COMMONGEN: Towards Generative Commonsense Reasoning via A Constrained Text Generation Challenge

Bill Yuchen Lin

Abstract

Given a set of common concepts like "{*apple (noun), pick (verb), tree (noun)*}", humans find it easy to write a sentence describing a grammatical and logically coherent scenario that covers these concepts, for example: "a boy *picks* an *apple* from a *tree*". The process of generating these sentences requires humans to use commonsense knowledge. We denote this ability as *generative commonsense reasoning*. Recent work in commonsense reasoning has focused mainly on *discriminating* the most plausible scenes from distractors via natural language understanding (NLU) settings such as multi-choice question answering. However, generative commonsense reasoning is a relatively unexplored research area, primarily due to the lack of a specialized benchmark dataset.

In this paper, we present a constrained natural language generation (NLG) dataset, named COMMONGEN, to explicitly challenge machines in generative commonsense reasoning. It consists of 30k concept-sets with human-written sentences as references. Crowdworkers were also asked to write the rationales (i.e. the commonsense facts) used for generating the sentences in the development and test sets. We conduct experiments on a variety of generation models with both automatic and human evaluation. Experimental results show that there is still a large gap between the current state-of-the-art pre-trained model, UniLM, and human performance. [1]

1 Introduction

Commonsense reasoning, the ability to make acceptable and logical assumptions about ordinary scenes in our daily life, has long been acknowledged as a critical bottleneck of artificial intelligence and natural language processing (Davis and Marcus, 2015). A distinct characteristic of commonsense reasoning problems is that they are trivial for humans but surprisingly challenging for machine models, especially in a generative setting. For instance, given a collection of concepts (or. concept-set) "{*apple* (noun), *bag* (noun), *pick* (verb), *place* (verb), *tree* (noun)}", what sentences could we come up that both use all the words in the concept-set and are general enough to be considered an everyday scenario?

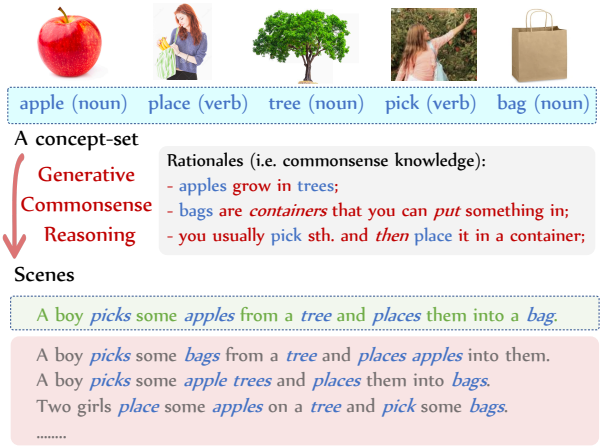

Figure 1: A motivating example for generative commonsense reasoning and the COMMONGEN task.

Humans can easily come up with sentences that fit this criteria, for example: "*a boy picks some apples from a tree and places them into a bag* ", as shown in Figure 1. Considering the size of the potential search space, all possible word sequences containing the words in the concept-set, as well as the cost of comparing each possible sequence's usage of commonsense logic, it becomes abundantly clear that this task is non-trivial for a machine. To effectively generate a sentence that incorporates all given concepts and describes a plausible everyday scenario, a reasoner needs to be able to discern which relations between the concepts in the set are logically sound, grammatically correct and will re-

[1] *Our data/code/appendix are submitted and will be public.*

sult in a common place scenario being described. We term this ability as "*generative commonsense reasoning*". Empowering machines with such reasoning ability is challenging, because it inherently requires complex compositions of various types of commonsense knowledge such as spatial relations, object properties, human behaviors, social conventions, temporal events, etc.

Most existing tasks targeting commonsense reasoning are framed as natural language understanding (NLU) tasks (Storks et al., 2019) in the form of multi-choice question answering (QA), such as the CSQA (Talmor et al., 2019) and SWAG (Zellers et al., 2018) datasets. These tasks ask reasoners to compare the possibilities of multiple given scenes (constructed by combining each answer choice with the question's context) and choose which would be most plausible in daily life. The major disadvantage of discriminative commonsense reasoners is that they are less practical in real life situations where no pre-defined choices are given.

As we seek to advance machine commonsense reasoners beyond the current success of *discriminative* reasoning and towards tasks that require more *generative* reasoning, a crucial step is conferring them with the ability to move from multi-choice QA to natural language generation (NLG). We argue that generative commonsense reasoning, which can be viewed as a way of modeling prior distributions of everyday scenes conditioned on the given concepts, can benefit many downstream applications. For example, recent research in the "language and vision" community finds that commonsense reasoning is a significant bottleneck for scene cognition (Zellers et al., 2019a). Tasks such as image and video captioning (Qiao et al., 2019; Wang et al., 2019), scene-based visual reasoning and question answering (Zellers et al., 2019a; Hudson and Manning, 2019), storytelling (Guan et al., 2018; Yang et al., 2019b), as well as dialogue systems (Zhou et al., 2018a,b)could all benefit from improved generative commonsense reasoning. Unfortunately, this important research direction is under-explored because the community lacks an appropriate problem statement and a large-scale benchmark dataset to experiment on.

