A Dog Is Passing Over The Jet? A Text-Generation Dataset for Korean Commonsense Reasoning and Evaluation

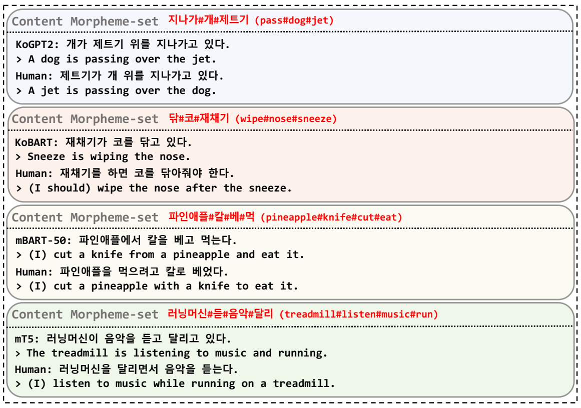
Jaehyung Seo[1*], **Seounghoon Lee**[2*†], **Chanjun Park**[1,3], **Yoonna Jang**[1]
Hyeonseok Moon[1], **Sugyeong Eo**[1], **Seonmin Koo**[1], **Heuiseok Lim**[1,2‡]
[1]Department of Computer Science and Engineering, Korea University
[2]Human Inspired Artificial Intelligence Research (HIAI), Korea University
[3]Upstage

Abstract

Recent natural language understanding (NLU) research on the Korean language has been vigorously maturing with the advancements of pretrained language models and datasets. However, Korean pretrained language models still struggle to generate a short sentence with a given condition based on compositionality and commonsense reasoning (i.e., generative commonsense reasoning). The two major challenges are inadequate data resources to develop generative commonsense reasoning regarding Korean linguistic features and to evaluate language models which are necessary for natural language generation (NLG). To solve these problems, we propose a text-generation dataset for Korean generative commonsense reasoning and language model evaluation. In this work, a semi-automatic dataset construction approach filters out contents inexplicable to commonsense, ascertains quality, and reduces the cost of building the dataset. We also present an in-depth analysis of the generation results of language models with various evaluation metrics along with human-annotated scores. The whole dataset is publicly available at (https://aihub.or.kr/opendata/korea-university).

1 Introduction

With the advent of Transformer (Vaswani et al., 2017) model, the importance of language resources and language modeling in natural language processing (NLP) has been heightened. Indeed, various studies on Korean language resources, such as Korean morpheme analysis (Matteson et al., 2018; Kim and Colineau, 2020; Moon and Okazaki, 2020), natural language understanding (NLU) tasks including KorNLI and KorSTS (Ham et al., 2020), KMRE (Lee et al., 2020), and KLUE (Park et al., 2021) are being conducted alongside the studies on off-the-shelf pretrained language models in Korean (Devlin et al., 2019; Radford et al., 2019; Lewis et al., 2020).



Content Morpheme-set 지나가#개#제트기 (pass#dog#jet)
KoGPT2: 개가 제트기 위를 지나가고 있다.
> A dog is passing over the jet.
Human: 제트기가 개 위를 지나가고 있다.
> A jet is passing over the dog.

Content Morpheme-set 닦#코#재채기 (wipe#nose#sneeze)
KoBART: 재채기가 코를 닦고 있다.
> Sneeze is wiping the nose.
Human: 재채기를 하면 코를 닦아줘야 한다.
> (I should) wipe the nose after the sneeze.

Content Morpheme-set 파인애플#칼#베#먹 (pineapple#knife#cut#eat)
mBART-50: 파인애플에서 칼을 베고 먹는다.
> (I) cut a knife from a pineapple and eat it.
Human: 파인애플을 먹으려고 칼로 베었다.
> (I) cut a pineapple with a knife to eat it.

Content Morpheme-set 러닝머신#듣#음악#달리 (treadmill#listen#music#run)
mT5: 러닝머신이 음악을 듣고 달리고 있다.
> The treadmill is listening to music and running.
Human: 러닝머신을 달리면서 음악을 듣는다.
> (I) listen to music while running on a treadmill.

Figure 1: Example results of the KoGPT2, KoBART, mBART-50, mT5, and humans on Korean generative commonsense reasoning task. The results are sentences that combine the given content morphemes (in **red boldface**).

Despite the prosperity of Korean NLP research, two critical problems remain: (i) an absence of a research base for natural language generation (NLG) and (ii) a deficient ability for models to generate commonsense knowledge. In other words, (i) there exists neither a dataset (Gehrmann et al., 2021) nor standards for evaluating (Celikyilmaz et al., 2020) the results generated by language models in Korean because most Korean language resources are focused on NLU tasks thereby making it difficult to accelerate the development of NLG research.

(ii) Korean language models encounter difficulties even in generating sentences using simple commonsense knowledge. Commonsense knowledge is a sociocultural knowledge shared by humans (Liu and Singh, 2004). It is not visible but it is melted in their words (Tandon et al., 2018). To make natural sentences using commonsense knowledge like those made by humans, comprehensive abilities

[*] Equally contributed
[†] Present affiliation : Institute for Infocomm Research, A*STAR
[‡] Corresponding author

Findings of the Association for Computational Linguistics: NAACL 2022, pages 2233 - 2249
July 10-15, 2022 ©2022 Association for Computational Linguistics

of generative commonsense reasoning (Lin et al., 2020) are needed. This requires holistic understanding of commonsense reasoning (Lake and Baroni, 2018; Keysers et al., 2020) and sentence compositionality (Hahm et al., 2020).

Though people acquire commonsense knowledge and effortlessly use it in their daily lives, it is challenging for language models to imitate this ability. As shown in Figure 1, off-the-shelf Korean and multilingual language models seem to lack the competence for generative commonsense reasoning. Some model-generated sentences *do not make sense* (e.g., "개가 제트기 위를 지나고 있다. A dog is passing over the jet.", "재채기가 코를 닦고 있다. Sneeze is wiping the nose."), *use inappropriate prepositions* (e.g., "파인애플에서 칼을 베고 먹는다. I cut a knife from a pineapple and eat it."), or *misplace parts of speech* (e.g., "러닝머신이 음악을 듣고 달리고 있다. The treadmill is listening to music and running.").

To address these issues and inspired by CommonGen (Lin et al., 2020), we develop a Korean CommonGen dataset for generative commonsense reasoning. The dataset is composed of commonly used daily-life concepts and sentences made by combining those concepts. Our dataset differs from the CommonGen dataset as follows: (i) We collect Korean corpus and label it to cover Korean sociocultural commonsense knowledge. For example, the sentence of the corpus contains the unique Korean sociocultural terms "귀농 (return to the farm)" and "곶감 (dried persimmons)". (ii) Although we adopt the image caption data from CommonGen, we add the summary data of daily conversations into our dataset thereby obtaining diverse sentences. (iii) Because Korean language models use segmented morphemes as vocabulary (Lee et al., 2020; Kim and Colineau, 2020), we construct the concept set with content morphemes that have linguistic features and lexical meanings. (iv) We analyze the evaluation metrics including the human-annotated score to demonstrate the validity of the evaluation criteria.

We reduce the cost of data construction significantly through an automated method and inspect the quality and unethical issue of Korean CommonGen by crowd-sourcing[1]. In addition, we conduct an in-depth study of our proposed dataset with various ablation experiments on morpheme segmentation and training methods. Furthermore, the model-generated sentences are compared and analyzed by quantitative, qualitative, and human scores. We disclose the dataset used in this paper to contribute to the development of Korean NLG research.

2 Related Works

Commonsense Knowledge Commonsense knowledge is knowledge about everyday life that all people possess, and it is arguably the most general and widely applicable knowledge (Liu and Singh, 2004). Compared to encyclopedic knowledge, which returns specific details about named entities on a modern search engine, commonsense knowledge includes elusiveness and context dependence (Tandon et al., 2018).

Compositionality Compositionality is an essential element that AI systems have to solve for given conditions. For example, the MS-COCO (Lake and Baroni, 2018) dataset is utilized for image caption task generating the description from an image data as an input data. The task demands the compositionality of the model as the model composes the natural description. Moreover, SCAN (Lake and Baroni, 2018) demonstrates mapping instructions to sequence an RNN model's ability to generate continuous behavior. However, these studies have shown that AI systems still struggle to generate complete results.

Commonsense Reasoning Commonsense reasoning is the ability to infer unrecognized commonsense knowledge or relations among given concepts. In a recent NLP research, various commonsense reasoning datasets have been disclosed. CommonsenseQA (Talmor et al., 2019) organizes the dataset with closed questions to commonsense based on the concept of ConceptNet and analyzes the skills of commonsense reasoning required for each question by categorizing them. Cosmos QA dataset (Huang et al., 2019) introduces a question answering dataset based on the fact that is not externally revealed in the context. CoS-E dataset (Rajani et al., 2019) attempts to strengthen the commonsense training of the model by adding a person's description of the commonsense QA. The XCOPA dataset (Ponti et al., 2020) mitigates the gaps in commonsense that could arise from linguistic and cultural differences, while building a dataset.

