Lexically-constrained Text Generation through Commonsense Knowledge Extraction and Injection

Yikang Li[1*], Pulkit Goel[1*], Varsha Kuppur Rajendra[1], Har Simrat Singh[1], Jonathan Francis[1,2*], Kaixin Ma[1*], Eric Nyberg[1], Alessandro Oltramari[2]

[1]Language Technologies Institute, School of Computer Science, Carnegie Mellon University
[2]Human-Machine Collaboration, Bosch Research Pittsburgh
{yikangli, pulkitgo, vkuppurr, harsimrs, jmf1, kaixinm, ehn}@cs.cmu.edu, alessando.oltramari@us.bosch.com

Abstract

Conditional text generation has been a challenging task that is yet to see human-level performance from state-of-the-art models. In this work, we specifically focus on the `CommonGen` benchmark, wherein the aim is to generate a plausible sentence for a given set of input concepts. Despite advances in other tasks, large pre-trained language models that are fine-tuned on this dataset often produce sentences that are syntactically correct but qualitatively deviate from a human understanding of common sense. Furthermore, generated sequences are unable to fulfill such lexical requirements as matching part-of-speech and full concept coverage. In this paper, we explore how commonsense knowledge graphs can enhance model performance, with respect to commonsense reasoning and lexically-constrained decoding. We propose strategies for enhancing the semantic correctness of the generated text, which we accomplish through: extracting commonsense relations from `ConceptNet`, injecting these relations into the Unified Language Model (UniLM) through attention mechanisms, and enforcing the aforementioned lexical requirements through output constraints. By performing several ablations, we find that commonsense injection enables the generation of sentences that are more aligned with human understanding, while remaining compliant with lexical requirements.

Introduction

Natural language generation is the backbone for a plethora of applications, where language models—such as GPT-2 (Radford et al. 2019), UniLM (Dong et al. 2019a), and BART (Lewis et al. 2019)—are leveraged for challenging tasks, such as dialogue generation, story-telling, text summarization, and descriptive question answering. While these language models have brought about significant performance improvements, largely due to their scale, these models are yet to reach human-level performance. Model performance worsens further for tasks, such as constrained text generation, where the generated text is expected to follow a set of pre-defined rules or "requirements." Examples of lexically-constrained text generation include natural language generation, conditioned on the content in structured tables, or caption generation, based on a list of words.

The `CommonGen` dataset (Lin et al. 2020) is an instance of the word-conditioned caption generation task. Here, the aim is to generate a syntactically- and semantically-coherent sentence from a given concept-set; these concepts are usually nouns and verbs and represent entities from day-to-day life. The authors observed that, although popular language models generate sentences with reasonable grammatical structure, they struggle with two major aspects of the task. Firstly, generated sentences did not completely adhere to a human's understanding of common sense. As an example, for the a given set of concepts, say {*dog, catch, throw, frisbee*}, GPT-2 generated the following sentence: *A dog throws a frisbee at a football player*, while UniLM generated "*Two dogs are throwing frisbee at each other*". Even though the generated sentences have syntactic integrity, the language models were far from grasping the essence of common sense. Secondly, these pre-trained language models also struggled to include all the given concepts in the concept-set, instead producing sentences with only partial coverage. In the examples, above, the concept *catch* was missing in text generations from both GPT-2 and UniLM. Another example is of text generation from the T5 language model (Raffel et al. 2019), where the generated sentence, *dog catches a frisbee and throws it to a dog*, not only lacks common sense but also sees a repetition of the concept *dog*. These phenomena of missing concepts, repetition, and lacking concept coverage limit the quality of generated text and, by extension, negatively affect models' task performance.

We hypothesize that addressing these common faults, directly, will lead to improvements in generated sentence quality and to increases in model performance. In this paper, we address these limitations in two parts. First, we ground the given concept-set on multi-hop knowledge primitives found in existing commonsense knowledge graphs: we bias models by extracting concept-specific relations from knowledge graphs and by injecting this additional context into early layers of popular language model classes. In this way, the model is able to expand the original concept set, to include additional concepts that, ultimately, make the generated text more realistic; through ablation studies, we show that this coupled extraction/injection process increases the semantic

*Correspondence

correctness of the generated text output. Next, we further enhance the models' adherence to task requirements by imposing lexical constraints on the models' output, through adjusted beam search decoding. In this manner, we significantly reduce issues related to repetitive concept phenomena. Our contributions include: (1) strategies to extract commonsense information from knowledge graphs; (2) attention mechanism to inject the commonsense knowledge into the language model; and (3) enforcing lexical constraints by modifying beam search decoding. Our strategy can thus be formulated as a three step process:

- **Knowledge Extraction**: For a given set of concepts, knowledge graphs can help identify relationships between a pair of concepts, as seen in a real-world scenario. In this paper, we focus on strategies to extract relations between concepts from the `ConceptNet` knowledge graph.
- **Knowledge Injection**: We explore different methodologies that aim to inject the extracted relations between concepts into the language model. We limit our experiments to UniLM, which is a transformer-based seq2seq language model, and had state of the art performance in many metrics on the `CommonGen` dataset. We discuss both attention-based and non attention-based approaches.
- **Constrained Decoding**: We modify the beam search decoding to assess the best-N beams for their adherence to the given lexical-constraints. Sentence with the highest beam score that includes all the given concepts as well as matches the given part-of-speech (POS) tags of the concepts is selected as the output.

