Template Filling for Controllable Commonsense Reasoning

Dheeraj Rajagopal♠, Vivek Khetan♣, Bogdan Sacaleanu♣,
Anatole Gershman♠ Andrew Fano♣, Eduard Hovy♠
♠ Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, USA
♣ Accenture Labs, San Francisco, USA
{dheeraj, anatoleg, hovy}@cs.cmu.edu
{vivek.a.khetan, andrew.e.fano, bogdan.e.sacaleanu}@accenture.com

Abstract

Large-scale sequence-to-sequence models have shown to be adept at both multiple-choice and open-domain commonsense reasoning tasks. However, the current formulations do not provide the ability to control the various attributes of the reasoning chain. To enable better controllability, we propose to study the *commonsense reasoning as a template filling task* (TemplateCSR) — where the language models fills reasoning templates with the given constraints as control factors. As an approach to TemplateCSR, we (i) propose a dataset of commonsense reasoning template-expansion pairs for *healthcare and well-being* domain and (ii) introduce ITO, an instruction fine-tuned sequence-to-sequence model that performs commonsense reasoning across concepts in the template. Our experiments show that our approach outperforms baseline both in generation metrics and factuality metrics. We also present a detailed error analysis on our approach's ability to reliably perform template based commonsense reasoning[1].

1 Introduction

Commonsense reasoning has been studied across both *multiple choice* (Tandon et al., 2018; Talmor et al., 2019; Lv et al., 2020), and *open-ended knowledge base* settings (Lin et al., 2021a). While multiple choice approaches require a list of answer options, open-ended KB approaches assume that the answer exists in an available knowledge base (KB). Such constraints often limit these systems' ability in practical applications where control is required (e.g. a web search query with specific conditions).

To complement the existing commonsense reasoning efforts, our work aims to enhance the commonsense reasoning capabilities of natural language processing (NLP) systems by studying **template commonsense reasoning** (TemplateCSR)
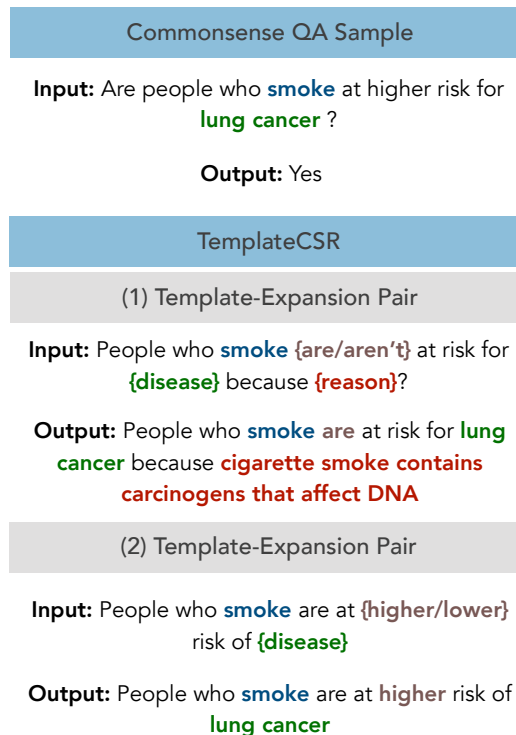

Figure 1: In this example, we show how a commonsense reasoning question can be formulated as two different template-expansion pairs, each focusing on different aspects of reasoning between the concepts *smoking* and *lung cancer*. While formulation (1) focuses on the explanation, (2) aims to understand the qualitative relationship between them.

— where reasoning is achieved by filling templates with restricted template slots, rather than selecting answers from a list of candidates or KB. TemplateCSR task is challenging as there are no available annotations and potentially multiple correct expansions for each template. Moreover, the task of designing templates with slots that satisfy arbitrary constrains and their expansion is still an open challenge. For example, for an example reasoning template **People who smoke are at a risk of {disease}** , a system needs to first constrain the slot to only *diseases*, and then use the addi-

[1] All code and data will be released publicly

tional constraint of *smoking* to arrive at the right answer in the slot. In comparison to Language Model (LM) probing approaches (Ribeiro et al., 2020) that test capabilities of LM that are already trained, we aim to propose a task and train a model for TemplateCSR task.

Figure 1 shows one such example from the *healthcare and well-being* domain, where we show how an existing commonsense reasoning query can be formulated as different template-expansion pairs with control over different aspects of the reasoning. In the first expansion, the reasoning chain focuses on the relationship between *smoking* and *cancer* with the corresponding explanation (`reason`), while the second chain solely focuses on the qualitative relationship between the *smoking* and *cancer*.

