NLP Model for English-Korean Translation with Formality Adaptation

Eliot Arntz, Michelle Cheung, Richard Zhang

Abstract

This project aims to develop an advanced Natural Language Processing (NLP) model to translate between Korean and English, with a particular focus on accurately handling the multiple levels of formality in the Korean language. The approach involves fine-tuning LLaMa's (Llama-2-7b model) weights using Parameter-Efficient Fine-Tuning (PEFT) with Quantized Low-Rank Adaptation (qLoRA) and incorporating formality level tags in the training data. The results demonstrate that incorporating formality levels enhances the model's translation quality, as evidenced by higher BERT and COMET scores.

1 Introduction

Translation models are crucial in various domains, such as business, education, and healthcare, enabling effective communication across languages. The demand for high-quality machine translation systems has been growing significantly, driven by globalization and the need for real-time, accurate communication (Ruder et al., 2019).

Despite the advancements in machine translation, handling the nuances of formality, especially in languages like Korean, remains a challenging task. Korean language features a complex system of honorifics and speech levels, which are used to convey varying degrees of respect and politeness depending on the social status and relationship between the speakers (Brown & Levinson, 1987; Sohn, 1999). There are seven primary levels of formality in Korean, ranging from highly formal to very casual, and each level requires different verb endings and vocabulary (Brown, 2015). Existing models like BERT (Devlin et al., 2019) and GPT-3 (Brown et al., 2020) have demonstrated impressive

Formality Levels	**Sentence Type**
Differential	-pnita
Polite	-a/eyo
Blunt	-(s)o
Familiar	-ney
Intimate	-a/e
Plain	-([nu]n)ta

Table 1: Korean Speech Formality levels and Sentence Types by Cho 2006; Sohn 1999.

capabilities in natural language understanding and generation. However, they often fall short in capturing the subtleties of formality and honorifics in translation tasks. Recent studies have highlighted the need for specialized models that can adapt to the formality context of the target language (Hu et al., 2021; Ziegler et al., 2019). In business applications, accurate translation with appropriate formality levels is essential for maintaining professionalism and cultural sensitivity. For instance, multinational companies often need to translate corporate documents, marketing materials, and customer service communications into multiple languages, ensuring that the translations are not only accurate but also culturally appropriate (Lewis et al., 2020). Misinterpretation due to incorrect formality can lead to misunderstandings, damage to brand reputation, and loss of customer trust. Moreover, in sectors like healthcare, accurate translation is critical for patient safety and effective communication between healthcare providers and patients. Miscommunication due to incorrect formality levels can lead to serious consequences, including medical errors and compromised patient care (Bahdanau et al., 2015). This project aims to address these challenges by developing an NLP model specifically designed to handle the formality levels in Korean language translation. The model is designed to provide English translations to Korean sentences at different formality levels. By incorporating formality level tags into the training data and fine-tuning the model using PEFT with LoRA, we aim to achieve translations that are both accurate and contextually appropriate.

2 Background

The Korean language features a highly developed honorific system, which is crucial for expressing formality and politeness. Research by Sohn (1999) indicates that Korean has the most systematic grammatical pattern for honorifics. According to Brown and Whitman (2015), Korean's honorific system, especially in addressee honorification, distinguishes between four and seven levels of politeness. This project will leverage insights from various studies to handle these complexities in translation models.

Honorifics in Korean are not merely linguistic artifacts but are embedded deeply in the culture. Sohn (1999) explains that honorifics in Korean involve morphological changes at various levels, including verb endings, noun particles, and pronouns. Brown and Whitman (2015) further delve into the intricacies of Korean honorifics, identifying multiple levels of politeness that depend on the social hierarchy, the relationship between the speaker and the listener, and the context of the conversation. This complexity makes Korean a challenging language for machine translation tasks, especially when trying to capture the nuances of formality and politeness.

Neural Machine Translation (NMT) has seen significant advancements over the past decade. The Transformer model by Vaswani et al. (2017) revolutionized NMT with its attention mechanisms, allowing models to focus on different parts of the input sentence. This model's architecture has become the foundation for many state-of-the-art translation models.

Various approaches have been proposed to handle formality in translation. Sennrich et al. (2016) introduced a method to control formality by adding tags to the training data, which guided the model to produce translations at different formality levels. Similarly, Johnson et al. (2017) demonstrated that multilingual NMT models could be fine-tuned to handle different formality levels by training on mixed datasets that include formal and informal text pairs.