[1] We employ AI & Human Resources Platform Crowd-Works

Basic Statistics	Train	Validation	Test
# Content morpheme-sets	43,188	1,000	2,040
- Set size less than 3	5,089	115	334
- Set size 4	10,810	241	604
- Set size 5	13,397	332	577
- Set size 6	12,811	292	513
- Set size more than 7	1,081	20	12
# Unique content morphemes	40,874	2,000	3,272
- # Unseen single	-	332	748
- # Unseen pair	-	5,305	10,461
- # Unseen triple	-	9,728	17,648
# Additional morphemes	-	-	4,682
# Sentences	43,188	1,000	6,120
- Average length	26.06	24.74	23.54
- Caption-based rate	45.58	50	48.99
- Dialogue-based rate	54.42	50	51.01

Table 1: Statistics for Korean CommonGen dataset. We construct a test set to generate sentences by reasoning unseen content morphemes in training. **# Additional morphemes** is the number of unseen single morphemes counted through extra 4,080 human references. **#Caption-based rate** and **#Dialogue-based rates** mean the ratio of each data among the total dataset.

Generative Commonsense Reasoning Based on compositionality and commonsense reasoning, we concentrate on generative commonsense reasoning, the ability required to generate sentences that conform to commonsense knowledge for given conditions. In the case of English-based language models, various studies have improved the ability of generative commonsense reasoning based on the CommonGen (Lin et al., 2020) dataset. KG-BART (Liu et al., 2021b) enhances the model allowing to capture the relationship between nodes in the graph, while including the ConceptNet knowledge graph in the attention calculation process for text input. RE-T5 (Wang et al., 2021) reinforces the input value by using a retriever to import sentences related to the concepts from external knowledge. KFCNet (Li et al., 2021) presents the state-of-the-art performance in CommonGen by removing low-quality sentences in external knowledge and applying contrastive learning, respectively.

However, there does not exist Korean dataset for generative commonsense reasoning and advanced research as well. Therefore, in this paper, we aim to create a new text-generation dataset inspired by CommonGen (Lin et al., 2020) and grant a direction for the future Korean NLG research.

3 Korean CommonGen

Korean CommonGen is a text-generation dataset for Korean commonsense reasoning and evaluation. As depicted in Figure 2, Korean CommonGen consists of concept sets typically used in daily life and sentences depicting those concepts. Language models are trained to generate a sentence by reasoning and combining the concepts based on human-generated sentences. The model-generated result should include all the given concepts, be grammatically correct and make sense.

Korean CommonGen includes 43,188 training examples and 2,040 testing examples as presented in Table 1. Corresponding to one concept set of content morphemes, the training example comprises one sentence, and the testing example contains three sentences. In the case of the training set, 45.58% of examples are English-Korean translated MS-COCO[2] (Lin et al., 2014) image caption data, and 52.42% are Korean dialogue summary data of AI-HUB[3]. To make the models learn the semantic role and relation of unseen content morphemes through commonsense reasoning, the test set has 748 single content morphemes unseen during training. Additionally, more than one pair of unseen content morphemes is included in each testing example. The test set has additional two reference sentences for every 2,040 content morpheme set annotated by crowd-sourcing. These additional references enable the consideration of the diverse possibilities in model-generated sentences (Chomsky, 1965). We allow additional references to include 4,682 newly annotated content morphemes that do not appear in the concept set. As a result, the model can obtain a higher score when the generated sentence has additional morpheme fitted on commonsense. The dataset is constructed through semi-automatic approach, and the details are described in §3.1 Automatic Dataset Construction and §3.2 Annotate and Refine by Crowd-sourcing.

3.1 Automatic Dataset Construction

We implement the automatic construction approach with a part-of-speech (POS) tagger, a named entity recognition (NER) tagger, a sentence level filtering, and Korean Hate Speech classifier. They extract the content morphemes, and screen out the sentences that include unethical expressions, violate commonsense or have unnatural sentence structures. With these automated modules, we reduce the cost for human annotation by $7,766.

[2] https://aihub.or.kr/keti_data_board/visual_intelligence

[3] https://aihub.or.kr/aidata/30714

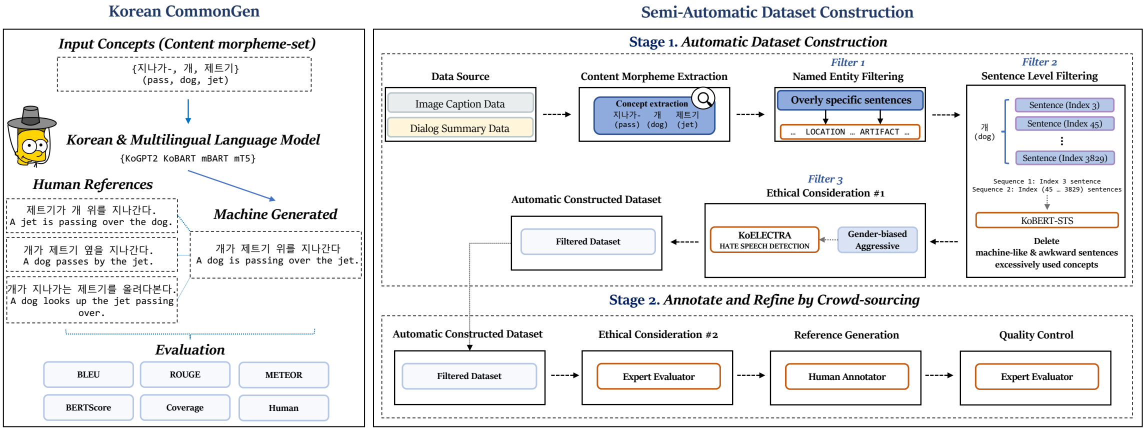

Korean CommonGen
Input Concepts (Content morpheme-set)
{지나가-, 개, 제트기}
(pass, dog, jet)
Korean & Multilingual Language Model
{KoGPT2 KoBART mBART mT5}
Human References
제트기가 개 위를 지나간다.
A jet is passing over the dog.
개가 제트기 옆을 지나간다.
A dog passes by the jet.
개가 지나가는 제트기를 올려다본다.
A dog looks up the jet passing over.
Machine Generated
개가 제트기 위를 지나간다
A dog is passing over the jet.
Evaluation
BLEU
ROUGE
METEOR
BERTScore
Coverage
Human
Semi-Automatic Dataset Construction
Stage 1. Automatic Dataset Construction
Data Source
Image Caption Data
Dialog Summary Data
Content Morpheme Extraction
Concept extraction
지나가- 개 제트기
(pass) (dog) (jet)
Filter 1
Named Entity Filtering
Overly specific sentences
… LOCATION … ARTIFACT …
Filter 2
Sentence Level Filtering
개
(dog)
Sentence (Index 3)
Sentence (Index 45)
Sentence (Index 3829)
Sequence 1: Index 3 sentence
Sequence 2: Index (45 … 3829) sentences
KoBERT-STS
Delete
machine-like & awkward sentences
excessively used concepts
Filter 3
Ethical Consideration #1
KoELECTRA
HATE SPEECH DETECTION
Gender-biased
Aggressive
Automatic Constructed Dataset
Filtered Dataset
Stage 2. Annotate and Refine by Crowd-sourcing
Automatic Constructed Dataset
Filtered Dataset
Ethical Consideration #2
Expert Evaluator
Reference Generation
Human Annotator
Quality Control
Expert Evaluator

Figure 2: Overview of Korean CommonGen and pipeline of semi-automatic dataset construction.

Data Sources In this study, we organize the dataset with the image caption and dialogue summary sentences describing daily life based on commonsense knowledge that does not correspond to the specific or professional domains.

For data construction, we utilize the Korean image caption dataset released by AI-HUB visual intelligence, which is translated from the original English MS-COCO dataset. The caption sentences in this dataset describe the scenes that occur in everyday life. It implies that universal commonsense knowledge implicitly lies in those sentences. As they are the combinations of objects, and relations exist in the corresponding (ground-truth) image, it is appropriate for models to learn how to compose sentences by considering all the given concepts.

The Korean dialogue summary dataset of AI-HUB is also adopted for data construction. These data contain conversations on everyday topics and include commonsense knowledge that has implicit Korean sociocultural content. We secure the diversity of sentences and commonsense knowledge by adding non-visual conversation contents not solely using image caption sentences based on visual information, such as conventional CommonGen.

We delete the sentences involving the unexpected foreign words and special symbol tokens in the phase of preprocessing. Both datasets consist of sentences that take the form of declarative statements. In this process, we unify the structure of sentences to be ended with '다.' (i.e., an ending word in Korean declarative statement) for minimizing the variances in performance evaluation according to the decoding hyperparameters or strategies.

Content Morpheme Extraction POS tagging is the task that assigns the grammatical group tag to the text based on the language's perspective and definition (Kanakaraddi and Nandyal, 2018). We utilize POS tagging to extract the content morphemes essential for making concept sets. We apply the morpheme segmentation of ko-mecab (KUDO, 2005) using the Korean morpheme analysis package KoNLPy (Park and Cho, 2014).

Korean is a highly inflected language with many inflectional morphemes and has multiple POS tag patterns. In addition, Korean is an agglutinative language. If the Korean language is tokenized with *eojeol* segmentation by white spaces, the number of vocabulary units increases exponentially, and the accuracy of correct tagging decreases. Korean embedded models also operate morpheme segmentation to avoid expensive computational costs caused by the exponential increase in new vocabularies with *eojeol* segmentation (Lee et al., 2020; Kim and Colineau, 2020). Therefore, we adopt the morpheme segmentation method to improve efficiency and suitability.

