StoryChatBuddy: An Interactive Story Understating System with Automated Commonsense Reasoning

ANONYMOUS AUTHOR(S)*

The recent breakthroughs of commonsense reasoning in Artificial Intelligence tremendously improve computer understanding of human thinking and behavior. Although there have been many efforts made by HCI researchers on propelling this progress, few works explore how to integrate the advanced commonsense reasoning AI into end user application. In this work, we present a novel interactive story understanding system called "StoryChatBuddy", which allows users to develop a proper story understanding in collaboration with a state-of-the-art commonsense reasoning model. Besides, we provide an user case study to illustrate the user participation in the collaborative story understanding process. We hope this practice can inform the future work aiming to enrich their application with commonsense reasoning AI. Plus, this system also lay a foundation for us to further investigate of how human participation can facilitate the online learning process for AI reasoning.

CCS Concepts: • **Computer systems organization** → **Embedded systems**; *Redundancy*; Robotics; • **Networks** → Network reliability.

Additional Key Words and Phrases: datasets, neural networks, gaze detection, text tagging

ACM Reference Format:
Anonymous Author(s). 2018. StoryChatBuddy: An Interactive Story Understating System with Automated Commonsense Reasoning. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Recent breakthroughs in commonsense reasoning within Artificial Intelligence (AI) offers new opportunities to improve computer understanding of human thinking and behaviors. For example, given the story "*Tina's boyfriend cooked her spaghetti, but she disliked the taste. Tina pretended it tasted good.*", and the question *why did Tina pretend the food was tasty?*, a state-of-the-art commonsense reasoning model could make an appropriate reasoning: "*Tina pretends the food is tasty is caused by Tina wants to spare her boyfriend's feelings*" [9]. HCI researchers have impelled this advance largely by designing effective crowdsourcing methods for commonsense knowledge (reasoning) [14, 17], reducing the cost of data collection and diversifying the collected knowledge scope. Despite the benefits that HCI has brought to AI commonsense reasoning, there is little research exploring how the state-of-the-art AI commonsense reasoning can enrich the user experiences in a wide range of application domains (e.g. early education, social-aware personal assistants). We believe integrating a commonsense reasoning model into HCI applications could bring unique benefits: first, it would make interactive applications more socially aware, developing a deep understanding of human behavior and intention; second, the exposure of model's thinking process can increase the transparency of the application and cultivate human trust.

Manuscript submitted to ACM

One particular domain that may benefit from AI commonsense reasoning is education technologies for socio-emotional skill development. When children listen to stories, they are naturally exposed to rich social situations. This experience will help them gradually obtain socio-emotional skills that allow them to identify a character's behavior as being linked to its mental states such as belief, goals and emotion, especially when children engage in meaningful discussion with others such as parents and peers. Inspired by early childhood story reading, we developed an initial prototype of an interactive commonsense-aware story understanding system called StoryChatBuddy. It utilizes the state-of-the-art AI commonsense reasoning engine [9] to explain underlying human needs involved in simple story plots, and the commonsense reasoning process behind through dialog-like interaction with lay users without a background in AI. The initial prototype serves two main purposes: first, it illustrates a meaningful use case of integrating AI commonsense reasoning with HCI applications. Second, acknowledging the space for improvement of human-level rationality with state-of-the-art AI commonsense reasoning, the prototype will lay the foundation for us to investigate how human participation can facilitate the online learning process for AI reasoning.

In this paper, we first introduce the design and development of StoryChatBuddy. We then provide a use case of the interactive story understanding experience supported by StoryChatBuddy. Finally, we discuss our future work.

2 DESIGN AND IMPLEMENTATION OF STORYCHATBUDDY

StoryChatBuddy aims to combine an intelligent agent and a human user's reasoning ability by establishing a human-AI team that collaboratively develops context-appropriate social commonsense reasoning within the setting of story understanding. StoryChatBuddy is designed as a web-based intelligent agent that: (1) reads a simple story with a human user; (2) answers users' questions to explain human needs involved in the narrative and what aspects of the commonsense knowledge are used for in the inference; (3) engage the user in a discussion about the social commonsense reasoning. During the discussion, StoryChatBuddy will assist users in assessing therationality of the agent's reasoning and collect human feedback to be used for improving to the agent. This collaborative process ends when users are satisfied with the modified reasoning. The remainder of this section will introduce the story source, reasoning engine, and interface design in detail.