Hu et al. (2021) introduced Low-Rank Adaptation (LoRA), a technique to adapt large language models efficiently. LoRA decomposes the weight matrices of neural networks into smaller, low-rank matrices, which significantly reduces the number of parameters to be fine-tuned. This approach is particularly useful in scenarios where computational resources are limited. The application of qLoRA in our project aims to fine-tune the LLaMa model to better handle the complexities of Korean formality levels.

Several pre-trained models, such as BERT (Devlin et al., 2019), GPT-3 (Brown et al., 2020), and MarianMT (Junczys-Dowmunt et al., 2018), have shown remarkable performance in various NMT tasks. However, these models often fall short in handling languages with complex honorific systems like Korean. Ott et al. (2019) with the fairseq toolkit, and Lewis et al. (2020) with BART, have provided frameworks for training and fine-tuning NMT models, but incorporating formality adaptation remains a challenge.

Our approach stands out by integrating insights from various studies and leveraging advanced fine-tuning techniques to address the nuances of Korean formality in translation. By incorporating formality level tags into the training data and using Parameter-Efficient Fine-Tuning (PEFT) with qLoRA, our model aims to achieve higher translation accuracy and better contextual understanding. This project not only aims to improve machine translation between English and Korean but also sets a precedent for handling other languages with complex honorific systems.

This project's innovative approach of using formality level tags and advanced fine-tuning techniques is significant as it addresses a gap in current NMT models' ability to handle complex honorific systems. The integration of these methodologies will contribute to the existing body of research, providing a robust framework for future studies in machine translation and formality adaptation.

3 Methodology

3.1 Data Preparation

In this work we utilized datasets from the OPUS-100 project, specifically, available on hugging-face platform at https://huggingface.co/datasets/Helsinki-NLP/opus-100/viewer/en-ko. OPUS-100 is an

English-centric multilingual corpus covering 100 languages.

OPUS-100 is English-centric, meaning that all training pairs include English on either the source or target side. The corpus covers 100 languages (including English). The languages were selected based on the volume of parallel data available in OPUS. The dataset is split into training, validation, and test partitions. Data was prepared by randomly sampled 1M sentence pairs of English-Korean language pair for training and 2000 each for validation and test. This dataset lacks formality level annotations for the Korean language. To address this, we have categorized the dataset into three formality levels: formal high, formal medium, and informal.

Formality Levels	**Sentence Ending**
Formal High	니다
Formal Medium	요
Informal	다 or 어

Table 2: Formality levels based on Sentence Ending (Contributor, 2022).

Sentences ending with “니다” were labeled as formal high, those ending with “요” were labeled as formal medium, and sentences ending with “다” or “어” were labeled as informal (Contributor, 2022). After incorporating these formality tags, the dataset was reduced to over 300,000 translations, as sentences with unknown formalities were excluded.

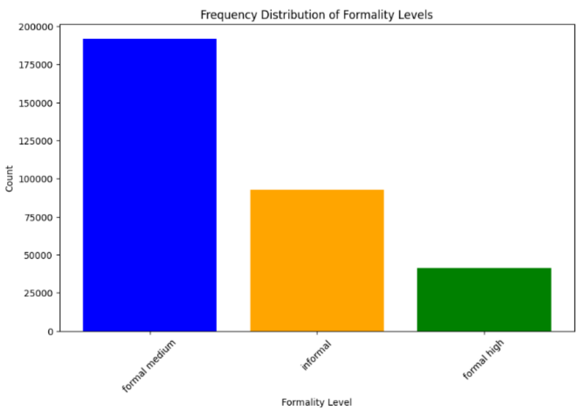

Figure1: Distrubution of sentences based on formality levels

Figure 1 shows the frequency distribution of sentences in the training dataset across different formality levels. The plot reveals that the "formal medium" formality level has the highest frequency, followed by "informal" and "formal high" levels.

3.2 Model Configuration

For this project, we selected the LLaMa 2-7B model as our baseline. The LLaMa (Large Language Model by Meta AI) series of models is renowned for its state-of-the-art performance across various natural language processing tasks. We initially loaded the baseline model using Hugging Face Transformers with default parameters and conducted several translations between English and Korean. The initial performance on translation tasks was suboptimal.