When creating the concept set, we employ the content morphemes. The content morphemes include the definite actions, states, or semantic information of the sentence. They involve the verb and adjective stems, and some adjectives are grammatically similar to the verbs in Korean. Therefore, it is possible to create a sentence with the content morphemes, considering both the semantic relations between them and their grammatical usage.

Based on the criteria of the content morpheme tagging, we classify the nouns (`NNG, NNP, NNB, NNBC, NR, NP`), determiners (`MM`), adverbs (`MAG, MAJ`), verbs (`VV, VA, VX, VCP, VCN`), radixes (`XR`), and interjections (`IC`) as the content morpheme[4]. However, the proper noun (NNP), numeral (NR), pronoun (NP) are only understandable in certain situations and perspectives among the content morpheme. Thus, we delete the sentences including NNP, NR, and NP in the process of morpheme segmentation. The detailed differences according to the segmentation are described in §6.

Named Entity Filtering NER is a subtask of Information Extraction, which attempts to recognize the named entities such as a person, location, and quantity from the unstructured text (Nadeau and Sekine, 2007). We utilize the NER to remove the sentences including the non-commonsense knowledge from the data sources.

As mentioned above, the Korean dialogue summary dataset of AI-HUB carries the daily conversations, which means it contains commonsense knowledge in abundance. However, there is a possibility that the sentences consist of specific domain knowledge that is shared only by the particular group or time. Named entities, such as specific names of persons, organizations, and locations, are not commonsense entities because most people do not comprehend them. For instance, "윤호는 착해 보이지만 연예인 기질은 아닌 것 같다. (Yoonho looks kind, but he is not talented for the entertainment.)", name of person '윤호(Yoonho)' is vague to be categorized as commonsense knowledge.

We vacate the sentences containing the non-commonsense knowledge with NER tagging in the phase of dataset construction. In the NER tagging, we adopt the neural network model in Pororo library[5] and remove the 119,355 sentences containing the non-commonsense named entities.

Sentence Level Filtering Through NER and POS tagging, a substantial number of sentences with non-commonsense knowledge can be filtered out. However, there remain several sentences that contain awkward translations or do not properly reflect commonsense knowledge. To filter out these residues, we selectively extract sentences that follow common Korean sentence structure and contain rich commonsense knowledge at the sentence level.

These are proceeded by the comparison between sentences within the same morpheme. For the selection process, we apply inverted index to every $s \in S$ that contains unique content morpheme $x \in X$. A set of inverted indexed sentences for each context morpheme x is denoted as $s^x = \{s_i^x\}_{i=1}^{N_s}$, where the maximum size of N_s is set to 100.

Then, based on the contextual representation embedding of the language model, we estimate the similarity score between every two diverse sentences in s^x using the KoBERT[6] fine-tuned with the KorSTS (Ham et al., 2020) dataset. More specifically, the similarity score $\hat{y}_{i,j}$ between two sequences s_i^x, s_j^x is estimated as shown in Equation 1.

$$\hat{y}_{i,j} = \sigma(W(h_{ij}) + b) \tag{1}$$

h_{ij} indicates the KoBERT encoded representation of the concatenated sequence s_i^x, s_j^x. σ denotes activation unit and W, b are trainable parameters of a linear pooling layer. To figure out descent sentences that fluently follow the common Korean sentence form and contain rich commonsense knowledge within the same content morpheme, we set up $score_i^x$ for each sentence s_i^x. $score_i^x$ is estimated by summing up all the similarity score $\hat{y}_{i,j}$ between s_i^x and other sentences, as shown in Equation 2.

$$score_i^x = \left(\sum_{j=1}^{N_s} \hat{y}_{i,j}\right) - \hat{y}_{i,i} \tag{2}$$

We evaluate $score_i^x$ for each sentence s_i^x and sort all the sentences in a descending order. According to their scores, the top-2 sentences are selected for each unique content morpheme.

Ethical Consideration Machine-translated MS-COCO can deviate from the intended purpose or have aggressive terms because of cultural differences. Also, a dialogue summary of daily conversations can include socially inappropriate or discriminatory content. Thus, we filter out these expressions so that models cannot unintentionally return inappropriate responses to some triggers.

To detect unethical expressions, we use the Korean Hate Speech Detection (Moon et al., 2020) dataset. As a model, KoELECTRA[7] is pretrained with the ELECTRA (Clark et al., 2019) structure and Korean corpora. KoELECTRA is trained to classify whether input sequences contain gender biased or aggressive representation. The classifier's predicted 1,083 results are potentially unethical

[4] https://www.korean.go.kr/front/onlineQna/onlineQnaView.do?mn_id=216&qna_seq=209597

[5] https://github.com/kakaobrain/pororo

[6] https://github.com/SKTBrain/KoBERT

[7] https://github.com/monologg/KoELECTRA

risk statements, including either gender bias or offensive. Among them, 172 sentences contain both gender biased and offensive problems.

However, we note that the classifier is not complete (Roller et al., 2021); thus, we reschedule those results with a secondary inspection through two expert annotators. The classifier tends to be overly sensitive to words that refer to a particular gender (e.g., male, female) or words that can be used aggressively (e.g., cut, bound, etc.). In addition, the published source data: MSCOCO and AI-HUB Dialogue Summary, have completed pre-validation and data curation. We find the problem that most of the predicted sentences are false-positive. Therefore, we conduct a second round evaluation to consider whether the sentences conform to the definition of commonsense knowledge and do not deviate from social norms. In the second round, two human annotators majoring in linguistics and computer science remove 124 sentences with criminal, drug, and excessive biased among predicted 1,083 results.

3.2 Annotate and Refine by Crowd-sourcing

We generate additional reference sentences using the extracted content morphemes via crowd-sourcing. Employing the expert human annotators, we also check the quality of the references and implement secondary inspection on the dataset to filter out sentences that contain ethical issues.

Reference Generation To evaluate the diversity of model-generated sentences, we produce additional 4,080 references based on the 2,040 concept set designated as test data. We employ 22 human annotators with bachelor's degrees through crowd-sourcing[8]. The working guidelines are as follows: First, we ensure the additional references are not similar or do not merely change the position of the subject/verb/object. Second, the given content morphemes are preserved, and the annotators can append extra modifiers conforming to commonsense. Third, references do not incorporate overly specific named entities or numerical expressions.

Quality Control 17 expert annotators holding bachelor's degrees in Korean language or linguistic secure the quality of the automatic constructed data via secondary inspection and assessment[9]. Since the references determine the model's performance, human correction is performed on the references to maintain high quality leading to 303 inappropriate sentences being removed.

4 Experiments Settings

4.1 Evaluation Metrics

The evaluation metrics consist of n-gram overlapping, semantic similarity, content morpheme coverage, and human score. We use automatic evaluation metrics based on n-gram overlapping such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and METEOR (Banerjee and Lavie, 2005). These metrics compute the token level similarity between model-generated candidate and reference sentences. Next, we utilize BERTScore (Zhang et al., 2019) as semantic similarity and evaluate the outputs using mBERT and KoBERT to identify differences between multilingual and monolingual models. We also indicate the concept coverage which is the average percentage of given concepts that exist in model-generated sentences.

To estimate human evaluation, we employ 17 expert annotators as per the conditions specified in §3.2 to evaluate four criteria as follows: (i) Grammar Correction: Is it a valid sentence for Korean grammar?; (ii) Factuality: Does it contain the given content morphemes as much as possible?; (iii) Commonsense: Is it following commonsense knowledge?; (iv) Fluency: Is it a natural sentence for a mother tongue speaker? The human annotators score each measure with 2 points for excellent, 1 point for regular, and 0 points for insufficient. Moreover, we estimate human annotator performance by considering their reference sentences in the test set. We develop a system of prediction by comparing each annotator's references to calculate inter-annotator agreement. We measure the inter-annotator agreement of 3 evaluators with Fleiss' Kapa (Fleiss, 1971). Overall Fleiss' Kapa coefficient correlations for each human evaluation are Commonsense 0.426, Factuality 0.478, Fluency 0.401, and Grammar Correction 0.344; therefore have moderate reliability among evaluators[10].

4.2 Baselines

The baselines include GPT2 (Radford et al., 2019) using only the decoder structure of Transformer (Vaswani et al., 2017) and BART (Lewis et al., 2020) using the encoder-decoder. Among

[8] The construction cost for one sentence is 0.13$, and the working period is three weeks.

[9] The inspection cost for one sentence is 0.04$, and the working period is two weeks.

[10] Each model-generated sentence is evaluated by randomly selected 3 among 17 human annotators.

Model	Size	BLEU 3	BLEU 4	ROUGE-2	ROUGE-L	METEOR	mBERTScore	KoBERTScore	Coverage
KoGPT2 (Radford et al., 2019)	125M	29.24	18.91	43.36	60.41	39.89	84.08	90.92	79.43
KoBART (Lewis et al., 2020)	124M	39.54	29.16	53.60	68.55	51.17	87.41	92.59	93.65
mBART (Liu et al., 2020)	610M	41.83	31.63	54.21	68.36	52.08	87.25	92.26	91.39
mBART-50 (Tang et al., 2020)	610M	40.51	30.20	53.50	68.18	50.90	87.31	92.26	91.71
mT5-small (Xue et al., 2021)	300M	34.18	23.29	49.48	66.46	46.10	87.39	92.28	92.02
mT5-base (Xue et al., 2021)	580M	40.87	30.22	54.87	70.21	51.76	88.15	92.77	94.83
mT5-large (Xue et al., 2021)	**1280M**	**46.33**	**35.90**	**58.91**	**72.78**	**56.52**	**88.54**	**92.92**	**95.07**
Human Performance		49.12	41.64	61.02	73.29	58.60	91.13	95.26	98.30

Table 2: Experimental results of various baselines on the Korean CommonGen test set.

