Kosmic: Korean Text Similarity Metric Reflecting Honorific Distinctions

Yerin Hwang[1] **Yongil Kim**[2] **Hyunkyung Bae**[3]
Jeesoo Bang[3] **Hwanhee Lee**[4†] **Kyomin Jung**[1,2,5†]
[1]IPAI, Seoul National University [2]Dept. of ECE, Seoul National University
[3]LG AI Research [4]Chung-Ang University [5]SNU-LG AI Research Center
{dpfls589, miles94, kjung}@snu.ac.kr
{hkbae, jeesoo.bang}@lgresearch.ai, hwanheelee@cau.ac.kr

Abstract

Existing English-based text similarity measurements primarily focus on the semantic dimension, neglecting the unique linguistic attributes found in languages like Korean, where honorific expressions are explicitly integrated. To address this limitation, this study proposes Kosmic, a novel Korean text-similarity metric that encompasses the semantic and tonal facets of a given text pair. For the evaluation, we introduce a novel benchmark annotated by human experts, empirically showing that Kosmic outperforms the existing method. Moreover, by leveraging Kosmic, we assess various Korean paraphrasing methods to determine which techniques are most effective in preserving semantics and tone.

Keywords: Evaluation Metric, Text Similarity Measurement, Korean language processing

1. Introduction

Text similarity measurement is a critical task in NLP and plays an essential role in diverse applications such as document retrieval, machine translation, recommendation systems, and document matching (Wang and Dong, 2020). Research in this area can be broadly categorized into two aspects: quantifying text distance using various methods (Deza et al., 2009; Nielsen, 2010) and determining text similarity through effective text representation (Pennington et al., 2014; Osman and Barukub, 2020).

However, although existing methods perform well in English (Jiang et al., 2019; Jeyaraj and Kasthurirathna, 2021), they may not adequately capture the unique linguistic characteristics of other languages that possess features distinct from English. This is particularly evident in languages such as Korean, where the relationship between speakers is explicitly reflected in the text using both honorific and casual speech forms (Kim et al., 2018). In Korean, the use of honorifics and informal speech varies depending on the situation and relationship, potentially reflecting vital societal dynamics such as hierarchical structures, intimacy, or courtesy (Hwang et al., 2021). On the other hand, in English, such societal dynamics are often not explicitly reflected within the language but are conveyed through contextual implications. Consequently, English-based methods might miss these nuances, potentially resulting in distorted context and the omission of significant aspects of the text. For example, as shown in Figure 1, two sentences may appear identical in English due to their shared semantics but vary in Korean formality and their usage. Relying solely on semantics may lead to the overlooking of these critical differences.

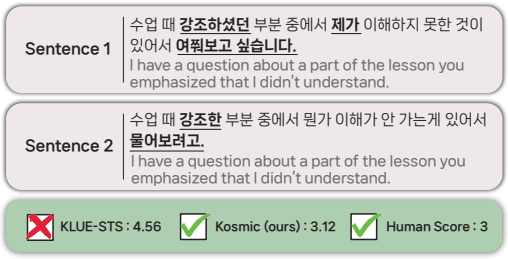

Figure 1: Although two Korean sentences are translated into identical English phrases, they differ in terms of formality, politeness, and usage. Unlike existing Korean STS metric, our Kosmic metric considers honorific distinctions in its scoring.

To address this issue, we propose **Kosmic**, a novel **Ko**rean text **s**imilarity **m**etr**ic** that captures both the semantic and tonal dimensions of a given text pair. As part of the Kosmic scoring framework, we develop a *semantic-similarity model* and *two tone-similarity models* tailored to specific datasets and tasks. The first model operates similarly to existing text-similarity measurement models and is designed to evaluate the semantic similarity of a pair of texts. The subsequent two models consist of one trained in a classification setting and another in a contrastive setting, designed for the evaluation of tone similarity. The scores from these three models are combined to obtain the Kosmic score.

Furthermore, to evaluate the effectiveness of the proposed metric, we introduce a novel benchmark, KTSEval1k. Existing Korean text similarity datasets (Ham et al., 2020; Park et al., 2021) are labeled with an exclusive focus on semantic aspects, making it impossible to measure the efficacy of a metric for distinguishing the tone of the text. Thus,

†Corresponding authors.