To this end, we create a large-scale benchmark dataset, named COMMONGEN, as a challenging constrained text generation task shown in Figure 1. We sample $29,559$ diverse concept-sets from several large corpora of image and video captions. Through additional crowd-sourcing via *Amazon Mechanical Turk*[2] (AMT), we obtain $49,129$ human-written sentences. For each example in the development and test set, we collect on average four sentences. The crowd-workers are also asked to explicitly write *rationale* sentences about the commonsense knowledge that they have used for generating the sentences. These additional sentences can not only be leveraged as further context in the task, but also act as a quality assurance measure, as the workers are forced to think about the assumptions they are making and whether those assumptions are indeed commonsense.

To understand the difficulty of COMMONGEN, we utilize the largest commonsense knowledge graph, ConceptNet (Speer et al., 2017), to analyze the connectivity and relation type distribution among the input concept-sets. We also investigate several sophisticated sequence generation methods on the task with extensive and carefully designed automatic and manual evaluation. We find that even though UniLM (Dong et al., 2019), the state-of-the-art language generation model pre-trained with BERT, achieves the best performance for many metrics, there is still a substantial gap between UniLM and human performance.

2 Problem Formulation

In this section, we formulate the proposed COMMONGEN task with mathematical notations and discuss its inherent challenges. The input to the task is an unordered set of k concepts $x = \{c_1, c_2, \ldots, c_k\} \in \mathcal{X}$ (i.e. a concept-set), where each concept $c_i \in \mathcal{C}$ is a common noun or verb word. Let $\mathcal{X}$ denote the space of all possible concept-sets and $\mathcal{C}$ denote the concept vocabulary, which is a subset of the ConceptNet vocabulary. The expected output of the task is a simple, grammatical sentence $y \in \mathcal{Y}$ that describing a common scene in our daily life, covering all given concepts in x. A scene sentence can depict either a static situation like an image caption or a short series of actions like the caption of a video clip. Note that other forms of given words are also accepted, such as plural forms of nouns and verbs. The task is to learn a structured predictive function $f : \mathcal{X} \rightarrow \mathcal{Y}$, which maps a concept-set x to an associated sentence y. Thus, it can be seen as a special case of *constrained text generation* (Hu et al., 2017). The unique challenges of this task come from two major

[2] https://www.mturk.com

aspects as follows.

Constrained Sentence Decoding. Lexically constrained sentence decoding has been studied in the machine translation community (Hokamp and Liu, 2017). They mainly focus on the situation when specific alignment of words or phrases (e.g. terminology) must present in target sentences. Simply ordering a bag of words for recovering a complete sentence based on syntactic information (Zhang and Clark, 2015) has also been investigated. However, it is still an open problem how to generate sentences given an *unordered* set of *multiple* keywords with potential *morphological changes* (e.g. "pick" $\rightarrow$ "picks"). Apart from that, the part-of-speech constraints also brings more difficulties (e.g. "place" can be verb/noun).

Relational Reasoning with Commonsense. Expected generative reasoners should prioritize the most plausible scenes over an infinite number of less plausible scenes, such as *"a boy picks an apple tree and places it into bags"* or *"a boy places some bags on a tree and picks an apple"*. This not only requires understanding commonsense relations between each pair of concepts individually, but also finding the best composition of them with global relation reasoning. The underlying reasoning chains are inherently based on a variety of background knowledge about 1) spatial relations (e.g. "apples are usually located on the trees"), 2) object properties and 3) physical rules (e.g. "bag is a container; apples are smaller than bags; containers are used for placing objects smaller than them"), 4) temporal event knowledge of human behaviors (e.g. "you pick something first and then place them in to a container"), 5) social conventions, etc. These commonsense facts or truisms may not be recorded in any existing knowledge bases, and thus the task can be also viewed as learning to querying them from textual corpora towards generating a sentence.

3 The COMMONGEN Dataset

We now introduce the construction and analysis of the proposed COMMONGEN dataset in this section. To ensure that the concepts in each input concept-set are likely to be present together in a everyday scene, we utilize a wide range of existing caption corpora for sampling frequent concept-sets (Section 3.1). We also carefully control the overlap between the training set and dev/test set, such that the task is more challenging in terms of generalization of the reasoning ability of interest. Afterwards, we employ workers on the crowd-sourcing platform AMT for collecting more human-written sentences (Section 3.2), and thus enrich the diversity of development (dev) and test set. Finally, we show the statistics of the COMMONGEN dataset, and utilize ConceptNet as an intermediate tool to investigate the concept connectivity and the distribution of various knowledge types (Section 3.3).

