Task Proposal for SemEval-2020: Commonsense Validation and Explanation

Anonymous NAACL submission

1 Overview

Summary. The task is to directly test whether a system can differentiate natural language statements that make sense from those that do not make sense. We designed three subtasks. The first task is to choose from two natural language statements with similar wordings which one **makes sense** and which one does **not make sense**; The second task is to find the key reason from three options why a given statement does not make sense; The third task asks machine to generate the reasons and we use BLEU to evaluate them.

Motivation. Natural Language Understanding (NLU) has received increasing research attention in recent years. With language models trained on large corpora (Peters et al., 2018; Devlin et al., 2018), algorithms show better performance than humans on some benchmarks (Group, 2017; Devlin et al., 2018). Compared to humans, however, most end-to-end trained systems are rather weak on common sense. For example, it is straightforward for a human to understand that *someone can put a turkey into a fridge* but *he can never put an elephant into a fridge* with basic commonsense reasoning, but it can be non-trivial for a system to tell the difference. Arguably, commonsense reasoning should be a central capability in a practical NLU system (Davis, 2017); it is, therefore, important to be able to evaluate how well a model can do for commonsense validation.

Existing datasets test common sense indirectly through tasks that require extra knowledge, such as co-reference resolution, like Winograd Schema Challenge (Levesque et al., 2012; Morgenstern and Ortiz, 2015), subsequent event prediction like Choice of Plausible Alternatives (Roemmele et al., 2011),the JHU Ordinal Common-sense Inference (Zhang et al., 2017) and Situations with Adversarial Generations (Zellers et al., 2018), or reading comprehension like Story Cloze Test and ROCStories Corpora (Mostafazadeh et al., 2016; Ostermann et al., 2018b) and MCScript (Ostermann et al., 2018a), or QA problems like SQUABU (Davis, 2016), CommonsenseQA (Talmor et al., 2018) and OpenBookQA (Mihaylov et al., 2018). They verify whether a system is equipped with common sense by testing whether it can give a correct answer where the input does not contain such knowledge. However, there are two limitations to such benchmarks. First, they do not give a direct metric to quantitatively measure commonsense validation capability. Second, they do not explicitly identify the key factor required in a commonsense validation process.

To our knowledge, our dataset is the first benchmark for direct linguistic commonsense validation and explanation. Note that there has been dataset which focuses on non-linguistic world knowledge plausibility (Wang et al., 2018) or only limited attributes or actions of physical knowledge like verbphysics (Forbes and Choi, 2017). They are related to our dataset but serve robotic research mainly. We hope this benchmark can be used for commonsense reasoning by the fNLP community. Besides, we also expect that this work could be instructive on enhancing interpretability on commonsense reasoning and other NLP tasks and on combining explanation with language generation.

2 Task

2.1 Formal Definition

Formally, each instance in our dataset is composed of 10 sentences: $\{s_1, s_2, o_1, o_2, o_3, r_1, r_2, r_3\}$. s_1 and s_2 are two similar statements which in the same syntactic structure and differ by only a few words, but only one of them makes sense while the other does not. They are used on our first subtask called **Validation**, which requires the model to

identify which one makes sense. For the against-common-sense statement s_1 or s_2, we have three optional sentences o_1, o_2 and o_3 to explain why the statement does not make sense. Our subtask 2, named **Explanation (Multi-Choice)**, requires that the only one correct reason be identified from two other confusing ones. For the same against-common-sense statement s_1 or s_2, our subtask 3 naming **Explanation (Generation)**, asks the participants to generate the reason why it does not make sense. The 3 referential reasons r_1, r_2, r_3 are used for evaluating subtask 3.

2.2 Example

Task 1: Validation

Task: Which statement of the two is against common sense?

Statement1: He put a turkey into the fridge.

Statement2: He put an elephant into the fridge.

Task 2: Explanation (Multi-Choice)

Task: Select the most corresponding reason why this statement is against common sense.

Statement: He put an elephant into the fridge.

A: An elephant is much bigger than a fridge.

B: Elephants are usually white while fridges are usually white.

C: An elephant cannot eat a fridge.

Task 3: Explanation (Generation)

Task: Generate the reason why this statement is against common sense and we will use *BELU* to evaluate it.

Statement: He put an elephant into the fridge.

