Data-Driven Approach for Formality-Sensitive Machine Translation: Language-Specific Handling and Synthetic Data Generation

Seugnjun Lee [1] **Hyeonseok Moon** [1] **Chanjun Park** [1 2] **Heuiseok Lim** [1]

Abstract

In this paper, we introduce a data-driven approach for Formality-Sensitive Machine Translation (FSMT) that caters to the unique linguistic properties of four target languages. Our methodology centers on two core strategies: 1) language-specific data handling, and 2) synthetic data generation using large-scale language models and empirical prompt engineering. This approach demonstrates a considerable improvement over the baseline, highlighting the effectiveness of data-centric techniques. Our prompt engineering strategy further improves performance by producing superior synthetic translation examples.

1. Introduction

Neural machine translation (NMT) models, despite their impressive progress, often overlook the role of style and pragmatic aspects in translation, such as formality or politeness (Britz et al., 2017; Stahlberg, 2020). This has given rise to the field of formality-sensitive machine translation (FSMT), which aims to control the level of formality in translated text across languages.

However, managing formality in MT is a challenging endeavor due to the lack of gold standard translations with different formality levels and the diverse formality markers across languages (Nădejde et al., 2022). For example, in many Indo-European languages, personal pronouns and verb agreement denote formality. Meanwhile, in Korean, formality control is complex due to the common use of morphological markers to express polite, respectful, and humble speech, making it an intriguing test case for FSMT.

In this paper, we propose a data-centric approach to FSMT for the English-Korean (EN-KO) and English-Vietnamese (EN-VI) language pairs. Our approach comprises two primary strategies: 1) a language-specific data-driven technique, and 2) synthetic data generation using large-scale language models and prompt engineering.

2. Proposed Method

2.1. Language Specialized Data-Centric Approach

We employ a language-specialized, data-centric approach that merges transfer learning techniques (Zoph et al., 2016) with language-specific subword methods, resulting in improved translation performance (Zoph et al., 2016; Bojanowski et al., 2017; Park et al., 2020; 2021). The pretrained model (PLM) is fine-tuned on the supervised train set for each language pair.

For both EN-KO and EN-VI translations, we adopt a data-centric approach emphasizing pre-training and fine-tuning on high-quality, language-specific datasets. For EN-KO, we use a Transformer model and a morpheme-aware subword tokenization method (Park et al., 2020), enhancing performance by addressing linguistic peculiarities of Korean. Similarly, for EN-VI, we utilize the specialized EnViT5 (Ngo et al., 2022) model, with training conducted on the expanded CC100 (Wenzek et al., 2020), MTet (Ngo et al., 2022), and PhoMT (Doan et al., 2021) datasets, improving translation in low-resource settings and underrepresented domains.

2.2. Synthetic Data Generation and Data-Centric Approach

To enhance translation quality, especially in low-resource settings, we utilize a data-centric approach by generating synthetic examples using ChatGPT with the GPT-4 engine (OpenAI, 2023). Our synthetic data are created through a conditioned translation generation task and refined using a formality classifier (Rippeth et al., 2022), thus ensuring accurate formality control.

Supervised Setting We leverage a prompt-based method, incorporating n randomly selected shots from the English training set of various language pairs for context. These prompts guide ChatGPT to produce translations in either informal or formal target language. The translated examples

[1]Department of Computer Science and Engineering, Korea University, Seoul 02841, Korea [2]Upstage, Gyeonggi-do, Korea. Correspondence to: Heuiseok Lim <limhseok@korea.ac.kr>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

		EN-KO				EN-VI			
	METHOD	BLEU	COMET	%M-ACC	%C-F	BLEU	COMET	%M-ACC	%C-F
Formal	Official Baseline	4.91	0.211	78.3	98.6	26.71	0.363	96.0	99.7
	ChatGPT	5.65	0.524	83.3	**100.0**	27.07	0.510	**100.0**	98.0
	Ours	**26.60**	**0.727**	**87.0**	**100.0**	**47.00**	**0.669**	99.4	**100.0**
	Ours + Augmentation	17.09	0.667	79.4	99.5	41.57	0.653	99.4	99.7
Informal	Official Baseline	4.85	0.170	97.6	99.5	25.28	0.345	96.0	98.2
	ChatGPT	5.60	0.564	**100.0**	**100.0**	25.83	0.482	**100.0**	**100.0**
	Ours	**27.10**	**0.715**	98.0	95.0	**45.60**	**0.637**	98.8	**100.0**
	Ours + Augmentation	20.35	0.621	98.5	98.8	40.46	0.484	98.7	**100.0**

Table 1. Results on the test set of Formality Dataset for formal and informal supervised settings.

		EN-PT				EN-RU			
	METHOD	BLEU	COMET	%M-ACC	%C-F	BLEU	COMET	%M-ACC	%C-F
Formal	Official Baseline	27.29	0.448	96.3	97.7	21.96	0.349	96.2	92.0
	ChatGPT	**31.25**	**0.655**	92.0	96.0	**31.25**	0.655	92.0	96.0
	Ours	31.00	0.525	**100.0**	**100.0**	25.80	0.445	**100.0**	**100.0**
Informal	Official Baseline	**30.93**	0.416	**93.2**	**90.8**	21.63	0.348	84.1	85.2
	ChatGPT	27.38	**0.512**	48.4	46.0	**31.25**	**0.655**	92.0	**100.0**
	Ours	19.90	0.249	68.0	90.0	26.30	0.418	**100.0**	**100.0**

Table 2. Results on the test set of Formality Dataset for formal and informal zero-shot settings.

are filtered using a formality classifier, and those meeting the formality criteria are integrated into the training sets for EN-KO and EN-VI fine-tuning. This data augmentation strategy and its impact are further evaluated through comparative experiments.