Related Work

External Knowledge Resources

Commonsense knowledge graphs seek to codify a human-like understanding of the relations between concepts and events that occur in the real world. Chief among these is `ConceptNet` (Speer, Chin, and Havasi 2016), which encodes commonsense knowledge as triples, of the form: $[C_1, r, C_2]$, where C_1 and C_2 represent commonly-used head and tail concepts and r denotes the type of relation between these concepts, such as RelatedTo, Synonym, etc. The ATOMIC knowledge graph (Sap et al. 2019) enables reasoning about *what*, *how*, and *why* a cause can lead to an effect: this resource models the interactions between concepts/entities as *if-then* relationships, as opposed to the taxonomic relations modeled by `ConceptNet`. The `DICE` knowledge graph (Chalier, Razniewski, and Weikum 2020) gives a multi-faceted nature to concept relations by providing scores for characterizing them as plausible, remarkable, salient, and typical. In other words, `DICE` comments on the circumstances under which two concepts are related and how. Other knowledge graphs incorporate task-specific commonsense knowledge, such as `SenticNet` (Cambria et al. 2020), which is custom built for concept-level sentiment analysis. Previous work has also consolidated and utilized a diverse set of commonsense knowledge graphs (including `WordNet` (Miller 1995), `ConceptNet`, `ATOMIC`, `Wikidata` (Ilievski, Szekely, and Schwabe 2020), and `VisualGenome` (Krishna et al. 2016)) into a unified framework (Ilievski et al. 2020; Ma et al. 2021), used for multiple-choice commonsense question answering. Regardless of the task *format*, we follow Ma et al. (2019, 2021) in asserting that the choice of external resource plays a significant role in the downstream performance—based on the alignment between the task semantics and the type of common sense encoded by the resource; consequently, we utilize `ConceptNet` in this work.

Knowledge Manipulation

Given knowledge primitives from an external resource, various approaches have been proposed for transforming the symbolic knowledge into a neural representation that can be easily consumed by language models. We call this process of infusing models with knowledge as "knowledge injection." Among the earliest works involving extraction and injection of knowledge from a KG are, (Ahn et al. 2017) and (Yang et al. 2017) where the authors propose architectures which combine symbolic knowledge provided by a KG with an RNN language model. Lin et al. (2020) propose concatenating human-generated hints ("rationale" tokens) to the input concepts, for conditional text generation. Bauer, Wang, and Bansal (2019); Mihaylov and Frank (2018) inject knowledge into the intermediate layers of neural models, but they focus on reading comprehension and multiple-choice question answering tasks, where the role of common sense is less defined. Ma et al. (2019); Oltramari et al. (2020); Ma et al. (2021); Wang et al. (2020) propose using attention mechanisms for commonsense knowledge injection, for multiple-choice commonsense question answering, by applying attention with respect to the question followed by an Option Comparison Network (OCN) cell. Inspired by Ma et al. (2019); Lin et al. (2020), we adapt and unify these methodologies for lexically-constrained, conditional text generation on commongen. Similar to the recent works (Lauscher et al. 2020; Liu et al. 2020), we inject `ConceptNet` relations in sentence form into the transformer layers. While these works use an adapter-based residual bottleneck and evidence generators for NLU tasks like classification or MCQA, respectively, we introduce a multi-linear attention distribution to the hidden representation of the encoder, to solve the problem of constrained text generation. Works like (Liu et al. 2019) and (Chen et al. 2020) also adopt a similar strategy for knowledge injection into a language model albeit the problem that they intend to solve is different.

The aforementioned works focus on the downstream prediction tasks, with less emphasis on analysing the ideal type and size of knowledge to be injected. Our work brings together the techniques of better knowledge extraction, commonsense knowledge manipulation and constrained text generation.

Constrained Text Generation

As a methodology, lexically-constrained text generation enjoys application to many real-world domains, such as dialogue systems, machine translation, and paragraph/story generation. Table-to-text is one such task, where, given a structured dataset such as a table, the aim is to generate

human-standard sentence(s) with the constraint that all the words specified in the table should be included in the generated text. In their work on Neural Template generation, Wiseman, Shieber, and Rush (2018) explore conditional text generation on the WikiBio and E2E datasets by learning latent, discrete templates using a neural Hidden Semi-Markov Model (HSMM) decoder; Lebret, Grangier, and Auli (2016) propose copy actions for the same task, in order to include all the given concepts in the generated text. In this paper, we focus on sentence generation from a given list of concepts, prescribed by the `CommonGen` dataset. In their work on this dataset, Lin et al. (2020) experiment with multiple baseline models and pre-trained language models, including GPT-2, BERT, UniLM, BART, and T5.