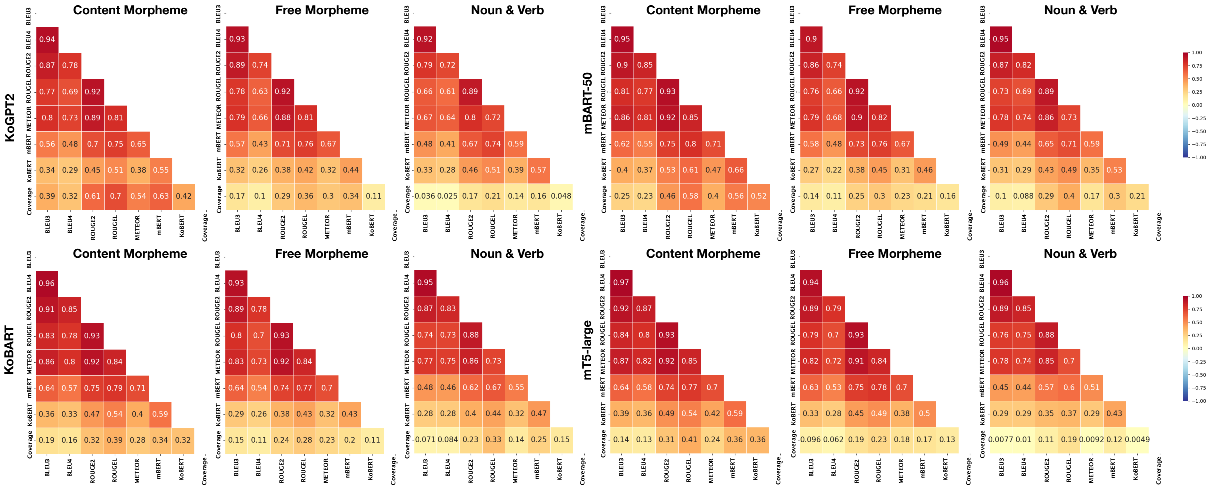


Figure 3: Correlation of automatic evaluation metrics among Korean and multilingual language models. The box's color is deep with reddish, and the score is the more extensive the white boldface, the higher correlation.

the models with the mentioned structure, we use KoGPT2[11] and KoBART[12] pretrained with Korean corpora. We also conduct experiments on mBART (Liu et al., 2020), mBART-50 (Tang et al., 2020), and mT5 (Xue et al., 2021) based on encoder-decoder as multilingual language models.

5 Quantitative Evaluation

We conduct a quantitative analysis of our dataset as shown in Table 2. First, encoder-decoder models exhibit higher performance than the decoder-only model. The decoder-only model is limited in reconstructing input content morphemes into acceptable quality sentences based on generative commonsense reasoning only with uni-directional prompt engineering (Liu et al., 2021a). Moreover, the encoder-decoder models with an encoder can formulate sentences based on bi-directive embedding information for given content morphemes.

Second, KoBART has a model parameter of 124M which is smaller than that of the mBART and mBART-50 models (whose model parameters are 610M) but exhibits partially competent performance. This result shows that if the models are of identical structure, the multilingual model with a high proportion of machine-translated data in the pre-training process may encounter difficulties in generating high-quality sentences based on generative commonsense reasoning.

Third, mBART, and mBART-50 show similar performance. In the case of mBART-50, low-resource languages have the effect of improving, but high-resource languages present partial degradation because of the curse of dimension. Korean is a medium-resource language, but its performance decreases like other high-resource languages.

Forth, mT5-large model has the most model parameters of 1.3B, and the pre-training method using sequence-to-sequence task form with prompt engineering appears to impact it positively. Additionally, the mT5-large shows the most comparable execution to human performance and higher coverage and semantic scores than BARTs. However, most of small size baselines still have tribulation generating sentences containing all of the given content morphemes, and overall performance is lower than humans.

[11] `https://github.com/SKT-AI/KoGPT2`
[12] `https://github.com/SKT-AI/KoBART`

6 Ablation Study

We attempt to demonstrate the validity of concept extraction and data sources. Thus, we perform ablation studies with respect to the concept set and data source configuration method.

6.1 Other Than The Content Morphemes

The first ablation study is conducted by tokenizing the set of concepts into noun and verb as suggested in CommonGen and free morphemes that can be used alone depending on the presence or absence of independence. According to the configuration method of the concept set, Figure 3 shows the correlation between automatic evaluation metrics and Table 3 exhibits the performance of the baselines.

The p-value for the correlation of all evaluation metrics is less than 0.05 (statistical significance). As depicted in Figure 3, content morpheme concepts have the highest overall correlation with other evaluation metrics. Table 3 also shows that this approach has the highest performance in most evaluation indicators. Through these results, we find that constructing concepts based on content morpheme has minimized the loss of information required for complete Korean sentences.

When it comes to free morpheme concepts, the concept set does not include the vocabulary of verbs and adjectives. As described in Figure 3, difficulty in constituting sentences increases dramatically, and correlations with the evaluation metrics are lowered by omitting information on verbs and adjectives, which are core components of a sentence. Consequently, free morpheme concepts overlook most of the valuable information, making it challenging to infer essential semantic components to compose sentences and exhibit inferior performance, as shown in Table 3.

Next, in the case of noun and verb concepts, Figure 3 shows the lowest overall correlation with other evaluation metrics. Table 3 also presents the performance gap between evaluation metrics which is considerable on account of omitting information on adjectives and adverbs according to the combination of concepts. In addition, it is tough to reproduce the relationship between the word root and ending according to the conjugation of Korean in forming sentences. Therefore, all baselines have difficulty with sentence composition considering inflectional units and indicate the lowest coverage on average.

Among the configuration methods of the concept set, content morpheme concepts have the highest overall correlation with other evaluation metrics. This method also shows that the loss of information required for complete Korean sentences is minimized and that sentence composition considering the Korean linguistic features is also possible.

Free Morph	**BLEU3/4**	**ROUGE2/L**	**METEOR**	**m/koBERTScore**	**Coverage**
KoGPT2	21.22/12.39	35.60/55.15	32.48	81.02/89.92	75.74
KoBART	24.02/14.78	39.65/57.91	39.01	82.71/90.98	84.27
mBART-50	23.31/13.78	40.49/58.63	39.36	83.79/90.73	84.63
mT5-large	**29.54/19.39**	**44.76/62.72**	**42.59**	**84.07/91.02**	**85.79**
Noun & Verb	**BLEU3/4**	**ROUGE2/L**	**METEOR**	**m/koBERTScore**	**Coverage**
KoGPT2	27.27/17.21	37.60/53.43	30.07	81.35/86.02	67.48
KoBART	42.27/31.53	52.83/66.32	42.83	85.10/88.22	80.22
mBART-50	44.70/33.82	54.18/67.16	43.73	85.39/88.11	80.08
mT5-large	**52.37/41.15**	**59.73/71.67**	**48.27**	**86.61/88.50**	**83.35**
Content Morph	**BLEU3/4**	**ROUGE2/L**	**METEOR**	**m/koBERTScore**	**Coverage**
KoGPT2	29.24/18.91	43.36/60.41	39.89	84.08/90.92	79.43
KoBART	39.54/29.16	53.60/68.55	51.17	87.41/92.59	93.65
mBART-50	40.51/30.20	53.50/68.18	50.90	87.31/92.26	91.71
mT5-large	**46.33/35.90**	**58.91/72.78**	**56.52**	**88.54/92.92**	**95.07**

Table 3: Ablation study for concept tokenization method. Baselines train with **Free Morph**: Free morpheme concepts, **Noun & Verb**: Noun and verb concepts, and **Content Morph**: Content morpheme concepts.

IC	**BLEU3/4**	**ROUGE2/L**	**METEOR**	**m/koBERTScore**	**Coverage**
KoGPT2	20.66/12.63	34.87/53.46	31.76	79.49/89.09	68.85
KoBART	33.19/23.62	48.31/65.15	44.84	86.32/92.19	90.38
mBART-50	33.61/24.21	47.36/64.41	43.83	86.13/92.13	87.20
mT5-large	**40.19/29.75**	**53.77/69.66**	**50.62**	**88.00/92.74**	**94.44**
DS	**BLEU3/4**	**ROUGE2/L**	**METEOR**	**m/koBERTScore**	**Coverage**
KoGPT2	17.13/9.22	30.45/49.03	27.70	80.26/88.70	71.84
KoBART	23.62/14.44	39.48/58.28	37.23	85.14/91.08	91.32
mBART-50	23.78/14.71	38.48/57.80	35.63	84.86/90.93	89.78
mT5-large	**33.96/23.12**	**48.11/65.95**	**45.72**	**86.91/92.14**	**94.67**
IC&DS	**BLEU3/4**	**ROUGE2/L**	**METEOR**	**m/koBERTScore**	**Coverage**
KoGPT2	29.24/18.91	43.36/60.41	39.89	84.08/90.92	79.43
KoBART	39.54/29.16	53.60/68.55	51.17	87.41/92.59	93.65
mBART-50	40.51/30.20	53.50/68.18	50.90	87.31/92.26	91.71
mT5-large	**46.33/35.90**	**58.91/72.78**	**56.52**	**88.54/92.92**	**95.07**

Table 4: Ablation study for the data source. Baselines train with **IC**: Image caption data, **DS**: Dialogue summary data, and **IC & DS**: Image caption and dialogue summary data.

