LyriKOR: English to Korean Song Translation with Syllabic Alignment

Korea University COSE461 Final Project

Eunbeen Hong
Department of Computer Science
Team 2
2021320120

Hyejin Jo
Department of Computer Science
Team 2
2020320159

Jeemin Oh
Department of Computer Science
Team 2
2021320150

Junghwan Park
Department of Cyber Defense
Team 2
2021350205

Abstract

With the rise of globalization and the growing mainstream popularity of cross-cultural tastes in music, it is important to ensure accessibility for overseas fans to understand and sing along to foreign songs. Our model LyriKOR aims to translate English songs to Korean in a way that not only captures the meaning but also somewhat aligns with the rhythmic flow of the original song, allowing Korean listeners to sing along in their native language. With LyriKOR, we introduce a method to train syllabically aligned translation models with limited paired data by using a two part model for translation and syllabic adjustment.

1 Introduction

With the increasing trend of globalism and the growing popularity of foreign music, many people in Korea regularly listen to overseas artists. Following this, there is an increasing demand for singable translated music. In the past, we have seen the successful popularization of Korean dubs for more "classic" music, such as Disney songs or children' s lullabies, but there have been limited attempts to translate popular music on a larger scale, and to the best of our knowledge, these translations have been done by human experts.

The problem of generating singable song translations is twofold. Firstly, the translation must accurately capture the meaning of the song in its original language. That is to say, we cannot change the meaning or the essence of the song in order to fit the rhythmic or melodic structure. Secondly, the translation must also have a similar rhythmic structure to the original song. If the original line was 5 syllables in English, but the Korean translation was 10 syllables, it would be almost impossible to successfully sing the translated output, indicating a poor song translation. It is widely agreed that there is a trade-off between phonetic similarity and semantic accuracy, making the task of balancing rhythmic alignment and semantic accuracy a difficult issue[7, 11]. Thus, in order to generate a successful translation of songs, we must pay close attention to both the meaning and syllable count of our translation.

Prior research has been done to solve similar problems in different media areas, such as in Korean dubs of foreign media such as TV shows or animated movies. However, song translation differs from this task because of the relatively shorter lines and higher importance of syllabic alignment, as well as lower necessity for multi-modal inputs like audio or visual data. Thus, for the problem of song translation, we decided to make a model that is based solely on syllable counts and the lyrics themselves, with no extraneous data other than text data.

We have two main contributions in our model, LyriKOR:

- We addressed the lack of paired data for song translation by creating a proxy dataset to train our syllabic adjustment model.
- We proposed a new, two-part approach to solve the problem of song lyric translation with syllabic bounds. This new model can successfully be trained without English-Korean paired data and perform well during test time on unseen data.

2 Related Work

2.1 Text Translation

There are a plethora of models that deal with machine translation, including English to Korean translation. Moreover, there are a number of pretrained models that deal with English-Korean translation, which is useful for our specific problem of syllabically aligned song translation. Given the prevalence and relative success of pretrained translation models, as well as the considerable amount of computation and data it would take to train a completely new translation model, we decided to use one of these models to solve part of our problem instead of training a new translation model. In particular, we used a KoBART English to Korean translation model [1], which is pretrained on a large corpus of Korean-English text pairs. Moreover, we also used a Korean to English translation model to generate our dataset of syllabically corrected phrases.

2.2 Syllabic Constraints

The more challenging part of our problem was implementing syllabic constraints into our translation. While most text generation problems are free text generation, meaning the generated output length does not matter, there has been some previous research done on constrained text generation.

Syllabic constraints have been explored in AD problems, with previous works enforcing soft constraints to modify output lengths during translation, using verbosity control directly in the translation model, with methods such as N-best rescoring[7]. This idea was previously also mentioned by Saboo and Baumann [11], who used a heuristic with a weighted average of subscores for phonetic synchrony and translation faithfulness for dubbing-optimized translation. This approach of using subscores to account for both translation faithfulness and translation length results in a model that does not tightly adjust for rhythm and incurs tradeoffs in length in order to preserve meaning accuracy.

Our approach differs from these approaches for several reasons. Firstly, the problem of AD, while closely related, differs from our problem of song translation in both necessary input data and goal flexibility. Moreover, phonetic sounds or tones do not matter as much, meaning there is less importance in audio and visual data, which is heavy and requires more computational resources. Furthermore, syllabic alignment requirements are less stringent in AD problems, so the models above have relatively flexible syllabic constraints. Secondly, there is a much wider availability of paired data in dubbing, which makes it easier to find paired training and testing data.