Approach

Knowledge Extraction Methodology

We refer to "knowledge extraction" as the process of obtaining relevant knowledge primitives (e.g., triples or multi-hop paths) from an external resource (`ConceptNet`), in order to satisfy a downstream task (question answering, text generation). In this section, we discuss multi-hop knowledge extraction with `ConceptNet`, knowledge selection strategies, and query expansion—all geared towards obtaining the best knowledge for downstream use with the model.

Multi-hop extraction We derive our multi-hop commonsense knowledge extraction procedure, as an extension of the single-hop case used in Ma et al. (2019, 2021); Oltramari et al. (2020). The `CommonGen` dataset contains sample concept-sets, with set lengths that range from 3-5 concepts. As a consequence of how the dataset was generated, most concept-sets are connected by either 1-hop or 2-hop relations in `ConceptNet` (Lin et al. 2020). For a given concept-set, our goal is to extract the commonsense relations that connect the concepts in the concept-set. The multi-hop extraction method is developed as follows:

- In order to capture salient relations between concept-set elements, we consider two hops by searching among all 1-hop neighbors of each concept-set element and setting these neighbors as the root concept to look for *their* relations with other concept-set elements. For example, given a concept-set ["broccoli", "cheese", "chicken", "pizza"], we extract 1-hop relations such as ["cheese", "AtLocation", "pizza"] and 2-hop relations such as ["broccoli", "AtLocation", "plate", "RelatedTo", "pizza"].
- Sometimes, there are low-connectivity concepts that have no 1-hop or 2-hop relations with other input concepts. For these, we use 3-hop extractions. However, searching among all possible 3-hop relations is time-consuming. Instead of using all neighbors, we use only the five "nearest" neighbors (i.e., those with highest `ConceptNet` relation weights, as a proxy for the strength by which the edge expresses the assertion) as the root concepts for the second-level search. Because 3-hop relations involve more than 3 components, we use the term "knowledge relations", in the following sections, to refer to all extracted relations.

Knowledge Selection One issue with multi-hop extraction is that the same pair of concepts can be linked by different relations, yielding a noisy inductive bias for training models. In fact, for the `CommonGen` task, an average of 9 knowledge relations are extracted for each concept-set. However, we recognize that some knowledge relations are more useful than others. While it is hard to automatically evaluate relation relevance, we propose heuristic selection mechanisms.

- *Relation types and POS-based selection.* Our knowledge extraction process excludes such relation types as: 'FormOf', 'DerivedFrom', 'Antonym' and 'DistinctFrom'. 'FormOf' and 'DerivedFrom' can be discarded, since they indicate purely syntactic relations, as in ['walk', 'FormOf', 'walking']. 'Antonym' and 'DistinctFrom' are designed to link concepts that have opposing semantic interpretations, which can harm the goal of linking concept-set elements together. We only extract relations following the given POS tags of concept-set elements.
- *Random subset selection*: Instead of filtering on the above-mentioned criteria, we can perform random subset selection over all knowledge relations by assigning a random selection probability to each of the relations and selecting the relations above a particular threshold. The selection can be constrained such that at least one relation for each input concept will be kept.
- *Subset selection using Prior probability*: Instead of assigning a random probability to all knowledge relations, we compute a prior probability over the relation types i.e. the number of occurrences of a relation type over total number of relations. We also have a random component similar to 'random subset selection' so that we retain some of the relations with less frequent relation type. The probability of choosing a knowledge relation is the sum of prior and random probabilities. We observe the distribution obtained on the training data and choose a threshold for relation selection. The average number of relations for each concept-set is now 3 to 4.

Query Expansion Apart from knowledge extraction, we perform query expansion on the given concept-set. The motivation is to make our model generate sentences with more contents by expanding our concept-set. We count the frequency for all single-hop neighbors of input concepts and select neighbors that are connected with more than half of the input concepts. For example, the given concept set ["drill", "field", "run", "team"] has ["baseball", "sport", "football"] as expansion concepts. The expanded concepts are obtained in order of their frequencies, so that the number of expanded concepts can be adjusted by setting a threshold. They serve as supplementary concepts to feed into the model.

Knowledge Injection Methodology

We use UniLM (Dong et al. 2019b) as the backbone architecture, adding improvisations with respect to commonsense injection within it. On a high-level the architecture of this Language Model (LM) is as described in Figure 1 consisting of multiple stacked layers of bidirectional Transformer encoder and a unidirectional decoder being opti-

mized for Masked LM loss. Originally, UniLM takes input in the form *(Concept-Set [SEP] TARGET_SENT)*. We fine-tune the UniLM model for the generation task in seq-2-seq mode. Using UniLM as the base network, we attempt to perform language generation requiring commonsense knowledge using the following methods:

Knowledge Concatenation As a baseline injection methodology, we concatenate the extracted tokens from the concept relations to our inputs. This method is based on rationale concatenated methods proposed by (Lin et al. 2020) in their work. The expanded list of input tokens is then fed to the language model for text generation.