LREC-COLING 2024, pages 9954–9960

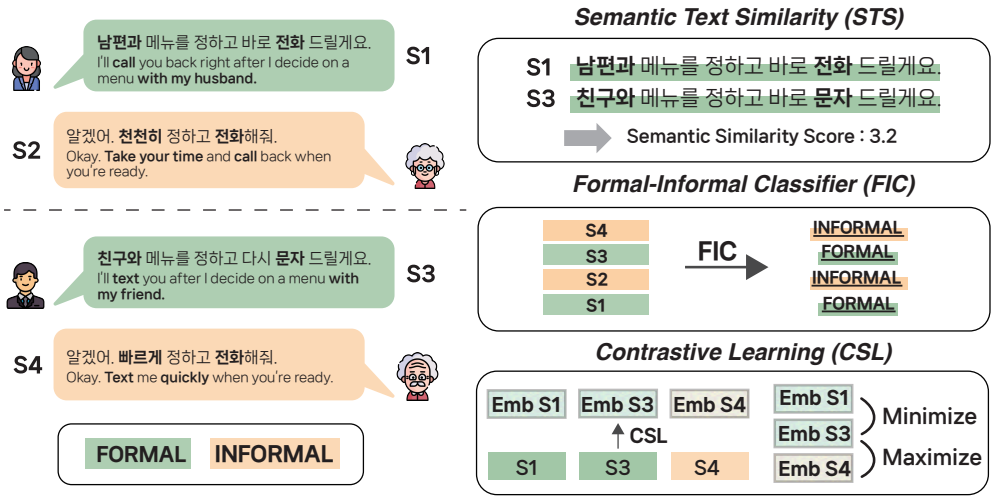

Figure 2: An overview of the training methods of three models for the Kosmic score: the Semantic Measurement Model (STS), the Formal-Informal Classification Model (FIC), and the Contrastive Learning Model (CSL).

we create an evaluation dataset in which humans annotated the similarity of given text pairs considering both semantic and tonal aspects. Utilizing KTSEval1k, we demonstrate that Kosmic shows a notably high correlation with human judgment for measuring Korean text similarity compared to the baseline metric. Furthermore, we highlight the utility of Kosmic by analyzing the efficacy of various Korean paraphrasing methods in preserving semantic and tonal aspects of Korean texts.

2. Backgrounds

2.1. Text Similarity Measurement

Measuring textual similarity extends beyond merely examining the lexical dimension using simple sequences of characters (Gomaa et al., 2013; Qurashi et al., 2020). It aims to encapsulate both semantic and contextual nuances to determine the degree of resemblance between text pairs (Benedetti et al., 2019). Current research on text similarity focuses on measuring distances, such as length and semantic distance (Rahutomo et al., 2012; Dice, 1945), and enhancing text representation through string-based, corpus-based, and graph-structure-based approaches (Islam and Inkpen, 2008; Tomita et al., 2004). However, conventional methods often fail to capture nuanced contextual similarities in languages such as Korean, which explicitly incorporate honorific expressions.

2.2. Korean Honorific Systems

While English text implicitly conveys formality, Korean does so explicitly, mainly by differentiating between formal and informal tones (Sohn, 2005). Among formal expressions, Korean honorifics stand out for their complexity and diversity (Ku et al., 2014), necessitating the careful selection of the appropriate formality level based on the situation and its counterparts (Brown, 2011; Lee et al., 2023). Recognizing and understanding these nuances is a significant aspect of Korean culture when engaging in conversation or writing (Brown, 2015). This paper introduces a new metric for text similarity that effectively captures both formality in Korean and its semantic aspects.

3. Methodology

In this study, we propose a new Korean text similarity metric called Kosmic, which is designed to effectively capture both semantic and tonal aspects. The Kosmic framework comprises three distinct models: a semantic-similarity model (§3.1) and two tone-similarity models (§3.2 and §3.3).

