A Dataset for Cross-Domain Reasoning via Template Filling

Anonymous ACL submission

Abstract

While several benchmarks exist for reasoning tasks, reasoning across domains is an underexplored area in NLP. Towards this, we present a dataset and a *prompt-template-filling* approach to enable sequence to sequence models to perform cross-domain reasoning. We also present a case-study with *commonsense* and *health and well-being* domains, where we study how *prompt-template-filling* enables pretrained sequence to sequence models across domains. Our experiments across several pretrained encoder-decoder models show that cross-domain reasoning is challenging for current models. We also show an in-depth error analysis and avenues for future research for reasoning across domains[1].

1 Introduction

Humans often need to reason across different domains for several day-to-day decisions. For instance, *Are leafy greens good for people with history of blood clots ?* Answering this question requires commonsense understanding that *leafy greens are high in vitamin-K* and a related health domain knowledge that *people with history of blood clots are prescribed blood thinners and vitamin-K inhibits blood thinner action, increasing blood clots*. Answering questions like these present an unique challenge - it requires knowledge in both *commonsense* and *health and well-being* domains as well as the ability to reason across them correctly and coherently.

We formally define this as the *cross-domain reasoning* task, as one where the reasoning chain spans across multiple domains. While humans are adept at reasoning across domains, research in cognitive science shows that they often have different processing preferences for individual domains, and it is dependent on domain specific expertise and their reliability of intuition for reasoning across domains (Pachur and Spaar, 2015; Oktar and Lombrozo, 2020). Whether machines can do such cross-domain reasoning is still an open challenge.

[1]All code and data will be released upon acceptance

Input: The first blank is an activity. The second blank is a disease. Person who often does [BLANK] is at a higher risk of [BLANK]

Output: Person who is on blood thinner and eats leafy vegetables is at a higher risk of blood clots

Figure 1: An example of *prompt-template-filling*. We propose an approach for cross-domain reasoning via filling templates guided by prompts. In this example, each prompt signifies a concept from a different domain (**`activity`** from *commonsense* domain and **`disease`** from *health and well-being* domain).

Our goal in this work to is explore whether we can train NLP models that can effectively reason across domains in a given situation. Cross-domain reasoning in NLP literature has been primarily addressed via knowledge bases (KB) (Mendes et al., 2012). Recently, pretrained NLP models have shown immense promise for reasoning applications in several tasks such as commonsense reasoning (Bosselut et al., 2019b; Shwartz et al., 2020b), defeasible reasoning (Madaan et al., 2021), procedural knowledge (Rajagopal et al., 2021) and rule-based reasoning (Clark et al., 2020). Inspired by findings in cognitive science and the current advances in reasoning systems, our work extends this line of investigation to study whether pretrained sequence-to-sequence models (SEQ-TO-SEQ) can be used to reason across knowledge that connects diverse domains.

We model the cross domain reasoning challenge as a prompt-based template filling task (*prompt-template-filling*) where a SEQ-TO-SEQ model is trained to fill a template that connects concepts

