Deriving Commonsense Reasoning Evaluation from Interactive Fiction Games

Mo Yu[*1], **Xiaoxiao Guo**[*1], **Yufei Feng**[*2], **Xiaodan Zhu**[2], **Michael Greenspan**[2], **Murray Campbell**[1]
[1] IBM Research [2] Queens University
yum@us.ibm.com xiaoxiao.guo@ibm.com feng.yufei@queensu.ca

Abstract

Commonsense reasoning simulates the human ability to make presumptions about our physical world, and it is an indispensable cornerstone in building general AI systems. We propose a new commonsense reasoning dataset based on human's interactive fiction game playings as human players demonstrate plentiful and diverse commonsense reasoning. The new dataset mitigates several limitations of the prior art. Experiments show that our task is solvable to human experts with sufficient commonsense knowledge but poses challenges to existing machine reading models, with a big performance gap of more than 30%. Our code and data will be released at https://github.com/Gorov/zucc.

1 Introduction

When playing an Interactive Fiction (IF) game, we explore and progress through a fantasy world by observing textual descriptions and sending text commands to control the protagonist. While in pure texts, we relate the implicit knowledge of these fantasy worlds with those in our physical world. For example, we explore unvisited regions by planning over the mentioned locations (spatial relations); we eat apples to recover health and attack the enemies with swords, but not vice versa (physical interaction relations); we retrospect the poor choice of breaking the lantern when we find the protagonist in a dangerous dark wood (cause and effects). Plentiful and diverse commonsense knowledge from our physical world is encoded in our game playing texts, which inspire this work of utilizing the IF games to build a new commonsense reasoning dataset.

There has been a flurry of recent datasets and benchmarks on commonsense reasoning [12, 23, 20, 14, 11, 16, 4, 10, 5, 17, 22]. All these existing benchmarks adopt a multi-choice form task. With the input query and an optional short paragraph of the background description, each candidate forms a statement. The statement that is consistent with a commonsense knowledge fact corresponds to the correct answer. We notice some common deficiencies in the construction of these benchmarks. First, nearly all these benchmarks focus on one specific facet and ask human annotators to write candidates related to the specific type of commonsense. As a result, the distribution of these datasets is not natural but biased to a specific facet. For example, most benchmarks focus on collocation, association or other relations (e.g., ConceptNet [19] relations) between words or concepts [12, 20, 14, 11]. Other examples include temporal commonsense [23], physical interactions between action and objects [5], emotions and behaviors of people under the given situation [17], and cause-effects between events and states [16, 4, 10]. Second, the task form makes them more likely commonsense validation, i.e., validation between a commonsense fact and a text statement, but neglecting hops among multiple facts.[1] The limitations above of previous works, namely limitations in *distributions of required*

[*]Equal contribution from the corresponding authors.

[1]These datasets do contain a portion of instances that require explicit reasoning capacity, especially [4, 10, 5, 17]. Still, many of the instances can be solved with standalone facts.

Wordplay: When Language Meets Games Workshop @ NeurIPS 2020,
https://wordplay-workshop.github.io/

commonsense knowledge types and *forms of tasks*, restricted their potentials. These tasks are naturally easy to be handled with pre-trained Language Models (LMs) such as BERT [6]. It is mainly because (1) the narrow reasoning types are easier to be fit by a powerful LM; (2) the dominating portion of commonsense validation instances are easier to be captured by pre-training if texts on these facts have presented in pre-training. Additionally, the above limitations naturally lead to discrepancies between practical NLP tasks that require broad reasoning ability on various facets.

Our Contributions To overcome these shortcomings, we derive *commonsense reasoning tasks from the model-based reinforcement learning challenge of text games*. Our work is inspired by recent advances in interactive fiction (IF) game playing [9, 2, 7]. Figure 1 illustrates sample gameplay of the classic game *Zork1*.

The research community has recognized several commonsense reasoning problems in IF game playing [9], such as detecting valid actions and predicting the effects of different actions. In this work, we derive a commonsense evaluation related to the latter problem, i.e., predicting which is the most likely observation when applying an action to a game state.

