Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 7122–7128
Marseille, 20-25 June 2022

StyleKQC: A Style-Variant Paraphrase Corpus for Korean Questions and Commands

Won Ik Cho[1], Sangwhan Moon[2], Jong In Kim[3], Seok Min Kim[1], Nam Soo Kim[1]
Department of Electrical and Computer Engineering and INMC, Seoul National University[1]
Department of Computer Science, Tokyo Institute of Technology[2]
Interdisciplinary Program in Cognitive Science, Seoul National University[3]
wicho@hi.snu.ac.kr,sangwhan@iki.fi,prows12@gmail.com
smkim@hi.snu.ac.kr,nkim@snu.ac.kr

Abstract

Paraphrasing is often performed with less concern for controlled style conversion. Especially for questions and commands, style-variant paraphrasing can be crucial in tone and manner, which also matters with industrial applications such as dialog systems. In this paper, we attack this issue with a corpus construction scheme that simultaneously considers the core content and style of directives, namely intent and formality, for the Korean language. Utilizing manually generated natural language queries on six daily topics, we expand the corpus to formal and informal sentences by human rewriting and transferring. We verify the validity and industrial applicability of our approach by checking the adequate classification and inference performance that fit with conventional fine-tuning approaches, at the same time proposing a supervised formality transfer task.

Keywords: Paraphrase, Style-variant, Korean, Spoken language, Directives

1. Introduction

Paraphrasing, the act of using different sentences with the same meaning (Bhagat and Hovy, 2013), is strongly related to the text style conversion or transfer (Yamshchikov et al., 2020). While prior studies often modify sentiment or offensiveness (Logeswaran et al., 2018; dos Santos et al., 2018), in view of paraphrasing, it should be well checked whether the core content of the sentence is maintained during the conversion process. If the sentence meaning stays the same while changing politeness or formality (Rao and Tetreault, 2018), we can call it paraphrasing or rewriting. Such styles can be represented in diverse ways across genre, domain, and language (Jhamtani et al., 2017; Fu et al., 2018; Yang et al., 2019).

Up to date, paraphrasing has been adopted as a useful strategy for text data augmentation. For instance, recent text augmentation schemes such as Dhole et al. (2021) exploit various automatic rewriting tools such as abbreviating, transliteration, lexical shift etc., while paraphrasing through text style transfer is one of them. The approach bases on unsupervised learning of English text styles, following the scheme of Krishna et al. (2020). However, automatic style transfer may not always guarantee the naturalness of the sentence and the preservation of core contents. Also, it is not easy to attain direct text style transfer pairs from unsupervised and automatic approaches. This challenge is visible in the languages with a comparably lower amount of resources, where substantial resources are not guaranteed for each desired text style.

In this light, we attempt to make up a solid scheme for the manual construction of text style transfer database, in a less studied language, Korean. We deal with the scheme of constructing a corpus of style-variant paraphrases for directive sentences such as questions and commands, targeting the Korean language where politeness (suffix) and honorifics play a significant role in conversation (Strauss and Eun, 2005). Here, we consider topic and speech act as attributes constituting the directive sentence (Cho et al., 2020a) and construct a formal style paraphrase set using the natural language queries displaying each topic and speech act. Finally, style-variant paraphrase pairs are obtained by manual conversion from formal to informal sentences in consideration of content preservation, and are to be released publicly as the first open text style transfer dataset in Korean. Our contribution is as follows:

- We present a corpus construction scheme capable of performing multiple tasks while enabling parallel sentence style transfer.
- We release a Korean corpus where sentence formality style is well defined, regarding the daily used questions and commands.[1]

2. Related Work

In general, sentence style[2] is handled regarding tone and manner in writing, though with a subtle difference (Brooks, 2020). However, previous research on content-preserving style transfer (Logeswaran et al., 2018; Tian et al., 2018) does not seem to be only about tone in that the change in sentiment may influence the core speaker intent. Furthermore, most approaches were from the perspective of unsupervised learning (dos Santos et

[1] https://github.com/cynthia/stylekqc

[2] In this paper, we view ‘formality’ in Korean as a style, while interchangeably using ‘conversion’ and ‘transfer’.