Attention Injection Method To better handle the amount and context of commonsense knowledge fed to the LM, we adopt attention based injection methodology following HyKAS (Ma et al. 2019). We extract commonsense knowledge from `ConceptNet` as described in the Knowledge Extraction Methodology section and provide that to the model in the form of knowledge relations.

We make use of commonsense embeddings (marked as *cs_embeddings* in Figure 1) which are sent to a bi-LSTM encoder to get commonsense encodings (*cs_encodings*). We consider the hidden representation of the concept-set segment after the first encoder layer of UniLM and compute Key-Value attention with the *cs_encodings* obtained from the bi-LSTM layer. We compute attention according to the given equations. Here, Q,K,V are the projections of *cs_encodings*, H_{hid} is the hidden representation from UniLM, M is the joint mask for *commonsense_encoding* and H_{hid}:

$$A = softmax(QK^T + M)$$
$$H_{ctxt} = A \cdot V$$
$$H_{attn} = W^T \cdot H_{ctxt} + H_{hid}$$

We use the obtained attention_scores along with the hidden representation to get *commonsense attended hidden representation*, H_{attn}. This is inserted back to the UniLM model and it propagates through the consecutive encoder layers.

- **Attention on Knowledge relations** The knowledge relations are given as additional inputs to the LM along with available input set, *[Concept-Set, Target Sentence]*. We convert the relations into regular sentences, e.g. *(football → RelatedTo → sport)* is converted to the form "football related to sport". We further tokenize and obtain BertEmbeddings (cs_embeddings) for such sentences, which are sent to the encoder described above. We compute attention between these relations and the input concept-net and inject it into the hidden representation of the encoder.
- **Attention for expansion concepts** In order to understand the context required for sentence generation we derive additional concepts from `ConceptNet` which tries to add details to the story (sentence) being built from concepts in the given concept-set as building blocks. We call this set as the expanded concept-set. The expanded concept-sets are given as input to the LM by concatenating them with provided concept-sets. In a general setting with injection, we try to capture the relation between the given concepts and the concept relations which correspond to them whereas for expanded concepts, there are no knowledge relations describing them. By considering the expanded concepts as input, we will be increasing the sparsity of the model and losing vital information that could be captured and this is essentially adding some amount of noise into the model. In order to avoid such scenario, we build masks to differentiate the given and expanded concept-sets within the injection mechanism. Attention is now computed only on the initial or given concept-set as is necessary.

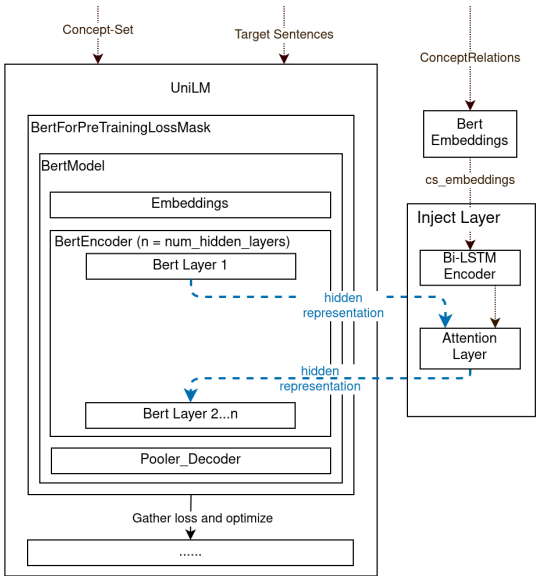

Figure 1: Integrating Attention Injection with UniLM

Imposing Lexical Constraints

Each concept in the dataset sample is of the format *concept_POS*, an example being *drill_N#field_N#run_V#team_N* where *drill* is a concept with expected POS Tag *Noun* while *run* is another concept with expected POS Tag *Verb*. We thus formulate the lexical constraint rules as follows:

- Each of the given concepts should appear at least once in the generated sentence
- Each of the given concepts should have the same POS tag in the generated sentence as given in the dataset

We experiment with concept and POS tag aware knowledge extraction as described previously (Knowledge Extraction Methodology section). The knowledge relations selected are constrained to select relations for each concept in the concept set, while rejecting any relations where the corresponding POS tag of the concept in the relation from `ConceptNet` does not match the given POS tag. We are thus able to cover 99.57% of all the unique concepts in the dataset. This constraint is also maintained during random subset selection, where subset of the relations is constrained to select at least one relation for each concept while determining the subset of relations.

