Investigating a Framework for Improving Knowledge-Aware Commonsense Reasoning with Human Intelligence

Richard Magnotti
University of Rochester
rmagnott@ur.rochester.edu

Zheng Zhang
University of Rochester
zzhang95@cs.rochester.edu

Abstract

Recent state-of-the-art commonsense reasoners are able to generate inferences with *sub-graph evidence paths* or paths from external structured commonsense knowledge graphs. However, such evidence tends to also posses a large amount of noisy knowledge which is detrimental to model prediction capability. In this project update, we aim to utilize human intelligence to refine the evidence that the commonsense reasoner relies on. We propose a *human-in-the-loop* framework which aims to improve the model's performance by eliciting human feedback on machine-generated evidence paths. Additionally, we conduct a case study to demonstrate how our framework could be applied to a specific commonsense reasoner.

1 Introduction

Commonsense reasoning is the ability to make "sound yet un-reflective" presumptions about how the world operates and how humans behave in their daily lives (Elio et al., 2002). Although such reasoning seems innate to humans, automating commonsense reasoning has been recognized as a vital bottleneck of artificial intelligence (Davis and Marcus, 2015). To encourage advances in this field, some commonsense question-answering tasks have been devised to test the different aspects of a model's ability to perform commonsense reasoning (Zellers et al., 2018; Sap et al., 2019b; Rashkin et al., 2018). In general, successfully addressing these tasks would require background knowledge that is not explicitly expressed in the question (Lv et al., 2020).

To do this, many recent state-of-the-art commonsense reasoning systems suggest to incorporate background knowledge from external commonsense knowledge graphs (CSKG) as relational inductive biases (Lin et al., 2019; Paul and Frank, 2019; Wang et al., 2020). In other words, they extract a topically related sub-graph from CSKG and utilize self-attention mechanisms to selectively aggregate the most important path vectors. The aggregated vector is presumed to reflect the holistic semantic relations between question context and candidate answers. Additionally, the attention weights are able to interpret the inference process by showing how much each knowledge path is taken into account as the reasoner makes a prediction. Consequently, most of these works claim that incorporating external knowledge could not only boost the model's reasoning capacity, but also increase the transparency of model behavior.

However, one apparent limitation of these knowledge incorporation methods is that those extracted sub-graphs and paths tend to contain numerous concepts that are irrelevant to the question at hand (Wang et al., 2020). Although a prior case study found that self-attention mechanisms will diminish noisy concepts while modeling (Lin et al., 2019), constructing noise-free external evidence will be undoubtedly beneficial to the model's reasoning.

In this work, we propose a human-in-the-loop framework, which is designed to improve the commonsense reasoning model by asking human participants to refine the evidence paths that the model relies on. In addition, we conduct a case study to demonstrate potential applications of our framework, through which we further identify cognitive obstacles faced by humans that future work should address.

2 Related Work

Recently, many existing neural-based commonsense reasoning systems (Lin et al., 2019; Paul and Frank, 2019; Wang et al., 2020; Lv et al., 2020) exploit external structured knowledge graphs such as ConceptNet (Speer et al., 2016) and ATOMIC (Sap

et al., 2019a) to "connect the dots" between question context and answers. Generally, they curate a local graph based on the concept pairs of questions and answers, with which the models could further produce multi-hop knowledge paths via either static path-finding algorithms (Paul and Frank, 2019) or dynamic path generators (Wang et al., 2020). Next, a self-attention layer is used to decide the importance of each path and aggregate the path vectors based on their attention weights.

One obvious drawback of directly using the local graph of a CSKG is that normally only a few facts contained in it are relevant to the question context. Moreover, the remaining knowledge is generally noisy and is thus unhelpful at best or at worst negatively impacts the model's learning ability. To address this issue, some works propose to leverage a pre-trained language model (PTLM) as a dynamic path generator (Wang et al., 2020), under the presumption that PTLM could supplement the raw knowledge paths with unstructured background knowledge stored in it. However, since the previous works did not supervise the path generator with context information, the whole path is still likely to deviate from the question context.

In contrast, humans engage in commonsense reasoning throughout daily life (Elio et al., 2002). Therefore it is reasonable to presume that humans could modify the evidence paths and rework them into more reasonable constructions based on their background knowledge and understanding of story.