Because of these differences, we elected to first translate the English lyrics to Korean, then input the translated lyrics with a goal length token to get the final output. This is similar to the idea used by Style Transformer[3], which takes input text and style token to generate modified text based on the style token. Our intention with this was to 1) create a model with stricter enforcement of syllabic alignment, which would be more appropriate for singable

song translation, 2) to train a model with limited data and 3) make use of existing pretrained models in order to lessen the computational burden and use of resources.

2.3 Song Translation

There has been some research in the area of singable song translation. Previous work on automatic song translation for Chinese [6] has proposed some baseline evaluation criteria and model structure. For model evaluation, three main criteria of meaning, singability, and intelligibility were proposed. We chose to adopt similar criteria, evaluating our model on two standards– meaning and singability. In terms of approach, Guo' s model requires more detailed input data, because the output language is Chinese, which includes extra conditions for pitch and tone.

We build on the syllable-to-syllable method of enforcing rhythmic alignment in our model and cut down on unnecessary data by removing the melody input, as Korean is not tonal and does not have such requirements.

3 Approach

Our model is a two part model that splits the task into two simpler subtasks: 1) unconstrained translation and 2) syllabic adjustment of the translated text to generate the final output of syllabically aligned translated text. A simple diagram of our model structure is provided in Figure 1.

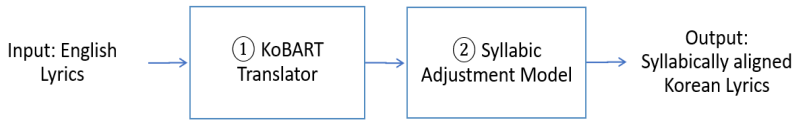

Figure 1: Model Structure Overview Diagram

Our first subpart uses the pretrained KoBART translator to take an input of English lyrics and output an unconstrained Korean translation that accurately captures the meaning. Our second subpart is the Syllabic Adjustment Model. This model takes the input of a Korean phrase and a goal length token and outputs a Korean phrase with the same meaning but syllables to fit the goal length. This second model is trained separately with our generated dataset, and the two models for each subpart are combined in a single pipeline during test time to create a single syllabically aligned English to Korean lyric translation model.

3.1 Unconstrained Machine Translation

The first section of our model simply uses a pretrained machine translation model to generate Korean text that accurately captures the meaning of the original English lyrics. We used pretrained KoBART translation model for the translation task[1] to generate unconstrained Korean translations of English lyrics. The reason why we chose the model is that it is light with good performance.

3.2 Syllabic Adjustment

The second section of our model takes the tokenized output from section 1 combined with a goal length token as input, and aims to output a modified phrase with the same meaning, but with adjustments to fit the goal syllabic length. A diagram of our data generation and model training process of provided in Figure 2.

The top figure shown in Figure 2 is our model for dataset generation, which is only used during training, and the bottom figure is our Syllabic Adjustment Model.

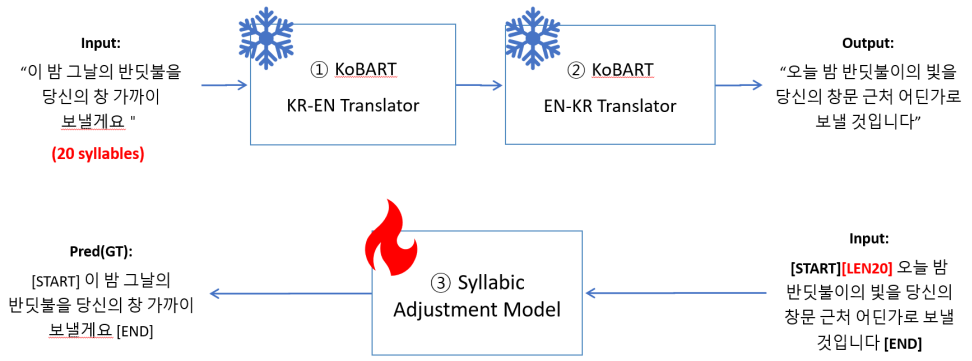

Figure 2: Model Structure of Syllabic Adjustment. ❄: Freeze the parameters, 🔥: Fine-tune the parameters of models.