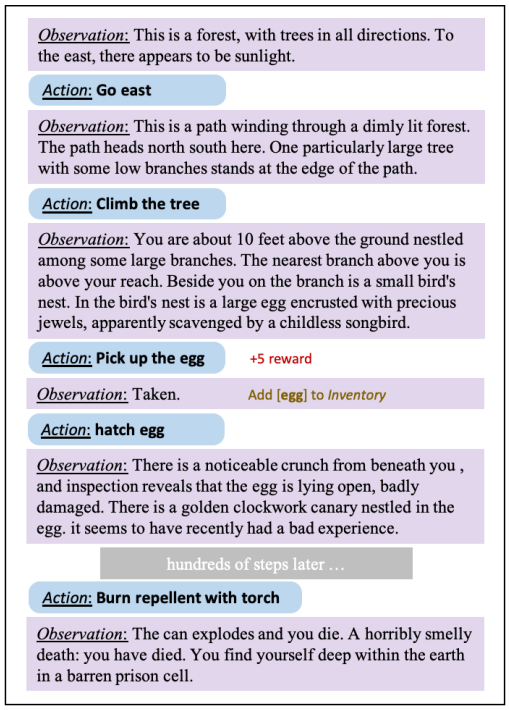

Figure 1: Classic dungeon game *Zork1* gameplay sample. The player receives textual observations describing the current game state. The player sends textual action commands to control the protagonist. Various types of commonsense reasoning are illustrated in the texts of observations and commands from the gameplay interaction, such as spatial relations, objective manipulation, and physical relations.

Our approach of commonsense benchmark construction has several advantages. Specifically, it naturally relaxes the restrictions in commonsense types and reasoning forms. First, we relax the limitation in commonsense types by noticing that predicting the next observation naturally requires various commonsense knowledge and reasoning types. As shown in Figure 1, a primary commonsense type is spatial reasoning, e.g., ‘‘`climb the tree`’’ makes the protagonist up on a tree. Another primary type is reasoning with object interactions, such as with relationships, like keys can open locks; with object's properties, such as ‘‘`hatch egg`’’ will reveal "things" inside the egg; with physical reasoning, like ‘‘`burn repellent with torch`’’ leads to an explosion and kills the player. The above interactions are much more comprehensive than the relationships defined in ConceptNet as used in previous datasets. Second, we enforce the task to have more commonsense reasoning steps over simple commonsense validation. A large portion of IF game observations are narrative, and the next observation is less likely to be a sole statement of the action effect, but an extended narrates about what happens because of the effect.[2]

Our benchmark designs based on the IF games support automatic data generation from multiple genres and domains, including dungeon crawl, Sci-Fi, mystery, comedy, and horror. From an RL perspective, our commonsense reasoning task formulation shares the essence of dynamics model learning for model-based RL solutions, especially those based to next state predictions [15, 8, 18, 1]. As a result, models developed on our benchmarks provide values to both commonsense reasoning and model-based reinforcement learning.

We introduce a new commonsense reasoning benchmark from four IF games in the *Zork Universe*, the Zork Universe Commonsense Comprehension task (**ZUCC**). Our experiments show that existing standard models perform poorly on the resulted benchmark, with a significant human-machine gap. Our way of construction is general and easy-to-implement, thus the dataset is easy to be scaled up with more text games as long as their simulators and walkthroughs are available.

[2]For some actions, like get and drop objects, the returns are simple statements. We removed some of these actions. Details can be found in Section 2.

2 ZUCC Dataset Construction

We pick games from the *Zork Universe* that are supported by the *Jericho* environment [9], namely *zork1*, *zork3*, *enchanter*, *sorcerer*,[3] to construct our **ZUCC** dataset. This section first reviews the necessary definitions in the IF game domain; then describes how we construct our **ZUCC** dataset as a forward prediction from the game walkthrough trajectories.

2.1 Interactive Fiction Game Background

Textual Observations and the POMDP Formulation An IF game-playing agent interacts with the game engine in multiple turns until the game is over or the maximum number of steps is reached. At the t-th turn, the agent receives a *textual observation* describing the current game state $o_t \in O$ and an additional reward scalar r_t indicating the game progress and it sends back a textual command $a_t \in A$ to control the protagonist.

Trajectories and Walkthroughs A *trajectory* in text game playing is a sequence of tuples (o_t, a_t, r_t, o_{t+1}) starting with the initial observation o_0. We define the *walkthrough* of a text game as a trajectory that completes the game progress. Our dataset construction is based on the walkthroughs since each of them correspond to an entire story and, hence, represents a natural distribution of commonsense tasks. In this work we use the walkthroughs provided by the Jericho environment for the selected games.