To address the above mentioned challenges, our contributions in this paper for TemplateCSR are two-fold. First, we present a dataset of commonsense reasoning templates for the *healthcare and well-being* domain and their corresponding expansions that are valid completions of the template, which we define as template-expansion pairs (Fass and Wilks, 1983). The slots in the templates are open-ended and are not restricted to any particular categories and enable controlling the reasoning chain. Given the recent focus on explainable models for reasoning (Wiegreffe and Marasović, 2021), we also augment templates with an optional free-form explanation slot that explains the reasoning connection between various commonsense concepts. Our TemplateCSR dataset comprises of about 3600 unique template-expansion pairs collected from diverse sources, and we hope to enable SEQ-TO-SEQ systems to effectively learn to fill commonsense reasoning templates.

Next, we present ITO, a model that formulates the TemplateCSR challenge as a SEQ-TO-SEQ task where given a template with slots for specific concepts, the goal of the model is to produce meaningful completed sentences for the template. The concept in each slot in the template is provided via an *instruction* (Wei et al., 2022a), which indicates an abstraction of the nature of the slot. The multiple choice *qualifier* slot is used to model the relationship between the concepts and the *explanation* slot generates a free-form text explanation for the reasoning chain. Specifying each slots in free-form text enables control allowing commonsense reasoning questions to specify concepts, the

qualitative relationship and the nature of explanation.

In our experiments for the TemplateCSR task, ITO outperforms baseline both in terms of generation metrics such as ROUGE and BERTSCORE, and factual correctness (factuality) metrics such as FACTCC. We also evaluate the factuality of the dataset using expert human judges and present a detailed analysis of model failures. While we still observe factual errors, our approach provides deeper understanding of the mistakes, enabling alternate ways to build commonsense reasoning systems via templates using SEQ-TO-SEQ models.

2 Dataset

For our use-case, we create dataset samples of commonsense reasoning templates related to *healthcare and well-being*. Incorporating NLP systems for aiding healthy lifestyle has been an active area of research in the past decade (Liberato et al., 2014; Fadhil and Gabrielli, 2017; Doustmohammadian and Bazhan, 2021; Ahne et al., 2022). Inspired by this line of research, we want to collect templates that describe a relation between lifestyle related commonsense concept and a corresponding health related concept. In comparison to existing datasets like *commonsenseqa* (Talmor et al., 2019) which relies on fixed set of relationships from a knowledge-base (Speer and Havasi, 2013), we do not restrict the relationship types or number of concepts or hops, making it close to open-vocabulary text. We believe that our dataset augments well with the existing commonsense reasoning datasets in the community, contributing to the diversity of the data.

Based on the efficacy assessment for NLP systems in health and lifestyle related settings (Laranjo et al., 2018; Abd-alrazaq et al., 2020; Hoermann et al., 2017), we designed our basic template structure. Our basic units for the TemplateCSR task is as follows:

1. *concept slot* : contains an abstract category of a concept. The concept's abstraction is provided in a natural language format in open-vocabulary, without fixed class constraints. In the example shown in figure 2, *people with habit* and *disease* are concept slots.

2. *multiple-choice qualifier slot* : a word or phrase that describes the nature of the relationship between the concepts. This slot
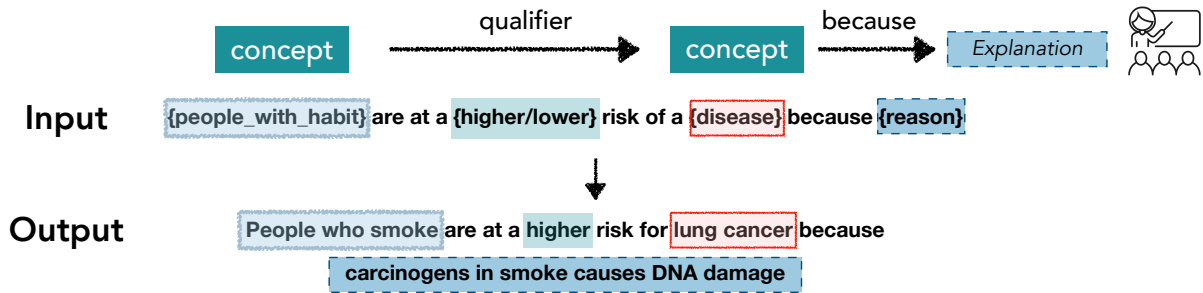
Figure 2: An overview of the overall template structure for our approach. Our goal is to reason across concepts for TemplateCSR. In this example template, concept slots are *people_with_habit* and *disease*, and the multiple choice qualifier slot - *higher/lower* describes their relationship and an explanation *reason* slot aims to get a free-form text explanation for how they are related.

is typically framed as a multiple-choice slot, where the goal is to pick an option from the choices rather than replacing the text in the template slot. Figure 2 shows an example where the slot *higher/lower* is one such multiple-choice qualifier slot.

3. *explanation slot* : this optional field consists of a free-form explanation that explains the reasoning between concepts, typically marked as *reason* slot.