2.1 Story source

We chose RocStories dataset [8] as our story source. The RocStories dataset is a collection of five-sentence stories that captures a rich set of stereotypical causal and temporal commonsense relations between daily activities. To fully understand a story's narrative, readers must identify relevant commonsense knowledge from the narrative and develop appropriate reasoning using it. In particular, we utilize a modified version of the RocStories dataset [13] that annotates the motivation (i.e., human need) and emotion of stories' characters on a sentence-level. With this dataset, we ask both parties to answer questions in the form of: "*What underlying human need motivates XXX to do XXX?*", the answer should be one of the human needs categories.

2.2 AI Commonsense Reasoning Engine

We use a state-of-the-art commonsense reasoning model called GLUCOSE [9] as the reasoning backbone of StoryChatBuddy. GLUCOSE fine-tuned the GPT-2 [11] with the collected commonsense knowledge that RocStories contains. Given an event about story, this model is able to make ten types of commonsense inference, such as "*event directly causes X*", "*emotion or human drives that motivate X*", etc. Specifically, we use the one-sided generation version of GLUCOSE, which is able to generate conclusion based on premise (story event) and inference type. For example, given the event

"*Karen made a pan of lasagna*" and inference type "*An event that X Causes/Enables*", the model can output "*Karen made lasagna Causes/Enables Karen ate lasagna*" (Figure 1). By default, GLUCOSE can only provide single-step inferences. To generate more sophisticated commonsense reasoning, StoryChatBuddy extends the default single-step inference into multi-step commonsense reasoning. To achieve so, for each sample story, we heuristically provide the model with inference type at each step, and treat the output of model at each step as the new input event in the next inference. This iterative process will end until one of the target human needs is reached.

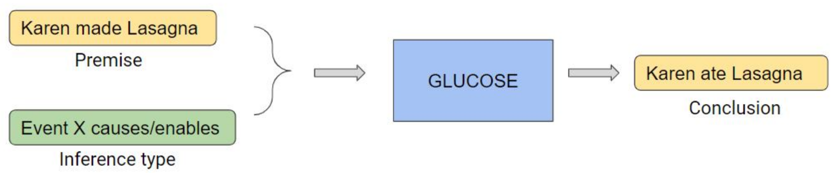

Fig. 1. An example of GLUCOSE inference, given event "Karen made Lasagna" and inference type "Event X causes/enables", GLUCOSE model concludes "Karen ate Lasagna"

2.3 Interface and Interaction Design

At the beginning, users are asked to choose a story from the story set. Once the user chooses a specific story, they are told to read this story together with the intelligent agent and think of the question about character's motivation. After finishing the reading, they could click the button "*Please show me AI's reasoning*". The system will present the reasoning engine's inferred answer as well as a corresponding commonsense reasoning path (Figure 2). This multi-step reasoning was presented in a semi-structured format consisting of a series of card elements, each representing a single reasoning step. The rationale of using argument card representation is that since each reasoning step can be seen as an argument, the argument structure can be explicitly visualized in this format; Plus, user can easily click the problematic components to fix. Each card element contains three fields of natural language sentence displaying the premise, conclusion and type of the reasoning step. The conclusion of the last reasoning step was the motivation of the story character that AI derived. Each subsequent reasoning step used the premise of the previous one as its conclusion.

With the presented reasoning, users were asked to fix the inappropriate steps that AI made. Users began the fault identification by clicking the checkbox of problematic cards (or card components). Once users made a selection, they would be asked to assess the quality of selected part from multiple dimensions, each dimension corresponded to a 5-Likert scale (Figure 3). These dimensions were designed to scaffold users' think process by reminding them of aspects of sound and convincing reasoning. An explanation of the dimension could be viewed, as a modal, by clicking the dimension label.