3.2.1 Generation of Training Data

We used Korean lyrics data translated to English and retranslated back to Korean. We decided to use KoBART Korean-English[2] and English-Korean translation models[1] when creating our proxy dataset in order to create the closest distribution to our predicted test data. To include syllable information in our data, We added [LEN<number of syllables>] token after the [START] token, creating input data with the following format:

[START][LEN<number of syllables>]our corpus(Korean)[END]

See 4.1 for more detail on dataset generation.

3.2.2 Syllabic Adjustment Model Architecture

Our Syllabic Adjustment Model, while simpler than a length-constrained translation model, still required significant pretraining to understand the relationship between the length token and goal output. Therefore, instead of building a vanilla transformer model from scratch, we elected to modify and fine-tune a pretrained KoBART model[13], which is a BART [8] based model pretrained on over 40GB of Korean text corpus.

Our model has encoder-decoder architecture, where the encoder and decoder are Sequence-to-Sequence Transformers with 6 layers each. There are 16 attention heads, with a feed-forward dimension of 3072 and a hidden dimension of 768. To train, we used Cross Entropy Loss, as used in BART[8], as shown below:

$$Loss(ours) = -\sum\sum p(x)log(q(x)) \tag{1}$$

p(x) is the true probability distribution, and q(x) is the predicted probability distribution, which is summed over all classes and tokens. Please refer to 4.1.2 for training details.

4 Experiments

4.1 Experimental Details

4.1.1 Training Data

To train our model to translate English lyrics to Korean, the most essential data is a paired English-Korean lyrics dataset, preferably with matched syllable counts, as the syllabic alignment is important for translated lyrics to be sung along with the original melody. However, there is no quality paired translation dataset with matched syllables. This is why we decided to take a different approach, and split our model into two parts, allowing us to create our own training dataset.

In order to get the Korean lyrics, we collected online lyric data from Melon, the most popular music service in Korea, by web crawling. We collected Korean lyrics in the monthly Top-100 chart starting from January 2000 to April 2023 and indie music. After removing duplicates, we were able to collect around 170,000 lines of lyrics from approximately 8,000 songs. We removed foreign songs that were not in Korean, such as Japanese or Chinese songs, but kept Korean songs that included English lyrics.

These original Korean lyrics were used as the ground truth during training, and we translated these lyrics into English with a pretrained language model, then retranslated back to Korean, which we used as the sample input data. This gave us a paired dataset where we had Korean lyrics which were translated from English as input data, and well written Korean lyrics as the goal output. This data was used to train our Syllabic Adjustment Model, which adjusts the syllables of translated Korean lyrics to match the necessary rhythm.

4.1.2 Training Details

Because the first half of our model is a pretrained translation model, we froze the parameters. The training details below pertain to our Syllabic Adjustment model.

We trained the Syllabic Adjustment Model with a batch size of 512 and ran 30 epochs. We used the AdamW optimizer[9] with 3e-5 learning rate. We also used a cosine scheduler with 0.1 warmup ratio, and gradient clipping of 1.0.

We trained and ran our model in the Google Colaboratory environment, using A100. On a dataset of 170,000 lines, training took approximately 4 minutes per epoch, and 2 hours total for all 30 epochs, which is quite efficient considering the task and results.

We achieved a validation loss of 0.36 when training with 110,000 lines of data, and a validation loss of 0.27 when training with 170,000 lines of data. In both cases, improvements in validation loss stagnated after approximately 15 epochs.

We trained our model in Google Colab and used NVDIA A100.

4.2 Evaluation Method

In order to evaluate the success of our model, we decided to use both quantitative and qualitative evaluation methods. Quantitative evaluation provides us with concrete assessments on task success in translation and syllabic alignment. However, due to the subjective and artistic nature of music, we assessed that qualitative evaluation by humans is also necessary to judge the singability and appropriateness of our results.

4.2.1 Quantitative Evaluation

We conducted quantitative evaluations of our model in two areas: syllabic alignment and semantic accuracy. Because there was no paired dataset to provide a ground truth goal for comparison, we decided to use two separate evaluation methods to assess our model.