Towards this, we collect a set of template (x) and its corresponding expansions (y) based on this overall schema for commonsense reasoning. In the example shown in figure 2, the template comprises of two concept slots, (*people with habit* and *disease*). The qualifier slot (*higher/lower*) specifies how one concept is connected to another concept in terms of their qualitative relationship. The template also includes an optional *explanation* slot that specifies in free-form text how smoking is connected to cancer. A valid output for the above-mentioned template is for instance, *people who smoke are at a higher risk for lung cancer because carcinogens in smoke causes DNA damage*, where *people with habit* is replaced by *people who smoke*, and the multiple choice qualifier slot *higher/lower* is replaced by *higher*, and *disease* slot replaced by *lung cancer* and finally the *reason* slot replaced by explanation of the qualitative relationship *carcinogens in smoke causes DNA damage*. In this example, we show how both the template-expansion pairs aim to uncover the relationship between *smoke* and *lung cancer*, while also providing the flexibility to additionally constrain the reasoning chain in any way.

Task Setup : To collect our dataset using crowd-sourcing, we use amazon mechanical turk platform [2]. Each datapoint took ∼120 seconds to annotate, and we paid an average of $15 per hour. Additionally, we used a filtering step to select master annotators with an approval rate of more than 90%. All the turkers were given specific instructions to input only factual information and not opinionated statements. Specifically, the turkers were instructed to use the following sources: *CDC*[3]*, WebMD*[4]*, Healthline*[5] and *Mayo Clinic*[6]. The annotators were also instructed to give a template, and at least two corresponding sentences that matches the template. The statistics of the data are as follows: the average sentence length was about *14.57* words, with mean 2.4 slots per template. Some qualitative examples from the dataset are given in the table 1. Overall, our dataset contains about 7000 template-sentence pairs with about 3600 unique templates. Once the templates are collected, the authors post-process the data to verify each template-expansion pair for correctness and validate that we do not have any identifying information like proper names. We then create a standard 70/10/20 train/val/test split.

3 Model

Early NLP systems have often relied on rule-based templatic systems (Riloff, 1996; Brin, 1998; Agichtein and Gravano, 1999; Craven et al., 2000) due to their simplistic nature. Compared to machine learning methods, they were often rigid (Yih, 1997). Despite their rigidity, template based

Concepts	Template Expansion Pairs	
	Sample Template	Valid Expansions for the Template
location, disease	**{person_at_location}** has a **{higher/lower}** risk of **{disease}** because **{reason_for_risk}**	Person who lives in a city has a higher risk - of depression because of stress due to noise Person who lives near a village has a lower risk - of respiratory illness because of lower pollution
prescription medication, disease	**{person_taking_prescription}** has a higher risk of **{disease}** due to **{reason}**	Someone on steroids have a higher risk for heart disease - because steroids compromise heart pumping People on insulin have a lower risk of hyperglycemia - because of lower glucose levels.
food item, substitute item	**{food_item_1}** should not be consumed with **{food_item_2}** because **{reason}**	Steak should not be consumed with mashed potatoes - because pairing fried foods increases the risk of diabetes. Pizza should not be consumed with French fries because - proteins require a much different stomach - environment than starches for proper digestion
behavior change, medical condition	A change in behavior such as **{behavior_change}** is often associated with **{a_medical_condition}** because **{reason_for_condition}**	A change in behavior such as becoming more sedentary is - often associated with obesity - because less activity leads to less calorie burning. A change in behavior such as no longer drinking coffee - is often associated with diminished insomnia - because less caffeine equals improved sleep.
symptom, medical condition, everyday action	When severe symptoms like **{a_symptom}** for a **{a_medical_condition}** shows up, immediately one should perform **{an_action}**	When severe symptoms like confusion or disorientation - for heatstroke show up, immediately, one should perform - cooling actions, such as applying cooling towels. When severe symptoms like unconsciousness for a - heart attack show up, immediately one should - call 911 and perform CPR while awaiting help.
lifestyle activity, disease	People often do **{an_activity}** before going to bed in night to prevent risk of **{disease}**. This is because **{reason_for_activity}**	People often do reading before going to bed in night - to prevent risk of insomnia. This is because - doing some light reading helps lull you to sleep. People often do teeth brushing before going to bed in night - to prevent risk of tooth decay. This is because - brushing removes cavity-causing plaque from teeth.

Table 1: Examples from our dataset. Each template has two corresponding sentences. **[concept]** is a concept, and **[text]** represents the explanation and **[text]** represents a qualifier. We show two sentences each for a template. Each template slot is given in free-form text without any restriction in vocabulary.

systems are often easy to comprehend, and lend themselves to easily incorporate domain knowledge (Chiticariu et al., 2013). Our goal is to combine the strengths of both template-based systems and recent advances in pretrained SEQ-TO-SEQ models for the task of commonsense reasoning via template expansion.