Based on users' evaluation, the system would show users multiple options of resolving the quality issue of the selected element. Three main options are "*request AI to make alternative reasoning*", "*request AI to make a different type of inference*" and "*direct modification*". The first option was for the situation when both the premise and inference type made sense but the conclusion was irrational. It would evoke the GLUCOSE model to make other attempts with the same premise and inference type, users could then decide if any of alternative reasoning looked reasonable to them. The second option could be selected if users thought the premise made sense but inference type was improper at this step. The GLUCOSE model would generate a reasoning with the old premise and new inference type, user could decide

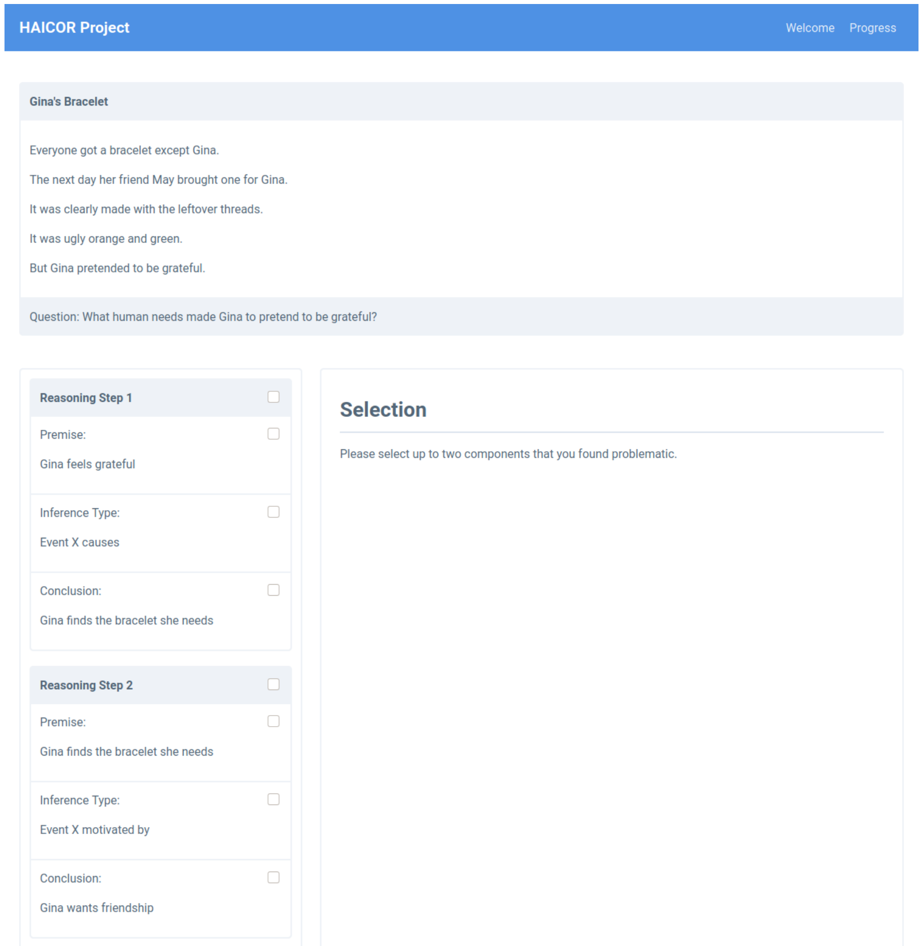

Fig. 2. Display of AI's commonsense reasoning. User is allowed to select problematic reasoning steps (components) to fix.

whether to accept this new inference. Lastly, if users found the first two options could not return a satisfying reasoning, they could directly modify the reasoning step into a rational one.

When user finished the fault resolution for a reasoning step, they would be asked to collaborate with AI to produce the intermediate reasoning from this step to a subsequent step of AI's reasoning that users thought reasonable (as shown in user case study). If no such step existed, users needed to demonstrate how the correct answer could be inferred from the resolved step.

The whole interaction process ended when there was no reasoning step labelled as problematic anymore. Users would be asked to submit the final reasoning after confirmation.