3.1 Collecting Concept-Sets from Captions

It is obviously nonsense if we ask a reasoner to generate a scene about an arbitrarily combined concept-set, which is even impossible for humans. The expected input concept-sets of our task are supposed to be very likely to co-occur in common, daily-life scenes. Such everyday scenarios are ubiquitous in images and video clips, and this leads to think about using image/video captioning datasets as a natural resource for collecting concept-sets and sentences. We therefore collect a large amount of caption sentences from all publicly available visual caption corpora, including *image captioning* datasets, such as *Flickr30k* (Young et al., 2014), *MSCOCO* (Lin et al., 2014), *Conceptual Captions* (Sharma et al., 2018), and also *video captioning* datasets such as *LSMDC* (Rohrbach et al., 2017), *ActivityNet* (Krishna et al., 2017), and *VATEX* (Wang et al., 2019)

We first conduct part-of-speech tagging over all sentences in the corpora such that words in sentences can be matched to the concept vocabulary of ConceptNet. Then, we compute the sentence frequency of concept-sets that consist of 3$\sim$5 concepts. That is, for each combination of three/four/five concepts in the vocabulary, we now know how many sentences are there in the caption corpora which cover all of them.

Towards building a more representative dataset, we expect our selected subset of concept-sets can reflect the distribution in the real world. A straightforward intuition is to directly treat the frequency as the measure of likelihood of concept-sets, and then conduct probabilistic sampling based on this distribution. However, this method tends to sample concept-sets that contain one or two single highly frequent concept, thus leading to corpus-dependent bias. Also, merely using the sentence number can be imprecise to measure the scene diversity since many images and videos were sampled interdependently. We therefore design a scoring function to weight a concept-set x to incorporate diversity and

Statistics	Train	Dev	Test
# Concept-Sets	**27,069**	**993**	**1,497**
Size=3	20,580	493	-
Size=4	4,207	250	747
Size=5	2,282	250	750
# Sentencens	**39,069**	**4,018**	**6,042**
Average Length	10.85	13.15	13.80
Concept Vocab Size	6,643	813	1,351
Intersection w/ Train	100%	88.43%	85.94%

Table 1: The basic statistics of the COMMONGEN data.

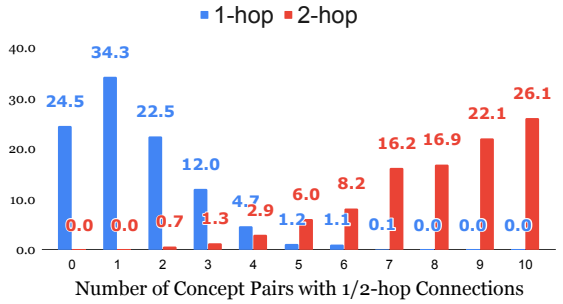

Figure 2: Connectivity on ConceptNet between concepts in 5-size-concepts in the COMMONGEN test set.

penalty of inverse set frequency:

$$\text{score}(x) = |S(x)| \frac{|\bigcup_{s_i \in S(x)} \{w | w \in s_i\}|}{\sum_{s_i \in S(x)} \texttt{Length}(s_i)} \rho(x).$$

We denote $S(x)$ as the set of different sentences that contain all its concepts $\{c_1, c_2, \ldots, c_k\} = x$, s_i as one of the sentences, and $|S(x)|$ to be the number of sentences. The second term is to divide the number of unique words in these sentences by the sum of the lengths of all the sentences, which can roughly represent the diversity of the scenes described in these sentences. Then, we times the result with the last term $\rho(x) = |\mathcal{X}|/(\max_{c_i \in x} |\{x' \mid c_i \in x' \text{ and } x' \in \mathcal{X}\}|)$. The idea is to find the concept in x that has the maximum set frequency (i.e. the number of different concept-sets (with non-zero weight) contains it), and then take the inverse with normalization of the number of all concept-sets. This penalty effectively controls the bias towards highly frequent concepts. With the distribution of such scores, we sample 100k concept-sets as candidate inputs.

3.2 Crowd-Sourcing References via AMT

Although the human-written sentences in the caption corpora can be seen as quality annotations for the COMMONGEN task as well, they were written with specific visual context (i.e. an image or a video clip). Toward better diversity of the scenes about sampled concept-sets and more rigorous evaluation for systems, crowd-sourcing additional human references is necessary that are written with only concept-sets as the context. We decide to use the AMT platform[3] for collecting such sentences for covered the top-ranked 2,500 concept-sets in the sampled results, due to the expensive cost of human efforts in writing sentences and the difficulty in verifying the quality of collected sentences. Each of them is assigned to at least three different workers. To force workers to write about everyday scenarios about given concept-sets, we ask them to write rationale sentences as well to explain what commonsense facts they have used.

We use these 2,500 concept-sets as the dev and test set examples for their higher weights and better diversity of human-written sentences. Furthermore, we use the remaining concept-sets as the training examples, for which we use the associated captions as the target outputs. Note that we explicitly control the overlap between the training and dev/test examples by filtering training concept-sets that have more than two overlapping concepts with any example in the dev/test set. The basic statistics of the final dataset is shown in Table 1. There are on average four sentences for each example in dev and test sets, which provide a richer and more diverse test-bed for further automatic and manual evaluation. Note that there are about 12/15% of concepts in dev/test set that are unseen in the training set, which can thus assess the generalization ability.

3.3 Analysis about Commonsense Knowledge

We here introduce deeper analysis of the dataset by utilizing the largest commonsense knowledge graph (KG), ConceptNet (Speer et al., 2017), as an tool to study connectivity and relation types.