Referential Reasons:

1. An elephant is much bigger than a fridge.

2. A fridge is much smaller than an elephant.

3. Most of the fridges aren't large enough to contain an elephant.

3 Data&Resources

Because there are no training data, participants are encouraged to use what they deemed appropriate for the tasks to train their model, and use our testset for evaluation, for example, the ConceptNet (Liu and Singh, 2004; Havasi et al., 2007; Speer and Havasi, 2013). And will release the inspirational corpora during SemEval-2020.

3.1 Annotation Process

We ask data annotators to write instances by themselves. And we also provide them some with sources to stimulate inspiration, such as the raw English sentences of ConceptNet5.5 (Speer et al., 2017). For example, *"he was sent to a (restaurant)/(hospital) for treatment after a car crash"* were inspired by the two sentences *"restaurants provide food"* and *"hospitals provide medical care"* However, those corpora may have incorrect and one-sided knowledge, we do not use those sentences. Besides, we also let them get inspiration from existing commonsense reasoning questions like WSC (Levesque et al., 2012; Morgenstern and Ortiz, 2015), COPA (Roemmele et al., 2011) and SQUABU (Davis, 2016). **Those existing corpora will never be used directly but only for inspiration, which means those sentences will not appear in our testset directly.**

Quality Control. First, researchers will examine each instance case by case. Then, there is also *Human Evaluation*. Each instance is answered by at least three testees. If more than half testees do one instance wrong (either *Validation* or *Explanation*), we will rewrite or abolish the instance. For subtask 3, if one answer is considered unsuitable by more than half testees, we will rewrite it.

3.2 Annotation guidelines

When writing instances, annotators were asked to follow several soft principles. First, two statements should be in the same syntactic structure and only differ by a few words. Second, try to avoid complex knowledge and focus on daily common sense, and they should make the questions as understandable as possible. Every literate person is able to give the right answers. Third, the confusing reasons should better contain more important words like entities and activities in the wrong statements and not deviate from the problem context, for example, the confusing reasons of "*he put an elephant into the fridge*" should better contain both "*elephant*" and "*fridge*". Otherwise, it may be easily captured by BERT (Talmor et al., 2018), which models the statement contexts explicitly. Next, the three option reasons should be only related to the wrong statement rather the right statement. Because we want further studies can estimate wrong statements without those right statements. Furthermore, confusing reasons should be correct themselves. Otherwise, the models may

	Validation	Explanation (Multi-Choice)
Random	50.0%	33.3%
ELMo	69.4%	33.4%
BERT	70.1%	45.6%
Human	99.1%	97.8%

Table 1: Pilot Task Experimental Results

select the right reasons without considering the causal relations between statement and reasons. This worry was raised from that models can perform well in the ROC Story Cloze Task when only looking at alternative endings ignoring stories (Schwartz et al., 2017). Last, we control the length of each sentence, make the right reason neither too long nor too short among the three reasons. However, those principles are all soft restrictions, some instances are still good ones even without obeying those principles.

3.3 Data Size & Resources

We plan to create at least 4,000 instances. Thus, task 1 has 4,000 against-common-sense statements, 4,000 right statements; Task 2 has 4,000 true reasons and 8,000 confusing reasons; Task 3 has 12,000 true reasons, with 4,000 true reasons in task 2 counting in.

We have already finished 2,000 instances, having 4,000 statements for task 1 and 6,000 reasons for task 2. We plan to spend about 4 months and 4,000 dollars to prepare for the rest.

4 Evaluation

Each participating team can only access to 100 instances. Then, the unlabelled test data will be released. After SemEval-2020, the standard answers for the test data will be released as well.

Evaluation Metric. We use the accuracy score to evaluate subtask 1 and subtask 2. For subtask 3, we use BELU score to evaluate the results.

When performing *Human Evaluation*, which stated in *Data&Resources* section, if the less than half testees do one instance wrong, the results will be counted in *Human Performance*.

5 Pilot Task

We perform the pilot experiments for task 1 and task 2 on the 2,000 instances. We choose state-of-the-art language models trained over large texts as our baselines, assuming that common sense knowledge is encoded over texts. For the validation task, we calculate perplexities of both statements, choosing the one with lower scores as the correct one. For task 2, we first concatenate the statement with each reason and then use the three concatenated sentences to calculate perplexities. For example, concatenate *"he put an elephant into the fridge"* with its optional reasons to be *" "he put an elephant into the fridge" is against common sense because an elephant cannot eat a fridge"* and so on. The results are shown in Table 1.