Table 1: Experimental results for knowledge injection through self-attention on knowledge relations

Experiment	BLEU (↑)				ROUGE (↑)				
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	ROUGE-2	METEOR (↑)	CIDEr (↑)	SPICE (↑)
UniLM Baseline	-	-	38.3	27.7	43.87	21.48	29.7	14.85	30.2
Attention + Knowledge selection	71.6	52.5	37.8	27.0	43.3	21.67	29.2	14.57	29.6
Attention + Multi-hop	71.4	53.3	38.8	28.1	**49.8**	**25.1**	29.3	14.78	29.5
Attention + Multi-hop + Best N Beam Scoring	71.7	53.3	38.7	27.9	44.2	23.1	29.8	15.11	30.1
Attention + Random subset selection	**72.4**	53.2	38.1	27.2	43.72	22.45	29.8	14.92	30.2
Attention + Prior subset selection	71.8	**53.5**	**39.0**	**28.4**	43.8	22.8	29.6	15.06	30.0
Attention + Prior subset selection + Best N Beam Scoring	71.9	53.4	38.7	27.9	44.4	23.43	**30.1**	**15.23**	**30.6**

Table 2: Experimental results for knowledge injection on Query Expansion

Experiment	BLEU (↑)				ROUGE (↑)				
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	ROUGE-2	METEOR (↑)	CIDEr (↑)	SPICE (↑)
UniLM Baseline	-	-	38.3	27.7	43.87	21.48	29.7	14.85	30.2
Concatenation + Query Expansion	68.6	49.1	34.6	24.5	40.5	19.4	27.6	13.26	27.7
Attention + Query Expansion + Multi-hop	68.2	50.2	36.2	26.2	41.3	20.74	27.7	13.65	28.2
Attention + Query Expansion + Knowledge Selection	67.9	49.3	35.3	25.3	41.27	20.22	27.5	13.50	27.6
Attention + Constrained Query Expansion + Multi-hop	69.6	51	36.5	26.1	42.0	20.5	28	13.82	28.3

Best-*N* Beam Scoring

We propose modifying constrained decoding by scanning the generations from the top N beams for their adherence to the constraints formulated above, N being a hyperparameter. We choose N = 4. For each of the N extracted sentences, we calculate the coverage score as product of % of given concepts present in the generation * % of concepts with correct POS Tag in the generation. The sentences with the highest coverage score are selected as generation. In case of a tie, the sentence with a higher beam score is selected.

Experiments

In this section, we evaluate the efficacy of our knowledge extraction and injection models on the `CommonGen` dataset with UniLM as our baseline language model and `ConceptNet` as the knowledge graph.

We evaluate the performance of the model for the task of lexically-constrained text generation w.r.t. human generations by calculating BLEU (Papineni et al. 2002), ROUGE (Lin 2004), METEOR (Denkowski and Lavie 2014), CIDEr (Vedantam, Lawrence Zitnick, and Parikh 2015) and SPICE (Anderson et al. 2016) scores.

As a baseline, we perform text generation using UniLM without any knowledge injection. We then experiment with two different strategies of knowledge injection: attention-based and concatenation based.

Concatenation based: This is a baseline injection methodology inspired by rationale concatenation method by (Lin et al. 2020). We obtain expansion concepts using query expansion methodology as described previously and the tokens corresponding to these expansion concepts are then concatenated with the input concepts. Thus, for a given concept set *run team drill field* and expansion set *baseball sport football*, we send as input a concatenated list of concepts, *run team drill field baseball sport football*. This method is referred to as *Concatenation + Query Expansion* in later sections of the paper.

Attention injection on knowledge relations: Evaluations are performed on two major knowledge relation extraction models: (1) Multi-hop *(Attention + Multi-hop)* and, (2) length and frequency-based Knowledge selection *(Attention + Knowledge selection)*. Separately, we also experiment with performing Knowledge selection by selecting a subset of relations from multi-hop relation relations. For this method, the subset selection can be done in a randomized fashion by selecting the relations randomly with a probability of 0.5 (*Attention + Random subset selection*) or by also associating a prior probability with each relation (*Attention + Prior subset selection*), as explained in Knowledge Extraction Methodology Section. In both these cases, the subset selection is forced to select at least one relation for each input concept, wherever possible.

Attention injection on query expansion: We experiment with attention injection while incorporating the expansion concepts obtained from query expansion. We first experiment with all expansion concepts for both multi-hop *(Attention + Query Expansion + Multi-hop)* and knowledge selection *(Attention + Query Expansion + Knowledge Selection)* strategies for obtaining the knowledge relations on which the attention weights are calculated. Then we also experiment with the constrained version of Query Expansion with multi-hop *(Attention + Constrained Query Expansion + Multi-hop)*, where each input sample contains at most 2 expansion concepts(with the highest frequencies). The threshold 2 is selected given that our target sentences across train/dev/test contain on average 2.5 additional verb/noun concepts.