Connectivity Distribution. Obviously, if the concepts inside a given concept-set is more densely connected with each other on the KG, then it is easier to write a scene about them. In each 5-concept-set (i.e. a concept-set consists of five concepts), there are 10 unique pairs of concepts, the connections of which we are interested in. As shown in Figure 2, if we look at the one-hop links on the KG, about 60% of the 5-concept-sets have less than one link among all concept-pairs. On the other hand, if we consider two-hop links, then nearly 50% of them are almost fully connected (i.e. each pair of

[3]The instruction and user interface are shown in *Appendix*.

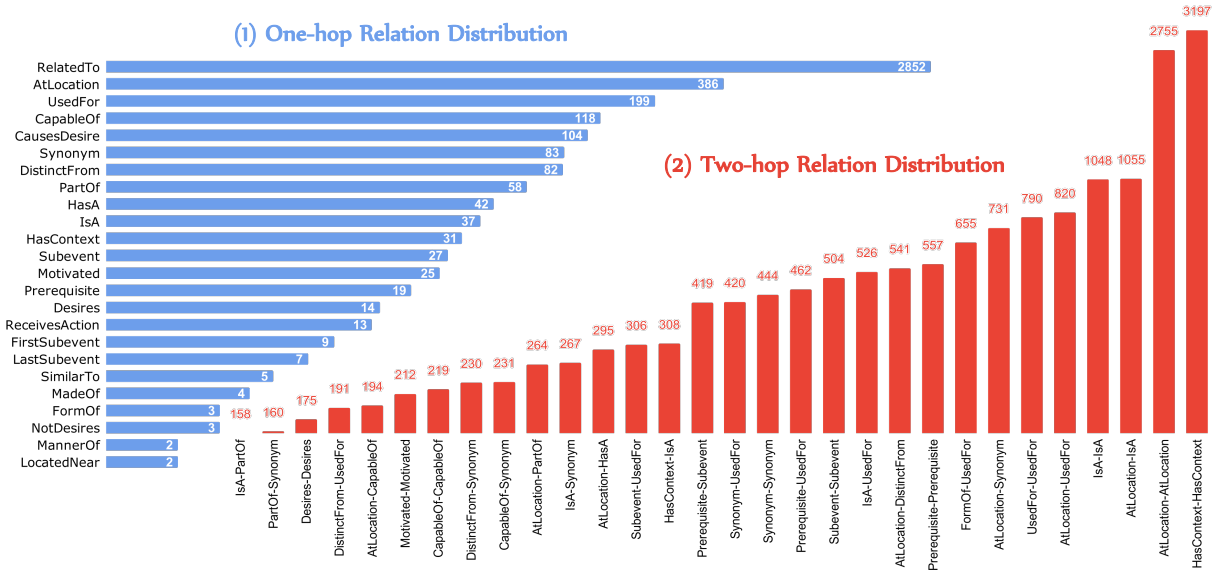

Figure 3: One/two-hop relation frequency in the COMMONGEN dev.&test sets on ConceptNet.

concepts has connections). These two observations together suggest that the COMMONGEN has a reasonable difficulty: the concepts are not too distant or too close, and reasoning about the associated scenes is thus neither too difficult nor too trivial.

Relation Distribution. Furthermore, the relation types of such connections can also tell us what kinds of commonsense knowledge are potentially useful for relational reasoning towards generation. We report the frequency of different relation types[4] of the one/two-hop connections among concept-pairs in the dev and test examples in Figure 3. In both cases, we find most frequent relation types are about 1) ***spatial*** knowledge(e.g. ATLOCATION, LOCATEDNEAR), 2) ***object properties*** (e.g. USEDFOR, CAPABLEOF, PARTOF, RECEIVEACTION), 3) ***human behavior and social conventions*** (e.g. CAUSEDESIRE, MOTIVATED), 4) ***temporal*** knowledge (e.g. [First/Last-]SUBEVENT, PERQUISITE), and 5) other ***general commonsense*** (e.g. RELATEDTO, HASCONTEXT, ISA).

4 Methods

In this section, we briefly introduce the adopted baseline methods that are tested on the proposed COMMONGEN task. As there is no principled approach for the proposed setting, to the best of our knowledge, we mainly consider it as a conditional sentence generation task that can be solved by many sequence-to-sequence (seq2seq) models.

Encoder-Decoder Models. Bidirectional RNN ("bRNN") and Transformer (Vaswani et al., 2017) ("Trans.") are two most popular architectures for seq2seq learning. We use them with the addition of attention mechanism (Luong et al., 2015) with copying ability (Gu et al., 2016), which are based on an open-source framework OpenNMT-py[5] (Klein et al., 2017). To alleviate the influence from the concept ordering in such sequential learning methods, we randomly permute them multiple times for training and decoding and then take take average performance. To explicitly eliminate the order-sensitivity of inputs, we replace the encoder with a mean-pooling based MLP network for the bRNN ("Mean Seq"). Similarly, we consider removing the position embedding module from the Transformer as well ("Set. Trans.").

Non-autoregressive generation. Recent advances (Lee et al., 2018; Stern et al., 2019) in conditional sentence generation have an embeding interest on (edit-based) non-autoregressive generation models, which iteratively refine generated sequences (usually bounded by a fixed length). We assume that these models potentially would have better performance because of explicit modeling on iterative refinements, and thus study the most recent such model *Levenshtein Transformer* (Gu et al., 2019) (LevenTrans.).