In this work, we present ITO (**I**nstruction Fine-tuning for **T**emplate Based Commonsense Reasoning), an approach that models the TemplateCSR task as a instruction fine-tuning task inspired by the recent advances in instruction based fine-tuning (Wei et al., 2022a).

Table 2 shows an example of our task setup for our ITO approach. In comparison to approaches such as Donahue et al. (2020), our approach does not strictly enforce that sentences

only fill missing spans of text. Rather, the expanded sentences are allowed to have additional modifications (token addition, deletion and rewrite). For instance, for the following input template - **{person_at_location}** has a **{higher/lower}** risk of **{disease}** because **{reason_for_risk}**, a valid expansion is *person who lives in the city has a higher risk of depression due to noise*.

3.1 Training

Given a template $x \in \mathcal{X}$ and its corresponding expansion $y \in \mathcal{Y}$, we can train any sequence-to-sequence model that models $p_\theta(y|x)$. Towards this, we use a pretrained sequence-to-sequence model $\mathcal{M}$ to estimate the filled template y for an input x. We model the conditional distribution

Input (Template)	Output (Expansion)
The first blank is **person_at_location**. The second blank is **higher/lower**. The third blank is **disease**. The fourth blank is a **reason_for_risk**. **[MASK]** has a **[MASK]** risk of **[MASK]** because **[MASK]**	Person who lives in a city has a higher risk of depression because of higher stress due to noise

Table 2: Overview of ITO approach for TemplateCSR. Each concept category is given as a instruction to the input and the slots are represented via the [MASK] token. The instruction describes each slot's abstraction and the task is to generate the *output*.

$p_\theta(y \mid x)$ parameterized by θ as

$$p_\theta(y \mid x) = \prod_{k=1}^{M} p_\theta(y^k \mid x, y^1, .., y^{k-1})$$

where M is the length of y.

3.2 Inference to Decode Template Expansions

The auto-regressive factorization of SEQ-TO-SEQ p_θ allows us to effectively cast the constrained decoding of filling the template as generating the sequence given the input x. For each expansion, we sample $y_j^1 \sim p_\theta(y \mid \mathbf{x}_j)$. Consequently, we sample $y_j^2 \sim p_\theta(y \mid \mathbf{x}_j, y_j^1)$, and the token generation process is repeated until we reach the end-symbol. For each symbol, the model has to decide between generating a token to replace the template slot or generate part of the template, while also ensuring the overall generated output sequence is consistent with the constraints given in the template.

4 Experiments

In this section, we describe the experimental setup, and baselines for our approach. Since our ITO approach is agnostic to the pretrained encoder-decoder architecture type, we perform experiments on two state-of-the-art SEQ-TO-SEQ models - BART and T5.

4.1 Experimental Setup

Metrics : We use the following evaluation metrics for evaluation for the TemplateCSR task: (i) ROUGE (Lin, 2004) and (ii) BERTSCORE (Zhang et al., 2019). N-gram metrics such as ROUGE are known to be limited, specifically for reasoning tasks. To mitigate this, we use BERTSCORE, which uses the similarity score between the reference and generated output using conceptual embeddings from BERT (Devlin et al., 2019) model,

which correlates better towards human judgements.

To perform the evaluation, we compare the generated sentence for the template against the gold annotations in the dataset. We remove the template words from the output and only compare the slot filler concepts to avoid score inflation due to copying. All the experiments were performed on a cluster of 8 NVIDIA V100 GPUs for about 32 GPU hours.

4.2 Models

We follow the same experimental settings across the baseline and our approach for all the models. We initialize all the models with their pretrained weights. We use commonly used encoder-decoder architectures for our experiments - BART-BASE, BART-LARGE, T5-BASE and T5-LARGE. The model settings are given in below:

(i) BART-BASE: This pretrained encoder-decoder transformer architecture is based on Lewis et al. (2020). It consists of 12 transformer layers each with 768 hidden size, 16 attention heads and overall with 139M params. (ii) BART-LARGE: Larger version of BART-BASE, with 24 transformer layers, 1024 hidden size, 16 heads and 406M params. (iii) T5-BASE: The T5 model is also a transformer encoder-decoder model based on Raffel et al. (2020) with 220M parameters with 12-layers each with 768 hidden-state, 3072 feed-forward hidden-state and 12 attention heads. (iv) T5-LARGE: T5-Large model version comprises of 770M parameters with 24-layers with 1024 hidden-state, 4096 feed-forward hidden-state and 16 attention heads [7].