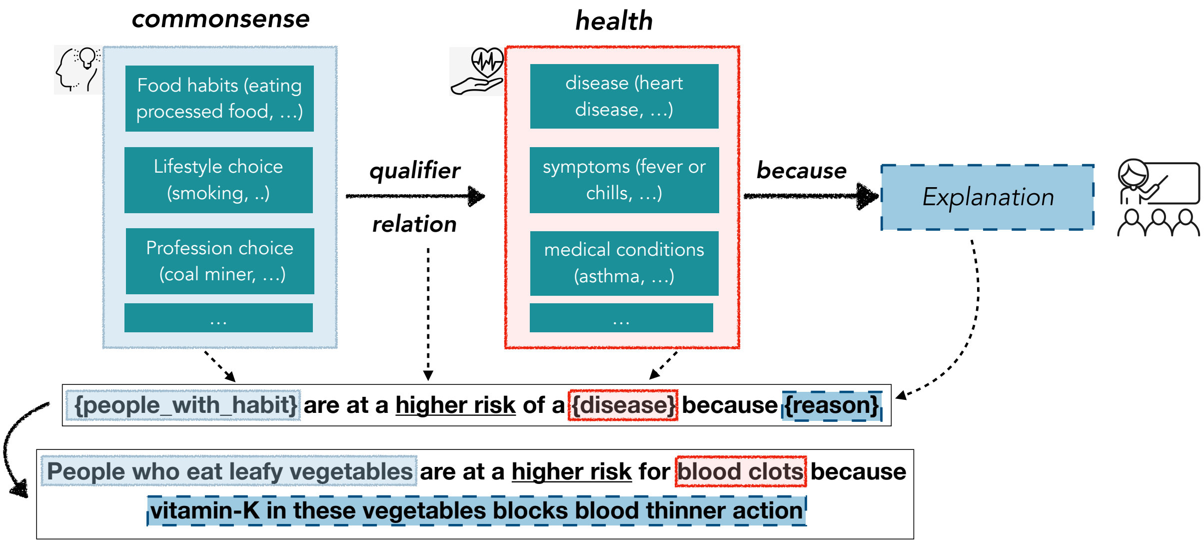

Figure 2: A sample from our cross-domain reasoning task. In this figure, `{people with habit}` is a *commonsense* concept slot, `higher` signifying the qualifier, and `{disease}` represents the *health and well-being* slot and `{reason}` for the explanation in the template. The sentence below is a valid expansion sentence for the template is given in the figure.

across domains. Figure 1 shows an example of our approach. In our use-case, we evaluate whether LMs can effectively reason across *commonsense* domain and *health and well-being* domain. Towards this, our contributions in this paper are twofold. First, we present a dataset of cross-domain cloze style templates and corresponding sentences that are valid completions of the template. The slots in the templates are open-ended and are not restricted to any particular vocabulary. The concept in each slot in the template is provided via a *prompt* indicates a category or an abstraction of a concept from a particular domain. Figure 1 shows an example, where the first prompt indicates a commonsense concept **`activity`** and the second slot indicates a health concept **`disease`**.

Next, our *prompt-template-filling* approach models the cross-domain reasoning challenge as a SEQ-TO-SEQ task, where given a template, the goal of the model is to produce meaningful completed sentences for the template. Our experiments on reasoning across commonsense and health domain shows that SEQ-TO-SEQ models show reasonable ability for cross-domain reasoning. We also present an in-depth error analysis along with our empirical analysis, leaving several open avenues for future research.

To summarize, (i) we present the first prompting based approach to enable SEQ-TO-SEQ perform cross-domain reasoning that uses prompts to specify domain specific concepts to fill templates (*prompt-template-filling*). (ii) For the use-case of reasoning across the *commonsense* and *health and well-being* domain, we present a dataset and a corresponding study on the ability of *prompt-template-filling* to enable SEQ-TO-SEQ to reason cross-domain.

2 Dataset

To investigate whether SEQ-TO-SEQ models are effective at cross domain reasoning, we collect a dataset of templates that are composed of cross-domain reasoning chains and corresponding sentences that match the template. Figure 2 shows an example of a sample from our dataset. Each template in our dataset is composed of the following basic units:

1. *concept slot* : contains an abstract category form of a concept from one of the domains.
2. *qualifier slot* : a word or phrase that describes the nature of the effect of concept of one domain on the other (e.g. higher, lower,...)
3. *explanation slot* : this optional field consists of a free-form explanation that explains the reasoning across the concepts from the different domains.

For our use-case, we use the *commonsense* domain and the *health and well-being* domain. In

Template	Sentences
{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
{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_1} should not be consumed with {food_item_2} because {reason}	Steak should not be consumed with mashed potatoes because - pairing fried foods with starchy carbohydrates 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
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.
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.
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 commonsense knowledge concept, [concept] is a health and well-being concept, and [text] represents the explanation and [text] represents a qualifier. We show two sentences each for a template.

reasoning, it is a long-standing challenge to address *commonsense* reasoning with approaches ranging from building commonsense knowledge bases (Matuszek et al., 2006; Speer and Havasi, 2013) and neural-network based approaches (Sap et al., 2019; Bosselut et al., 2019a). There has also been specialized knowledge resources for reasoning in the *health and well-being* domain (Bodenreider, 2004; Schmidt and Gierl, 2000). Due to their significant impact over the years, we chose these domains to collect corpus for our use-case.