In order to measure our syllabic alignment, we performed a quantitative evaluation comparing our output syllable count to the original English syllable count. To calculate syllabic alignment score, we calculated the absolute value of the syllabic differences over the goal syllable count for each line as shown in equation (1). This score gave us an idea of how well our model adjusts phrases to fit the rhythm.

$$\frac{|syll(English) - syll(translated\ Korean)|}{syll(English)} \tag{2}$$

In order to evaluate translation accuracy, the standard evaluation method used by previous works [11][7] seems to be a Bilingual Evaluation Understudy (BLEU) score. However, evaluation with BLEU requires us to have a ground truth input in Korean to compare our results with. This presents an issue because there is no acceptable scale paired dataset that would allow us to evaluate our translation accurately using BLEU. Therefore, we used the BERTScore [15] which we found to be the most appropriate evaluation model as it evaluates semantic similarity between phrases using the cosine similarity of token embeddings

rather than n-gram overlap or other strategies that are not as effective for comparing the similarity of phrases in different languages. We translated N songs from an online dataset of unpaired English song lyrics, then assessed the BERTScore using our predicted output and the original input.

4.2.2 Qualitative Evaluation

Because the quantitative evaluation is not able to fully assess more subjective aspects of our results, we decided to also perform a qualitative evaluation. Our qualitative evaluation aimed to assess translation accuracy and general singability more holistically, taking into account human preferences and interpretation. We performed our qualitative evaluation in two parts: meaning accuracy and singability. In order to perform our qualitative evaluation, we asked a random pool of 20 people to be a part of the evaluation panel. The panel was given a random pop song, and the translated Korean version generated by our model then asked to rate the translation on two metrics: 1) translation accuracy and 2) singability. Furthermore, we asked the panel for feedback and opinions on the strengths and weaknesses of the translation.

4.3 Results

To measure our results during test time, we used an unseen, unpaired test dataset of English lyrics, and ran it through our model to generate syllabically aligned Korean lyrics. Below is an example of a randomly sampled translated English lyric from test time:

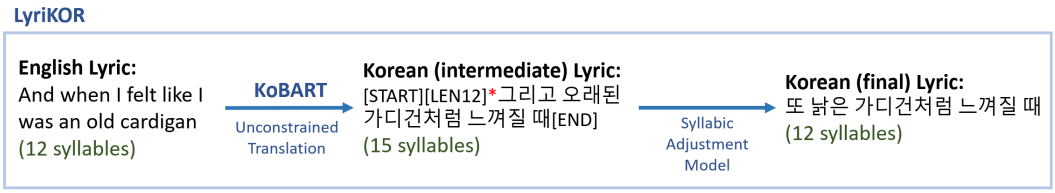

Figure 3: Sample Test Time Generated Result. *[LEN20] token refers to length token 12

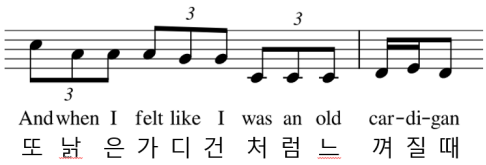

Figure 4: Sample lyrics matched with music score

4.3.1 Quantitative Evaluation

Figure 5 shows the box graphs of our evaluation results. In BERT Score[15], over 75 percent of our prediction scored over 0.65. "Look at this stuff" and "이것 좀 봐" receives a Bert Score of 0.7, suggesting a score over 0.65 can be interpreted as a highly accurate result, which means our model preserves semantic meanings. In Syllable Evaluation (4.2.1), most of our outputs are nearly zero. This indicates our Syllabic adjustment model can generate accurate outputs. There are some outliers, with errors up to 15 syllables long, which may be due to an error in translation, or because the model could not successfully adjust for length, resulting in a blank output.

4.3.2 Qualitative Evaluation

While the quantitative evaluation above provides us with a rough idea of how well our model performs rhythmically, we believe that ultimately, qualitative evaluation is a better indicator

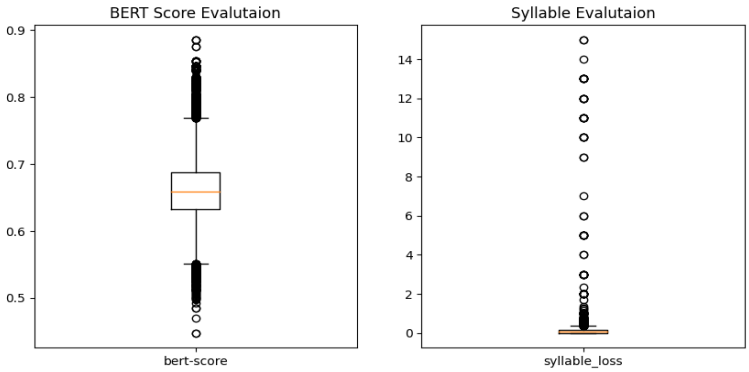

Figure 5: (Left) The result of BERTScore, (Right) The result of Syllable Evaluation.

of performance as it is better nuanced. Of the 20 people in our panel, 15 stated that the translation was of acceptable quality, 3 were unsure, and 2 judged the translation to be of low quality. The average rating of our model was 9/10 for translation accuracy and 8.7/10 for singability. Both of these ratings show a decent standard of success for the criteria of meaning accuracy and singability.