Model	Type	ROUGE-1	ROUGE-2	ROUGE-L	BERTSCORE
T5-BASE	SPL TOKEN	14.00	2.71	12.58	2.2
T5-BASE	ITO	14.01	2.60	12.57	6.1
T5-LARGE	SPL TOKEN	13.74	3.11	13.74	4.8
T5-LARGE	ITO	16.74	4.33	15.37	6.7
BART-BASE	SPL TOKEN	17.17	5.60	16.32	3.9
BART-BASE	ITO	18.89	5.87	17.96	6.3
BART-LARGE	SPL TOKEN	19.54	7.57	18.49	7.0
BART-LARGE	ITO	20.58	7.32	19.58	7.6

Table 3: Overview of the results compared to baselines. The table shows that BART-BASE performs better than T5-BASE model and BART-LARGE outperforms both. Both in terms of ROUGE and BERTSCORE, we also observe that our INSTRUCTION approach outperforms SPL TOKEN approach. All experiments were done with 5 seeds, and reported are the average.

Baseline	Template	Output
SPL TOKEN (Donahue et al., 2020)	[S]person_at_location[/S] has a [S]higher/lower[/S] risk of [S]disease[/S] because [S]reason_for_risk[/S]	Person who lives in a city has a higher risk of depression because of stress due to noise

Table 4: Baseline Setup: We use Donahue et al. (2020) and use special tokens to indicate the start and end of each slot. The model is trained to predict the output, which is a valid expansion for the template.

4.3 Baseline Method

SPL TOKEN: In this approach, we use the special token approach (SPL TOKEN) (Donahue et al., 2020), where we indicate the start and end of each template slot in the input and generate the output sentence. Table 4 shows the baseline setup.

4.4 Results

The results across various pretrained encoder-decoder approaches are shown in table 3. In this table, we see that on average, BART models perform better than T5 models on average. We hypothesize this might be an effect of their pre-training task choices and corresponding datasets. We also observe that INSTRUCTION based models outperform the SPL TOKEN based approach. For all of the models and baselines, we used the greedy decoding strategy. Across all the experiments, we found that the ITO approach outperforms SPL TOKEN approach across both ROUGE and BERTSCORE scores for all models.

5 Factual Correctness Evaluation

Expert Evaluation of Dataset : Since our dataset contains health related commonsense knowledge, we additionally verified our data from domain experts. In our case, we selected a subset of 100 data samples from our validation set, and we verified the factual claims by recruiting two annotators with a Doctor of Medicine (M.D.) degree [8]. Each expert annotator labeled either *correct* or *incorrect* for a template-expansion pair. Overall, they found an average of 70% of the samples to be factually correct with an inter-annotator agreement of 80%.

Model Output Evaluation : To assess the quality of generated output, we perform additional factuality evaluation towards our best performing models - SPL TOKEN and ITO approach using BART-LARGE. Towards this we use the FACTCC factuality metric (Kryscinski et al., 2020), which uses entailment classification to predict a binary factuality label between the source document and generated output.

Computing factuality using FACTCC metric requires an input source document; (i.e.) the gener-

---

[7]Implementation adapted from Huggingface (Wolf et al., 2020)

Model	Type	FACTCC
BART-LARGE	SPL TOKEN	65.27
BART-LARGE	ITO	79.88

Table 5: Factual consistency results. In this experiment, we show that our `ITO` approach outperforms the `SPL TOKEN` approach in terms of factuality metric `FACTCC`, showing its relative effectiveness

ated output is compared against the source document for factual correctness. For this evaluation setup, we augmented each generated output y with a source document. Towards this, we use a large scale retrieval corpus based on Nguyen et al. (2016), and retrieve the top similar document D (Lin et al., 2021b) to a generated template expansion. Using the (D, y) pairs, we compute the factual correctness of our best performing models. From the table 5, we observe that our `ITO` approach outperforms the `SPL TOKEN` approach for factual correctness by ∼14 points in accuracy.

Additionally, we also perform human evaluation of factual correctness. For this experiment, three human judges annotated 100 unique samples for *correctness* - that indicates how many samples were correct from a human perspective. We used our best performing BART-BASE-ITO model for this evaluation. In this experiment, a sentence generated by the model for a given template was given to each human judge and they were asked to evaluate whether the sentence was correct, given the template. The inter-annotator agreement on sentence expansion correctness was substantial with a Fleiss' Kappa score (Fleiss and Cohen, 1973) of 0.73. From our evaluation, we found that human judges rated about 69% of the sentences to be correct given a template, comparable to our `FACTCC` evaluation metric numbers. Both the automated and human evaluation suggests that our `ITO` approach has better factual consistency.

6 Error Analysis

In this section, we analyze in detail how well language models perform template-expansion task for multihop reasoning. To understand the errors in depth, we complement our automated evaluation with manual error analysis. For this analysis, we randomly select 100 samples from the validation set predictions where the `ROUGE` scores were low. We observe the following categories of errors that language models exhibit. Table 6 shows the

common type of errors and a corresponding example for each type.