Zero-shot Setting In the zero-shot scenarios (EN-PT and EN-RU), we generate synthetic examples using the Gopalakrishnan et al. (2019). As in the supervised setting, prompts guide the model to produce translations in either formal or informal target language. The examples are filtered for accurate formality before being used to fine-tune the pre-trained multilingual translation model. This approach maximizes the model's generalization ability across languages and formality levels, demonstrating the utility of synthetic data in expanding pre-trained language models' capabilities.

3. Experiments

3.1. Experimental Settings

We conducted experiments using the Formality dataset (Nădejde et al., 2022) for the language pairs EN-{KO, VI} in the supervised setting and EN-{PT, RU} in the zero-shot setting. Prompt engineering was applied for EN-{KO, VI}, and synthetic examples were generated for fine-tuning on EN-{PT, RU}. Training details varied for each language pair, with EN-KO utilizing morpheme-aware tokenization and pre-training with a Transformer model. EN-{VI, PT, RU} pairs were fine-tuned using mBART-50 (Liu et al., 2020) model.

3.2. Experimental Results

Our data-centric approach yielded promising results, as evidenced in Table 1 and Table 2 for supervised and zero-shot settings, respectively. Our model, trained on the Formality Dataset, demonstrated near-perfect formality control, with high translation accuracy for most tasks, especially in the EN-KO and EN-VI language pairs. However, data augmentation with ChatGPT sometimes led to subpar performance, hinting at the requirement for more elaborate prompts considering formality control. Notably, the zero-shot EN-PT task results were significantly low, suggesting a need for specialized techniques for formality control per language pair and revealing a potential training data bias in ChatGPT.

4. Conclusion

We propose a data-centric approach for FSMT, incorporating language-specific techniques and synthetic data generation. Our approach achieves superior performance in EN-KO and EN-VI translations, delivering high-quality formality translations. While EN-PT informal exhibits lower performance, other pairs surpass the baseline, showcasing the translation capabilities of ChatGPT. For future work, we suggest exploring larger translation models, analyzing shot examples in more depth, employing linguistic-based data augmentation, and further investigating zero-shot transfer to enhance FSMT performance.

Acknowledgements

This work was supported by Institute for Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (No. 2022-0-00369, (Part 4) Development of AI Technology to support Expert Decision-making that can Explain the Reasons/Grounds for Judgment Results based on Expert Knowledge) and supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT)(No. 2022R1A5A7026673).

References

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5: 135–146, 2017.

Britz, D., Goldie, A., Luong, M.-T., and Le, Q. Massive exploration of neural machine translation architectures. *arXiv preprint arXiv:1703.03906*, 2017.

Doan, L., Nguyen, L. T., Tran, N. L., Hoang, T., and Nguyen, D. Q. PhoMT: A high-quality and large-scale benchmark dataset for Vietnamese-English machine translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 4495–4503, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.369. URL https://aclanthology.org/2021.emnlp-main.369.

Gopalakrishnan, K., Hedayatnia, B., Chen, Q., Gottardi, A., Kwatra, S., Venkatesh, A., Gabriel, R., and Hakkani-Tür, D. Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations. In *Proc. Interspeech 2019*, pp. 1891–1895, 2019. doi: 10.21437/Interspeech.2019-3079. URL http://dx.doi.org/10.21437/Interspeech.2019-3079.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., and Zettlemoyer, L. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8: 726–742, 2020.

Nădejde, M., Currey, A., Hsu, B., Niu, X., Federico, M., and Dinu, G. Cocoa-mt: A dataset and benchmark for contrastive controlled mt with application to formality. *arXiv preprint arXiv:2205.04022*, 2022.

Ngo, C., Trinh, T. H., Phan, L., Tran, H., Dang, T., Nguyen, H., Nguyen, M., and Luong, M.-T. Mtet: Multi-domain translation for english and vietnamese. *arXiv preprint arXiv:2210.05610*, 2022.

OpenAI. Gpt-4 technical report, 2023.

Park, C., Eo, S., Moon, H., and Lim, H.-S. Should we find another model?: Improving neural machine translation performance with one-piece tokenization method without model modification. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Papers*, pp. 97–104, 2021.

Park, K., Lee, J., Jang, S., and Jung, D. An empirical study of tokenization strategies for various korean nlp tasks. *arXiv preprint arXiv:2010.02534*, 2020.

Rippeth, E., Agrawal, S., and Carpuat, M. Controlling translation formality using pre-trained multilingual language models. *arXiv preprint arXiv:2205.06644*, 2022.

Stahlberg, F. Neural machine translation: A review. *Journal of Artificial Intelligence Research*, 69:343–418, 2020.

Wenzek, G., Lachaux, M.-A., Conneau, A., Chaudhary, V., Guzmán, F., Joulin, A., and Grave, E. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 4003–4012, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.494.

Zoph, B., Yuret, D., May, J., and Knight, K. Transfer learning for low-resource neural machine translation. *arXiv preprint arXiv:1604.02201*, 2016.