The SOTA Pre-trained NLG model. We employ a novel unified pre-trained language generation model, named UniLM (Dong et al., 2019), which

[4]More explanations are at https://github.com/commonsense/conceptnet5/wiki/Relations.

[5]https://github.com/OpenNMT/OpenNMT-py

uses pre-trained BERT as the encoder and then fine-tunes its whole architecture with many different generation objective. It achieved the *state-of-the-art* performance on many generation tasks including summarization, question generation, etc.

Imposing Commonsense Knowledge. Recent work (Lv et al., 2019) finds that the OMCS corpus (Singh et al., 2002), which has derived the ConceptNet, is a valuable resource for retrieving relevant commonsense facts for discriminative reasoning about questions. We follow the same steps to retrieve related facts by querying input concepts. Then, we concatenate them with the original concept-sets as the final input sequence to the above-mentioned methods ("w/ omcs"), mimicking abstractive summarization tasks.

5 Evaluation

In this section, we first introduce our metrics for automatic evaluation, then analyze the performance of tested systems, and finally provide carefully designed human evaluation analysis and case study.

5.1 Automatic Metrics

Following other conditional language generation tasks, we use several widely used automatic metrics to automatically assess the performance, such as *BLEU*-3/4 (Papineni et al., 2001), *ROUGE*-2/L (Lin, 2004), *METER* (Banerjee and Lavie, 2005), which mainly focus on measuring surface similarities. In addition, we can also regard COMMONGEN task as also a captioning task, where context are concept-sets instead of real visual signals. Therefore, it is more suitable to use specialized caption evaluation metrics such as *CIDEr* (Vedantam et al., 2014) and *SPICE* (Anderson et al., 2016) . Towards more task-specific evaluation, we first report the concept coverage in system predictions. "PosCov." stands for the average percentage of input concepts that are present in the system predictions with correct *part-of-speech* requirement, while "LemCov." only requires matches after lemmatization.

We also propose a novel metric specially designed for the COMMONGEN, named **PivotBERT**, on the top of *BERTScore* (Zhang et al., 2019) which is a recently proposed embedding-based metric. Concretely, we first utilize a dependency parser to parse both system-generated sentences and human references, then compute the recall score in terms of "shortest paths" between given concepts, and finally times *BERTScore*. For example, given a reference sentence "a man picks some apples from a tree and puts them into a bag ", we can find all valid shortest paths between any pair of given concepts (underlined) as pivots in its dependency parse tree, such as the path between "tree" and "bag" will be "*tree* $\xrightarrow{\text{nmod}}$ *pick* $\xrightarrow{\text{conj}}$ *put* $\xleftarrow{\text{nmod}}$ *bag*" (✓). A failed prediction like "a boy picks some bags from a tree and puts apples into them" thus cannot recall the previous path because it produces "*tree* $\xrightarrow{\text{nmod}}$ *pick* $\xrightarrow{\text{dobj}}$ *bag*" (✗). Then, with such computed path-level recall score (i.e. *PivotScore*), we can focus on the the relations between given concepts in system predictions and human references. We further times *PivotScore* with *BERTScore* for its better assessment the semantic similarity.

To estimate *human performance* within each metric, we iteratively treat each reference sentence in dev/test data as a "system" prediction to be compared with all other references. Thus, systems that have equivalent reasoning ability as crowd workers on average should exceed this "**Human Bound**".

5.2 Experimental Results

Table 2 presents the experimental results[6] of all methods in terms of different metrics. The order-insensitive method "Mean Seq." outperforms its order-sensitive counterparts like "bRNN", but such a marginal improvement is not seen in the comparison between "Trans." and "Set. Trans.". We assume that for short sequences the order sensitivity does not harm Transformer encoders too much, but positional embeddings are quite necessary to ensure its self-attention mechanism.

We find that vanilla Transformer architectures are not outperforming simpler models like bRNN, which probably is because of its complex structure needs more carefully tuning or a specially copying attention in this setting. The better performance in edit-based Transformer model, LevenTrans, also suggests that, yielding the best performance among models without pre-training. We argue that the non-autoregressive models with iterative refinement on previously decoded sentences are more promising in our setting, as our inputs can be naturally viewed as incomplete sentence prototypes.

The best model is UniLM, which makes sense for its powerful pre-trained encoder BERT. We be-

[6]Implementation details like hyper-parameters about training/decoding are detailed in the *reproducibility instructions* within the **submitted code**. The dev results are in **Appendix**.