3.1. Semantic Text Similarity Model (STS)

The **s**emantic **t**extual **s**imilarity Model (**STS**) employs regression to learn the semantic similarities between text pairs. As depicted in Figure 2, it considers two Korean sentences as inputs and outputs the semantic similarity between them. The loss is:

$$L(\theta_{_{STS}}) = \frac{1}{N}\sum_{i=1}^{N}(y_i - h_{\theta_{_{STS}}}(x_i))^2,$$

where N is the batch size, x_i is the input text pair, y_i is the ground-truth label, and $h_{\theta_{_{STS}}}(x_i)$ is the prediction score.

Subsequently, during the evaluation phase, the STS model outputs the semantic similarity between

two sentences as a score on a scale of zero to five.

$$STS(x_i, x_j) = w * h_{\theta_{STS}}((x_i, x_j)),$$

where x_i and x_j represent the two Korean sentences, and (x_i, x_j) signifies the pair. w is set to 5.0 to ensure the score scale.

3.2. Formal-Informal Classifier (FIC)

To learn the tonal aspects of the sentences, we devise two distinct models: The first model, as depicted in Figure 2, is the **f**ormal-**i**nformal **c**lassifier (**FIC**) that learns to classify each Korean sentence as either formal or informal through binary-label classification. This model is designed to gather tone information of individual sentences before being considered as pairs. Consequently, it is trained using binary cross-entropy loss:

$$L(\theta_{FIC}) = -\left(y \log(\hat{y}) + (1-y) \log(1-\hat{y})\right),$$

where y represents the true binary label, taking values of either 0 or 1, $\hat{y}$ represents the predicted probability $h_{\theta_{FIC}}(x)$, which takes values between 0 and 1.

During subsequent evaluations, the trained parameter θ_{FIC} is used to calculate the absolute difference between the logit values of the two input Korean sentences. The scoring method employed in the FIC model is detailed below:

$$FIC(x_i, x_j) = w * (1 - |h_{\theta_{FIC}}(x_i) - h_{\theta_{FIC}}(x_j)|),$$

where x_i and x_j represent the two Korean sentences and w is set to 5.0.

3.3. Contrastive Learning Model (CSL)

To measure the similarity in honorific tones between two Korean sentences, a **c**ontra**s**tive **l**earning model (**CSL**) is trained via contrastive loss. As illustrated in Figure 2, sentences with matching tones are treated as positive samples, whereas those with differing tones are treated as negative samples. Through contrastive learning, the model learns the distances between the respective embeddings (Le-Khac et al., 2020). The contrastive loss employed to train the CSL model is described as follows (Chen et al., 2020):

$$L(\theta_{CSL}) = \frac{1}{2N} \sum_{i=1}^{N} \left[(1 - Y_i) \cdot \frac{1}{2} \cdot D_i^2 + Y_i \cdot \frac{1}{2} \cdot (\max(0, m - D_i))^2 \right],$$

where N represents the batch size, Y_i is 1 for positive samples and 0 for negative samples, D_i represents the distance, and m is the margin value. The cosine similarity is used for the distance metric D.

Subsequently, during the evaluation phase, the cosine similarity between the embeddings of the two input Korean sentences is utilized to determine the score. Therefore, the CSL score is expressed as follows:

$$CSL(x_i, x_j) = w * max(cos(h_{\theta_{CSL}(x_i)}, h_{\theta_{CSL}(x_j)}), 0),$$

where x_i and x_j represent the two Korean sentences and w is set to 5.0.

Finally, the Kosmic score is a combination of the three aforementioned scores:

$$Kosmic = \lambda_{STS} * STS + \lambda_{FIC} * FIC + \lambda_{CSL} * CSL.$$

4. KTSEval1K

We introduce a novel Korean text similarity benchmark, KTSEval1k, to determine whether the Kosmic metric adequately captures semantic and tonal nuances. Most existing Korean text similarity datasets (Ham et al., 2020; Park et al., 2021) are constructed primarily using back translation, which does not guarantee the preservation of tonal nuances. However, these datasets overlook this aspect, concentrating solely on semantic attributes when assigning text similarity scores. Consequently, they may not effectively evaluate whether a Korean text similarity metric measures similarity by considering the contextual aspects of paired texts.