al., 2018; Bao et al., 2019), with less explored fields of parallel style-variant corpus for supervised learning, which might provide robust guidance for the generative pre-trained models nowadays (Radford et al., 2019).
This trend was similarly revealed in previous studies on Korean. Since the early approaches follow the studies in English and other languages, sentiment or stance-based style transfer has been predominantly suggested (Lee et al., 2019; Choi and Na, 2019).[3] In Hong et al. (2018), the transfer regarding politeness suffix of the sentence enders was considered at the same time maintaining the sentence meaning, mainly regarding '*hay-yo*' and '*hap-syo*' enders which differ in the degree of formality. However, it dealt only with the syntactic change, not the modification in the lexicon, adverbs, or tone and manner of the speech, which are all considered influential for the honorific system (Strauss and Eun, 2005). In this regard, we thought that formality style transfer should be well-defined along with content preservation. Furthermore, there is no open dataset for Korean style transfer that can be utilized for research and commercial purposes. We aim to resolve the above issues by proposing a straightforward and effective building scheme.

3. Proposed Scheme

We construct a corpus of Korean directives, namely questions and commands, where the question consists of an alternative question (Alt. Q) or wh-question (wh-Q), and the command consists of prohibition (PH) and requirement (REQ), following Cho et al. (2020a). In other words, we target four types of speech acts and assume sentences that can be uttered to humans or artificial intelligent (AI) agents. There are six topics involved in this: *messenger, calendar, weather and news, smart home, shopping,* and *entertainment*, which come from a recent survey on customers' usage (Lee et al., 2020).
Twelve workers from different backgrounds were recruited. In detail, there were six researchers/students with linguistics background, three researchers/students with non-linguistic background, and three participants working in an industry not related to the linguistics domain. We required specifying two likes and one dislike on the topic, and these preferences were taken into account when creating a total of 6 subgroups with two people each. Here, to help participants interact with each other's strategies and at the same time proceed in a way that is more linguistically feasible, we placed a researcher with the linguistics background to each group.
We created a construction scheme that goes through the following three steps to check its reliability while generating utterances of 5,000 per topic and 7,500 per speech act.

1. Writing natural language queries
2. Rewriting queries in a sentence with the formal tone
3. Converting the formal sentences to informal ones

Query generation First, query generation is a process in which participants directly suggest the core content of directives which are to be rewritten in a formal style. In this process, participants were asked to write a natural language query for each of the given two speech acts on the assigned topic.[4] Since the query structure differs by speech act type as in Cho et al. (2020a), the created queries did not overlap across the workers. The queries were checked for their suitability, to avoid personally identifiable stuff or those that can cause social harm. 125 queries were generated for each (***topic, act***) pair. The example of queries per some (***topic, act***) is shown below. All the queries are generated in Korean, but described here in English for demonstrative purposes.

- (*Shopping, Alt. Q*) *The one that has better A/S between Samsung and Apple*
- (*Entertainment, Wh-Q*) *The TV channel number where the news is on at 8:00 p.m.*
- (*Messenger, PH*) *Not to turn on WeChat automatic update*
- (*Smart home, REQ*) *To recharge the wireless vacuum cleaner in the multi-room*

No particular principle was considered in the query generation, but the workers were asked to make diverse expressions that fit with colloquial context and daily life. Too knowledge-intensive questions or queries with multiple contents were asked for a modification.

Writing formal sentences The next is a process in which the workers of subgroups exchange queries generated by each other and rewrite them into formal style sentences.[5] We primarily asked for the formal style because there are more diverse expressions for formal utterances in the Korean language regarding indirect speech and honorifics (Byon, 2006), so that the paraphrasing is easier compared to informal ones that might not come to the worker's mind at the first place. The formal utterances were required to fit with the conversation with senior or elderly addressees rather than friends or juniors.
Rewriting was required for a total of 5 sentences. To make the paraphrases as diverse as possible, the asking strategies in Byon (2006) and Cho (2008) were requested. We display some excerpts:

- Softening the commands to requests
- Indirectly mentioning the addressee's obligation
- Mentioning the addressee's responsibility

[3] Most of the work are not in an internationally readable format; thus, we note here the methods used in the papers.

[4] These were readily provided by the process managers in Cho et al. (2020a), but here we let them be created by the workers to make the contents more diverse and to benefit from the preferences. Also, 'query' here does not only apply to the phrase for questions, but also the nominalized phrase for commands. Refer to Cho et al. (2020b) for further information.