3 Framework Design

In this section, we propose a generalizable framework that allows the incorporation of human feedback during the inference process of our knowledge-aware commonsense reasoning model. We work under the assumption that the human-improved evidence paths have the potential to enhance model performance in real time.

Our human-in-the-loop framework consists of four steps:

1. The commonsense reasoning model makes a prediction on the instance of interest. Then the model returns the top-N evidence paths as ranked by attention weights, as well as predicted answers back to the human participants.

2. Users then evaluate the plausibility of each path. If the prediction is not plausible, they are asked to modify each questionable evidence path by performing addition, revision, and deletion operations.

3. The model then updates its prediction with human-modified paths, and reports the new predictions and top-N paths to user.

4. The user repeats steps 2 and 3 until the prediction sounds correct to him/her.

4 Case Study I: Generalizable Framework, 'Human-In-The-Loop'

In this pilot study, we aim to demonstrate how to apply our proposed framework to a specific commonsense reasoning model that utilizes external evidence paths.

4.1 Study Setup

We base our study on the human needs prediction task proposed by Rashkin et al (Rashkin et al., 2018), which requires the reasoner to infer the underlying human needs of the character in the story through commonsense reasoning. Prior works indicate that it is difficult for machine learning models to predict and explain implicit human motivation (Phan et al., 2016). Therefore, we expect subjects to assist the machine to construct reasoning paths that otherwise a model could not generate on its own.

We adopt the reasoning model developed by Paul et al. (Paul and Frank, 2019) which integrates multi-hop relation paths from ConceptNet to enhance its prediction on human needs. Like most knowledge-aware reasoning models, the attention score reveals how the paths are ranked by importance during the model's inference.

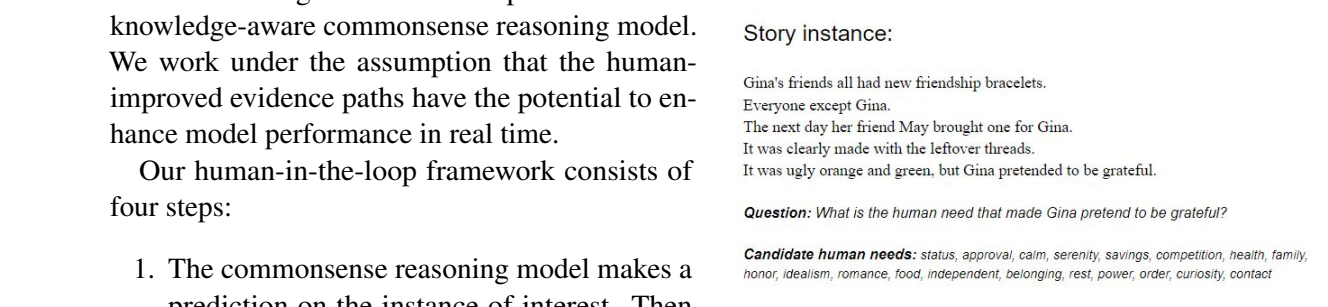

Figure 1: The instance of human needs prediction task that is used in our first case study

4.2 Study Design

We conducted a case study within our research group to establish the degree to which human and

Story:
Gina's friends all had new friendship bracelets.
Everyone except Gina.
The next day her friend May brought one for Gina.
It was clearly made with the leftover threads.
It was ugly orange and green, but Gina pretended to be grateful.

ConceptNet path: leftover related to reserve related to ration related to food

Modified path: leftover related to reserve related to ration related to ~~food~~ the last related to compromise related to status

Figure 2: An example of human modification

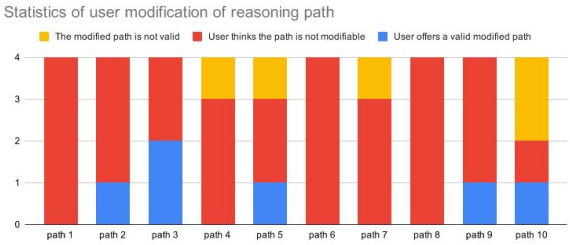

Figure 3: Statistics of human modification on each path

machine responses agree. To do this, first one instance (see in Figure 1) was selected from the dataset of human need tasks, and we ran Paul's model to make the predictions. After which, we presented group members with the top 10 ConceptNet evidence paths as well as predicted answers.