Therefore, to address this limitation, we create a benchmark that takes into account contextual aspects. Crowd workers are instructed to craft new sentences by adjusting both semantics and tone using reference sentences, thereby creating text pairs. Other crowd workers annotate the similarities between pairs on a scale of 1 to 5 by considering both semantic and tonal nuances. Furthermore, each sentence is labeled for formality, either formal or informal. The dataset consists of 1,000 text pairs, and we design it to ensure a uniform distribution of formality between the reference sentences and their paraphrased counterparts.[1]

5. Experiments

5.1. Model Implementations

The STS, FIC, and CSL models employ KLUE-BERT (Park et al., 2021) as their backbone, a pre-trained model optimized for Korean sentence language modeling. For the STS model, the KLUE-STS training set (Park et al., 2021) is used, which comprises 50k sentences. Furthermore, for the FIC and CSL models, the training datasets are sourced

[1]The examples, statistics, and annotation details of KTSEval1k will be included in the appendix.

	KTSEval1k		KTSEval1k			Smile		
	ρ	τ	$Acc.$	$F1$	P	$Acc.$	$F1$	P
KLUE$_{STS}$	0.184	0.139	0.523	0.671	0.512	0.507	0.688	0.504
Kosmic$_{FIC}$	0.753	0.652	0.795	0.812	0.752	0.798	0.821	0.769
Kosmic$_{CSL}$	0.681	0.589	0.747	0.782	0.705	0.775	0.801	0.728
Kosmic	**0.755**	**0.653**	**0.798**	**0.817**	**0.756**	**0.806**	**0.824**	**0.778**

Table 1: The experimental results to validate Kosmic's efficacy in measuring Korean text similarity: Human correlation results (left) and classification setting results (middle and right).

from XPersona (Lin et al., 2020) and AI Hub [2]. The formal and informal sentence datasets each comprise 120k samples.

5.2. Experimental Setup

Using the human-annotated text similarity score in KTSEval1k, which considers both the semantic and tonal similarities of Korean sentences, we evaluate the correlation between the scores output by each metric using Pearson correlation (ρ) (Cohen et al., 2009) and Kendall tau correlation (τ) (Sen, 1968) metrics. The Pearson correlation quantifies linear connections, while the Kendall Tau correlation, which relies on rankings, evaluates the magnitude of monotonic relationships. In addition, to specifically assess the ability of each metric to accurately detect the formality of a text pair, we conduct experiments in a classification setting using two test datasets. Both the KTSEval1k and Smile datasets (Kim, 2022) contain labels indicating whether each sentence in a text pair is formal or informal. Setting a classification threshold of 2.5 since each metric outputs a formality score between 0 and 5, we evaluate each metric using accuracy, F1 score, and precision. For the two experimental setups, a comparative analysis of four models is conducted: KLUE$_{STS}$, Kosmic $_{FIC}$, Kosmic $_{CSL}$, and Kosmic. For Kosmic $_{FIC}$, λ_{CSL} is set to 0, while conversely, λ_{FIC} is set to 0 for Kosmic $_{CSL}$. Kosmic is a model optimized through hyperparameter tuning, utilizing the three best-performing λ values.

5.3. Results

When analyzing the human correlation results from KTSEval1k, the baseline metric exhibits a notably low human correlation due to its inability to detect honorific distinctions, as shown in Table 1. For both Kosmic $_{FIC}$ and Kosmic $_{CSL}$, the human correlation scores increased owing to the measurement of tone similarity. Furthermore, Kosmic, which leverages both the FIC and CSL models to measure tone similarity, achieves the highest score. In addition, the evaluation of classification setups exhibits a similar tendency. The STS-exclusive model performs at near-random guess levels, but both Kosmic $_{FIC}$ and Kosmic $_{CSL}$ showcase improvements. Moreover, Kosmic demonstrates the most outstanding performance. These findings suggest that Kosmic effectively reflects semantic and tonal aspects, positioning it as a robust metric for assessing Korean text similarity.

[2]https://www.aihub.or.kr/

Method	Model	KLUE$_{STS}$	Kosmic
Back translation	**GoogleAPI**	4.00	**3.04**
	GPT-3.5	3.77	**3.27**
	GPT-4	3.87	**3.16**
Direct paraphrase	**GPT-3.5**	3.91	3.81
	GPT-4	3.98	3.90

Table 2: The experimental results of various Korean paraphrasing methods on semantic and tonal preservation.