[5] In this process, the workers check the validity of the query created by each other, that the incompleteness of the queries overlooked by the moderator (author) can be pointed out.

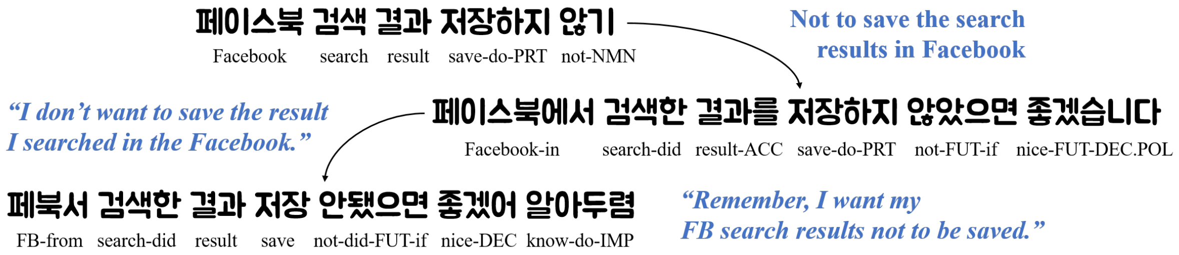
페이스북 검색 결과 저장하지 않기
Facebook search result save-do-PRT not-NMN
Not to save the search results in Facebook
페이스북에서 검색한 결과를 저장하지 않았으면 좋겠습니다
Facebook-in search-did result-ACC save-do-PRT not-FUT-if nice-FUT-DEC.POL
"I don't want to save the result I searched in the Facebook."
페북서 검색한 결과 저장 안됐으면 좋겠어 알아두렴
FB-from search-did result save not-did-FUT-if nice-DEC know-do-IMP
"Remember, I want my FB search results not to be saved."

Figure 1: An example of query generation-formal sentence writing-informal transferring, along with the gloss and translation (*PRT* particle, *NMN* nominalizer, *ACC* accusative, *FUT* futuristic, *DEC* declarative, *POL* politeness suffix, *IMP* imperative). Though not reflected in the English translation, the transferring preserves the overall structure of the formal sentence as well as the core content.

- Alleviating the addressee's burden with polarity items such as *please* or *bit*
- Asking the availability of the addressee

Some of these characteristics are shared across the culture (Brown et al., 1987). It may also be exhibited similar in the East Asian society (Gu, 1990) and within a similar syntax such as Japanese (Okamoto, 1999; Fukada and Asato, 2004). However, we faced language-specific considerations regarding functional and lexical expressions and asked the workers to reflect them in the construction. Simultaneously, to fit with the naturalness within colloquial context, written-style or outdated phrases/words were avoided.

Converting to informal style The final process is modifying directive sentences written in formal style into informal sentences. Here, the workers convert the other person's formal sentences, created from the original query they had generated, checking the typos and misunderstandings once again. 'Informality' defined here is slightly different from being rude or impolite, but instead means that the conversation moves towards a more comfortable and personal relationship. (Rao and Tetreault, 2018).

In this process, we asked the workers to maintain the overall sentence structure, of which the diversity was already obtained owing to policies in writing formal sentences. With this, we could prevent the potential overlap between the converted sentences and also guarantee the 'parallelness' of the created data. This can be more effective in the Korean language where indirectness is often distinguished from formality; for instance, a cautious request to a younger brother can be informal but indirect.

Style conversion was performed in various aspects such as change in sentence enders, honorifics, and lexicons (such as *nation* to *country*). The workers were encouraged to insert or delete some phrases depending on the naturalness of the content, and to perform at least two word-level modifications. The detailed guideline[6] for the whole process was provided to the workers with example query-sentence tuples, and we exhibit one of them (Figure 1).

Refinement The corpus was refined by three native speakers with corpus construction experience for Korean directive sentences. In this process, typos, awkward sentences, and paraphrases that are not sufficiently diverse were inspected, and the reviews were reflected by the moderator.

4. Experiment

4.1. Task Setting

Through the experiment, we display that the proposed construction scheme provides a corpus that enables creating multiple task sets simultaneously, which can bring advantages from a practical viewpoint.