2.2 Data Construction from the Forward Prediction Task

The Forward Prediction Task We represent our commonsense reasoning benchmark as a next-observation prediction task, given the current observation and action. The benchmark construction starts with all the tuples in a walkthrough trajectory, and we then extend the tuple set by including all valid actions and their corresponding next-observations conditioned on the current observations in the walkthrough. Specifically, for a walkthrough tuple $(o_t, a_t, r_t, o_{t+1},)$, we first obtain the complete valid action set A_t for o_t. We sample and collect one next observation o^j_{t+1} after executing the corresponding action $a^j_t \in A_t$. The next-observation prediction task is thus to select the next observation o^j_{t+1} given (o_t, a^j_t) from the complete set of next observations $O_{t+1} = \{o^k_{t+1}, \forall k\}$.[4]

Data Processing We collect tuples from the walkthrough data provided by the Jericho environments. We detect the valid actions via the Jericho API and the game-specific templates. Following previous work [9], we augmented the observation with the textual feedback returned by the command [*inventory*] and [*look*]. The former returns the protagonist's objects, and the latter returns the current location description. When multiple actions lead to the same next-observation, we randomly keep one action and next-observation in our dataset. We leave all the tuples from the *zork3* game for evaluation. We split the walkthrough of *zork3*, keeping the first 136 tuples as a development set and the rest 135 tuples as a test set. We remove the `drop OBJ` actions since it only leads to synthetic observations with minimal variety. For each step t, we keep at most 15 candidate observations in O_t for the evaluation sets. When there are more than 15 candidates, we select the candidate that differs most from o_t with Rouge-L measure [13].

Table 1 summarizes statistics of the resulted **ZUCC** dataset. The number of tuples is much larger in the test set because there are actions that do not have the form of `drop OBJ` but have the actual effects of dropping objects. Through the game playing process, more objects will be collected in the inventory at the later stages. The test data will be much easier as long as these non-standard drop actions have been recognized. A similar problem happens to actions like `burn repellent` that can be performed at every step once the object is in the inventory. To deal with such problems in the test set, we finally down-sample these biased actions to achieve similar distributions in development and

[3] There is an excluded game, *spellbreaker*, in Jericho that belongs to the Zork Universe. As the study in their paper shows, the game contains a large portion of non-standard actions that are usages of spells, and handling its non-standard vocabulary is beyond this paper's scope.

[4] Similarly, we can use the backward prediction, i.e., predicting a^j_t given o_t and o^j_{t+1}. Our preliminary study shows that the backward prediction does not introduce extra challenges compared the forward one. Therefore we only focus on the latter in the paper.

Sets	#WT Tuples	#Tuples before Proc	#Tuples after Proc
Train	913	17,741	10,498
Dev	136	1,982	1,276
Test before down-sampling	135	2,087	1,573
Test (final)	-	-	822

Table 1: Data statistics of our **ZUCC** task. **WT** is short for walkthrough. Train set is from the game *Zork1*, *Enchanter*, and *Sorcerer*. Both dev and test sets are from the game *Zork3*.

Method	Dev Acc	Test Acc
Random Guess	10.66	16.42
Match LSTM	57.52	62.17
BERT-siamese	49.29	53.77
BERT-concat	64.73	64.48
Human Average Performance*	86.80	-
Human Expert Performance*	96.40	-

Table 2: Evaluation on our **ZUCC** data. Human performance (*) is computed on a subset of data.

test sets. We perform down-sampling with rule-based methods. The resulted final version of the test set is denoted as *Test (final)* in the table.

Remark on Further Impacts Our benchmark design opens opportunities beyond commonsense evaluation. For example, the form of our tasks, compared to the relevant tasks of next-sentence generation, such as the SWAG [22], introduces actions as intervention, thus encourage causal reasoning. Therefore it has a potential impact on causal knowledge acquisition. On the other hand, the *partial observability* nature of IF games makes o_t and a_t^j not sufficient for predicting o_t^j sometimes. Therefore our task encourages the development of structured abstract representations to summarize the history [2, 3].

3 Experiments

We first benchmark the state-of-the-art models for natural language inference on our dataset, with and without pre-trained Language Models (LMs). Then we conduct a human study on a sub-set of our development data to quantitatively measure the human performance and the human-machine gap.

Baselines We compare the following baselines on the **ZUCC** dataset. In the model descriptions, the notations o_t, a_t of observations and actions represent their word sequences.