Model	ROUGE-2	/-L	BLEU-3	/-4	METER	CIDEr	SPICE	PosCov.	LemCov.	PivotBERT
Mean Seq.	3.30	19.35	6.60	2.40	13.50	4.34	13.00	31.26	44.05	1.47
bRNN	2.90	19.25	5.50	2.00	12.70	3.99	10.60	30.30	42.25	1.22
bRNN w/ OMCS	4.15	21.74	7.70	2.80	15.10	5.22	14.30	37.20	50.67	1.19
Set. Trans.	1.59	12.96	3.20	1.20	8.60	2.02	7.00	17.71	24.10	0.92
Trans.	2.28	14.04	4.30	2.00	9.10	2.31	7.50	19.00	24.19	0.86
Trans. w/ OMCS	1.73	12.81	4.20	1.40	9.20	2.48	8.10	14.93	20.87	0.32
LevenTrans.	5.74	21.24	8.80	4.00	13.30	3.72	14.00	32.93	36.80	2.91
UniLM	21.57	41.96	38.30	27.50	29.40	14.92	29.90	86.71	90.13	29.05
UniLM w/ OMCS	20.77	41.25	36.40	25.70	29.30	14.08	29.20	86.32	89.42	27.87
HumanBound	48.88	63.79	48.20	44.90	36.20	43.53	63.50	95.97	99.31	54.55

Table 2: Experimental results of different baseline methods on the COMMONGEN test set.

	Mean	Set.	Trans.	bRNN	Leven	UniLM	Human	Overall
Mean.	N/A	65.6	43.8	69.1	41.1	5.1	2.2	11.82
Set.	34.4	N/A	6.2	79.2	45.8	0.0	0.0	4.89
Trans.	56.2	93.8	N/A	75.0	75.0	5.4	0.0	9.67
bRNN	30.9	20.8	25.0	N/A	25.3	2.9	0.5	5.32
Leven	58.9	54.2	25.0	74.7	N/A	5.4	0.0	11.98
UniLM	94.9	100.0	94.6	97.1	94.6	N/A	17.5	72.07
Human	**97.8**	**100.0**	**100.0**	**99.5**	**100.0**	**82.5**	N/A	92.74

Figure 4: Pair-based human evaluation results.

lieve the masked language modeling (MLM) task of BERT, which aims to predict missing words, is similar to our task. Also, the further fine-tuning tasks of language generation enable UniLM to better rearrange words for completing sentences. This result also suggests that further modifying over pre-trained models is a promising direction for generative commonsense reasoning.

The use of the OMCS corpus helps the bRNN method, but decreases the performance of Transformer-based methods. We think this is mainly because the order of OMCS sentences. Transformer-based models explicitly use position embedding to deal with the forgetting issue in encoding long sequences, but this becomes a harmful way when the ordering of words or sentences does not have meaningful patterns. Instead, RNNs do not have explicit assumptions on the word/sentence ordering when encoding long sequences. We argue that imposing commonsense knowledge with additional graph structures (Lin et al., 2019) is a more promising future direction for the COMMONGEN task as graphs are naturally order-insensitive.

5.3 Human Evaluation

To better compare the tested models, we conduct specialized human evaluation. For each instance in the test set, we build *anonymized* pairs of sentences that come from two different models, and then we ask human to judge which scene is more plausible about given concepts. Note that we only compare sentence-pairs that *have the same covered words*, such that human judges[7] can focus on evaluating the plausibility. Judges can also choose "equal" when they think a pair of sentences are equally plausible. We also regard human references as a kind of model predictions here.

From the matrix in Figure 4, we can clearly compare every pair of models. Each entry means the average percentage that a row-model (green names) produce better sentences than a column-model (red names). For instance, the left-bottom cell "97.80" means that 97.8% of the sentence pairs, judges prefer the human references over the result from the "Mean Seq." model. These numbers only reflect the plausibility under the situations where both models have the same covered concepts, so we further weight them by their "PosCov." scores for the final weighted overall performance.

5.4 A Case study

Table 3 shows the predictions of different models and human reference about an input concept-set[8]. We can find that bRNN and MeanSeq can hardly cover all of the given concepts, and the sentences are not quite grammatical. The LevenTrans model performs better in covering more concepts with correct preposition words but produces repetition of words. The two UniLM model both miss a concept but are significantly better than other baseline methods. However, "diving into an object" does not make sense, and it is much less plausible that "someone retrieves an object and *then* throws it away in a pool" than the scenarios of human references. This suggests that their main drawback is the lack of more thorough and comprehensive

[7] We recruit five college students who are English speakers. The Pearson-correlation among their final scores is 0.915, which indicates a high inter-annotator agreement.

[8] More studies is in the submitted *Appendix*.

Concept-set	{**dive** (verb), **object** (noun), **pool** (noun), **retrieve** (verb), **throw** (verb)}
bRNN	A man is holding a **retrieve** and then flings the **object** .
MeanSeq.	A man stands on stage *picks up* an **object** from his hands while he watch .
bRNN*	A **pool** of people on floor with their mustache clearing an **object** .
Trans.	A person **retrieves** a table **object** to **threw** breast.
LevenTrans	A man is **diving** down at an **object** and an **object** .
UniLM*	A man **retrieves** an **object** and **throws** it into a **pool** .
UniLM	A man **dives** into an **object** in a **pool** and **retrieves** it .
References	The woman threw an **object** in the **pool** and the dog **dove** to **retrieve** it .
	A man **dived** into the **pool** to **retrieve** the **object** he **threw** in it .

Table 3: A qualitative case study. ($*$ = "w/ omcs").

relational reasoning with commonsense knowledge (e.g. temporal events/human behavior here).