- Topic classification
- Speech act classification
- Paraphrase detection
- Sentence style transfer

4.2. Implementation

For each of the total 24 [topic, act] chunks where we have 125 queries each, we set aside 80% (100 queries) for training, 4% (5 queries) for validation, and 16% (20 queries) for the test. From the whole dataset of volume 30,000, the training set contains 24,000 sentences and 1,200/4,800 for dev/test each. The queries were chosen randomly, and all the sets have an equal rate of topic and speech act ratio.

Topic (TOPIC) and speech act (ACT) classification are intuitively formulated. There are 5,000 utterances for each topic and 7,500 utterances for each speech act, where six topics and four speech act types are set as labels.

Paraphrase detection (PARA) requires a sentence pair. In Cho et al. (2020a), the sentence similarity was defined 5-fold, checking if the topic or speech act overlaps between the two input sentences, with the highest similarity if the queries are identical (the paraphrases). The paraphrase detection task was derived by formulating the multi-class problem into a binary task. See

[6] https://docs.google.com/document/d/1gjyEMCcp0mxmdzSKdd5OrLFVikyq22OsxXHisSr2THY written in Korean.

	TOPIC	ACT	PARA	STYLE
Input	Sentence	Sentence	Pair	Sentence
Class #	6	4	2	-
Volume	30,000	30,000	270,000	15,000
F1 Score	92.68	97.75	99.93	-
Accuracy	92.83	97.75	99.93	99.58†
CED	-	-	-	0.451

Table 1: Experiment results on four subtasks.

`DATA-GEN/mkdata.py` in the supplementary material[7] for further detail.

Finally, we checked whether sentence style transfer (STYLE) works using the pairs within; 12,000 pairs for training, 600 for validation and 2,400 for the test. The training was done in the way of converting the formal sentences to informal ones.

Both sentence classification and paraphrase detection tasks were implemented based on a BERT-based (Devlin et al., 2019) KcBERT[8] (Lee, 2020), and for sentence style transfer, KoGPT2[9] that bases on GPT2 (Radford et al., 2019) was adopted. F1 (macro) and accuracy were used for the classification tasks, and for style transfer, we checked character edit distance (CED). The accuracy for style transfer (†) denotes the precision obtained with the model learned upon the train set (Pang, 2019). Experimental settings are provided as supplementary.

4.3. Results

In classification and inference, we have the evaluation results that show consistency between the train and test dataset (Table 1). Considering that queries in each set are distinguished from each other, we claim that our dataset displays the extensibility to wider world problems, also providing the comprehensive coverage of topics and acts that are of interest in usual conversation and smart speaker dialogues. Though the baseline score is quite high for ACT and PARA, it does not harm one of our goals to provide a solid scheme for corpus construction that suffices practical, real-world applicability.

On STYLE, we adopted CED since our 'style' more regards the change in suffix and some lexicons rather than the whole word order and phrase usage.[10] Nonetheless, we found the transfer task still challenging in view of the objective measure. Instead, we observed the practical validity using a style classifier learned upon train and valid set, which displays sufficiently high accuracy.[11] We qualitatively checked that the seq2seq (Sutskever et al., 2014) approach with a pre-trained generative model guarantees the intended style transfer.

4.3.1. Error Analysis

For style transfer, some errors have occurred in the following forms:

1. Unknown stop in the decoding session
2. Repetition of some phrases
3. Appearance of irrelevant terms

Unknown stops We first assumed OOV for a reason, but it turned out not since it happened for the text cases where all the tokens exist in the training set. Another analysis suggests that the change of word order (which is tolerated in Korean for being scrambling) which makes it challenging for the language understanding module to comprehend a full sentence, might have caused the decoding module to fall in collapse and finish the decoding just by facing the end of usual sentences. For instance, an in-out pair

(a) 국내 브랜드가 더 많이 들어가 있는 곳을 알아봐주세요 지마켓과 신세계 중에 ("Please find out where more domestic brands are located, among G-Market and Shinsegae.")
(b) *[12]국내 브랜드가 더 많이 들어가 있는 곳 좀 알아봐줘 지마켓 (*"G-Market, find out where more domestic brands are in.")

shows that the scrambling, which preserves sentence acceptability in Korean, might confuse the trained module.