• **Match LSTM** The neural attention model was proposed in [21], commonly used in natural language inference as baselines. Specifically, we concatenate o_t and a_t separated by a special split token as the premise and use the o_{t+1}^j as the hypothesis. The matching scores for all o_{t+1}^j are then fed to a softmax layer for the final prediction.

• **BERT Siamese** The Siamese model uses a pre-trained BERT model to encode the current observation-action pair (o_t, a_t) and candidate observations $\tilde{o}_{t+1}^j, j = 1, ..., N$. All inputs to BERT start with the "[CLS]" token. o_t and a_t are concatenated by a "[SEP]" token:

$$\boldsymbol{h}_t = \text{BERT}([o_t, a_t]), \quad \tilde{\boldsymbol{h}}_{t+1}^j = \text{BERT}(\tilde{o}_{t+1}^j),$$
$$l_j = f([\boldsymbol{h}_t, \tilde{\boldsymbol{h}}_{t+1}^j, \boldsymbol{h}_t - \tilde{\boldsymbol{h}}_{t+1}^j, \boldsymbol{h}_t * \tilde{\boldsymbol{h}}_{t+1}^j]),$$

where $[\cdot, \cdot]$ denotes concatenation. $\boldsymbol{h}_t$ and $\tilde{\boldsymbol{h}}_{t+1}^j$ are last layer hidden state vectors that correspond to the "[CLS]" token. Each candidate next-observation is scored by an output function f, and the logits are normalized by the softmax function. We use the cross-entropy loss as the training objective.

• **BERT Concat** It represents the standard pairwise prediction mode of BERT. We concatenate o_t and a_t with a special split token as the first segment and treat $\tilde{o}_{t+1}^j$ as the second. We then concatenate the

two with the "[SEP]" token. We have a matching score for each o_{t+1}^j with a linear mapping from the hidden state of the "[CLS]" token, and then feed the scores to a softmax layer for the final prediction. This model is much less efficient than the former two; thus, it is not practical in IF game playing. Here we report its results for reference.

Implementation Details We experimented with training the three baselines on both full training tuples (biased training) and the processed training set (de-biased training). We reported the best development set performance for each model.

Results Table 2 summarizes the model performance. All three baselines manage to learn decent models, i.e., significantly better than a random guess. For both Match LSTM and BERT-Siamese, the best development performance was found with de-biased training because this training setting is more consistent with the evaluation scenarios.

There is an exception for the BERT-Concat because the model is not learning in the de-biased training setting, i.e., the training accuracy stays around 10%, the level of a random guess. A possible reason is that the BERT-Concat model works directly on a complicated concatenated string of multiple types of inputs. Therefore it is challenging for it to distinguish the structures of input/output observations and actions. For example, it may not learn which parts of the inputs correspond to inventories. To make the model work, we first pre-train the BERT-concat model on the biased training data until converging, then fine-tune the model on the de-biased data. This procedure finally gives the best performance on our **ZUCC**.

Although the baselines are making progress, as shown in our human evaluation, the best development accuracy (64.73%) is still far from human-level performance. Compared to the human expert's near-perfect performance, the substantial performance gaps confirms that our **ZUCC** captures challenging commonsense understanding problems.

Human Evaluation We present to the human evaluator each time a batch of tuples starting from the same observation o_t, together with its shuffled valid actions A_{t+1} and next observations O_{t+1}. The evaluators are asked to read the start observation o_t first, then to align each $o \in O_{t+1}$ with an action $a \in A_{t+1}$. Besides, for each observation o, besides guessing the action's alignment, the subjects are asked to answer a secondary question: whether the provided o_t, o pair is sufficient for them to predict the action. If they believe there are not enough clues and their action prediction is based on a random guess, they are instructed to answer "UNK" to the second question.

We collect two sets of human predictions on 250 samples. The first set is annotated by one of the co-authors who have experience in interactive fiction game playing (but have **not** played *Zork3* before). We denote the corresponding result as *Human Expert Performance*. The second set is annotated by three of our co-authors who have never played IF games. The corresponding result is denoted as *Human Average Performance*. The corresponding accuracy is shown in Table 2. The human expert performs more than 30% higher compared to the machines. It is also interesting to see that even the human annotators who do not play IF games much can outperform the machine with more than 20%. Since these annotators have not been trained for this task, their performance could represent human-level domain transferability with commonsense knowledge.