6.2 Image Caption or Dialogue Summary

Table 4 shows the results of baselines by dividing data source into image caption and Korean dialogue summary. The evaluation data includes both image captions and dialogue summarized sentences.

The image caption is straightforward but still contains awkward expressions owing to machine-translated results. Moreover, the Korean dialogue summary is relatively elaborated and incorporates a natural expression of mother-tongues to a wide range of everyday conversation topics.

As described in Table 4, the entire data combining the attributes of the two sources demonstrate the highest performance. These results show that our dataset requires more diverse sentence composition and commonsense reasoning processes than solely using image caption sources offered by CommonGen (Lin et al., 2020). Additionally, the dialogue summary includes various sociocultural commonsense knowledge of mother-tongues. It is challenging to feed all contexts with only a small amount of data. Thus, models trained only with dialogue summary show lower performance than those trained only with image caption. Considering experimental results of Table 4, we organize the training such that the model first learns image captions and then highly complicated dialogue summarized sentences.

7 Human Evaluation

We further conduct a human evaluation for model-generated sentences, as shown in Table 5. The experimental results show that the mT5 model achieves the highest score and tendency of the score distribution parallel to the automatic evaluations. This point indicates that the estimation results with automatic metrics obtained from earlier experiments are comparable to human agreements. Closely examining the four evaluation criteria of human scores reveals that the baselines show the lowest score on average in fluency compared to other evaluation criteria and the highest score on average in factuality. This means that the models are relatively well-trained to use content morphemes in complete sentences, but their ability to generate spontaneous sentences is insufficient.

The correlation score between each evaluation criteria of human evaluation and automatic evaluation metrics is described in Figure 4. The heatmap shows that recall-based metrics have high correlations with human scores. This result corresponds to the demonstrations in (Lin, 2004) and (Banerjee and Lavie, 2005). In the case of coverage, a different trend showing low correlations with other human scores except for factuality is observed. It can be predicted that the criteria we have suggested and the basis for human judgment are somewhat consistent. The lower correlation in factuality with automatic evaluation metrics shows that containing all the given concepts does not necessarily constitute a well-crafted sentence.

Moreover, human-annotated scores have statistical significance in correlation with automatic evaluation metrics (p.value < 0.05), and show considerably similar values to Figure 3 in BERT-based semantic scores. However, there are weak positive correlations with other automatic evaluation metrics. These results show that it is difficult to evaluate the model's generative commonsense reasoning in the traditional estimation approach, and advanced research is needed to improve it.

Human evaluation	GC	CS	FC	FL	TT
KoGPT2	0.85	0.74	1.27	0.64	3.50
KoBART	1.32	1.25	1.74	1.15	5.49
mBART-50	1.31	1.24	1.71	1.15	5.40
mT5-large	1.44	1.36	1.80	1.28	5.89

Table 5: Human evaluation for model-generated outputs including **GC**: Grammar Correction, **CS**: Commonsense, **FC**: Factuality, **FL**: Fluency, and **TT**: Total

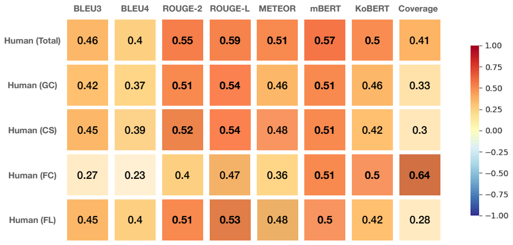


Figure 4: Correlation heatmap between automatic evaluation metrics (X axis) and human scores (Y axis). We experiment dividing Y axis into human-annotated **GC**: Grammar Correction, **CS**: Commonsense, **FC**: Factuality, **FL**: Fluency, and total score.

8 Conclusion

In this paper, we propose a Korean CommonGen dataset including Korean sociocultural commonsense knowledge and morpheme-based linguistic features. The dataset heeds the semi-automatic dataset construction method based on automatic construction and crowd-sourcing annotation with quality assessment. We perform a comparative analysis and an ablation study to demonstrate the validity of the dataset and evaluation metrics for generative commonsense reasoning. Moreover, we conduct a Korean and multilingual language models' standard performance experiment to investigate the dataset's problems and competencies. In future work, the dataset will enhance evaluation metrics regarding syntax and diversity of sentences to improve the interpretation of the model-generated results. We believe that our proposed dataset will serve as a fundamental resource to Korean NLG and commonsense reasoning research.

Acknowledgments

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2018-0-01405) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation). This work was supported by Institute of Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (No. 2020-0-00368, A Neural-Symbolic Model for Knowledge Acquisition and Inference Techniques). This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(NRF-2021R1A6A1A03045425).

References

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2020. Evaluation of text generation: A survey. *arXiv preprint arXiv:2006.14799*.

Noam Chomsky. 1965. Aspects of the theory of syntax.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2019. Electra: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Sebastian Gehrmann, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Anuoluwapo Aremu, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna-Adriana Clinciu, Dipanjan Das, Kaustubh Dhole, Wanyu Du, Esin Durmus, Ondřej Dušek, Chris Chinenye Emezue, Varun Gangal, Cristina Garbacea, Tatsunori Hashimoto, Yufang Hou, Yacine Jernite, Harsh Jhamtani, Yangfeng Ji, Shailza Jolly, Mihir Kale, Dhruv Kumar, Faisal Ladhak, Aman Madaan, Mounica Maddela, Khyati Mahajan, Saad Mahamood, Bodhisattwa Prasad Majumder, Pedro Henrique Martins, Angelina McMillan-Major, Simon Mille, Emiel van Miltenburg, Moin Nadeem, Shashi Narayan, Vitaly Nikolaev, Andre Niyongabo Rubungo, Salomey Osei, Ankur Parikh, Laura Perez-Beltrachini, Niranjan Ramesh Rao, Vikas Raunak, Juan Diego Rodriguez, Sashank Santhanam, João Sedoc, Thibault Sellam, Samira Shaikh, Anastasia Shimorina, Marco Antonio Sobrevilla Cabezudo, Hendrik Strobelt, Nishant Subramani, Wei Xu, Diyi Yang, Akhila Yerukola, and Jiawei Zhou. 2021. The GEM benchmark: Natural language generation, its evaluation and metrics. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 96–120, Online. Association for Computational Linguistics.

Younggyun Hahm, Youngbin Noh, Ji Yoon Han, Tae Hwan Oh, Hyonsu Choe, Hansaem Kim, and Key-Sun Choi. 2020. Crowdsourcing in the development of a multilingual framenet: A case study of korean framenet. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 236–244.

Jiyeon Ham, Yo Joong Choe, Kyubyong Park, Ilji Choi, and Hyungjoon Soh. 2020. Kornli and korsts: New benchmark datasets for korean natural language understanding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 422–430.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos qa: Machine reading comprehension with contextual commonsense reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2391–2401.

Suvarna G Kanakaraddi and Suvarna S Nandyal. 2018. Survey on parts of speech tagger techniques. In *2018 International Conference on Current Trends towards Converging Technologies (ICCTCT)*, pages 1–6. IEEE.

Daniel Keysers, Nathanael Schärli, Nathan Scales, Hylke Buisman, Daniel Furrer, Sergii Kashubin, Nikola Momchev, Danila Sinopalnikov, Lukasz Stafiniak, Tibor Tihon, Dmitry Tsarkov, Xiao Wang, Marc van Zee, and Olivier Bousquet. 2020. Measuring compositional generalization: A comprehensive method on realistic data. In *International Conference on Learning Representations*.

Myung Hee Kim and Nathalie Colineau. 2020. An enhanced mapping scheme of the universal part-of-speech for korean. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3826–3833.

T KUDO. 2005. Mecab: Yet another part-of-speech and morphological analyzer. *http://mecab. sourceforge. net/*.

Brenden Lake and Marco Baroni. 2018. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *35th International Conference on Machine Learning, ICML 2018*, pages 4487–4499. International Machine Learning Society (IMLS).

Young-Jun Lee, Chae-Gyun Lim, and Ho-Jin Choi. 2020. Korean-specific emotion annotation procedure using n-gram-based distant supervision and korean-specific-feature-based distant supervision. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 1603–1610.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Haonan Li, Yeyun Gong, Jian Jiao, Ruofei Zhang, Timothy Baldwin, and Nan Duan. 2021. KFCNet: Knowledge filtering and contrastive learning for generative commonsense reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2918–2928, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Bill Yuchen Lin, Wangchunshu Zhou, Ming Shen, Pei Zhou, Chandra Bhagavatula, Yejin Choi, and Xiang Ren. 2020. Commongen: A constrained text generation challenge for generative commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 1823–1840.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Hugo Liu and Push Singh. 2004. Conceptnet—a practical commonsense reasoning tool-kit. *BT technology journal*, 22(4):211–226.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021a. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*.