For the use-case to reason across *commonsense* and *health and well-being*, we collect a set of template (x) and its corresponding expansions (y) based on this overall schema of reasoning across *commonsense* and *health and well-being* domain. An example is shown in figure 2. Each template has atleast one *concept* slot, one from each domain (*people eating leafy vegetables* from commonsense domain and *blood clot* from the medical domain in the example shown in the figure). A qualifier slot optionally specifies *how* the concept in a domain interacts with the concept from other domain. In the example in figure 2, *higher risk* indicates the qualifier. The template also includes an optional *explanation* slot that specifies in free-form text how leafy vegetable intake is connected to blood clots.

2.1 Task Setup

To collect our dataset, we use amazon mechanical turk platform [2]. The interface is shown in figure 3. 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 instructed to give a template, and atleast two corresponding sentences that matches the template. The statistics of the data are shown in table 2 and 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.

[2] https://www.mturk.com/
[3] https://www.cdc.gov/
[4] https://www.webmd.com/
[5] https://www.healthline.com/
[6] https://www.mayoclinic.org/

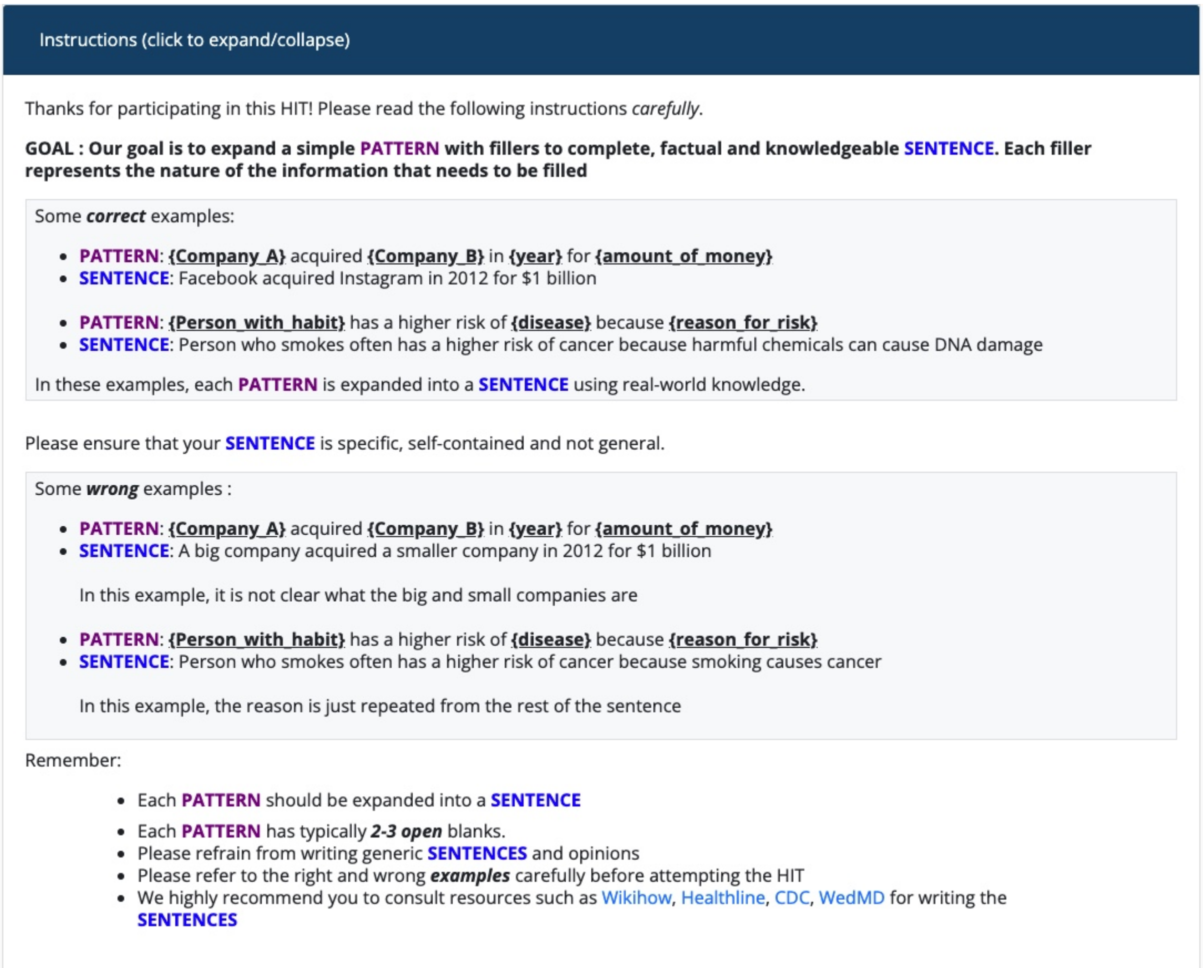

Figure 3: The mechanical turk interface for data collection. The human annotators were given instructors and examples to introduce them to the task.

Category	Statistic
#sent len	14.57
#datapoints	6909
# avg slots per template	2.4

Table 2: Dataset Statistics

Once the templates are collected, we post-process the data to validate that we do not have any identifying information like proper names. We then create a standard 70/10/20 train, validation test split with this dataset.

3 *Prompt Template-Filling Framework*

Early NLP systems have often relied with templated rule-based 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 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 pretrained SEQ-TO-SEQ models for the task of cross-domain reasoning.

In our *prompt-template-filling* formulation, we setup the template filling task as a prompt-tuning task inspired by the recent advances in prompt-tuning. Prompt-based approaches have achieved state-of-the-art performance in several few-shot learning experiments (Brown et al., 2020; Gao et al., 2021; Le Scao and Rush, 2021). Table 3 shows an example of our task setup. The template filling task takes an input template x, containing one or more template slots represented as spans (`[MASK]`) as input, and produce an expanded sequences y as output. Given a template x, the task is to model $p(y|x)$. Since there could be multiple sentences in the output y, we concatenate these