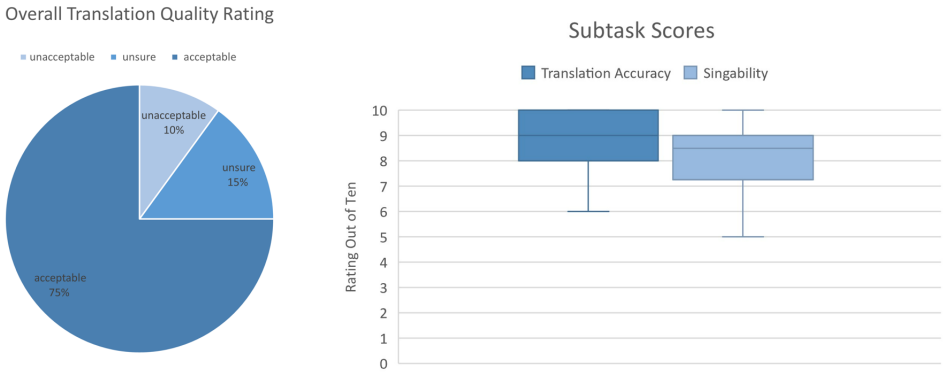

Figure 6: (left) Responses of 20 Panel Members When Asked to Judge the Overall Quality of the Translation as acceptable, unsure, or unacceptable. (right) Score Distribution on Criteria of Translation Accuracy and Singability, Judged by Panel Members.

When we asked for more qualitative feedback from the panel, the general consensus was that while the syllable constraints made the song easier to sing along to the melody on a line-by-line scale, some awkwardness resulted from the placement of notes in the melody. If certain notes were held for longer, the singability of the line was hindered as our model did not consider note length or other audio information. In addition, sometimes Korean words were repeated consecutively in order to match the syllable counts. We suspect that this occurred because of our emphasis on syllabic alignment. Moreover, while the translation meanings were overall mostly accurate, they were sometimes awkward or grammatically incorrect. This seems to be a common limitation for machine translation problems that are more artistic and have a less defined syntax to follow.

5 Analysis

5.1 Comparisons to Past Approaches

Many previous works stated that there is a direct tradeoff between length accuracy and translation accuracy, which is explicitly coded into loss functions, which use weighting hyperparameters to control and prioritize one or the other. While this tradeoff is unavoidable

for constrained translation problems, even when done by human experts, by separating the two parts so that there are two separate models that run sequentially, we could reduce the inevitability of this tradeoff. Because we used one model for accurate translation, and another model simply for length adjustment of a phrase, the task of length adjustment was simpler than translation with length constraints. By successfully accomplishing simpler tasks rather than attempting one complex one, we were able to maintain better quality translations and reduce sacrifices in both meaning and length accuracy.

Secondly, because we split the task into two subtasks, one of which is the extremely well-explored task of machine translation, we were able to shrink the burden on our computational resources by using a pretrained model for half of our task, allowing us to reduce training to a comparatively simpler task.

5.2 Limitations

However, our approach has some limitations as well. First, because we rely on two separate models for translation and length adjustment, with the output of the first model being inputted into our second model, there is a strong reliance on the first part of the model. As a result, for lyrics that have partial meaning, the translation model modifies the original meaning, causing the final lyric output to be inaccurate. While this is a potential issue, it was not critical to our model performance, because the pretrained translation model used for the first subpart showed reliably good performance.

Moreover, there is a disparity between our training data and our test data. Our generated ground truth for the Syllabic Adjustment Model was Korean lyrics and sample input data was Korean-English-Korean re-translated lyrics, to preserve similar meaning but create different syllable counts. While our aim was to create data that had as close a distribution to the real testing data as possible, the data distribution and characteristics for translated English lyrics used during training would inevitably have been different from those of human-written English lyrics. This difference, if significant, would have a significantly negative impact on our model performance.