For each path, every group member needed to first decide whether or not it was rational by their own standards. There were no metric guidelines provided in order to avoid influencing the users' answers. If the user deemed the path irrational, they were further asked to modify the ConceptNet path by: (1) deleting the part that doesn't make sense to them; (2) adding their new content in ConcpetNet format so that the entire path possesses reasonable multi-hop connections between context keyword and human need responses. If the members deemed the path so unreasonable that it is not modifiable, they were encouraged to give a brief explanation for their judgement.

We collected 6 valid human modified paths[1] in total out of all participants. One participant modification example is shown in Figure 2. Note that human members tend to think of a majority of paths as not modifiable (as shown in Figure 3). Through examining their explanations, we found that modification was most likely precluded by the fact that most if not all the concepts contained within the path were irrelevant to the story context.

[1]A valid path means that the path should start with a concept existing in story context and end with one of the human needs

4.3 Updating Model Predictions with Human-Modified Paths

Due to the relatively small pool of valid human-modified paths, instead of separately updating the model prediction with the paths produced by different participants, we utilized all of the paths simultaneously. We compared performance (via negative cross entropy loss) of the model enriched with between 1 and 6 of the human-modified paths, against 6 ConceptNet evidence paths generated by the un-enriched model [2]. We found that the model performance was improved marginally with human-modified paths. Specifically, we found that the model's cost decreases as we incorporate more human-modified paths into it. The Spearman correlation coefficient r between the model's negative cross entropy and number of human paths used for retraining is 0.893 ($p = 0.0068$). This demonstrates that human-modified paths have the potential to improve the performance of the commonsense reasoner.

5 Future work

Through these case studies, we found that most of the reasoning paths generated using ConceptNet are not directly modifiable due to their poor quality from a human perspective evaluated empirically. However, it is likely that with proper facilitation and response re-implementation, human-modified paths have the potential to improve model commonsense inferences. Therefore, it is necessary to extend our framework to be able to handle these undesirable cases. One possible strategy is to construct a scaffolding to allow users to create new sensible evidence paths over the knowledge graph.

6 Conclusion

In this paper, we propose a human-in-the-loop framework that aims to improve a machine learning commonsense reasoning system with human feedback using external evidence paths. We also provide a case study to illustrate how to apply our framework to a specific knowledge-aware commonsense reasoning model and the potential of human-cooperation to improve model reasoning ability. Additionally, we identify one of the cognitive obstacles that human participants are confronted with

[2]To obtain a meaningful performance comparison of the human-enriched versus raw AI evidence paths, we generate evidence paths using both models for the same sample task 100 times and average the performance.

as modifying the raw evidence paths. Future work is needed to investigate how to enable humans to create sensible external evidence paths with machine assistance.

References

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

Renée Elio et al. 2002. *Common sense, reasoning, & rationality*. Oxford University Press on Demand.

Bill Yuchen Lin, Xinyue Chen, Jamin Chen, and Xiang Ren. 2019. Kagnet: Knowledge-aware graph networks for commonsense reasoning. *arXiv preprint arXiv:1909.02151*.

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*, pages 8449–8456.

Debjit Paul and Anette Frank. 2019. Ranking and selecting multi-hop knowledge paths to better predict human needs. *arXiv preprint arXiv:1904.00676*.

Nhathai Phan, Dejing Dou, Brigitte Piniewski, and David Kil. 2016. A deep learning approach for human behavior prediction with explanations in health social networks: social restricted boltzmann machine (srbm+). *Social network analysis and mining*, 6(1):79.

Hannah Rashkin, Antoine Bosselut, Maarten Sap, Kevin Knight, and Yejin Choi. 2018. Modeling naive psychology of characters in simple commonsense stories. *arXiv preprint arXiv:1805.06533*.

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

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. 2019b. Socialiqa: Commonsense reasoning about social interactions. *arXiv preprint arXiv:1904.09728*.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2016. Conceptnet 5.5: An open multilingual graph of general knowledge. *arXiv preprint arXiv:1612.03975*.

Peifeng Wang, Nanyun Peng, Pedro Szekely, and Xiang Ren. 2020. Connecting the dots: A knowledgeable path generator for commonsense question answering. *arXiv preprint arXiv:2005.00691*.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. Swag: A large-scale adversarial dataset for grounded commonsense inference. *arXiv preprint arXiv:1808.05326*.