6. Analysis

6.1. Benchmarking Paraphraing Methods

We demonstrate the utility of Kosmic by analyzing the effectiveness of various Korean paraphrasing methods in preserving both semantics and tone during the paraphrasing process. While preserving tone is not always crucial during paraphrasing, there are numerous situations in which it is essential for upholding the text's full context. Therefore, the detection of tone preservation is vital.

We compare two widely used Korean paraphrasing methods: back translation (Sennrich et al., 2015) and direct paraphrasing using large language models (LLMs) (Brown et al., 2020). Back translation has proven particularly effective, especially in the dialogue domain, as it captures context more adeptly than other techniques (Kulhánek et al., 2021). Furthermore, the integration of LLM capabilities for downstream tasks has recently attracted significant attention, given its proven effectiveness across diverse tasks, including paraphrasing (Witteveen and Andrews, 2019; Tang et al., 2023).

As shown in Table 2, we observe that although the back translation methods effectively paraphrase semantic aspects, they tend to lose information con-

cerning tone. Conversely, the direct paraphrasing methods utilizing LLMs not only perform well in preserving semantics but also prove effective in maintaining tone during paraphrasing.

7. Conclusion

This paper introduces Kosmic, a Korean text similarity metric designed to effectively encapsulate the semantic and tonal nuances in the Korean language. Utilizing the newly crafted dataset, KTSEval1k, we demonstrate that Kosmic exhibits a higher correlation with human judgment compared to the baseline that assesses only the semantic dimension. Moreover we compare Korean paraphrasing techniques using Kosmic and provide directions for future research leveraging Korean text similarity.

Ethic Statements

To ensure that the created dataset is free from ethical concerns, crowd workers are instructed to check that there weren't any offensive, sexist, or racist remarks; harmful language; or references to sexual conduct. These workers are fairly compensated, with wages exceeding $12 per hour in USD. Moreover, we utilize the translation models and LLMs from the official sites[3][4]. All models and datasets leveraged in our studies are obtained from publicly available websites or GitHub repositories. Our code and dataset will be made public.

Acknowledgements

This work was supported by LG AI Research. This work was partly supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) [NO.2021-0-01343, Artificial Intelligence Graduate School Program (Seoul National University) & NO.2021-0-02068, Artificial Intelligence Innovation Hub (Artificial Intelligence Institute, Seoul National University)], the BK21 FOUR program of the Education and Research Program for Future ICT Pioneers, Seoul National University in 2023, and the National Research Foundation of Korea (NRF) grant funded by the Korea government (No. 2021R1A2C2008855). K. Jung is with ASRI, Seoul National University, Korea.

Bibliographical References

Fabio Benedetti, Domenico Beneventano, Sonia Bergamaschi, and Giovanni Simonini. 2019. Computing inter-document similarity with context semantic analysis. *Information Systems*, 80:136–147.

Lucien Brown. 2011. Korean honorifics and 'revealed','ignored'and 'suppressed'aspects of korean culture and politeness. In *Politeness across cultures*, pages 106–127. Springer.

Lucien Brown. 2015. Expressive, social and gendered meanings of korean honorifics. *Korean Linguistics*, 17(2):242–266.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

[3]https://cloud.google.com/translate/
[4]https://openai.com/

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Israel Cohen, Yiteng Huang, Jingdong Chen, Jacob Benesty, Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. 2009. Pearson correlation coefficient. *Noise reduction in speech processing*, pages 1–4.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Elena Deza, Michel Marie Deza, Michel Marie Deza, and Elena Deza. 2009. *Encyclopedia of distances*. Springer.

Lee R Dice. 1945. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302.

Wael H Gomaa, Aly A Fahmy, et al. 2013. A survey of text similarity approaches. *international journal of Computer Applications*, 68(13):13–18.

Jiyeon Ham, Yo Joong Choe, Kyubyong Park, Ilji Choi, and Hyungjoon Soh. 2020. Kornli and korsts: New benchmark datasets for korean natural language understanding. *arXiv preprint arXiv:2004.03289*.

Yongkeun Hwang, Yanghoon Kim, and Kyomin Jung. 2021. Context-aware neural machine translation for korean honorific expressions. *Electronics*, 10(13):1589.