Further, we also integrate Best-*N* Beam scoring with the two best performing models and evaluate the results.

Assessing Effect of Lexical Constraints To evaluate the generated sentences w.r.t. the given lexical constraints, we calculate and report the following metrics:

- % of given concepts missing in the text generated
- % of samples with any mismatching POS tag in the given input concept-set and generated text

Table 3: Analysis of lexical constraints

Model	Missing Concepts (↓)	Mismatching POS (↓)
UniLM Baseline	42.9	22.45
Attention + Knowledge Selection	42.1	22.37
Attention + Multi-hop	44.87	21.1
Attention + Multi-hop + Best N Beam Scoring	28.12	21.64
Attention + Random subset selection	39.7	21.81
Attention + Prior subset selection	42.32	21.66
Attention + Prior subset selection + Best N Beam Scoring	**27.04**	21.69
Concatenation + Query Expansion	61.2	21.4
Attention + Query Expansion + Multi-hop	60.8	**19.31**
Attention + Query Expansion + Knowledge Selection	59.5	20.42
Attention + Constrained Query Expansion + Multi-hop	59.5	20.9

Results

We discuss the results obtained for the different proposed knowledge extraction and injection models in Tables 1 and 2. We find that *Attention + Prior subset selection* model achieves the highest BLEU score. However, constrained decoding on this model using the Best-*N* beam scoring (***Attention + Prior subset selection + Best N Beam Scoring***) not only improves metrics such as METEOR, CIDEr and SPICE, but also achieved the least % of missing concepts across all models, and is thus the best performing model. On the other hand, concatenation based methods perform poorly with decrease in model performance across metrics, as seen in Table 2. Our results show that attention based methods outperform other injection methodologies by huge margins.

Analyzing the sentences generated by baseline UniLM, we found that although the sentences appear correct grammatically, they lackred the required common sense knowledge to be meaningful. These sentences are still able to get a high BLEU score with a human generated sentence since the score is independent of the order of the n-grams. Although the scores for the proposed methods do not show much improvements, the generated sentences have a visible improvement in the commonsensical aspect of the sentence. We discuss this further with examples in the Discussion section.

Poor performance of concatenation and query expansion method proves that feeding more inputs to the model might actually result in noise addition as the model has no way to differentiate between given and expanded concepts. These results further justify the need of an attention based injection mechanism.

To do so, we incorporate the expansion concepts in the inputs as described in section: *Injection methodologies - Attention for expansion concepts*. The use of expansion concepts provide extra knowledge while encoding but do not interfere with the knowledge injection. We tabulate the results of the masked injection model with and without knowledge selection. A small improvement in scores is seen, but the sentences seem to be much more natural and sensible. We also experiment with constraining the number of expansion concepts and see minor improvements in the performance scores, thus underlining the fact that too many expansion concepts may result in noise injection in the model. Instead, using a subset of expansion concepts, which are selected preferring high frequencies, can achieve better results.

Constrained Decoding

As can be seen in Table 3, all our models perform better than the baseline UniLM model in terms of generating sentences with the correct POS tag, with more than 2 % points improvement at a maximum. This can be directly attributed to knowledge selection, where relation selection was done in accordance with the given POS tag, wherever possible.

We also see that Best-*N* Beam Scoring is very effective in reducing the percentage of missing concepts, with upto 15% improvements observed, whenever the method was applied. This can be attributed to the fact that beam search is biased towards choosing shorter sentences as product of token probabilities for a longer sentence is bound to result in a lower beam score. Larger number of concepts are naturally expected to result in longer sentences and exploring more beams enables choosing sentences with more concepts included, without compromising significantly on beam score.

In the absence of beam search, we see that the percentage of missing concepts worsens drastically for concatenation based methods but shows slight improvements for attention based methods. In case of concatenation-based injection methods, this can be attributed to noise addition by expansion concepts, where the generated sentences often picked expansion concepts instead of the given concepts, hinting at model's inability to differentiate between the original and expanded concepts.

We also observed that across all the models, the repetition of concepts increased. Although this might not necessarily result in any metric/generation get better/worse, it is not in alignment with the usual human spoken English, where over-repetition of words is not expected.