Error Type - Correct but not in gold (17%) : In several cases, we observe that the output produced by the language models are correct despite not matching the gold answer. This phenomenon is evident when the input template contains multiple possible answers. While the gold answer in the example shown in Table 6 (first row) fills the template using **smoking**, the model generates an answer related to **kidney damage**. While correct, the automated generation metrics such as ROUGE and BERTSCORE score such answers lower.

Error Type - Wrong commonsense concept (8%) : In this category of error, the model generates the wrong specification for the given slot. For instance (second row in table 6), the model mistakenly assumes **person taking less medication** as a **socioeconomic condition**. This error type gives a more nuanced understanding on which concept categories the model makes the most mistakes.

Error Type - Generic Explanation (53%): In several cases, the model resorts to generic explanation that are *obvious*. A generic explanation repeats the same information as the rest of sentence as an explanation, thereby not providing any new information compared to the rest of the sentence. In the example shown in Table 6 (row 3), the explanation **because of the strain of the heart** is already clear from the concept **chest pain**. A generic explanation is often unreliable in explainable NLP systems since it does not provide any insight into the reasoning capability of the model (Ye and Durrett, 2022).

Error Type - Factually Incorrect (22%) : Factual correctness is one of the biggest challenges in NLP applications (Petroni et al., 2020; Pagnoni et al., 2021). The incorrect factual information is also acute for cross-domain reasoning applications as well. As shown in the example (row 4 in table 6), the model incorrectly generates that **people with flu diagnosis** should do **exercise**. Factual correctness in generation models is an active area of research and we believe that template-based approaches can provide additional insight into this phenomenon.

Overall, TemplateCSR remains a challenging

Error Type	Template	Gold Answer	Generated Answer
Correct but not in gold	Children who are exposed to {environmental_factor} are often at a higher risk for {disease} because {reason}	Children exposed to second hand smoke are at a higher risk for lung disease because of breathing in the cigarette smoke	Children who are exposed to lead paint are often at a higher risk for kidney failure because lead causes kidney damage
Wrong commonsense concept	People with {certain_socioeconomic _condition} are at higher risk of {disease} as they are more exposed to {reason}	Person who often inhale a lot of dirt are at a higher risk of hay fever because of allergen content.	Person who often take less medications are at a higher risk of hay fever because of the drug ineffectiveness
Generic Explanation	When people with {certain_co-morbidities} show {symptoms}, this is because of {reason_for_patient_state}	When people with diabetes show lethargy, this is because of high glucose levels.	When people with heart disease show chest pain, this is because of the strain on the heart
Factually Incorrect	People with a {health_condition} should do {an_activity} because {reason}	People with a cardiovascular disease should do exercise since exercise burns excess fat	People with a flu diagnosis should do exercise because to stay active

Table 6: Error Analysis based on the BART-BASE-ITO model. We select 100 samples from the validation set and each row shows an example of each class of error.

task for SEQ-TO-SEQ models, specifically on their factual correctness and we believe it opens several avenues for progress in this research direction.

7 Related Work

Knowledge Bases : Knowledge Bases (KBs) have been the predominant approach to perform commonsense reasoning in the past (Speer and Havasi, 2013). Some of the prominent knowledge bases for commonsense reasoning include DBPedia (Mendes et al., 2012), YAGO (Suchanek et al., 2007) and NELL (Mitchell et al., 2018) or extending KBs with domain knowledge (Khetan et al., 2021, 2022). In this work, we focus on TemplateCSR using LM, which can be viewed as a complementary using KBs for commonsense.

Language Models for Reasoning: Using pretrained language models to generate knowledge has been studied for commonsense reasoning tasks. (Sap et al., 2019; Bosselut et al., 2019; Shwartz et al., 2020; Bosselut et al., 2021). Our work closely aligns with Bosselut et al. (2019, 2021). Compared to Bosselut et al. (2019), where our goal is to extend towards more controllable commonsense reasoning. Our work is also related to recent chain-of-thought prompting approach (Dalvi et al., 2021; Wei et al., 2022b), where a reasoning chain is first generated before the final solution. Compared to chain-of-thought prompting, our approach focuses on controllability of the reasoning process from input, via tem-

plate slots using instructions, similar to Wei et al. (2022a).

Language Model Infilling : Our work also closely relates to the language model infilling work in the literature such as Fedus et al. (2018) and Donahue et al. (2020). Compared to these works which only look at cloze-test infilling, our work aims to expand templates that cannot be directly modeled as cloze-style. Our work is also related to the story generation efforts such as Yao et al. (2019); Fan et al. (2018); Ippolito et al. (2019); Rashkin et al. (2020) but our application differs from them in that we focus on commonsense reasoning instead of content planning for stories.