Template	Output
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 stress due to noise

Table 3: Task Setup. Each concept category is given as a prompt to the input and the slots are represented via the [MASK] token. The task for SEQ-TO-SEQ is to generate the *output*

sentences as one for model training.

In comparison to approaches such as Donahue et al. (2020), our approach does not strictly enforce that that sentences only fill missing spans of text. Rather, the expanded sentences can have additional modifications. For instance, for the following input template - {person_at_location} has a {higher/lower} risk of {disease} because {reason_for_risk}, a valid sentence is *person who lives in the city has a higher risk of depression due to noise*. In this example, the word *because* does not match the output sentence phrase "*due to*" but it is considered a valid output for the template.

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 $p_\theta(y \mid x)$ 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.

4 Experiments

In this section, we describe the experimental setup, baselines for our approach. Since our approach is agnostic to the pretrained encoder-decoder architecture type, we perform experiments on several state-of-the-art seq-to-seq models.

4.1 Experimental Setup

Following experimental setup for similar reasoning tasks (Rudinger et al., 2020), we use the ROUGE metric (Lin, 2004) [7] as our automatic metric. To perform the evaluation, we compare the generated sentence for the template against the gold annotations in our dataset. We remove the template words from the output and only compare the slot filler concepts for ROUGE to avoid score inflation due to copying. All the experiments were performed on a cluster of 8 NVIDIA V100 GPUs for a total of 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. The model settings are given below:

- 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 parameters.
- BART-LARGE: Larger version of BART-BASE, consisting of 24 transformer layers, 1024 hidden size, 16 heads and 406M parameters.
- 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. [8].

[7] https://pypi.org/project/rouge-score/

[8] We use the implementation of all the models from the huggingface (Wolf et al., 2020) repository

Model	Template	Output
BERT [MASK]	[MASK] has a [MASK] risk of [MASK] because [MASK]	Person who lives in a city has a higher risk of depression because of stress due to noise
SPL TOKEN	[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: Task Setup for baselines. In the first baseline, we query the BERT MLM model to check if cross-domain knowledge is already present. In our second baseline, we use special tokens to indicate the start and end of each slot. In both the case, the SEQ-TO-SEQ is trained to generate the output.