Ye Liu, Yao Wan, Lifang He, Hao Peng, and S Yu Philip. 2021b. Kg-bart: Knowledge graph-augmented bart for generative commonsense reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 6418–6425.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Andrew Matteson, Chanhee Lee, Youngbum Kim, and Heui-Seok Lim. 2018. Rich character-level information for korean morphological analysis and part-of-speech tagging. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2482–2492.

Jihyung Moon, Won Ik Cho, and Junbum Lee. 2020. BEEP! Korean corpus of online news comments for toxic speech detection. In *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media*, pages 25–31, Online. Association for Computational Linguistics.

Sangwhan Moon and Naoaki Okazaki. 2020. Jamo pair encoding: Subcharacter representation-based extreme korean vocabulary compression for efficient subword tokenization. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3490–3497.

David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Eunjeong L Park and Sungzoon Cho. 2014. Konlpy: Korean natural language processing in python. In *Annual Conference on Human and Language Technology*, pages 133–136. Human and Language Technology.

Sungjoon Park, Jihyung Moon, Sungdong Kim, Won Ik Cho, Ji Yoon Han, Jangwon Park, Chisung Song, Junseong Kim, Youngsook Song, Taehwan Oh, et al. 2021. Klue: Korean language understanding evaluation. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020.

Xcopa: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4932–4942.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325, Online. Association for Computational Linguistics.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. 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 4149–4158.

Niket Tandon, Aparna S Varde, and Gerard de Melo. 2018. Commonsense knowledge in machine intelligence. *ACM SIGMOD Record*, 46(4):49–52.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv preprint arXiv:2008.00401*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30:5998–6008.

Han Wang, Yang Liu, Chenguang Zhu, Linjun Shou, Ming Gong, Yichong Xu, and Michael Zeng. 2021. Retrieval enhanced model for commonsense generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3056–3062.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

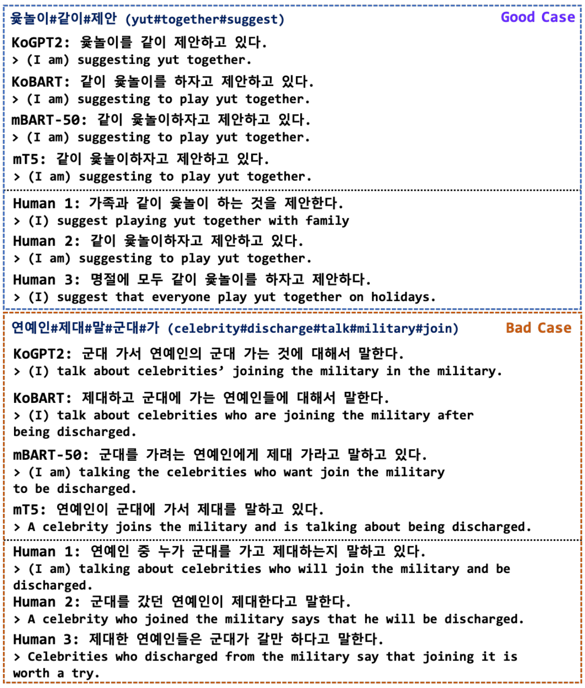
A Qualitative Evaluation

We conduct a qualitative analysis on the machine-generated sentences as illustrated in Figure 5. Qualitative evaluation is executed by composing the content morphemes of the test set to generate sentences based on commonsense. In the analysis, we separate the model-generated results embracing the Korean commonsense knowledge into the good case and the bad case.

Good Case In the case that requires low complicated compositionality and simple commonsense reasoning, the models tend to generate identical sentences. When KoGPT2 generates the sentences, the result shows prominent differences with the outputs of other models, and it has an inconsistent grammatical correction. On the contrary, KoBART, mBART-50, and mT5 yield brief sentences that adopt relatively stable grammar rules and postpositional expressions.

Bad Case In the case that requires complicated compositionality and struggling commonsense reasoning, the gap of qualities between the model-generated sentences is more significant than the opposite case. KoGPT2 generates incoherent sentences that exploit the given content morphemes repetitively or contain many errors in commonsense reasoning. KoBART induces more natural Korean sentences than the other models, but the results are still harsh to understand owing to their ambiguous expressions. From the multilingual language models, mBART-50 yields unnatural sentences like machine-translated sentences. In addition, the models show semantically incorrect results as if the models misunderstand the meaning of some words like '제대(discharge).' The mT5 employs more manageable words and brief sentences, including all given content morphemes as much as possible.

Although the ability of the Korean monolingual language model's compositionality still seems insufficient, the results show that the models partially understand the content morpheme with a cultural specialty such as the conscription system. In contrast, the multi-language model shows substantial compositionality and abnormal usage of the content morpheme with cultural specialty as its limitation.



윷놀이#같이#제안 (yut#together#suggest) Good Case

KoGPT2: 윷놀이를 같이 제안하고 있다.
> (I am) suggesting yut together.

KoBART: 같이 윷놀이를 하자고 제안하고 있다.
> (I am) suggesting to play yut together.

mBART-50: 같이 윷놀이하자고 제안하고 있다.
> (I am) suggesting to play yut together.

mT5: 같이 윷놀이하자고 제안하고 있다.
> (I am) suggesting to play yut together.

Human 1: 가족과 같이 윷놀이 하는 것을 제안한다.
> (I) suggest playing yut together with family

Human 2: 같이 윷놀이하자고 제안하고 있다.
> (I am) suggesting to play yut together.

Human 3: 명절에 모두 같이 윷놀이를 하자고 제안하다.
> (I) suggest that everyone play yut together on holidays.

연예인#제대#말#군대#가 (celebrity#discharge#talk#military#join) Bad Case

KoGPT2: 군대 가서 연예인의 군대 가는 것에 대해서 말한다.
> (I) talk about celebrities' joining the military in the military.

KoBART: 제대하고 군대에 가는 연예인들에 대해서 말한다.
> (I) talk about celebrities who are joining the military after being discharged.

mBART-50: 군대를 가려는 연예인에게 제대 가라고 말하고 있다.
> (I am) talking the celebrities who want join the military to be discharged.

mT5: 연예인이 군대에 가서 제대를 말하고 있다.
> A celebrity joins the military and is talking about being discharged.

Human 1: 연예인 중 누가 군대를 가고 제대하는지 말하고 있다.
> (I am) talking about celebrities who will join the military and be discharged.

Human 2: 군대를 갔던 연예인이 제대한다고 말한다.
> A celebrity who joined the military says that he will be discharged.

Human 3: 제대한 연예인들은 군대가 갈만 하다고 말한다.
> Celebrities who discharged from the military say that joining it is worth a try.

Figure 5: A case study comparing commonsense generation in the test set. We categorize the good and bad cases based on the model-generated sentences' quantitative evaluation. The two cases include Korean commonsense knowledge, 'Yut' is related to traditional Korean games, and 'military' and 'discharge' are related to the conscription, which expresses Korean sociocultural characteristics.

B Further Experiments

We conduct experiments to verify the training method of commonsense reasoning, assessment on the data dependency, and expandability within the commonsense domain.

B.1 High-Level Commonsense Reasoning

Table 6 shows the evaluation results on the models trained by randomly deleting one concept to evaluate the high-level commonsense reasoning. The seeds for the random deletion are set to $\{42, 52, 62, 72, 82\}$. Because of the random deletion, the training data get more challenging as if the model composed the sentence reasoning with at least one content morphemes not given. The performance of the models decreases compared with the model trained with the entire concept set. Deleted content morphemes affect the disappearance of the commonsense knowledge rather than the enhancement of the commonsense reasoning during the training phase. In the case of mBARTs, the extent of the performance decline is the most substantial, and mBARTs show inferior performance than KoBART on average. In addition, compared with the other models, mT5-base is sensitive to a given situation of the source input. This result means that a BART model pretrained with monolingual corpora is significantly robust, regardless of its small model size. The models trained with the identical architecture and corpora show a considerably different extent of optimization according to the model size in text generation.

B.2 Reformulated CommonGen Test set

In the following process, we introduce a new evaluation dataset that utilizes the raw data sources that are of little relevance to the training dataset.

First, based on the knowledge graph, we extract the concepts at the one-hop links ConceptNet (Liu and Singh, 2004). The top 20 ranked concepts with high weights are selected among the 25 according to the categories suggested in CommonGen. Second, the concepts are broadened based on the ConceptNet knowledge graph tagged with the grammatical role and translated as vocabulary units through NAVER dictionary[13] crawling. Third, we employ two professional annotators who possess high comprehension of the task and are proficient in Korean as a native language[14]. The annotators create a new Korean evaluation dataset combining the broadened concepts based on the CommonGen test set and ConceptNet knowledge graph. Fourth, we construct the dataset with the created sentences using the reverse extraction method used in §3.1. Therefore, the new evaluation dataset consists of 1,083 sentences embracing Korean commonsense knowledge with flawless syntax. Furthermore, utilizing the concepts of the knowledge graph ConcepNet representative in the commonsense domain, we evaluate the dataset's dependency on the raw data source of the training dataset.