8 Conclusion and Future Work

In this paper, we present the ITO approach that adapts language models to perform the TemplateCSR task via prompting. We collect a dataset for the same, and show that such an approach allows higher control over the reasoning process by enabling practitioners to specify the nature of the template slots. Through both automated and human metrics, we find that our ITO approach outperforms the baselines while also maintaining high factuality. For future work, we hope to extend this line of work towards other controllable generation tasks such as story generation and summarization.

9 Limitations and Ethics Considerations

Our work only tests single sentence template-expansion pairs. One of the limitation of our work is that we do not try on longer template sequences, which we want to explore for future work. While we present our dataset and corresponding modeling approaches, we acknowledge the limitations of the system and potential risks if it was used for real-world use-cases due to its factual errors. While we worked diligently with the annotators, and real domain experts to ascertain the quality of the dataset, we believe that there is immense room to be potentially improved and scaled further. We also did not test our dataset efficacy in large-scale models such as GPT-3 due to budget constraints, which we consider another limitation of this work.

As our results show, common sense reasoning and its health related knowledge reasoning is far from solved and we hope this dataset starts a research direction towards addressing this reasoning challenge. In no way, we support using this system for real-world commonsense related advice. The system, dataset and the accompanying publication is intended only for research purposes and ability to test current NLP systems' capabilities.

References

Alaa A. Abd-alrazaq, Asma Rababeh, Mohannad Alajlani, Bridgette M. Bewick, and Mowafa Said Househ. 2020. Effectiveness and safety of using chatbots to improve mental health: Systematic review and meta-analysis. Journal of Medical Internet Research, 22.

Eugene Agichtein and L. Gravano. 1999. Extracting relations from large plain-text collections.

Adrian Ahne, Vivek Khetan, Xavier Tannier, Md Imbesat Hassan Rizvi, Thomas Czernichow, Francisco Orchard, Charline Bour, Andy E. Fano, and Guy Fagherazzi. 2022. Extraction of explicit and implicit cause-effect relationships in patient-reported diabetes-related tweets from 2017 to 2021: Deep learning approach. JMIR Medical Informatics, 10.

Antoine Bosselut, Ronan Le Bras, and Yejin Choi. 2021. Dynamic neuro-symbolic knowledge graph construction for zero-shot commonsense question answering. In Proceedings of the 35th AAAI Conference on Artificial Intelligence (AAAI).

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Çelikyilmaz, and Yejin Choi. 2019. Comet: Commonsense transformers for automatic knowledge graph construction. In ACL.

S. Brin. 1998. Extracting patterns and relations from the world wide web. In WebDB.

Laura Chiticariu, Yunyao Li, and Frederick R. Reiss. 2013. Rule-based information extraction is dead! long live rule-based information extraction systems! In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 827–832, Seattle, Washington, USA. Association for Computational Linguistics.

M. Craven, Dan DiPasquo, Dayne Freitag, A. McCallum, Tom Michael Mitchell, K. Nigam, and Seán Slattery. 2000. Learning to construct knowledge bases from the world wide web. Artif. Intell., 118:69–113.

Bhavana Dalvi, Peter Jansen, Oyvind Tafjord, Zhengnan Xie, Hannah Smith, Leighanna Pipatanangkura, and Peter Clark. 2021. Explaining answers with entailment trees. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 7358–7370, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In NAACL 2019, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Chris Donahue, Mina Lee, and Percy Liang. 2020. Enabling language models to fill in the blanks. In ACL.

Azam Doustmohammadian and Marjan Bazhan. 2021. Social marketing-based interventions to promote healthy nutrition behaviors: a systematic review protocol. Systematic Reviews, 10.

Ahmed Fadhil and Silvia Gabrielli. 2017. Addressing challenges in promoting healthy lifestyles: the ai-chatbot approach. Proceedings of the 11th EAI International Conference on Pervasive Computing Technologies for Healthcare.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Dan Fass and Yorick Wilks. 1983. Preference semantics, iii-formedness, and metaphor. Am. J. Comput. Linguistics, 9:178–187.

William Fedus, Ian Goodfellow, and Andrew M. Dai. 2018. MaskGAN: Better text generation via filling in the _____. In International Conference on Learning Representations.

Joseph L Fleiss and Jacob Cohen. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. Educational and psychological measurement, 33(3):613–619.

Simon Hoermann, Kathryn L. McCabe, David N. Milne, and Rafael Alejandro Calvo. 2017. Application of synchronous text-based dialogue systems in mental health interventions: Systematic review. Journal of Medical Internet Research, 19.

Daphne Ippolito, David Grangier, Chris Callison-Burch, and Douglas Eck. 2019. Unsupervised hierarchical story infilling. In Proceedings of the First Workshop on Narrative Understanding, pages 37–43, Minneapolis, Minnesota. Association for Computational Linguistics.

Vivek Khetan, Annervaz K M, Erin Wetherley, Elena Eneva, Shubhashis Sengupta, and Andrew E. Fano. 2021. Knowledge graph anchored information-extraction for domain-specific insights.