4.3 Baseline Methods

- BERT [MASK]: To understand whether pre-trained models contain the knowledge already, we try a masked language modeling baseline where we query the template using [MASK] tokens[9].

- 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 of the models for our task with a corresponding example.

4.4 Results

The results across various pretrained encoder-decoder approaches are shown in table 5. 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 pretraining task choices and corresponding datasets. We also observe that PROMPT based models outperform the SPL TOKEN based approach. For all of the models and baselines, we used the greedy decoding strategy.

N-gram metrics such as ROUGE are known to be limited, specifically for reasoning tasks. To assess the quality of generated output, 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 model for this evaluation. In this experiment, a sentence generated by the SEQ-TO-SEQ for a given template was given to a human judge and they were asked to evaluate whether the sentence was correct, given the template. The judges were asked to refer to the same sources as the human annotators to verify the correctness. The inter-annotator agreement on graph 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. Both the automated and human evaluation suggests that there is ample room for further improving cross-domain reasoning ability of SEQ-TO-SEQ models.

5 Error Analysis

In this section, we analyze in detail how well language models perform cross-domain reasoning. Automated metrics such as ROUGE are restrictive in terms of understanding the reasoning abilities and 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 language models generates an answer that relates to **kidney damage**.

[9] Since mask tokens in BERT needs to be predetermined for this experiment, we try different variations with number of [MASK] tokens and report the best results.

Model	Type	ROUGE-1	ROUGE-2	ROUGE-L
BERT-BASE	[MASK]	5.33	0.72	4.94
BERT-LARGE	[MASK]	8.05	0.63	7.85
T5-BASE	SPL TOKEN	14.00	2.71	12.58
T5-BASE	PROMPT	14.01	2.60	12.57
BART-BASE	SPL TOKEN	17.17	5.60	16.32
BART-BASE	PROMPT	18.89	5.87	17.96
BART-LARGE	SPL TOKEN	19.54	7.57	18.49
BART-LARGE	PROMPT	20.58	7.32	19.58

Table 5: Overview of the results compared to baselines. The table shows that BART-BASE performs better than T5-BASE model and BART-LARGE outperforms both.

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_condiiton} are at higher risk of {disease} as they are more exposed to {reason}	Person who often inhales a lot of dirt is at a higher risk of hay fever because of allergen content.	Person who often does less medications is at a higher risk of hay fever because of the drug can help clear it up
Generic Explanation	When people with {certain_co-morbidities} shows {symptoms}, this is because of {reason_for_patient_state}	When people with diabetes shows lethargy, this is because of high glucose levels.	When people with heart disease shows 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

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

While correct, the automated metrics score this answer 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`.

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`.

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`.

Our errors highlight the difficulty of the task for language models. This leaves room for several research questions that requires future work. Overall, cross-domain reasoning is still an uphill task for language models with promising directions.

6 Related Work

Knowledge Bases : Knowledge Bases (KBs) have been the predominant approach to perform cross-domain reasoning in the past. Some of the prominent cross domain knowledge bases include DBPedia (Mendes et al., 2012), YAGO (Suchanek et al., 2007) and NELL (Mitchell et al., 2018). Most of these knowledge bases despite being cross-domain, the focus is primarily on the encyclopedic knowledge. In our work, we focus on ability of SEQ-TO-SEQ for cross-domain reasoning, which can be viewed as a complementary approach to KBs.

Language Models for Knowledge Generation: Using pretrained language models to generate knowledge has been studied for commonsense reasoning tasks. (Sap et al., 2019; Bosselut et al., 2019b; Shwartz et al., 2020a; Bosselut et al., 2021). Our work closely aligns with Bosselut et al. (2019b, 2021). Compared to Bosselut et al. (2019b), our focus in this work to extend this line of work from only commonsense reasoning to perform reasoning cross domain.

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 cross-domain reasoning instead of content planning for stories.

There has also been efforts to transfer knowledge cross-domain via transfer learning (Min et al., 2017; Wiese et al., 2017; Deng et al., 2018) but our work focuses on cross-domain reasoning in the same input sample unlike transfer learning based approaches.