To enhance the model's performance, we fine-tuned the baseline model using Parameter-Efficient Fine-Tuning (PEFT) and Quantized Low-Rank Adaptation (qLoRA) configurations. During this process, we encountered significant training delays with the large dataset, resulting in approximately 7 hours of training time on an A100 Google Colab High-RAM GPU. To address this, we reduced the training dataset size by removing sentences exceeding 100 tokens, which reduced the training time to 4 hours (Marie B., 2023). Additionally, we optimized various training parameters, including batch size per device, maximum token length, qLoRA parameters, prompts, and learning rates, to further decrease training time without compromising accuracy.

Model Parameters	**Value**
Batch size per device	96
Max-token length	120
Prompt	f'{{"text": "{ko_sentence} ###>{en_sentence} ###>{formality}"}}' (Marie B., 2023)
Learning rates	0.0001

Table 3: Final Model Parameters

The final model training time is approximately 3.5 hours, achieving better assessment metrics compared to other trained models. Therefore, this configuration can be considered optimal.

3.3 Model Assessment

We assessed the performance of the baseline model using BLEU, ROUGE-1, ROUGE-2, ROUGE-L, BERT and COMET scores on 100 translations from the test dataset. The results are presented in section four below. The BLEU score is widely used to evaluate translation models, and we will use it to compare the performance of our fine-tuned model to the baseline model. The BLEU score is calculated as follows:

$$BLEU = BP \times \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

Where, BP measures brevity Penalty, and p_n and w_n measures precision.

The BERT score is another powerful metric used in the assessment of language models performing translation tasks. It measures the similarity of the embeddings of the translated text and the reference text, providing a robust evaluation by capturing semantic meaning rather than just surface form similarity. Higher BERT scores indicate that the translated text is semantically closer to the reference text.

The ROUGE (Recall-Oriented Understudy for Gisting Evaluation) metrics include ROUGE-1, ROUGE-2, and ROUGE-L, which measure the overlap of unigrams, bigrams, and longest common subsequences between the generated translations and the reference translations, respectively. These metrics are useful for evaluating the quality of the translations by assessing how much of the reference text's content is preserved in the translations.

- **ROUGE-1** measures the overlap of unigram (single word) between the system and reference translations.
- **ROUGE-2** measures the overlap of bigrams (two consecutive words) between the system and reference translations.
- **ROUGE-L** measures the longest common subsequence between the system and reference translations, focusing on the fluency and coherence of the translations.

The COMET (Cross-lingual Optimized Metric for Evaluation of Translation) score assesses human judgment on translations by evaluating the quality of translations based on semantic similarity and adequacy. It provides a more nuanced understanding of translation quality by incorporating human-like judgment, which is crucial for translations involving complex formality levels and cultural nuances.

By using these diverse metrics, we ensure a comprehensive evaluation of our model's performance. The combination of precision-focused metrics like BLEU and ROUGE with semantic similarity metrics like BERT and human judgment metrics like COMET provides a holistic view of how well the model performs in real-world translation tasks.

4 Results and Discussions

4.1 Model Evaluation

We evaluated the baseline model using BLEU, ROUGE, BERT, and COMET scores. The baseline model was assessed using the same prompt as the fine-tuned model to ensure a fair comparison. This approach also allowed the baseline model to accommodate the various formality levels during the assessment process. Additionally, we considered testing the fine-tuned model both with and without formality levels by adjusting the final prompt during the evaluation process. The prompts were adjusted as follows:

- **Testing with formality levels:** my_text = input_text + "###>" + formality_level

  Prompt = "text:" + f"{my_text}" + " ###>" + "translation:"

- **Testing without formality levels:** "text:" + f"{my_text}" + " ###>" + "translation:"

  Prompt = "text:" + f"{my_text}" + " ###>" + "translation:"

This allowed us to assess the model translations from Korean to English while capturing formality levels and without considering formality. Furthermore, we evaluated the final model's performance across the different formality levels to understand how well it translates from Korean to

English at each formality level. The results of all model evaluations are presented below:

Metrics	Baseline Model with formality levels	Fine-tuned model without Formality levels	Fine-tuned model with Formality levels
BLEU	0.0000	0.0231	0.0152
Rouge-1 (F1)	0.000385	0.229	0.161
Rouge-2(F1)	0.000	0.093	0.079
Rouge-L(F1)	0.000385	0.206	0.147
BERT F1	0.764	0.836	0.817
COMET (Human Judgement)	0.277	0.397	0.414