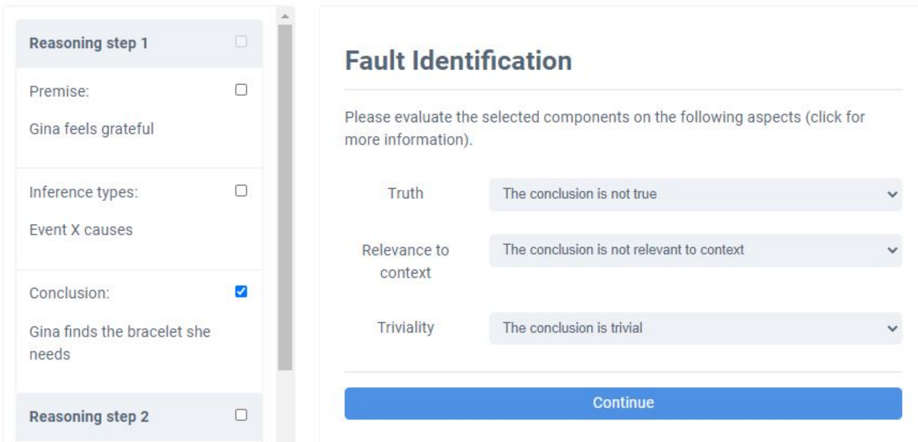

Fig. 3. User interface of evaluating issues of problematic step from multiple dimensions

3 USER CASE STUDY

We provide a user case of demonstrating how users can collaborate with AI on developing appropriate story understanding with our system. At first, the user chose a story titled "Gina's bracelet": "*Everyone got a bracelet 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.*". Both user and AI were asked to answer the question: "*What human needs does drive Gina to pretend to be grateful?*". Given this question, intelligent agent presented his commonsense reasoning: "*Gina pretends to be grateful enables Gina finds the bracelet she needs. Gina finds the bracelet she needs is motivated by Gina wants friendship. Gina wants friendship because Gina wants social contact (social contact is one of the candidate answers)*". The AI's reasoning will be represented as a sequence of argument cards as shown in Figure 2.

User perceived the first step of the reasoning didn't sound reasonable, he clicked the checkbox of the corresponding card and selected a scale for each dimension. Afterwards, user was prompted to do fault resolution. The user decided to choose the second option "*request AI to make a different type of inference*" and provided AI with a new inference type "*Event X is caused by*". Agents returned three candidate inferences: "*Gina pretend to be grateful is caused by: (1) May showed her how to make a Jim acorn pie; (2) May brought Gina the sweater; (3) May makes a shirt with Gina's hair*". User found none of them makes sense, therefore, he chose to directly provide a right inference by altering the conclusion to "*May bought Gina a bracelet*". After that, system asked user "*Does any of the remaining reasoning steps look reasonable and should be kept?*". User labeled the step "*Gina wants friendship because Gina wants social contact.*" as reasonable. Then AI system asked user "*Could you guide me to reach 'Gina wants friendship' from 'May bought Gina a bracelet'?*". The user prompted AI with the inference type of next step, and asked: "*What does cause 'May bought Gina a bracelet'?*". AI returned three candidate answers: "*May bought Gina a bracelet is caused by: (1) May buy out Gina a bracelet; (2) May liked Gina; (2) May wants to buy Gina a gift*". User thought the second answer made sense, thus he kept asking: "*What does May liked Gina cause?*". AI answered: "*May liked Gina causes: (1) Gina want a new bracelet for Christmas; (2) Gina wants to steal back Gina's bracelet; (3) Gina wants the bracelet as badly as May does.*". Since user found none of them sound reasonable, he provided a proper inference: "*May liked Gina causes Gina wants to keep friendship with May*", which reached the premise of the desriable reasoning step "*Gina wants friendship because Gina wants social contact*".

Therefore, the resulting reasoning after collaboration was "*Gina pretends to be grateful is caused by May bought Gina a bracelet. May bought Gina a bracelet is caused by May liked Gina. May liked Gina causes Gina wants to keep friendship with May. Gina wants to keep friendship with May is caused by Gina wants social contact*".