The performance of the models evaluate with the reformulated test set, in which two annotators create the sentences using the ConceptNet and CommonGen test set, are shown in Table 7. We train the models involving the same data and hyperparameters used in the experiment of Table 2. The models show the maintained performance with the new dataset constructed by a human using the other data source within the commonsense domain. Therefore, this result demonstrates that the models are not substantially dependent on the specific

[13] https://dict.naver.com/

[14] The construction cost for one sentence is 0.3$, and the working period is two weeks.

Model	Seed	BLEU 3	BLEU 4	ROUGE-2	ROUGE-L	METEOR	mBERTScore	KoBERTScore	Coverage
KoGPT2 (Radford et al., 2019)	42	24.23	15.24	39.73	56.56	38.09	82.29	89.09	78.41
KoGPT2	52	22.39	13.75	37.54	54.44	35.87	81.49	88.80	75.84
KoGPT2	62	24.39	15.19	40.02	56.70	38.77	83.10	90.16	79.85
KoGPT2	72	23.53	14.74	38.82	55.44	37.74	82.40	89.31	79.00
KoGPT2	82	26.14	16.67	40.73	57.78	38.48	82.53	89.71	76.93
KoGPT2 Mean (± Stdev)		24.14(± 1.37)	15.12(± 1.05)	39.37(± 1.23)	56.18(± 1.28)	37.79(± 1.14)	82.36(± 0.58)	89.41(± 0.53)	78.01(± 1.61)
KoBART (Lewis et al., 2020)	42	30.11	19.53	47.97	64.29	47.63	85.22	91.78	92.95
KoBART	52	30.31	20.05	47.55	63.76	47.42	85.37	91.65	92.67
KoBART	62	30.34	20.10	47.71	63.61	47.81	87.23	91.73	92.58
KoBART	72	30.32	19.82	47.70	63.92	47.69	85.20	91.69	92.68
KoBART	82	30.63	20.19	47.97	64.19	47.75	85.16	91.70	92.86
KoBART Mean (± Stdev)		30.34(± 0.19)	19.94(± 0.27)	47.78(± 0.18)	63.95(± 0.29)	47.66(± 0.15)	85.24(± 0.08)	91.71(± 0.05)	92.75(± 0.15)
mBART (Liu et al., 2020)	42	28.61	17.95	45.53	62.49	45.42	84.64	91.11	90.47
mBART	52	28.97	18.22	45.97	62.87	45.29	84.73	91.18	90.23
mBART	62	29.05	18.48	46.07	62.83	46.21	84.90	91.32	90.95
mBART	72	28.69	18.28	45.97	62.70	45.89	84.53	91.35	90.75
mBART	82	29.24	18.56	46.06	62.77	45.72	84.65	91.24	90.51
mBART Mean (± Stdev)		28.91(± 0.26)	18.30(± 0.24)	45.92(± 0.22)	62.73(± 0.15)	45.71(± 0.37)	84.69(± 0.14)	91.24(± 0.10)	90.58(± 0.28)
mBART-50 (Tang et al., 2020)	42	29.16	18.58	45.84	62.61	45.80	84.90	91.25	90.72
mBART-50	52	29.20	18.64	46.02	62.91	46.02	84.78	91.28	90.59
mBART-50	62	28.12	17.76	45.52	62.59	45.24	84.73	91.29	90.76
mBART-50	72	29.60	19.05	46.22	62.82	45.90	84.85	91.21	90.64
mBART-50	82	30.17	19.64	46.37	62.78	45.62	84.71	91.21	90.21
mBART-50 Mean (± Stdev)		29.25(± 0.75)	18.73(± 0.69)	45.99(± 0.33)	62.74(± 0.14)	45.72(± 0.30)	84.79(± 0.08)	91.25(± 0.04)	90.58(± 0.22)
mT5-small (Xue et al., 2021)	42	31.58	21.01	45.80	62.87	43.20	85.58	91.88	87.93
mT5-small	52	31.33	20.81	45.85	62.87	43.26	85.69	91.98	88.44
mT5-small	62	30.71	20.34	45.42	62.80	42.93	85.59	91.96	88.42
mT5-small	72	31.01	20.66	45.63	62.82	43.04	85.62	91.91	87.93
mT5-small	82	31.32	20.78	45.75	63.11	43.11	85.70	91.94	88.53
mT5-small Mean (± Stdev)		31.19(± 0.34)	20.72(± 0.25)	45.69(± 0.17)	62.89(± 0.12)	43.11(± 0.13)	85.64(± 0.06)	91.93(± 0.04)	88.25(± 0.30)
mT5-base (Xue et al., 2021)	42	30.86	20.31	47.75	64.51	46.42	86.29	91.98	93.71
mT5-base	52	33.26	23.46	48.90	64.93	47.39	85.99	91.82	93.08
mT5-base	62	30.42	20.70	47.16	63.95	46.10	85.15	91.68	93.06
mT5-base	72	31.29	21.56	47.50	64.05	46.30	85.36	91.81	93.10
mT5-base	82	29.32	19.76	45.67	62.45	44.92	85.10	91.54	93.87
mT5-base Mean (± Stdev)		31.03(± 1.45)	21.16(± 1.44)	47.40(± 1.17)	63.92(± 0.95)	46.23(± 0.88)	85.58(± 0.53)	91.77(± 0.17)	93.36(± 0.39)
mT5-large (Xue et al., 2021)	42	36.94	25.73	53.28	68.97	52.08	86.89	92.19	94.80
mT5-large	52	36.67	25.80	52.89	68.81	52.89	86.86	92.17	94.90
mT5-large	62	37.07	25.85	53.26	68.86	52.40	86.96	92.22	94.70
mT5-large	72	36.86	25.67	53.05	68.74	52.31	86.79	92.24	94.68
mT5-large	82	37.48	26.45	53.51	68.96	52.54	86.85	92.16	94.64
mT5-large Mean (± Stdev)		37.00(± 0.30)	25.90(± 0.31)	53.20(± 0.24)	68.87(± 0.10)	52.44(± 0.30)	86.87(± 0.06)	92.20(± 0.03)	94.74(± 0.11)

Table 6: Performance of generative language models in high-level commonsense reasoning test. **Mean** refers to an average value from sampled score using designated seeds, and **Stdev** is a standard deviation from sampled score using designated seeds.

data source. Moreover, KoBART is robust on the domain transfer, and the performances of mBARTs are decreased to a great extent in the case of experiment setting transition, revealing similar consequences with the experiment in Table 6. On the contrary, the performance of KoBART and mT5 models is enhanced indicating the expandability of the downstream task within the commonsense domain utilizing our dataset.

B.3 Human Evaluation

We perform human evaluation on the sentences generated via the model using the new test dataset irrelevant to the training dataset, as exhibited in Table 8. Compared to the results of the human evaluation using the test dataset relevant to the training dataset depicted in Table 5, each model has proximate results in the expected distribution and tendency, in which mT5-large achieves the best score. Despite the results in Table 8 evaluated with the low relevance in test dataset, models achieve an only marginal lower score than Table 5. This implies that the models not only learn with the semantic and syntactic role of the content morpheme in the concept set but also their ability to acquire unseen commonsense. Moreover, factuality shows the slightest gap between Table 8 and Table 5. This result indicates that factuality is more independent of the training dataset than the other evaluation metrics.

C Implementation Details

C.1 Training

Generative language models are trained to generate a reference sentence containing m tokens $s_{ref} = \{r_1, r_2, ..., r_m\}$ by referring a content morpheme-set containing l morphemes $x_{set} = \{x_1, x_2, ..., x_l\}$. Training is implemented by optimizing the objective conditional probability of given tokens in an

Model	BLEU 3	BLEU 4	ROUGE-2	ROUGE-L	METEOR	mBERTScore	KoBERTScore	Coverage
KoGPT2	25.08	16.14	38.59	56.67	31.06	86.40	93.84	77.10
KoBART	36.12	26.55	50.63	66.64	44.81	91.23	95.61	95.14
mBART	34.58	25.21	48.14	65.15	42.28	90.78	95.29	94.55
mBART-50	33.67	24.21	47.14	63.92	40.56	90.66	95.44	93.09
mT5-small	28.80	19.08	44.54	62.26	37.34	90.45	95.34	94.18
mT5-base	34.50	24.10	50.25	66.30	43.55	91.67	96.05	96.45
mT5-large	**41.86**	**31.72**	**55.51**	**70.06**	**49.55**	**92.27**	**96.23**	**96.80**

Table 7: Performance of generative language models in translated and reformulated CommonGen test.

Human evaluation	GC	CS	FC	FL	TT
KoGPT2	0.74	0.57	1.31	0.47	3.08
KoBART	1.21	1.13	1.85	0.99	5.18
mBART-50	1.17	1.08	1.78	0.96	4.99
mT5-large	1.34	1.26	1.89	1.15	5.65

Table 8: Human evaluation for model-generated outputs in reformulated CommonGen test set including **GC**: Grammar Correction; **CS**: Commonsense; **FC**: Factually; **FL**: Fluency; **TT**: Total

auto-regressive generation. As formulated in Equation 3, the conditional probability is configured with model parameters for maximizing the likelihood. The pretrained model parameter is initialized to θ by training with dataset D.

$$\max_{\theta} \frac{1}{|D|} \sum_{(s_{ref}, x_{set}) \in D} \log \left[\prod_{t=1}^{m} p_\theta(r_t \mid r_{<t}; x_{set}) \right] \quad (3)$$

C.2 Hyperparameter Settings

We implement Huggingface[15] framework for language modeling in a single NVIDIA Quadro RTX 8000 GPU with 48GB and 18-core Intel Xeon Gold 6230 CPU. For KoGPT2 training, parameters are trained by batch size 4 with gradient accumulation, seed 42, learning rate 5×10^{-5}, warmup steps 400, AdamW optimizer (Loshchilov and Hutter, 2019) ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1e - 8$), and block size 128. In the case of encoder-decoder models, key hyper-parameters are also initialized by default settings suitable for the model architectures. We set the hyper-parameters in training stage as batch size 16 with gradient accumulation, seed 42, initial learning rate 5×10^{-5}, warmup steps 400, AdamW optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1e-8$), max source length 64, max target length 256, and source prefix *"summarize"* (only for mT5-large).

