Gemini 1.0 pro and GPT-4 used temperature = 0.1 and top_p = 0.95. The prompt used for Gemini 1.0 pro and GPT-4 is as follows:

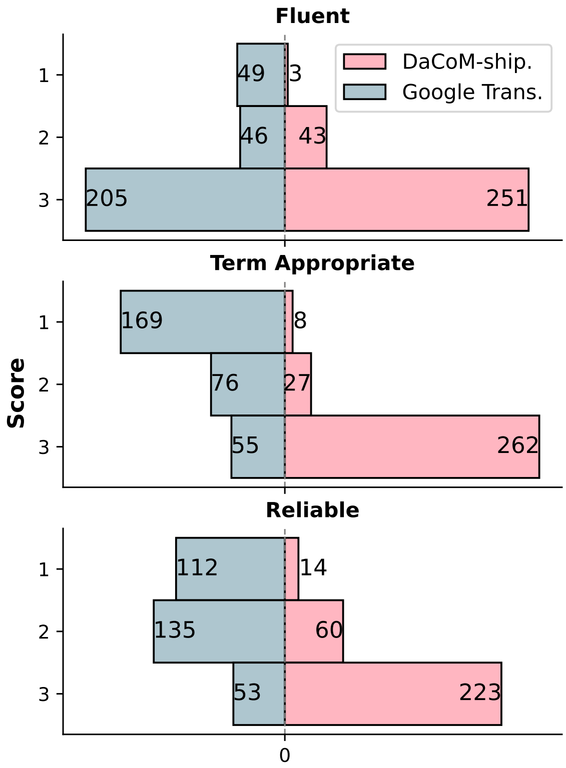

Figure 3: Score distribution of human evaluations

"You are a Language Translator. Translate from ‘Korean’ to ‘{TARGET LANGUAGE} . Always just return the translation of the prompt. prompt: {TEXT}”

D Details on Human Evaluation

The human evaluation is conducted with the metrics ‘Fluent’, ‘Term Appropriate’, and ‘Reliable’. Referring to He and Yiu (2022); He et al. (2024), we set questions and scoring criteria for each metric and asked three annotators to score following the criteria. We show the distribution of the results of GNMT and the DaCoM-created dataset in Figure 3.

Prompt
Text: {TEXT} ### Instruction ### Among the words in text, please change the words in the glossary considering the context. The glossary may be empty or contain the same words with different meanings. Please change naturally while preserving the context and meaning of the changed sentences/words. There may be typos, so if there is a word similar to the one in the glossary, please replace it with that word. Please write it well in Korean so that it can be translated well. Terminology: {TERMINOLOGY} Text: {TEXT} ### Example ### input: {SHOT-1 INPUT} output: {SHOT-1 OUTPUT} input: {SHOT-2 INPUT} output: {SHOT-2 OUTPUT} . . . input: {SHOT-N INPUT} output: {SHOT-N OUTPUT}

Table 9: Example of prompt to rewrite text using domain-specific terms in DaCoM

	Text
Source	족장 위에 공구 올려놓지마.
Target	Do not place tools on scaffolding.
GNMT	Don’t put tools on top of the pole.
GPT-4	Don’t put a tool on the tribal chief.
Gemini 1.0 pro	Don’t put tools on the workbench.
MADLAD-10b	Don’t put tools on the chief.
NLLB-54b	Don’t put the ball on the chief.
Source	저기에 있는 빼끼들 구루마에 싣고 1번 블럭으로 가세요.
Target	Put the paint on the cart over there and go to block 1.
GNMT	Put the hit and run guys over there on the cart and go to block 1.
GPT-4	Take those boxes over there and load them into the truck, then go to block 1.
Gemini 1.0 pro	Load the truck with the pigs over there and take them to Block 1.
MADLAD-10b	Get thoseguys in the truck and get them to block one.
NLLB-54b	Put the bags in the basket and go to Block 1.

Table 10: Qualitative examples from the models on DaCoM-created

D.1 Fluent

Annotators are asked to score each target text on a scale from 1 to 3 based on its fluency. To focus solely on fluency, the source text was not provided.
1: The text is incomprehensible and not fluent.
2: The text is comprehensible but not fluent or contains grammatical errors.
3: The text is fluent and there aren’t any grammatical errors.

D.2 Term Appropriate

Given source text, target text, and glossaries, annotators are asked to score the appropriateness of the translated domain terms in each instance on a scale from 1 to 3.
1: The translation of the domain-specific term is incomprehensible and inaccurate.
2: The translation of the domain-specific term is comprehensible but does not use appropriate words or expressions.
3: The translation of the domain-specific term uses appropriate words or expressions.

D.3 Reliable

Annotators are asked to score the target text on a scale from 1 to 3 based on how accurately it has the meaning of the source text.
1: The target text has a completely different meaning from the source text.
2: The target text has the intention of the source text but may be interpreted differently.
3: The target text accurately has the same meaning as the source text.

E Qualitative analysis

Table 10 presents qualitative examples from various translators. To achieve this, we randomly extracted two Korean-English pair sentences that contained at least two frequent domain-specific terms. In the examples, domain-specific terms from DaCoM-created and correctly translated terms are highlighted in blue, while incorrect ones are in red. Table 10 qualitatively demonstrates the difficulties translators face in translating domain-specific terms and shows that translation quality in specific domains depends on the accurate translation of these terms.