Discussion

We study the sentences generated by the baseline UniLM model and our proposed models and give few examples in Table 4. For the concept set *bed_N comb_V hair_N sit_V*, we can see that the Unilm generated sentence covers all the given concepts but uses the word *comb* as a noun instead of a verb, along with concept repetition. All our attention injec-

Table 4: Generated Sentences

Experiment	Generated Sentences		
	bed_N comb_V hair_N sit_V	cover_V front_N mountain_N short_N wear_V	board_N boat_N ride_V water_N
UniLM Baseline	A woman with a comb and hair comb sits on a bed	person wearing shorts and long covers to the front of the mountain	People board a boat and board a ride in the water
Attention + Knowledge Selection	A man with long hair sitting on a bed combing his hair	man wearing short shorts to the front of a mountain covered in snow	A man rides a boat in the water to a ride on a boat .
Attention + Multi-hop	A man sitting on a bed combing his hair with a comb .	man wearing a long sleeve shirt and shorts to the front of the mountain	A man rides a boat in the water .
Attention + Multi-hop + Best N Beam scoring	**A man sits on a bed and combs his hair .**	man wearing a long sleeve shirt and shorts to cover the front of the mountain	A man rides a boat in the water with a boy on board .
Attention + Random Subset selection	A man with combs sits on a bed and combs his hair .	A man wearing a long sleeved shirt and shorts covered in snow covered mountains .	A man rides a boat down a river and rides it into the water .
Attention + Prior Subset selection	A man sits on a bed combing his hair with a comb .	A man wearing short shorts covered in snow covered mountains	A man rides a boat in the water .
Attention + Prior Subset selection + Best N Beam scoring	A man sits on a bed combing his hair with a comb	**person wearing a short coat covered in snow on the front of mountain**	**A man rides a boat in the water with a boy on board**
Concatenation + Query Expansion	A man with comb and hair comb sits on a bed	A man wearing a long sleeved shirt is wearing a snow covered jacket	A man sits on a boat and is riding a wooden boat in the water
Attention + Query Expansion + Multi-hop	A woman sits on a bed and combs her hair with a comb .	A man is wearing a long sleeved shirt to cover the front of a mountain	A man rides a boat on the water
Attention + Query Expansion + Knowledge Selection	A woman sits on a bed and combs her hair with a comb .	A man is wearing a long sleeved shirt with a mountain covered in snow	A man boards a boat and rides it down the water
Attention + Constrained Query Expansion + Multi-hop	A woman sitting on a bed combing her hair with a comb.	A man is wearing a long coat to cover his face .	A man rides a boat in the water

tion models were able to successfully maintain the lexical constraint of including all the concepts with the proper POS tag. We also see that the issue of repetition was rectified in the best sentences generated.

In case of concept set *cover_V front_N mountain_N short_N, wear_V*, the sentence generated by baseline UniLM does not miss any concept but also does not make any sense. With the use of attention mechanism and external knowledge, the generated sentences improve on the commonsense nature of the sentences. In case of the concept set *board_N boat_N ride_V water_N*, the baseline sentences not only lacked commonsense but also missed the concept *board_V*. We observe that Best-*N* beam scoring on an attention-based model enabled including all the given concepts without compromising on the quality of the sentence generated.

Future Work

In our work, we experiment with various knowledge extraction methods and injection techniques. We observe that knowledge injection enables language models to perform better at text generation tasks that are lexically constrained. The improvement is visible in only a few examples and this motivates us to work towards improving our approach. As a future improvement, we plan to experiment further with constrained decoding where we plan to explore alternate methods to modify the beam score and give weightage to lexical constraints during the decoding process. We also wish to explore different attention mechanisms for knowledge injection. A self attention head based injection on the extracted concepts seems like a natural next step. Adding to that, defining a pre-training objective and pre-training the injection layer could help reduce the noise and generate much more meaningful sentences. `ConceptNet` relations and the input concepts from the dataset have a PoS tag associated with them: thus, we also plan to explore PoS based encoding and decoding, where the unimodal latent representational of text takes into account a POS Tag based embedding. Moving aside from `ConceptNet`, we wish to see how the extraction techniques differ on various other knowledge graphs and come up with a generalized extraction mechanism.

References

Ahn, S.; Choi, H.; Pärnamaa, T.; and Bengio, Y. 2017. A Neural Knowledge Language Model.

Anderson, P.; Fernando, B.; Johnson, M.; and Gould, S. 2016. SPICE: Semantic Propositional Image Caption Evaluation. In *ECCV*.

Bauer, L.; Wang, Y.; and Bansal, M. 2019. Commonsense for Generative Multi-Hop Question Answering Tasks.

Cambria, E.; Li, Y.; Xing, F.; Poria, S.; and Kwok, K. 2020. Sentic-

Net 6: Ensemble Application of Symbolic and Subsymbolic AI for Sentiment Analysis. *Proceedings of the 29th ACM International Conference on Information & Knowledge Management* .

Chalier, Y.; Razniewski, S.; and Weikum, G. 2020. Joint Reasoning for Multi-Faceted Commonsense Knowledge.

Chen, W.; Su, Y.; Yan, X.; and Wang, W. Y. 2020. KGPT: Knowledge-Grounded Pre-Training for Data-to-Text Generation.

Denkowski, M.; and Lavie, A. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, 376–380.