6 Task Organizers

Cunxiang Wang is a Ph.D. student at Westlake University, working under the supervision of Prof.Yue Zhang. His research focuses on commonsense reasoning and natural language understanding. He is also interested in introducing knowledge into current NLP systems.

Yue Zhang currently works as a tenured associate professor at Westlake University. His research interests include fundamental NLP and its application to the capital markets. Yue Zhang has been active in the research community, serving as area chairs for ACL(2017,18,19), EMNLP(2015, 17, 19), NAACL(2015) and COLING (2014, 18). He has given tutorials at NAACL 2010, ACL 2014, EMNLP 2016 and EMNLP 2018 on relevant subjects.

Xiaodan Zhu has previously co-organized **SemEval-2016 Task 6**. He has also competed for **SemEval-2013** and **SemEval-2014** as a main contributor for several top-ranking systems. Xiaodan Zhu is currently an assistant professor of Queen's University, Canada, and his recent interests include sentiment analysis, natural language inference, and financial text analytics. He serves as area chairs for ACL (2019, 18), NAACL (2019), and COLING (2018); publication co-chair for COLING-2018; workshop co-chair for COLING-2020; co-chair for SemEval-2019 and 2020.

Shuailong Liang is a Ph.D. student at Singapore University of Technology and Design (https://istd.sutd.edu.sg/people/phd-students/liang-shuailong) under the supervision of Prof Yue Zhang. His research interests are deep learning based natural language processing, such as NLP applications in social science, text retrieval, question answering etc. He is also interested in applying common sense and structured knowledge into current QA or other NLP systems.

References

Ernest Davis. 2016. How to write science questions that are easy for people and hard for computers. *AI Magazine*, 37:13–22.

Ernest Davis. 2017. Logical formalizations of commonsense reasoning: a survey. *Journal of Artificial Intelligence Research*, 59:651–723.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Maxwell Forbes and Yejin Choi. 2017. Verb physics: Relative physical knowledge of actions and objects. In *ACL*.

Natural Language Computing Group. 2017. R-net: Machine reading comprehension with self-matching networks.

Catherine Havasi, Robert Speer, and Jason Alonso. 2007. Conceptnet 3: a flexible, multilingual semantic network for common sense knowledge. In *Recent advances in natural language processing*, pages 27–29. Citeseer.

Hector J. Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*, KR'12, pages 552–561. AAAI Press.

Hugo Liu and Push Singh. 2004. Conceptnet—a practical commonsense reasoning tool-kit. *BT technology journal*, 22(4):211–226.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391. Association for Computational Linguistics.

Leora Morgenstern and Charles L. Ortiz. 2015. The winograd schema challenge: Evaluating progress in commonsense reasoning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, pages 4024–4025. AAAI Press.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James F. Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *HLT-NAACL*.

Simon Ostermann, Ashutosh Modi, Michael Roth, Stefan Thater, and Manfred Pinkal. 2018a. Mcscript: A novel dataset for assessing machine comprehension using script knowledge. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*. European Language Resource Association.

Simon Ostermann, Michael Roth, Ashutosh Modi, Stefan Thater, and Manfred Pinkal. 2018b. Semeval-2018 task 11: Machine comprehension using commonsense knowledge. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 747–757. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of NAACL*.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S. Gordon. 2011. Choice of Plausible Alternatives: An Evaluation of Commonsense Causal Reasoning. In *AAAI Spring Symposium on Logical Formalizations of Commonsense Reasoning*, Stanford University.

Roy Schwartz, Maarten Sap, Ioannis Konstas, Leila Zilles, Yejin Choi, and Noah A. Smith. 2017. The effect of different writing tasks on linguistic style: A case study of the roc story cloze task. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 15–25. Association for Computational Linguistics.

Robert Speer and Catherine Havasi. 2013. Conceptnet 5: A large semantic network for relational knowledge. In *The People's Web Meets NLP*, pages 161–176. Springer.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2018. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *CoRR*, abs/1811.00937.

Su Wang, Greg Durrett, and Katrin Erk. 2018. Modeling semantic plausibility by injecting world knowledge. *arXiv preprint arXiv:1804.00619*.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. Swag: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Sheng Zhang, Rachel Rudinger, Kevin Duh, and Benjamin Van Durme. 2017. Ordinal common-sense inference. *Transactions of the Association for Computational Linguistics*, 5:379–395.