6 Related Work

Machine Common Sense. There are many emerging datasets for testing machine commonsense from different angles, such as commonsense extraction (Xu et al., 2018; Li et al., 2016), next situation prediction (SWAG (Zellers et al., 2018), CODAH (Chen et al., 2019), HellaSWAG (Zellers et al., 2019b)), cultural and social understanding (Lin et al., 2018; Sap et al., 2019a,b), visual scene comprehension (Zellers et al., 2019a), and general commonsense question answering (Talmor et al., 2019; Huang et al., 2019).

Recent studies have shown that simply fine-tuning large pre-trained language models, e.g. RoBERTa (Liu et al., 2019), can yield near-human, or even exceeding-human, performance in these *discriminative reasoning* scenarios such as the SWAG dataset. We argure that the underlying reasons are two-fold: 1) The creation of distractor choices has *annotator bias* (Geva et al., 2019) which can be easily detected by NLU models. 2) Self-supervised training objectives in BERT-like models (Devlin et al., 2019) align well with the multi-choice QA setting; the SWAG task shares almost the same scenario with the Next Sentence Prediction (NSP) task, and because the CSQA task can be viewed as learning to recover missing words that are masked by "wh-words", it can be distantly learned using Masked Language Modeling (MLM). Therefore, these success does not necessarily mean machine reasoners can produce novel assumptions in an open, realistic, generative setting.

There are also a few works that incorporate commonsense knowledge in language generation tasks like story and essay generation (Guan et al., 2018; Yang et al., 2019a) as well as video captioning (Yang et al., 2019b), which suggest that generative commonsense reasoning has a great potential to benefit downstream applications. Towards this direction, our proposed CommonGen, to the best of our knowledge, is the very first constrained sentence generation dataset for assessing and conferring generative machine commonsense.

Constrained Text Generation. Constrained text generation aims to decode sentences with expected attributes such as sentiment (Luo et al., 2019a; Hu et al., 2017), tense (Hu et al., 2017), template (Zhu et al., 2019), style (Fu et al., 2018; Luo et al., 2019b; Li et al., 2018), topics (Feng et al., 2018), etc. A most similar scenario with our task is lexically constrained encoding, which has been mainly studied in the machine translation community (Hasler et al., 2018; Dinu et al., 2019; Hokamp and Liu, 2017). One recent work in this line is the CGMH (Miao et al., 2018) method, which can sample sentences with an *ordered* sequence of keywords from language models but cannot be trained and adopted in our case. We have also investigated some topical story generation models (Fan et al., 2018), however, their performance are significantly worse even than simple bRNN methods. Additionally, the CommonGen task brings some more challenges motioned in Section 2. Prior constrained generation methods cannot address these issues together in a unified model, and thus we expect CommonGen to be also a benchmark dataset for future works in this direction.

7 Conclusion

We present CommonGen, a new large-scale dataset targeting generative commonsense reasoning via a constrained sentence generation task. The dataset contains about 30k concept-sets as input with a diverse concept vocabulary. The task is inherently challenging as the search space of plausible answers is large and many objectives must be satisfied at once. Additionally the task requires complex relational commonsense reasoning to generate an expected sentence, a bottleneck in many current state of the art natural language systems. Most widely used baseline methods cannot produce desirable sentences, even when given a set of relevant commonsense facts as additional context. Even the best baseline model UniLM, a BERT-based NLG model, still performs significantly worse than human performance. We believe the proposed task and benchmark dataset can benefit future research in generative commonsense reasoning and downstream NLG applications that require commonsense knowledge and complex reasoning.

References

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In ECCV.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In IEEvaluation@ACL.

Michael Chen, Mike D'Arcy, Alisa Liu, Jared Fernandez, and Doug Downey. 2019. Codah: An adversarially authored question-answer dataset for common sense. ArXiv, abs/1904.04365.

Ernest Davis and Gary Marcus. 2015. Commonsense reasoning and commonsense knowledge in artificial intelligence. Commun. ACM, 58(9):92–103.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In NAACL-HLT.

Georgiana Dinu, Prashant Mathur, Marcello Federico, and Yaser Al-Onaizan. 2019. Training neural machine translation to apply terminology constraints. In ACL.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. ArXiv, abs/1905.03197.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In ACL.

Xiaocheng Feng, Ming Liu, Jiahao Liu, Bing Qin, Yibo Sun, and Ting Liu. 2018. Topic-to-essay generation with neural networks. In IJCAI.

Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2018. Style transfer in text: Exploration and evaluation. In AAAI.

Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. In EMNLP-IJCNLP.

Jiatao Gu, Zhengdong Lu, Hang Li, and Victor O.K. Li. 2016. Incorporating copying mechanism in sequence-to-sequence learning. In ACL.

Jiatao Gu, Changhan Wang, and Jake Zhao. 2019. Levenshtein transformer. ArXiv, abs/1905.11006.

Jian Guan, Yansen Wang, and Minlie Huang. 2018. Story ending generation with incremental encoding and commonsense knowledge. In AAAI.

Eva Hasler, Adrià de Gispert, Gonzalo Iglesias, and Bill Byrne. 2018. Neural machine translation decoding with terminology constraints. In NAACL-HLT.