Vivek Khetan, Md Imbesat Rizvi, Jessica Huber, Paige Bartusiak, Bogdan Sacaleanu, and Andrew Fano. 2022. MIMICause: Representation and automatic extraction of causal relation types from clinical notes. In Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland. Association for Computational Linguistics.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 9332–9346, Online. Association for Computational Linguistics.

Liliana Laranjo, Adam G. Dunn, Huong Ly Tong, Ahmet Baki Kocaballi, Jessica A. Chen, Rabia Bashir, Didi Surian, Blanca Gallego, Farah Magrabi, Annie Y. S. Lau, and Enrico W. Coiera. 2018. Conversational agents in healthcare: a systematic review. Journal of the American Medical Informatics Association : JAMIA, 25:1248 – 1258.

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, Online. Association for Computational Linguistics.

Selma Coelho Liberato, Ross Stewart Bailie, and Julie K. Brimblecombe. 2014. Nutrition interventions at point-of-sale to encourage healthier food purchasing: a systematic review. BMC Public Health, 14.

Bill Yuchen Lin, Haitian Sun, Bhuwan Dhingra, Manzil Zaheer, Xiang Ren, and William W. Cohen. 2021a. Differentiable open-ended commonsense reasoning. In NAACL.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In Text summarization branches out, pages 74–81.

Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021b. Pyserini: A Python toolkit for reproducible information retrieval research with sparse and dense representations. In Proceedings of the 44th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2021), pages 2356–2362.

Shangwen Lv, Daya Guo, Jingjing Xu, Duyu Tang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, and Songlin Hu. 2020. Graph-based reasoning over heterogeneous external knowledge for commonsense question answering. In AAAI.

Pablo Mendes, Max Jakob, and Christian Bizer. 2012. DBpedia: A multilingual cross-domain knowledge base. In Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), pages 1813–1817, Istanbul, Turkey. European Language Resources Association (ELRA).

T. Mitchell, W. Cohen, E. Hruschka, P. Talukdar, B. Yang, J. Betteridge, A. Carlson, B. Dalvi, M. Gardner, B. Kisiel, J. Krishnamurthy, N. Lao, K. Mazaitis, T. Mohamed, N. Nakashole, E. Platanios, A. Ritter, M. Samadi, B. Settles, R. Wang, D. Wijaya, A. Gupta, X. Chen, A. Saparov, M. Greaves, and J. Welling. 2018. Never-ending learning. Commun. ACM, 61(5):103–115.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. Ms marco: A human generated machine reading comprehension dataset.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 4812–4829, Online. Association for Computational Linguistics.

Fabio Petroni, Patrick Lewis, Aleksandra Piktus, Tim Rocktäschel, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. 2020. How context affects language models' factual predictions. In Automated Knowledge Base Construction.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research, 21(140):1–67.

Hannah Rashkin, Asli Celikyilmaz, Yejin Choi, and Jianfeng Gao. 2020. PlotMachines: Outline-conditioned generation with dynamic plot state tracking. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 4274–4295, Online. Association for Computational Linguistics.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with Check-List. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4902–4912, Online. Association for Computational Linguistics.

E. Riloff. 1996. Automatically generating extraction patterns from untagged text. In AAAI/IAAI, Vol. 2.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, pages 3027–3035.

Vered Shwartz, Peter West, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Unsupervised commonsense question answering with self-talk. arXiv preprint arXiv:2004.05483.

Robyn Speer and Catherine Havasi. 2013. Conceptnet 5: A large semantic network for relational knowledge. In The People's Web Meets NLP.

Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: A core of semantic knowledge. In Proceedings of the 16th International Conference on World Wide Web, WWW '07, page 697–706, New York, NY, USA. Association for Computing Machinery.

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

Niket Tandon, Bhavana Dalvi, Joel Grus, Wen-tau Yih, Antoine Bosselut, and Peter Clark. 2018. Reasoning about actions and state changes by injecting commonsense knowledge. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 57–66, Brussels, Belgium. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022a. Finetuned language models are zero-shot learners. In International Conference on Learning Representations.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022b. Chain of thought prompting elicits reasoning in large language models.

Sarah Wiegreffe and Ana Marasović. 2021. Teach me to explain: A review of datasets for explainable natural language processing. In NeurIPS Datasets and Benchmarks.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 38–45, Online. Association for Computational Linguistics.

Lili Yao, Nanyun Peng, Ralph Weischedel, Kevin Knight, Dongyan Zhao, and Rui Yan. 2019. Plan-and-write: Towards better automatic storytelling. Proceedings of the AAAI Conference on Artificial Intelligence, 33(01):7378–7385.

Xi Ye and Greg Durrett. 2022. The unreliability of explanations in few-shot in-context learning.

S. Yih. 1997. Template-based information extraction from tree-structured html documents.

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