Table 4: Model Evaluation

Korean Sentence	Baseline Translation	Final model Translation	Formality Level
아, 진짜 힘들어.	sierpni 26, 2020	Oh, it's really hard.	Informal
안녕하세요 안젤라에요 만나서 반갑습니다	paÅºdziernik 23, 2020 2020-10-23 00:00:00 2020-10-23 00:00:00 America/Mexico_City 2020-10-23 00:00:00 2020-10-23 00:00:00	Hello Angela, it's nice to meet you	Formal High
어쨌든, 당신의 임신에 대한 세부사항이 사건의 열쇠인 것 처럼 보이네요	paÅºdziernik 23, 2020 2020-10-23 00:00:00 2020-10-23 00:00:00 America/Mexico_City 901 Marquette Ave, Minneapolis, MN 55402 901 Marquette Ave, Minneapolis, MN 55402	however, the details of your pregnancy seem to be the key to the case	Formal Medium

Table 5: Baseline and Final model translations for three different formality levels.

Metrics	Formal High	Formal Medium	Informal
BLEU	0.02872	0.010285	0.017483
Rouge-1 (F1)	0.1801	0.172318	0.046966
Rouge-2(F1)	0.1282	0.076924	0.079
Rouge-L(F1)	0.1716	0.155399	0.105095
BERT F1	0.783651	0.829586	0.811626
COMET (Human Judgement)	0.405556	0.415000	0.418182

Table 6: Final model evaluation across all formality levels

4.2. Discussion of Findings

The evaluation of our models using BLEU, ROUGE, BERT, and COMET scores provides a comprehensive view of their performance, especially in handling formality levels in Korean language translations.

4.2.1. Baseline with Formality Levels

The baseline model, assessed with formality levels, performed poorly across all metrics. The BLEU score was 0.0000, and the ROUGE scores were also negligible, with ROUGE-1(F1) and ROUGE-L(F1) both at 0.000385 and ROUGE-2(F1) at 0.000. However, the BERT F1 score was relatively high at 0.764, and the COMET score, reflecting human judgment, was 0.277. This indicates that while the baseline model could capture some semantic similarity (as reflected by BERT F1), it failed to generate coherent and accurate translations.

4.2.2. Fine-tuned Model with Formality Levels

The fine-tuned model with formality levels demonstrated a nuanced performance. The BLEU score was slightly lower than the model without formality levels at 0.0152, and the ROUGE scores were also somewhat reduced (0.161 for ROUGE-1, 0.079 for ROUGE-2, and 0.147 for ROUGE-L). However, the BERT F1 score remained high at

0.817, and the COMET score improved to 0.414, indicating better human judgment alignment. This suggests that while incorporating formality levels might introduce some complexity affecting surface-level accuracy (as measured by BLEU and ROUGE), it enhances the model's ability to generate translations that are more contextually and culturally appropriate, as evidenced by higher COMET scores.

Finally, table 6 presents the final model evaluation across three formality levels: Formal High, Formal Medium, and Informal, using metrics including BLEU, ROUGE-1 (F1), ROUGE-2 (F1), ROUGE-L (F1), BERT F1, and COMET (Human Judgement). The BLEU score is highest for Formal High (0.02872), indicating superior performance in translating highly formal sentences compared to medium (0.010285) and informal (0.017483) levels. The ROUGE scores follow a similar trend, with Formal High achieving the best results (ROUGE-1: 0.1801, ROUGE-2: 0.1282, ROUGE-L: 0.1716), followed by Formal Medium and significantly lower scores for Informal. The BERT F1 scores demonstrate high semantic similarity across all formality levels, with Formal Medium scoring the highest (0.829586), and the COMET scores, reflecting human judgment, are slightly better for Informal (0.418182) compared to Formal Medium (0.415000) and Formal High (0.405556). Overall, the model shows robust performance across different formality levels, excelling in Formal High translations according to BLEU and ROUGE metrics, while BERT and COMET scores indicate high semantic similarity and human satisfaction across all levels.

Our findings align with previous research that highlights the challenges and benefits of incorporating formality in translation models. For instance, Sohn (1999) and Brown and Whitman (2015) emphasize the complexity of the Korean honorific system and the importance of accurately capturing formality in translations. Our work extends this by demonstrating that formality-aware models, while initially showing lower BLEU and ROUGE scores, achieve better human-judgment-aligned translations as reflected by COMET scores.

Hu et al. (2021) discussed the effectiveness of Low-Rank Adaptation (LoRA) in fine-tuning large language models efficiently. Our results corroborate this, showing substantial improvements in translation quality with the use of PEFT and qLoRA techniques. Furthermore, Ziegler et al. (2019) highlighted the importance of human feedback in fine-tuning models, which our use of COMET scores strongly supports.