C.3 Decoding Strategy

The decoding strategy is restricted to identical conditions in text generation. We set the beam size to 10, max sequence length to 30, min sequence length to 10, and no-repeat n-gram size to 3 for the imposition of a penalty on duplicate token generation. We re-rank the generated sentences sorted in descending order based on five candidate sentences that cover the number of the corresponding morphemes as the given content morpheme-set and select the highest rank as a concluding outcome.

D Error Analysis

This section investigates several errors found in our semi-automatic construction method.

The NER system has difficulty filtering out non-commonsense knowledge considering every person's perspectives. In this paper, we define commonsense knowledge as most people in the same society understand implicitly. However, it is challenging for commonsense knowledge to be defined with concrete scope and explanation. Therefore, some named entities can be viewed as commonsense knowledge to certain people. In our dataset construction method, the sentences holding the possibility of non-commonsense knowledge are safely deleted to maintain the rate of commonsense knowledge. Nevertheless, as the raw data sources pass through the automatic method, a proportion of sentences among the deleted sentences have the potential of reusability owing to the flexible definition of commonsense knowledge.

In the content morpheme extraction process, we solely use one POS tagger, ko-mecab, as a pilot study. Rather than discussing the performance of the various POS tagger, we mention the issues on the adoption process of the tagger. The process should include the sub-process of verification on the candidates of taggers or comparative study on the multiple datasets constructed with multiple taggers, respectively. The latter sub-process can cause

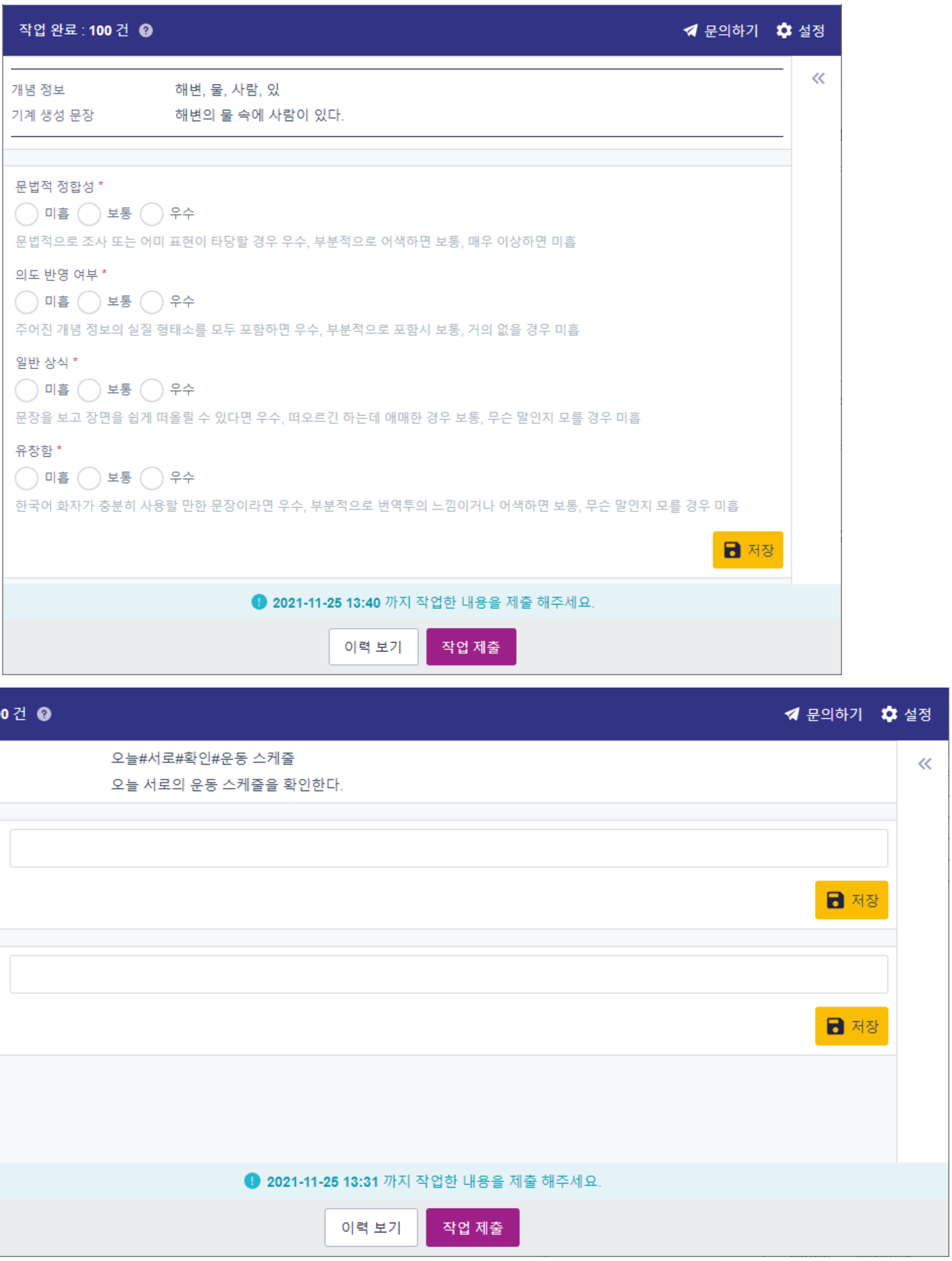
[15] https://github.com/huggingface/transformers

costly issues and require a lot of time to implement because new datasets for taggers should be reconstructed.

The usage of content morphemes can bring out uncertainty about whether the concept in the concept set should play a role as a verb or a noun. This feature embraces both pros and cons. The advantage is that the concept is granted unrestrained parts to endow the generated sentences diversity. The disadvantage is that the generation model may struggle to choose the semantic role of the concept. For example, the concept '이야기(story)', which plays a role as a noun, holds the probability of a role as a verb by combining with the verb derivative suffix. As a consequence, the generated sentence mismatches with the answer, conversely, increasing the diversity of the outputs.

To evaluate high-level commonsense reasoning, we set the seed and randomly delete one concept of the concept set. However, this deletion method is not the best way to assess commonsense reasoning. In selecting a concept to be removed, it is necessary to delete the concept deeply related to the remaining concepts so that the model can infer the meaning of deleted concept using the remaining concepts. For example, suppose the knowledge graph with more numerous data based on Korean commonsense knowledge such as ConceptNet is established. In that case, we would delete the more proper concept from the commonsense knowledge perspective.

E Crowd-sourcing Template



작업 완료 : 100 건

문의하기 설정

개념 정보 해변, 물, 사람, 있

기계 생성 문장 해변의 물 속에 사람이 있다.

문법적 정합성 *

○ 미흡 ○ 보통 ○ 우수

문법적으로 조사 또는 어미 표현이 타당할 경우 우수, 부분적으로 어색하면 보통, 매우 이상하면 미흡

의도 반영 여부 *

○ 미흡 ○ 보통 ○ 우수

주어진 개념 정보의 실질 형태소를 모두 포함하면 우수, 부분적으로 포함시 보통, 거의 없을 경우 미흡

일반 상식 *

○ 미흡 ○ 보통 ○ 우수

문장을 보고 장면을 쉽게 떠올릴 수 있다면 우수, 떠오르긴 하는데 애매한 경우 보통, 무슨 말인지 모를 경우 미흡

유창함 *

○ 미흡 ○ 보통 ○ 우수

한국어 화자가 충분히 사용할 만한 문장이라면 우수, 부분적으로 번역투의 느낌이거나 어색하면 보통, 무슨 말인지 모를 경우 미흡

저장

2021-11-25 13:40 까지 작업한 내용을 제출 해주세요.

이력 보기 작업 제출

작업 완료 : 100 건

문의하기 설정

제시 단어 오늘#서로#확인#운동 스케줄

예시 문장 오늘 서로의 운동 스케줄을 확인한다.

문장 생성1 *

저장

문장 생성2 *

저장

2021-11-25 13:31 까지 작업한 내용을 제출 해주세요.

이력 보기 작업 제출

Figure 6: The interfaces employed for human annotators experiments on CrowdWorks AI & Human Resources Platform. **Upper box** displays content morpheme concepts and a model-generated sentence to request tagging human evaluation score. 17 expert annotators assess 문법적 정합성(Grammar Correction), 의도 반영 여부(Factuality), 일반상식(Commonsense), and 유창함(Fluency) as described in §4.1. **Lower box** also exhibits content morpheme concepts and an example answer to ask generating two additional human references. 22 human annotators combine given concepts and produce two references following guidelines in §3.2.