COPA-SSE: Semi-structured Explanations for Commonsense Reasoning

Ana Brassard,[1,2] Benjamin Heinzerling,[1,2] Pride Kavumba,[2,1] Kentaro Inui[2,1]

[1]Riken AIP, [2]Tohoku NLP Lab

{ana.brassard, benjamin.heinzerling}@riken.jp,
kavumba.pride.q2@dc.tohoku.ac.jp, inui@tohoku.ac.jp

Abstract

We present Semi-Structured Explanations for COPA (COPA-SSE), a new crowdsourced dataset of 9,747 semi-structured, English common sense explanations for Choice of Plausible Alternatives (COPA) questions. The explanations are formatted as a set of triple-like common sense statements with ConceptNet relations but freely written concepts. This semi-structured format strikes a balance between the high quality but low coverage of structured data and the lower quality but high coverage of free-form crowdsourcing. Each explanation also includes a set of human-given quality ratings. With their familiar format, the explanations are geared towards commonsense reasoners operating on knowledge graphs and serve as a starting point for ongoing work on improving such systems. The dataset is available at https://github.com/a-brassard/copa-sse.
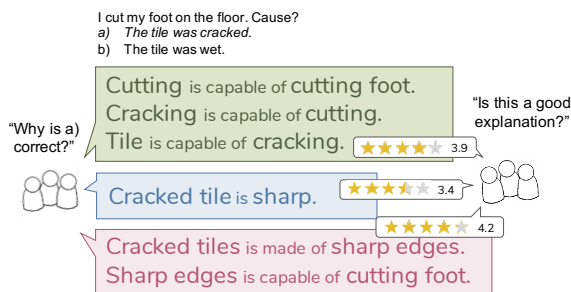
Keywords: Collaborative Resource Construction & Crowdsourcing, Corpus (Creation, Annotation, etc.), Knowledge Discovery/Representation, Question Answering

1. Introduction

While there are many datasets for question answering and commonsense reasoning (Rogers et al., 2021), models are known to exploit shortcuts such as superficial cues in these datasets, which leads to artificially high evaluation scores (Gururangan et al., 2018). One way to ensure models are reasoning as intended is to require explanations for their predictions (Bowman and Dahl, 2021). A prominent example of such a setting is the Commonsense Explanations Dataset (CoS-E) (Rajani et al., 2019), which provides crowdsourced justifications of the correct answers expressed in free text. While free-form crowdsourcing allows representing natural and diverse human reasoning, quality control is notoriously difficult (Daniel et al., 2018). At the other end of the spectrum are explanations that are fully grounded in a knowledge graph (KG), i.e., each element of the explanation corresponds to a node or edge in a KG. However, this structured approach is limited by the coverage of the KG, i.e., the explanation will be sub-optimal or impossible when the situation to explain is not covered by the KG. Here, we adopt a semi-structured approach aiming to combine the best of both worlds—the coverage potential of open-ended crowdsourcing and quality control of structured data.

Specifically, we introduce Semi-Structured Explanations for COPA (COPA-SSE), a new explanation dataset for the Choice of Plausible Alternatives (COPA) dataset (Roemmele et al., 2011).[1] Each explanation consists of a set of English statements, which, in turn, consist of a head text, a selected predicate, and tail text, mimicking ConceptNet (Speer et al., 2017) triples (Figure 1). The head and tail texts are free-form, allowing an open concept inventory. Each explanation

(a) Collecting and rating explanations for Balanced COPA.
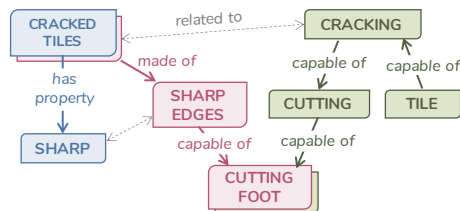
(b) The explanations in graph form.

Figure 1: Constructing COPA-SSE. Crowdworkers gave one or more triple-like statements explaining the correct answer which were then rated by different workers (a). Each statement consists of head and tail text linked by a ConceptNet relation. The statements can be aggregated into an explanation graph (b) (§3.5).

also includes quality ratings. Note that COPA-SSE is not meant to extend existing commonsense knowledge graphs, but rather to be used as examples of extraction and/or generation *results* based on a specific prompt (question).