Finally, the annotators recognized 10.0% cases with insufficient clues, indicating an upper-bound of methods without access to history observations.[5]

4 Conclusion

Interactive Fiction (IF) games encode plentiful and diverse commonsense knowledge of the physical world. In this work, we derive a commonsense reasoning benchmark **ZUCC** from IF games in the *Zork Universe*. Taking the form of predicting the most likely observation when applying an action to a game state, our automatically generated benchmark covers comprehensive commonsense reasoning types such as spatial reasoning and object interaction, etc. Our experiments on **ZUCC** show that current popular neural models have limited performance compared to humans.

[5]Humans can still make a correct prediction by first eliminating most irrelevant options and then making a random guess.

References

[1] A. Adhikari, X. Yuan, M.-A. Côté, M. Zelinka, M.-A. Rondeau, R. Laroche, P. Poupart, J. Tang, A. Trischler, and W. L. Hamilton. Learning dynamic knowledge graphs to generalize on text-based games. *arXiv preprint arXiv:2002.09127*, 2020.

[2] P. Ammanabrolu and M. Hausknecht. Graph constrained reinforcement learning for natural language action spaces. *arXiv*, pages arXiv–2001, 2020.

[3] P. Ammanabrolu, E. Tien, M. Hausknecht, and M. O. Riedl. How to avoid being eaten by a grue: Structured exploration strategies for textual worlds. *arXiv preprint arXiv:2006.07409*, 2020.

[4] C. Bhagavatula, R. Le Bras, C. Malaviya, K. Sakaguchi, A. Holtzman, H. Rashkin, D. Downey, W.-t. Yih, and Y. Choi. Abductive commonsense reasoning. In *International Conference on Learning Representations*, 2019.

[5] Y. Bisk, R. Zellers, R. LeBras, J. Gao, and Y. Choi. Piqa: Reasoning about physical commonsense in natural language. In *AAAI*, pages 7432–7439, 2020.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019.

[7] X. Guo, M. Yu, Y. Gao, C. Gan, M. Campbell, and S. Chang. Interactive fiction game playing as multi-paragraph reading comprehension with reinforcement learning. *arXiv preprint arXiv:2010.02386*, 2020.

[8] D. Ha and J. Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018.

[9] M. Hausknecht, P. Ammanabrolu, M.-A. Côté, and X. Yuan. Interactive fiction games: A colossal adventure. *arXiv preprint arXiv:1909.05398*, 2019.

[10] L. Huang, R. Le Bras, C. Bhagavatula, and Y. Choi. Cosmos qa: Machine reading comprehension with contextual commonsense reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2391–2401, 2019.

[11] M. Jiang, J. Luketina, N. Nardelli, P. Minervini, P. H. Torr, S. Whiteson, and T. Rocktäschel. Wordcraft: An environment for benchmarking commonsense agents. *arXiv preprint arXiv:2007.09185*, 2020.

[12] H. Levesque, E. Davis, and L. Morgenstern. The winograd schema challenge. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*. Citeseer, 2012.

[13] C.-Y. Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[14] J. Mullenbach, J. Gordon, N. Peng, and J. May. Do nuclear submarines have nuclear captains? a challenge dataset for commonsense reasoning over adjectives and objects. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6054–6060, 2019.

[15] D. Pathak, P. Agrawal, A. A. Efros, and T. Darrell. Curiosity-driven exploration by self-supervised prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 16–17, 2017.

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

[17] M. Sap, H. Rashkin, D. Chen, R. Le Bras, and Y. Choi. Social iqa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4453–4463, 2019.

[18] J. Schrittwieser, I. Antonoglou, T. Hubert, K. Simonyan, L. Sifre, S. Schmitt, A. Guez, E. Lockhart, D. Hassabis, T. Graepel, et al. Mastering atari, go, chess and shogi by planning with a learned model. *arXiv preprint arXiv:1911.08265*, 2019.

[19] R. Speer, J. Chin, and C. Havasi. Conceptnet 5.5: an open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pages 4444–4451, 2017.

[20] A. Talmor, J. Herzig, N. Lourie, and J. Berant. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, 2019.

[21] S. Wang and J. Jiang. Machine comprehension using match-lstm and answer pointer. *arXiv preprint arXiv:1608.07905*, 2016.

[22] R. Zellers, Y. Bisk, R. Schwartz, and Y. Choi. Swag: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 93–104, 2018.

[23] B. Zhou, D. Khashabi, Q. Ning, and D. Roth. "going on a vacation" takes longer than "going for a walk": A study of temporal commonsense understanding. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3354–3360, 2019.