Dong, L.; Yang, N.; Wang, W.; Wei, F.; Liu, X.; Wang, Y.; Gao, J.; Zhou, M.; and Hon, H.-W. 2019a. Unified Language Model Pre-training for Natural Language Understanding and Generation. In Wallach, H.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 32, 13063–13075. Curran Associates, Inc. URL https://proceedings.neurips.cc/paper/2019/file/c20bb2d9a50d5ac1f713f8b34d9aac5a-Paper.pdf.

Dong, L.; Yang, N.; Wang, W.; Wei, F.; Liu, X.; Wang, Y.; Gao, J.; Zhou, M.; and Hon, H.-W. 2019b. Unified Language Model Pre-training for Natural Language Understanding and Generation.

Ilievski, F.; Szekely, P.; Cheng, J.; Zhang, F.; and Qasemi, E. 2020. Consolidating Commonsense Knowledge.

Ilievski, F.; Szekely, P.; and Schwabe, D. 2020. Commonsense Knowledge in Wikidata.

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; Bernstein, M. S.; and Li, F.-F. 2016. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations.

Lauscher, A.; Majewska, O.; Ribeiro, L. F. R.; Gurevych, I.; Rozanov, N.; and Glavaš, G. 2020. Common Sense or World Knowledge? Investigating Adapter-Based Knowledge Injection into Pretrained Transformers.

Lebret, R.; Grangier, D.; and Auli, M. 2016. Neural Text Generation from Structured Data with Application to the Biography Domain. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 1203–1213. Austin, Texas: Association for Computational Linguistics. doi:10.18653/v1/D16-1128. URL https://www.aclweb.org/anthology/D16-1128.

Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461* .

Lin, B. Y.; Zhou, W.; Shen, M.; Zhou, P.; Bhagavatula, C.; Choi, Y.; and Ren, X. 2020. CommonGen: A Constrained Text Generation Challenge for Generative Commonsense Reasoning.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 74–81.

Liu, W.; Zhou, P.; Zhao, Z.; Wang, Z.; Ju, Q.; Deng, H.; and Wang, P. 2019. K-BERT: Enabling Language Representation with Knowledge Graph.

Liu, Y.; Yang, T.; You, Z.; Fan, W.; and Yu, P. S. 2020. Commonsense Evidence Generation and Injection in Reading Comprehension.

Ma, K.; Francis, J.; Lu, Q.; Nyberg, E.; and Oltramari, A. 2019. Towards Generalizable Neuro-Symbolic Systems for Commonsense Question Answering. In *Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing*, 22–32. Hong Kong, China: Association for Computational Linguistics. doi:10.18653/v1/D19-6003. URL https://www.aclweb.org/anthology/D19-6003.

Ma, K.; Ilievski, F.; Francis, J.; Bisk, Y.; Nyberg, E.; and Oltramari, A. 2021. Knowledge-driven Data Construction for Zero-shot Evaluation in Commonsense Question Answering. *In Proceedings of the 35th AAAI Conference on Artificial Intelligence (AAAI-21)* .

Mihaylov, T.; and Frank, A. 2018. Knowledgeable Reader: Enhancing Cloze-Style Reading Comprehension with External Commonsense Knowledge. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 821–832. Melbourne, Australia: Association for Computational Linguistics. doi:10.18653/v1/P18-1076. URL https://www.aclweb.org/anthology/P18-1076.

Miller, G. A. 1995. WordNet: A Lexical Database for English. *Commun. ACM* 38(11): 39–41. ISSN 0001-0782. doi:10.1145/219717.219748. URL https://doi.org/10.1145/219717.219748.

Oltramari, A.; Francis, J.; Henson, C.; Ma, K.; and Wickramarachchi, R. 2020. Neuro-symbolic Architectures for Context Understanding.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2019. Language Models are Unsupervised Multitask Learners.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683* .

Sap, M.; Le Bras, R.; Allaway, E.; Bhagavatula, C.; Lourie, N.; Rashkin, H.; Roof, B.; Smith, N. A.; and Choi, Y. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 3027–3035.

Speer, R.; Chin, J.; and Havasi, C. 2016. Conceptnet 5.5: An open multilingual graph of general knowledge. *arXiv preprint arXiv:1612.03975* .

Vedantam, R.; Lawrence Zitnick, C.; and Parikh, D. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4566–4575.

Wang, P.; Peng, N.; Ilievski, F.; Szekely, P.; and Ren, X. 2020. Connecting the Dots: A Knowledgeable Path Generator for Commonsense Question Answering.

Wiseman, S.; Shieber, S.; and Rush, A. 2018. Learning Neural Templates for Text Generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 3174–3187. Brussels, Belgium: Association for Computational Linguistics. doi:10.18653/v1/D18-1356. URL https://www.aclweb.org/anthology/D18-1356.

Yang, Z.; Blunsom, P.; Dyer, C.; and Ling, W. 2017. Reference-Aware Language Models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 1850–1859. Copenhagen, Denmark: Association for Computational Linguistics. doi:10.18653/v1/D17-1197. URL https://www.aclweb.org/anthology/D17-1197.