Aminul Islam and Diana Inkpen. 2008. Semantic text similarity using corpus-based word similarity and string similarity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 2(2):1–25.

Manuela Nayantara Jeyaraj and Dharshana Kasthurirathna. 2021. Mnet-sim: A multi-layered semantic similarity network to evaluate sentence similarity. *arXiv preprint arXiv:2111.05412*.

Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Tuo Zhao. 2019. Smart: Robust and efficient fine-tuning for pre-trained natural language models through principled regularized optimization. *arXiv preprint arXiv:1911.03437*.

Minkyung Kim, Hakyoon Lee, and Y Kim. 2018. Learning of korean honorifics through collaborative tasks. *Task-based approaches to teaching and assessing pragmatics*, pages 28–54.

Jeong Yoon Ku et al. 2014. Korean honorifics: a case study analysis of korean speech levels in naturally occurring conversations.

Jonáš Kulhánek, Vojtech Hudecek, Tomáš Nekvinda, and Ondrej Dušek. 2021. Augpt: Dialogue with pre-trained language models and data augmentation. *arXiv preprint arXiv:2102.05126*, 26:532–535.

Phuc H Le-Khac, Graham Healy, and Alan F Smeaton. 2020. Contrastive representation learning: A framework and review. *Ieee Access*, 8:193907–193934.

Seugnjun Lee, Hyeonseok Moon, Chanjun Park, and Heui-Seok Lim. 2023. Improving formality-sensitive machine translation using data-centric approaches and prompt engineering. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 420–432.

Zhaojiang Lin, Zihan Liu, Genta Indra Winata, Samuel Cahyawijaya, Andrea Madotto, Yejin Bang, Etsuko Ishii, and Pascale Fung. 2020. Xpersona: Evaluating multilingual personalized chatbot. *arXiv preprint arXiv:2003.07568*.

Frank Nielsen. 2010. A family of statistical symmetric divergences based on jensen's inequality. *arXiv preprint arXiv:1009.4004*.

Ahmed Hamza Osman and Omar Mohammed Barukub. 2020. Graph-based text representation and matching: A review of the state of the art and future challenges. *IEEE Access*, 8:87562–87583.

Sungjoon Park, Jihyung Moon, Sungdong Kim, Won Ik Cho, Jiyoon Han, Jangwon Park, Chisung Song, Junseong Kim, Yongsook Song, Taehwan Oh, et al. 2021. Klue: Korean language understanding evaluation. *arXiv preprint arXiv:2105.09680*.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Abdul Wahab Qurashi, Violeta Holmes, and Anju P Johnson. 2020. Document processing: Methods for semantic text similarity analysis. In *2020 International Conference on INnovations in Intelligent SysTems and Applications (INISTA)*, pages 1–6. IEEE.

Faisal Rahutomo, Teruaki Kitasuka, and Masayoshi Aritsugi. 2012. Semantic cosine similarity. In

The 7th international student conference on advanced science and technology ICAST, volume 4, page 1.

Pranab Kumar Sen. 1968. Estimates of the regression coefficient based on kendall's tau. *Journal of the American statistical association*, 63(324):1379–1389.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709*.

Ho-min Sohn. 2005. *Korean language in culture and society*. University of Hawaii press.

Tianyi Tang, Hongyuan Lu, Yuchen Eleanor Jiang, Haoyang Huang, Dongdong Zhang, Wayne Xin Zhao, and Furu Wei. 2023. Not all metrics are guilty: Improving nlg evaluation with llm paraphrasing. *arXiv preprint arXiv:2305.15067*.

Junji Tomita, Hidekazu Nakawatase, and Megumi Ishii. 2004. Calculating similarity between texts using graph-based text representation model. In *Proceedings of the thirteenth ACM international conference on Information and knowledge management*, pages 248–249.

Jiapeng Wang and Yihong Dong. 2020. Measurement of text similarity: a survey. *Information*, 11(9):421.

Sam Witteveen and Martin Andrews. 2019. Paraphrasing with large language models. *arXiv preprint arXiv:1911.09661*.

Language Resource References

Seonghyun Kim. 2022. *SmileStyle: Parallel Style-variant Corpus for Korean Multi-turn Chat Text Dataset*.