6 Conclusion

In our paper, we have proposed a new method to translate English songs to Korean. While other methods in the past combined translation and verbosity control, we proposed a method to divide and conquer, translating first and then adjusting syllables to fit constraints. The largest obstacle to overcome was the lack of paired dataset that with accurate syllable counts, and we worked around this challenge and were able to create a paired training dataset by translating Korean songs to eliminate the need for syllabically matched English-Korean paired lyrics. While we successfully generate tighter syllabic bounds with preserving meaning for lyrics, our model also had several limitations, such as dependence on the quality of the pretrained translation model and potential disparities between our training dataset and testing data distribution. In future works, we would like to train our model with a larger dataset, and collect at least some human-translated English-Korean dataset to minimize distributional differences from input data for better performance. Additionally, as our model currently translates the song line by line, it loses the general context of the entire song and has difficulty translating lyrics with partial meanings. In the future, we could address this by modifying our model to receive the whole lyrics to utilize the global context over the full song when translating.

References

[1] Circulus. circulus/kobart-trans-en-ko-v2. `https://huggingface.co/circulus/kobart-trans-en-ko-v2/tree/main`, 2023.

[2] Circulus. circulus/kobart-trans-ko-en-v2. `https://huggingface.co/circulus/kobart-trans-ko-en-v2/tree/main`, 2023.

[3] N. Dai, J. Liang, X. Qiu, and X. Huang. Style transformer: Unpaired text style transfer without disentangled latent representation, 2019.

[4] M. Ghazvininejad, Y. Choi, and K. Knight. Neural poetry translation. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2* (*Short Papers*), 2018.

[5] M. Ghazvininejad, X. Shi, Y. Choi, and K. Knight. Generating topical poetry. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016.

[6] F. Guo, C. Zhang, Z. Zhang, Q. He, K. Zhang, J. Xie, and J. Boyd-Graber. Automatic song translation for tonal languages. *Findings of the Association for Computational Linguistics: ACL 2022*, 2022.

[7] S. M. Lakew, M. Federico, Y. Wang, C. Hoang, Y. Virkar, R. Barra-Chicote, and R. Enyedi. Machine translation verbosity control for automatic dubbing. *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), 2021.

[8] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, 2019.

[9] I. Loshchilov and F. Hutter. Decoupled weight decay regularization, 2019.

[10] J. Lu and M. Eirinaki. Can a machine win a grammy? an evaluation of ai-generated song lyrics. In *2021 IEEE International Conference on Big Data* (*Big Data*), pages 4896–4905, 2021.

[11] A. Saboo and T. Baumann. Integration of dubbing constraints into machine translation. In *Proceedings of the Fourth Conference on Machine Translation* (*Volume 1: Research Papers*), pages 94–101, Florence, Italy, Aug. 2019. Association for Computational Linguistics.

[12] Z. Sheng, K. Song, X. Tan, Y. Ren, W. Ye, S. Zhang, and T. Qin. Songmass: Automatic song writing with pre-training and alignment constraint, 2020.

[13] skt ai. skt-ai/kobart. `https://github.com/SKT-AI/KoBART`, 2020.

[14] S.-H. Son, H.-Y. Lee, G.-H. Nam, and S.-S. Kang. Korean song lyrics generation by deep learning. In *Proceedings of the 2019 4th International Conference on Intelligent Information Technology*, ICIIT '19, page 96–100, New York, NY, USA, 2019. Association for Computing Machinery.

[15] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi. Bertscore: Evaluating text generation with bert, 2020.

A Appendix: Team contributions

All team members participated in the initial brainstorming and research. Active collaboration was key to analyzing the characteristics of our problem and deciding which models and approaches to pursue. Ultimately, we decided to explore two different approaches: the more traditional approach of modifying the translation to consider output length, and our more original idea of dividing the problem into two subtasks in order to make use of pretrained models and simplify each task. Individual contributions are as follows:

Eunbeen Hong: main contributor to dataset creation, collected the lyrics dataset by web crawling through melon, and contributed code for test time evaluation

Hyejin Jo: main contributor for training model, processed data to translate corpus to English and Korean, and combined individual members' code contributions for final training dataset

Jeemin Oh: main contributor to writing the final report, analysis, related works, code for syllable counter for Korean and English lyrics, processed data to include tokens for input into syllabic adjustment model, and contributed to evaluation model

Junghwan Park: researched loss functions for reordering translation candidates, trained model that attempted to utilize alternative end-to-end approach, and contributed to model code and test time evaluation code