In this paper, we introduce COPA-SSE (§2), detail its construction (§3), demonstrate a simple application (§4), and discuss future use cases (§5). COPA-SSE is available for download at https://github.com/a-brassard/copa-sse.

---

[1]We use Balanced COPA (Kavumba et al., 2019), a superset of COPA.

2. Semi-Structured Explanations for COPA

Design goals. Our goal is to add high-quality explanations to Balanced COPA. Since the nature of a good explanation is subject of debate (Miller, 2019), we adopt a working definition: A good explanation is a minimal set of relevant common sense statements that coherently connect the question and the answer. For example, the fact *Opening credits play before a film.* connects the question *The opening credits finished playing. What happened as a result?* and its answer *The film began.* Commonsense KGs such as ConceptNet provide such statements but have limited coverage (Hwang et al., 2021). For example, even if question and answer concepts are found in the KG, the paths between them can degenerate into long chains of statements that are neither minimal nor relevant (Figure 2).

In contrast to structured approaches, unstructured free-form text is not limited by KG coverage. Previous work has elicited such free-form explanations from crowdworkers, but suffers from low quality. For example, in a preliminary manual inspection of a random sample of 1,200 CoS-E explanations, one of the authors judged only a small fraction to be acceptable explanations in terms of relevance and thoroughness.

Aiming for a golden middle, we devise a semi-structured explanation scheme comprising a set of triple-like statements. Each statement consists of open-ended head text and tail text connected with a ConceptNet relation. In practice, crowdworkers created explanations by selecting a predicate from a list while providing free text for the two concept slots. This format encouraged workers to provide explanations close to our definition without being restricted to a pre-defined inventory of concepts. We refer to this combination of free text and ConceptNet predicates as *semi-structured explanations*.

Among related prior work, our approach is most similar to the explanation graphs by (Saha et al., 2021), which also combine ConceptNet relations and free text concepts, but differ in task, domain, and crowdsourcing protocol.

Dataset statistics. Table 1 shows examples of COPA-SSE explanations. COPA-SSE contains 9,747 commonsense explanations for 1,500 Balanced COPA questions. Each question has up to nine explanations given by different crowdworkers. We provide the triple-format described above, as well as a natural language version obtained by replacing ConceptNet relations with more human-readable descriptions. 61% of explanations are only one statement while the other 39% comprise two or more, with the longest explanation being ten statements (Figure 3). Each explanation has a quality rating on a scale of 1 to 5 as given by crowdworkers. Figure 4 shows the rating distribution after initial collection (original data). To guarantee that each Balanced COPA instance is explained by high-quality explanations, we collected additional ex-
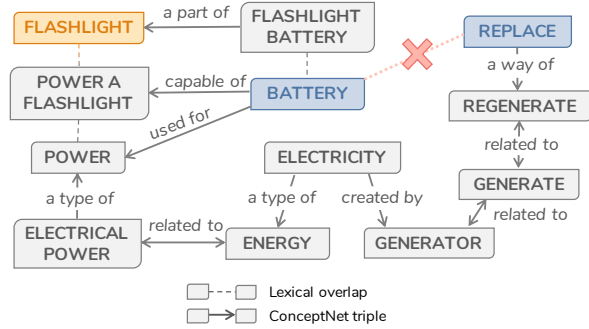

Figure 2: A manually extracted ConceptNet subgraph to illustrate the caveats of only using existing resources. One author attempted to find paths connecting concepts from the question *The flashlight was dead. Effect?* and the answer *I replaced the batteries.* and was unable to find a meaningful path between *battery* and *replace.* The two concepts are connected but the path contains irrelevant facts to the point of being meaningless.
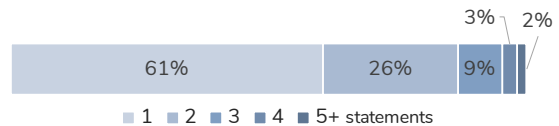

Figure 3: Number of statements per explanation.

planations until most Balanced COPA instances (98%) had at least one explanation rated 3.5 or higher (final data). In other words, 98% of the questions have at least one highly-rated explanation. Initially, 38% of all explanation were over this threshold, which increased to 44% after the additional collection run. We kept the lower-quality explanations as they can be useful negative samples, e.g., in contrastive learning settings where they can act as sub-optimal examples in terms of thoroughness or relevance. Finally, we created additional aggregated versions of the explanations by merging or connecting similar concepts. We now describe each step in more detail.