Repitition The repetition of phrases bursts out when the model is confused about what to transfer, sometimes because it misunderstood the act of the utterance. For instance, in an in-out pair

(c) 내일 재고확인하세요 모레 재고확인하세요 (''Will you check stock tomorrow or the day after tomorrow?")
(d) *내일 재고확인 좀 해 모레 재고확인해야겠어 모레 재고확인해야겠어 (*''Check stock tomorrow. I will check it the day after tomorrow. I will check it the day after tomorrow.")

the transfer model fails to understand that the input sentence is an alternative question and transfers it as a command (due to a seemingly ambiguous sentence ender 요 (yo) - whose role is clear at this circumstance), finally displaying a repetition, failing to emit an acceptable sentence.

Irrelevant terms The appearance of irrelevant terms happened rarely, but mainly seemed to be owing to the knowledge within the generative pre-trained models. It would be our future work to lessen this kind of malfunction where the pre-trained bias negatively affects the fine-tuned model.

[7] `https://www.dropbox.com/s/ju53oan78u2nfkx/supple-data.zip?dl=0`

[8] `https://github.com/Beomi/KcBERT`

[9] `https://github.com/SKT-AI/KoGPT2`

[10] On using other objective measures, the morpheme-level tokenization is not yet unified for Korean sentences, to make evaluation harder. More explanation is available in Appendix A.1.

[11] More explanation on using classifier accuracy in style checking is available in Appendix A.2.

[12] Wrong sentence.

4.3.2. Discussion

We have some notes on the validity of the created dataset. Primarily, though the dataset is first suggested open corpus for Korean style transfer, the granularity of the style difference within the pair is not provided here as in Rao and Tetreault (2018). Also, since our dataset provides the style transfer that maintains the overall sentence structure, some sentence pairs show minor differences, which is sufficient for spoken language processing but less robust to digitized online texts. Finally, since the formality conversion regards morpho-syntactic and lexical changes rather than the paraphrasing done in writing the formal sentences, the style diversity of expressions is limited to the sentence formats that are not awkward to utter.

Despite the limitations, we want to emphasize that our approach can suggest a reliable and efficient scheme for the service providers or task managers aiming at a particular style transfer for various types of sentences. For instance, if one replaces input queries with some structured query language (SQL) or canonical forms of statements and use 'rudeness' or 'twitter-likeness' as a style, the parallel dataset can be created in the same way, with a slightly different guideline. This kind of pair generation has been done with rule or back translation in Rao and Tetreault (2018), but we believe that human-aided construction is more reliable and eventually reduces the necessity of additional human checking. Also, see Appendix B to see how our manual construction process has considered the ethical sides of human factors.

5. Conclusion

In this paper, we construct and disclose the first style-variant Korean paraphrase corpus. Topic, speech act, and paraphrase are simultaneously considered in evaluating the final corpus, where the consistent composition is assumed to be guaranteed by the evaluation results. The entire guideline is currently specific to the formality transfer in Korean, but can be utilized in making up other parallel style transfer corpus with an extended pool of topics, speech acts, queries, and style. All the resources are available online[13], and we provide another implementation for politeness transfer using a Korean public PLM[14] to facilitate the future research on Korean text style transfer.

6. Acknowledgements

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2021-0-00456, Development of Ultra-high Speech Quality Technology for Remote Multi-speaker Conference System). We appreciate Seonghyun Kim for providing a codebase for KoBART implementation. Also, the corpus construction was possible thanks to the help of twelve passionate participants, namely Kyung Seo Ki, Dongho Lee, Yoon Kyung Lee, Hee Young Park, Yulhee Kim, Seyoung Park, Jiwon An, Jeonghwa Cho, Kihyo Park, Kyuhwan Lee, Soomin Lee, and Minhwa Chung.

7. Bibliographical References

Bao, Y., Zhou, H., Huang, S., Li, L., Mou, L., Vechtomova, O., Dai, X., and Chen, J. (2019). Generating sentences from disentangled syntactic and semantic spaces. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6008–6019.

Bhagat, R. and Hovy, E. (2013). What is a paraphrase? *Computational Linguistics*, 39(3):463–472.

Brooks, C. (2020). *Building Blocks of Academic Writing*. BCcampus.

Brown, P., Levinson, S. C., and Levinson, S. C. (1987). *Politeness: Some universals in language usage*, volume 4. Cambridge University Press.

Byon, A. S. (2006). The role of linguistic indirectness and honorifics in achieving linguistic politeness in Korean requests. *Journal of Politeness Research*, 2(2):247–276.