5 Conclusion

This project focused on developing an advanced Natural Language Processing (NLP) model for translating between English and Korean, with a particular emphasis on accurately handling multiple levels of formality in the Korean language. By leveraging the LLaMa 2-7B architecture and fine-tuning it using Parameter-Efficient Fine-Tuning (PEFT) and Quantized Low-Rank Adaptation (qLoRA), we sought to improve the model's performance in translation tasks that require nuanced understanding of formality levels.

We began by preparing a dataset from the OPUS-100 project and labeled the data with three distinct formality levels: formal high, formal medium, and informal. This preprocessing step was crucial for ensuring that the model could learn to recognize and appropriately handle different levels of formality in Korean sentences. The dataset was then filtered to include only sentences with a token length of less than 100, to optimize training efficiency and reduce computational load.

During the model configuration phase, we loaded the baseline LLaMa 2-7B model using Hugging Face Transformers and performed initial translations. The baseline model's performance was suboptimal, prompting us to fine-tune it using PEFT and qLoRA techniques. This fine-tuning process included adjusting various parameters such as batch size, token length, learning rates, and prompt configurations. By iteratively refining these parameters, we were able to significantly reduce training time while maintaining or improving model accuracy.

The final model, trained with the reduced dataset and optimized parameters, achieved a training time of approximately 3.5 hours. We evaluated the model using BLEU, ROUGE, BERT, and COMET scores on 100 translations from the test dataset. The results demonstrated that incorporating formality levels into the training process led to higher BERT and COMET scores, indicating better semantic similarity and human judgment alignment in the translations.

Comparing our findings with previous research, our approach aligns with the insights provided by Sohn (1999) and Brown and Whitman (2015) regarding the complexities of the Korean honorific system. By effectively integrating formality level tags and utilizing advanced fine-tuning techniques, our model addresses the challenges highlighted in previous studies and sets a new benchmark for handling formality in machine translation.

Overall, this project contributes to the field of machine translation by providing a robust framework for incorporating formality levels into translation models. Our results show that formality-aware models can achieve more contextually appropriate translations, which is essential for applications in business, healthcare, and other domains where cultural sensitivity and accuracy are paramount. Future work can build on this foundation by exploring additional formality levels, expanding the dataset, and refining the fine-tuning techniques to further enhance translation quality.

References

Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural Machine Translation by Jointly Learning to Align and Translate. Proceedings of the International Conference on Learning Representations (ICLR). https://arxiv.org/abs/1409.0473

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language Models are Few-Shot Learners. arXiv preprint arXiv:2005.14165. https://arxiv.org/abs/2005.14165

Brown, L., & Whitman, J. (2015). Honorifics and Politeness in Korean. In *Korean Syntax and Semantics* (pp. 154-178). Cambridge University Press.

Contributor, H. H. G. (2022b, October 17). Korean Honorifics and Speech Levels: Why, When, & How to use. Retrieved from https://www.habbihabbi.com/blogs/bilingual-resources/korean-honorifics-speech-levels

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 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), 4171-4186. https://www.aclweb.org/anthology/N19-1423/

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... & Rajpurkar, P. (2021). LoRA: Low-Rank Adaptation of Large Language Models. arXiv preprint arXiv:2106.09685. https://arxiv.org/abs/2106.09685

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., ... & Zettlemoyer, L. (2020). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv preprint arXiv:1910.13461. https://arxiv.org/abs/1910.13461

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2020). RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692. https://arxiv.org/pdf/1907.11692.pdf

Marie, B. (2023, November 2). Llama 2 MT: Turn Llama 2 into a Translation System with QLoRA. Retrieved from https://kaitchup.substack.com/p/llama-2-mt-turn-llama-2-into-a-translation

Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., ... & Auli, M. (2019). fairseq: A Fast, Extensible Toolkit for Sequence Modeling. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations), 48-53. https://www.aclweb.org/anthology/N19-4009/

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research, 21(140), 1-67. http://jmlr.org/papers/v21/20-074.html

Sohn, H. M. (1999). The Korean Language. Cambridge University Press.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30. https://arxiv.org/abs/1706.03762

Tiedemann, J. (2012). Parallel Data, Tools and Interfaces in OPUS. Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), 2214-2218. http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf

Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., ... & Christiano, P. F.

(2019). Fine-tuning language models from human preferences. arXiv preprint arXiv:1909.08593. https://arxiv.org/abs/1909.08593