3. Crowdsourcing Protocol

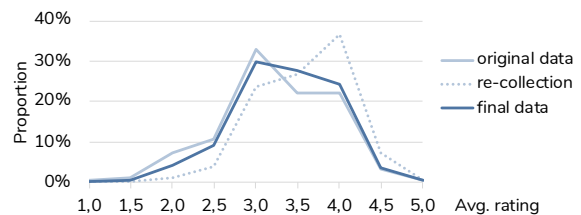Crowdworkers were asked to provide one or more statements that connect the question and the answer in

Figure 4: Average rating distribution before (original data) and after the re-collection round (final data). Values are rounded to the nearest half-star.

The documents were loose. Effect?	
✓ I paper clipped them together.	✗ I kept them in a secure place.

★★★★·	Paper clip is used for loose documents.
★★★★	Paper clips is used for keeping documents together. Paper clipping can be done to have the documents together.
★★★★	Paper clip is used for clipping paper together.
★★★◢	Paper clip is used for organizing papers.
★★★◢	Paper clip can be done to keep papers together.
★★★·	The paper clipped is a way of holding the papers together.

They lost the game. Cause?	
✗ Their coach pumped them up.	✓ Their best player was injured.

★★★★·	Game is a team work. Player is a part of a team. Player injured causes team not working properly. Team not working properly causes lose the game.
★★★★	Best player is a part of the team. Injury of the best player causes the team to lose.
★★★★	Their best player being injured causes the team to lose.
★★★★	Teams is made of players. Injuries is capable of causing losses.
★★★◢	Injury is capable of causing loss.
★★·	The team causes the injury.

Table 1: Examples of collected and rated explanations for Balanced COPA questions.

a triple format: a free-form head text, a selection of ConceptNet relations, and a free-form tail text, together forming a commonsense statement (§3.1). Each set of statements was then rated by five different workers (§3.2). To gather more high-quality explanations, we invited workers whose explanations were highly rated to provide additional explanations (§3.3). Section 3.4 lists worker qualifications and compensation, followed by further post-processing in Section 3.5.

3.1. Collecting Explanations

Figure 5 shows our collection form. Workers were given a COPA question and two answer choices with the correct one marked. The input row below consists of two text fields for inputting concepts and a drop-down box for selecting the relation between them. Workers could increase the number of rows to provide explanations with multiple statements, as they were encouraged (but not forced) to do. The relations are a subset of ConceptNet predicates which we selected and translated into human-readable English for easier understanding by non-experts.[2] For example, the input *an apple* is a *fruit* corresponds to the statement *"An apple is a fruit."* and the triple (*"an apple"*, IsA, *"fruit"*).

Free-form text guarantees neither consistent granularity nor chains of statements connected by matching concepts. For example, a phrase such as *"the act of eating a sweet fruit"* can be given as tail text, even though the next statement might not include that same phrase.

We opted to leave this freedom as longer statements can still form coherent explanations, and, as we found in preliminary runs, introducing strict constraints might lead to unnatural and/or less informative explanations. Overly long statements were rare, as most workers followed the simple examples we provided.

3.2. Rating Explanations

Figure 6 shows our form for rating explanations. Each explanation was rated by five workers. Workers were shown a COPA instance and five explanations to rate with up to five stars. As a control, workers had to rate the first explanation again at the end of the HIT, totaling six ratings per HIT. We disregarded (but did not reject) ratings by workers who had more than a one-star difference in this control.[3] Workers were instructed to give a higher rating to explanations containing relevant and more detailed statements and low ratings to uninformative or nonsensical explanations. We observed that detailed, related statements were also low-rated if they did not explain why the answer is correct. Examples of high-rated and low-rated explanations are shown in Table 2. While these ratings serve as generic estimate of quality, we recommend against using them as measurements of any single characteristic such as relevance or thoroughness since they were not defined as such.