Chris Hokamp and Qun Liu. 2017. Lexically constrained decoding for sequence generation using grid beam search. In ACL.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P. Xing. 2017. Toward controlled generation of text. In ICML.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos QA: Machine reading comprehension with contextual commonsense reasoning. In EMNLP-IJCNLP.

Drew A. Hudson and Christopher D. Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In CVPR.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. 2017. Opennmt: Open-source toolkit for neural machine translation. In ACL.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-captioning events in videos. In ICCV.

Jason D. Lee, Elman Mansimov, and Kyunghyun Cho. 2018. Deterministic non-autoregressive neural sequence modeling by iterative refinement. In EMNLP.

Juncen Li, Robin Jia, He He, and Percy Liang. 2018. Delete, retrieve, generate: a simple approach to sentiment and style transfer. In NAACL-HLT.

Xiang Li, Aynaz Taheri, Lifu Tu, and Kevin Gimpel. 2016. Commonsense knowledge base completion. In ACL.

Bill Yuchen Lin, Xinyue Chen, Jamin Chen, and Xiang Ren. 2019. Kagnet: Knowledge-aware graph networks for commonsense reasoning. In EMNLP-IJCNLP.

Bill Yuchen Lin, Frank F. Xu, Kenny Q. Zhu, and Seung won Hwang. 2018. Mining cross-cultural differences and similarities in social media. In ACL.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In ACL.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In ECCV.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar S. Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke S. Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. ArXiv, abs/1907.11692.

Fuli Luo, Peng Li, Pengcheng Yang, Jie Zhou, Yutong Tan, Baobao Chang, Zhifang Sui, and Xu Sun. 2019a. Towards fine-grained text sentiment transfer. In ACL.

Fuli Luo, Peng Li, Jie Zhou, Pengcheng Yang, Baobao Chang, Zhifang Sui, and Xu Sun. 2019b. A dual reinforcement learning framework for unsupervised text style transfer. In IJCAI.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In EMNLP.

Shangwen Lv, Daya Guo, Jingjing Xu, Duyu Tang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, and Songlin Hu. 2019. Graph-based reasoning over heterogeneous external knowledge for commonsense question answering. arXiv preprint arXiv:1909.05311.

Ning Miao, Hao Zhou, Lili Mou, Rui Yan, and Lei Li. 2018. Cgmh: Constrained sentence generation by metropolis-hastings sampling. In AAAI.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. Bleu: a method for automatic evaluation of machine translation. In ACL.

Tingting Qiao, Jing Zhang, Duanqing Xu, and Dacheng Tao. 2019. Mirrorgan: Learning text-to-image generation by redescription. ArXiv, abs/1903.05854.

Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Chris Pal, Hugo Larochelle, Aaron Courville, and Bernt Schiele. 2017. Movie description. International Journal of Computer Vision.

Maarten Sap, Ronan LeBras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019a. Atomic: An atlas of machine commonsense for if-then reasoning. In AAAI.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019b. Social IQa: Commonsense reasoning about social interactions. In EMNLP-IJCNLP.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In ACL.

Push Singh, Thomas Lin, Erik T. Mueller, Grace Lim, Travell Perkins, and Wan Li Zhu. 2002. Open mind common sense: Knowledge acquisition from the general public. In CoopIS/DOA/ODBASE.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In AAAI.

Mitchell Stern, William Chan, Jamie Ryan Kiros, and Jakob Uszkoreit. 2019. Insertion transformer: Flexible sequence generation via insertion operations. In ICML.

Shane Storks, Qiaozi Gao, and Joyce Yue Chai. 2019. Commonsense reasoning for natural language understanding: A survey of benchmarks, resources, and approaches. ArXiv, abs/1904.01172.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In NAACL-HLT.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In NIPS.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2014. Cider: Consensus-based image description evaluation. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 4566–4575.

Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. 2019. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In The IEEE ICCV.

Frank F. Xu, Bill Yuchen Lin, and Kenny Q. Zhu. 2018. Automatic extraction of commonsense locatednear knowledge. In ACL.

Pengcheng Yang, Lei Li, Fuli Luo, Tianyu Liu, and Xu Sun. 2019a. Enhancing topic-to-essay generation with external commonsense knowledge. In ACL.

Pengcheng Yang, Fuli Luo, Peng Chen, Lei Li, Zhiyi Yin, Xiaodong He, and Xu Sun. 2019b. Knowledgeable storyteller: A commonsense-driven generative model for visual storytelling. In IJCAI.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. TACL, 2.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019a. From recognition to cognition: Visual commonsense reasoning. In CVPR.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. Swag: A large-scale adversarial dataset for grounded commonsense inference. In EMNLP.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019b. Hellaswag: Can a machine really finish your sentence? In ACL.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. ArXiv, abs/1904.09675.

Yue Zhang and Stephen Clark. 2015. Discriminative syntax-based word ordering for text generation. Computational Linguistics, 41:503–538.

Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018a. Commonsense knowledge aware conversation generation with graph attention. In IJCAI.

Jie Zhou, Ganqu Cui, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, and Maosong Sun. 2018b. Graph neural networks: A review of methods and applications. arXiv preprint arXiv:1812.08434.

Wanrong Zhu, Zhiting Hu, and Eric P. Xing. 2019. Text infilling. ArXiv, abs/1901.00158.