7 Conclusion and Future Work

In this paper, we present a novel *prompt-template-filling* approach that adapts language models to perform cross-domain reasoning via prompting. To study this, we present a dataset via a use-case of reasoning across *commonsense* and *health and well-being* domain. Through both automated and human metrics, we find that there is immense room for progress towards improving language models' capability for cross-domain reasoning. For future work, we want to extend this work for multiple other cross-domain scenarios and understand the nature of cross-domain reasoning in depth.

8 Ethics Statement

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. As our results show, cross-domain 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 health related or 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

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

Olivier Bodenreider. 2004. {The Unified Medical Language System (UMLS): integrating biomedical terminology}. Nucleic Acids Research, 32(suppl_1):D267–D270.

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 Celikyilmaz, and Yejin Choi. 2019a. COMET: Commonsense transformers for automatic knowledge graph construction. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 4762–4779, Florence, Italy. Association for Computational Linguistics.

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

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

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and

Dario Amodei. 2020. Language models are few-shot learners. In Advances in Neural Information Processing Systems, volume 33, pages 1877–1901. Curran Associates, Inc.

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.

Peter Clark, Oyvind Tafjord, and Kyle Richardson. 2020. Transformers as soft reasoners over language. In Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20, pages 3882–3890. International Joint Conferences on Artificial Intelligence Organization. Main track.

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.

Yang Deng, Ying Shen, Min Yang, Yaliang Li, Nan Du, Wei Fan, and Kai Lei. 2018. Knowledge as a bridge: Improving cross-domain answer selection with external knowledge. In Proceedings of the 27th International Conference on Computational Linguistics, pages 3295–3305, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

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

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.

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.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 3816–3830, Online. Association for Computational Linguistics.

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.

Teven Le Scao and Alexander Rush. 2021. How many data points is a prompt worth? In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 2627–2636, Online. Association for Computational Linguistics.

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.

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

Aman Madaan, Dheeraj Rajagopal, Niket Tandon, Yiming Yang, and Eduard Hovy. 2021. Could you give me a hint ? generating inference graphs for defeasible reasoning. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 5138–5147, Online. Association for Computational Linguistics.

Cynthia Matuszek, John Cabral, M. Witbrock, and John DeOliveira. 2006. An introduction to the syntax and content of cyc. In AAAI Spring Symposium: Formalizing and Compiling Background Knowledge and Its Applications to Knowledge Representation and Question Answering.

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).

Sewon Min, Minjoon Seo, and Hannaneh Hajishirzi. 2017. Question answering through transfer learning from large fine-grained supervision data. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 510–517, Vancouver, Canada. Association for Computational Linguistics.

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.

Kerem Oktar and T. Lombrozo. 2020. You should really think this through: Cross-domain variation in preferences for intuition and deliberation. In CogSci.

Thorsten Pachur and Melanie Spaar. 2015. Domain-specific preferences for intuition and deliberation in decision making. Journal of applied research in memory and cognition, 4:303–311.

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.

Dheeraj Rajagopal, Aman Madaan, Niket Tandon, Yiming Yang, Shrimai Prabhumoye, Abhilasha Ravichander, Peter Clark, and Eduard Hovy. 2021. Curie: An iterative querying approach for reasoning about situations.

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.

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

Rachel Rudinger, Vered Shwartz, Jena D. Hwang, Chandra Bhagavatula, Maxwell Forbes, Ronan Le Bras, Noah A. Smith, and Yejin Choi. 2020. Thinking like a skeptic: Defeasible inference in natural language. In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 4661–4675, Online. Association for Computational Linguistics.

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.

Rainer Schmidt and Lothar Gierl. 2000. Case-based reasoning for medical knowledge-based systems. Studies in health technology and informatics, 77:720–5.

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

Vered Shwartz, Peter West, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020b. Unsupervised commonsense question answering with self-talk. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 4615–4629, Online. Association for Computational Linguistics.

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.

Georg Wiese, Dirk Weissenborn, and Mariana Neves. 2017. Neural domain adaptation for biomedical question answering. In Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017), pages 281–289, Vancouver, Canada. Association for Computational Linguistics.

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.

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