4 FUTURE WORK

Although our current interactive story understanding system is designed to accommodate all categories of user, our chief long-term trajectory is to focus on children with learning disabilities - namely autism. The motivating idea is to allow a child user to gradually improve their understanding of proper social conduct as interacting with the intelligent agent. Because motivation to complete activities may vary between adults and children, we propose a system reconstruction called gamifying. The process of gamification facilitates user-agent interaction in the form of a video game or digital puzzle to implicitly clue users to solutions. One of our design concepts is an exploratory text-based adventure game which allows users to navigate simplistic social interactions. Users will be guided through hypothetical scenarios by an intelligent interface towards appropriate communicative behaviors requiring social commonsense. It will encourage users to choose their avatar's behavior, and the interactive agent(s) will respond according to its level of social acceptability. With the long-term plan of tailoring to those with learning disabilities, our future user study will be conducted with children who fit that criteria between the ages of 9 to 14. Initially, we plan to use our current system to test on adults and gauge its efficacy, then pivot our efforts. Part of this process involves the interdisciplinary study of human behavior analysis, child developmental disorders, and child education.

Besides, we also plan to design a domain-independent interaction framework to facilitate lay users to understand, analyze, and correct the underlying flawed commonsense reasoning of AI-infused applications, which could guide the design of the advanced story understanding system described above. We expect that as users contribute their alterations to the model-generated inferences, the system could collect the user-improved reasoning as a dataset. This dataset can be used to improve the capacity of the commonsense reasoning model.

We aim to submit these planned works to UIST 2021.

5 RELATED WORK

5.1 Commonsense knowledge and reasoning in AI

Commonsense knowledge is considered the 'dark matter' of the field of AI inasmuch as it is vitally important to a variety of AI tasks, but challenging to track due to its subtlety and volume [3]. Additionally, relative to symbolic logical representations commonsense knowledge is difficult to manipulate due to its broad coverage and contextually implicit nature. To overcome these fundamental barriers, recent advances in Machine Learning (ML) allow improved inference capabilities by acquiring commonsense knowledge from structured knowledge bases or unstructured textual corpora [1, 6, 10, 18]. Despite the increased predictive accuracy on many commonsense-intensive benchmarks, these models tend to improperly justify their answers with irrational commonsense reasoning, exposing flawed cognitive processes and consequently reducing human trust in model reasoning capacity [12, 15]. There are several possible explanations which could account for this. First, these models struggle with contextual relevance and semantic ambiguity of knowledge [9]. They tend to wrongly draw conclusion depending on improper knowledge of contextual concepts. Second, due to the incomplete nature of existing knowledge bases, the required commonsense knowledge is likely beyond the scope of knowledge sources that AI uses [18].

There have been great efforts made on conferring capacity of generating rational reasoning to AI. The existing methods can be roughly grouped into unsupervised and supervised. At the one end of the spectrum, unsupervised approaches intend to learn an explanation generator that is able to synthesize the existing structured and unstructured evidence given reasoning context, which can alleviate the issues of sparsity and contextualization that a sole knowledge source usually face [15, 18]. The generated explanation will then be used to support model's prediction and serve as explanations for its decision. In contrast, supervised methods leverage the power of crowdsourcing, where researchers collect a dataset of human explanation and learn an explanation module to automatically generate human-like explanation for unseen situations [2, 12]. Prior works find that although the generated explanations can boost the accuracy of prediction module, there is still a large room for improving the rationality of these explanations on the testing dataset [12]. Besides, human workers have inclination to provide short and unsophisticated explanation for AI to learn, which often neglect the important commonsense knowledge and implicit reasoning flows that AI may need to know.

In our future work, we want to investigate how to enable users to improve the AI's flawed commonsense reasoning through their everyday use. We expect this approach could bring unique advantages over traditional methods for the following reasons: first, this approach allows users to provide rich feedback to agents, allowing influence over its reasoning process and final prediction. For instance, users can also highlight fallacies, missing knowledge, and the modification process. Diverse methods of feedback could possibly be used to enrich AI's knowledge bases, enable online learning approaches (e.g. imitation learning), and enhance the supervision signals of traditional learning methods. Furthermore, collaboration encourages users to produce more thorough and less biased reasoning compared to reasoning in isolation. Lastly, collecting reasoning feedback through daily use could reduce the cost of data collection and accelerate the update circle of AI models.