Cho, W. I., Kim, J. I., Moon, Y. K., and Kim, N. S. (2020a). Discourse component to sentence (DC2S): An efficient human-aided construction of paraphrase and sentence similarity dataset. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 6819–6826.

Cho, W. I., Moon, Y., Moon, S., Kim, S. M., and Kim, N. S. (2020b). Machines getting with the program: Understanding intent arguments of non-canonical directives. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 329–339, Online, November. Association for Computational Linguistics.

Cho, Y. (2008). Strategic use of Korean honorifics functions of 'partner-deference sangdae-nopim'. *Dialogue and Rhetoric*, 2:155.

Choi, H.-J. and Na, S.-H. (2019). Delete and generate: Korean style transfer based on deleting and generating word n-grams. In *Annual Conference on Human and Language Technology*, pages 400–403. Human and Language Technology.

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

Dhole, K. D., Gangal, V., Gehrmann, S., Gupta, A., Li, Z., Mahamood, S., Mahendiran, A., Mille, S., Srivastava, A., Tan, S., Wu, T., Sohl-Dickstein, J., Choi, J. D., Hovy, E. H., Dusek, O., Ruder, S., Anand, S., Aneja, N., Banjade, R., Barthe, L., Behnke, H.,

[13] https://github.com/cynthia/stylekqc

[14] https://colab.research.google.com/drive/1YjU1wlwl26X49hQLr6ZQvOKQm2yiR4sg

Berlot-Attwell, I., Boyle, C., Brun, C. D., Cabezudo, M. A. S., Cahyawijaya, S., Chapuis, E., Che, W., Choudhary, M., Clauss, C., Colombo, P., Cornell, F., Dagan, G., Das, M., Dixit, T., Dopierre, T., Dray, P.-A., Dubey, S., Ekeinhor, T., Giovanni, M. D., Gupta, R., Hamla, L., Han, S., Harel-Canada, F., Honoré, A., Jindal, I., Joniak, P. K., Kleyko, D., Kovatchev, V., Krishna, K., Kumar, A., Langer, S., Lee, S. R., Levinson, C. J., Liang, H., Liang, K., Liu, Z., Lukyanenko, A., Marivate, V., de Melo, G., Meoni, S., Meyer, M., Mir, A., Moosavi, N. S., Muennighoff, N., Mun, T. S. H., Murray, K. W., Namysl, M., Obedkova, M., Oli, P., Pasricha, N., Pfister, J., Plant, R., Prabhu, V. U., Pais, V. F., Qin, L., Raji, S., Rajpoot, P. K., Raunak, V., Rinberg, R., Roberts, N., Rodriguez, J. D., Roux, C., VasconcellosP.H., S., Sai, A. B., Schmidt, R. M., Scialom, T., Sefara, T. J., Shamsi, S., Shen, X., Shi, H., Shi, Y., Shvets, A. V., Siegel, N., Sileo, D., Simon, J., Singh, C., Sitelew, R., Soni, P., Sorensen, T. M., Soto, W., Srivastava, A., Srivatsa, K. V. A., Sun, T., MukundVarma, T., Tabassum, A., Tan, F. A., Teehan, R., Tiwari, M., Tolkiehn, M., Wang, A., Wang, Z., Wang, G., Wang, Z. J., Wei, F., Wilie, B., Winata, G. I., Wu, X., Wydma'nski, W., Xie, T., Yaseen, U., Yee, M.-H., Zhang, J., and Zhang, Y. (2021). Nl-augmenter: A framework for task-sensitive natural language augmentation. *arXiv preprint arXiv:2112.02721*.

dos Santos, C., Melnyk, I., and Padhi, I. (2018). Fighting offensive language on social media with unsupervised text style transfer. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 189–194.

Fu, Z., Tan, X., Peng, N., Zhao, D., and Yan, R. (2018). Style transfer in text: Exploration and evaluation. In *AAAI*, pages 663–670.

Fukada, A. and Asato, N. (2004). Universal politeness theory: application to the use of Japanese honorifics. *Journal of Pragmatics*, 36(11):1991–2002.

Gu, Y. (1990). Politeness phenomena in modern Chinese. *Journal of Pragmatics*, 14(2):237–257.

Hong, T., Xu, G., Ahn, H., Kang, S., and Seo, J. (2018). Korean text style transfer using attention-based sequence-to-sequence model. In *Annual Conference on Human and Language Technology*, pages 567–569. Human and Language Technology.