5.2 Human participation in fixing AI's task failure

Prior work has been conducted to enable end users to fix task failures of agents caused by inappropriate understanding or lack of knowledge. For example, SUGILITE [4] allows users to demonstrate the procedure of performing everyday tasks through a multi-modal interface when an intelligent agent fail to execute human command correctly; SOVITE [5] repairs a task-oriented agent's conversational breakdown by asking users to inspect its understanding of their intent and demonstrate the proper intent through natural language instruction and UI manipulation. In spite of managing to instruct intelligent agents to perform tasks properly, most existing approaches merely focus on supporting human inspection and instruction on AI's procedural knowledge rather than its cognitive reasoning process. To correctly understand human needs, intent or emotion, intelligent agents must perform context-specific commonsense reasoning [7, 16]. This knowledge gap also motivates us to investigate method of facilitating end users to improve flawed commonsense reasoning of AI system.

6 CONCLUSION

In this paper, we presented an novel interactive story understanding system based on a state-of-the-art commonsense reasoning model. We provided an user case to illustrate how users can collaborate with AI to develop a proper story understanding using our system. We wish this attempt can promote HCI communities to integrate the advanced commonsense reasoning AI into their interactive applications, which can promisingly enrich user experience and increase the system's transparency. In the future work, we will customize this system for accommodating child users with autism in order to help their social-emotional development. Besides, we will design an interaction framework to allow users improve AI's flawed commonsense reasoning through daily use.

REFERENCES

[1] Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: Commonsense Transformers for Automatic Knowledge Graph Construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 4762–4779.

[2] Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. In *Advances in Neural Information Processing Systems*. 9539–9549.

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

[4] Toby Jia-Jun Li, Amos Azaria, and Brad A Myers. 2017. SUGILITE: creating multimodal smartphone automation by demonstration. In *Proceedings of the 2017 CHI conference on human factors in computing systems*. 6038–6049.

[5] Toby Jia-Jun Li, Jingya Chen, Haijun Xia, Tom M Mitchell, and Brad A Myers. 2020. Multi-Modal Repairs of Conversational Breakdowns in Task-Oriented Dialogs. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*. 1094–1107.

[6] Bill Yuchen Lin, Xinyue Chen, Jamin Chen, and Xiang Ren. 2019. KagNet: Knowledge-Aware Graph Networks for Commonsense Reasoning.. In *Proceedings of EMNLP-IJCNLP*.

[7] John McCarthy et al. 1960. *Programs with common sense*. RLE and MIT computation center.

[8] Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 839–849.

[9] Nasrin Mostafazadeh, Aditya Kalyanpur, Lori Moon, David Buchanan, Lauren Berkowitz, Or Biran, and Jennifer Chu-Carroll. 2020. GLUCOSE: GeneraLized and COntextualized Story Explanations. In *The Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

[10] Debjit Paul and Anette Frank. 2019. Ranking and Selecting Multi-Hop Knowledge Paths to Better Predict Human Needs. In *Proceedings of NAACL-HLT*. 3671–3681.

[11] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. [n.d.]. Language models are unsupervised multitask learners. ([n. d.]).

[12] Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain Yourself! Leveraging Language Models for Commonsense Reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 4932–4942.

[13] Hannah Rashkin, Antoine Bosselut, Maarten Sap, Kevin Knight, and Yejin Choi. 2018. Modeling Naive Psychology of Characters in Simple Commonsense Stories. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2289–2299.

[14] Christos T Rodosthenous and Loizos Michael. 2014. Gathering background knowledge for story understanding through crowdsourcing. In *2014 Workshop on Computational Models of Narrative*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

[15] Vered Shwartz, Peter West, Ronan Le Bras, Chandra Bhagavatula, , and Yejin Choi. 2020. Unsupervised Commonsense Question Answering with Self-Talk. In *EMNLP*.

[16] Shane Storks, Qiaozi Gao, and Joyce Y Chai. 2019. Commonsense reasoning for natural language understanding: A survey of benchmarks, resources, and approaches. *arXiv preprint arXiv:1904.01172* (2019), 1–60.

[17] Luis Von Ahn, Mihir Kedia, and Manuel Blum. 2006. Verbosity: a game for collecting common-sense facts. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*. 75–78.

[18] 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* (2020).