Jhamtani, H., Gangal, V., Hovy, E., and Nyberg, E. (2017). Shakespearizing modern language using copy-enriched sequence to sequence models. In *Proceedings of the Workshop on Stylistic Variation*, pages 10–19.

Krishna, K., Wieting, J., and Iyyer, M. (2020). Reformulating unsupervised style transfer as paraphrase generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 737–762, Online, November. Association for Computational Linguistics.

Lee, J., Oh, Y., Byun, h., and Min, K. (2019). Controlled Korean style transfer using BERT. In *Annual Conference on Human and Language Technology*, pages 395–399. Human and Language Technology.

Lee, J. H., Seon, H. J., and Lee, H. J. (2020). Positioning of smart speakers by applying text mining to consumer reviews: Focusing on artificial intelligence factors. *Knowledge Management Research*, 21(1):197–210.

Lee, J. (2020). Kcbert: Korean comments bert. In *Annual Conference on Human and Language Technology*. Human and Language Technology.

Logeswaran, L., Lee, H., and Bengio, S. (2018). Content preserving text generation with attribute controls. In *Advances in Neural Information Processing Systems*, pages 5103–5113.

Okamoto, S. (1999). Situated politeness: Manipulating honorific and non-honorific expressions in Japanese conversations. *Pragmatics*, 9(1):51–74.

Pang, R. Y. (2019). The daunting task of real-world textual style transfer auto-evaluation. *arXiv preprint arXiv:1910.03747*.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Rao, S. and Tetreault, J. (2018). Dear sir or madam, may i introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 129–140.

Strauss, S. and Eun, J. O. (2005). Indexicality and honorific speech level choice in Korean. *Linguistics*, 43(3):611–651.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, pages 3104–3112.

Tian, Y., Hu, Z., and Yu, Z. (2018). Structured content preservation for unsupervised text style transfer. *arXiv preprint arXiv:1810.06526*.

Yamshchikov, I., Shibaev, V., Khlebnikov, N., and Tikhonov, A. (2020). Style-transfer and paraphrase: Looking for a sensible semantic similarity metric. *arXiv preprint arXiv:2004.05001*.

Yang, Z., Cai, P., Feng, Y., Li, F., Feng, W., Chiu, E.-Y., and Yu, H. (2019). Generating classical chinese poems from vernacular Chinese. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2019, page 6155. NIH Public Access.

A. Analysis on STYLE Results

In STYLE, CED and accuracy are used, where each is defined a little bit different from usual cases.

A.1. Using CED

First, on CED, since the output length may differ from GT, we normalize the CED with length so that it comes between 0 and 1, such as defined in `SOURCE/EVAL-STYLE/eval_style.py` among the supplementary material. In Korean, *character* denotes a morpho-syllabic block that corresponds to the subword in English, thus CED can have a role as a subword-level edit distance. This was considered more appropriate than BLEU or METEOR, which are usual for other Latin alphabet-based style transfer studies, since 1) ours aims at a structure-conservative style-variant paraphrasing, and 2) morphological decomposition schemes are not solidly unified in the empirical studies. Besides, we did not choose semantic-level measures such as BERTScore since most of the outputs would record a high score because the paraphrasing was guaranteed.

A.2. Using Accuracy

On the accuracy, which is defined differently from TOPIC, ACT, or PARA since the aim of STYLE is not originally in making a classifier, we check if the style classifier trained with the samples of the training set can precisely classify the transferred test sentences as informal ones. Thus, only the accuracy, which equals precision in this scenario, is calculated. Achieving a high performance here indirectly shows that the style transfer is adequately performed for a large portion of scenarios.

B. Ethical Considerations

In the corpus construction procedure which bases upon the documented approval of the workers, adequate compensation was paid to each of them, in all the processes of query generation, writing formal sentences, and transferring them to the informal one.

The participants, recruited from social media and the web, are familiar with smart speakers, and some of them had experience in corpus construction processes. For 12 participants, 250 WON (≈$0.22) was provided in writing each query and 200 WON (≈$0.18) for making up the sentences. Thus, each participant was paid 600,000 WON (≈$540) to make up 250 queries and write 2,500 sentences.

Our resource is free from license issues since all the materials were created according to the guideline (a kind of template) and checked for post-processing. The outcome of our project does not contain any personally identifiable information, nor the contents that can induce social harm.