3.3. Re-collection

To increase the number of higher-rated explanations, we invited workers who provided high-quality expla-

---

[2]E.g., A HasSubevent B is shown as A *happens during* B. The text-form explanations retained the original surface form, while in the triple format they are changed back to match ConceptNet.

[3]We allowed a 1-star difference as one could change their opinion on the first seen explanation after seeing other examples. In case of such a difference, we only retain the last rating.

Figure 5: Form for collecting explanations.

Figure 6: Form for rating Explanation.

nations to provide additional explanations for a higher fee. We collected four new explanations for questions that had all five explanations rated below 3.5-stars, two new explanations if one was above this threshold, and one new explanation if two were above this threshold. New explanations were then rated in the same way as the original ones.

3.4. Compensation and qualifications

Workers received $0.30 per explanation in the first collection round and $0.40 in the re-collection round. In the rating rounds, workers received $0.30 for six ratings (five unique and one control). We restricted all our rounds to workers in GB or the US with a HIT approval rate of 98% or more and 500 or more approved HITs.

For re-collection, we invited workers whose explanations averaged more than 3.5 stars over ten or more explanations. The total cost, including Amazon Mechanical Turk fees and excluding trial runs, was $8,651.16.

3.5. Post-processing: Aggregation

Free-form nodes occasionally contain very similar concepts expressed with different surface forms without being explicitly connected. Multiple explanations may also offer diverse information which, combined, results in a higher-quality explanation graph in terms of coverage. To aggregate the explanations, we scored the similarity between each node and merged similar nodes or connected them with a RelatedTo edge. Specifically, we computed the cosine similarity s of the node

The woman sensed a pleasant smell. Effect? ✓ She was reminded of her childhood.	
★ ★ ★ ★ ★	Pleasant smell is a way of bring happiness. Happiness causes nostalgia. Nostalgia is related to a smell. Smell causes her to think her childhood.
The flashlight was dead. Effect? ✓ I replaced the batteries.	
★ ★ ★ ★ ★	Batteries is used for flashlights. Power is created by batteries. Replacing batteries is a way of restoring power.
The car looked filthy. Effect? ✓ The owner took it to the car wash.	
★ ★ ★ ★ ★	The owner desires clean car. Car wash is used for washing cars.
My favorite song came on the radio. Effect? ✓ I sang along to it.	
★	This is a symbol of simple.
The rain subsided. Effect? ✓ I went for a walk.	
★	The rain has a fresh smell.
The girl was not lonely anymore. Cause? ✓ She made a new friend.	
★	Making is motivated by loneliness.

Table 2: Examples of top-rated and bottom-rated explanations. Highly rated explanations tend to be detailed and explicitly connect the question and answer. Low rated ones are incoherent, completely irrelevant, or related facts but irrelevant as an explanation.

texts using Sentence-BERT (Reimers and Gurevych, 2019) and merged if $s > 0.85$ or connected if $0.60 > s \geq 0.85$. The thresholds were manually determined by the authors with respect to the scores and resulting graphs.[4] Each edge also includes a weight calculated as the sum of average human ratings of the explanation the edge came from. Intuitively, these can be considered as the importance or relevance of the edge according to humans, at least in relation to all other given explanations for the sample. Post-processed versions of the graphs are also available in the repository.

4. Experiment

One use case of COPA-SSE is the creation of systems that automatically score explanations. To demonstrate, we present baseline results on the task of outputting a quality rating given an explanation and, optionally, the question and the correct answer as additional context. We evaluated the performance by measuring the Pearson correlation coefficient with human ratings and compared fine-tuned T5 (Roberts et al., 2019) implementations of various sizes ranging from 60M to 11B parameters. Each was tested with the following input format: "*Rate this explanation:* {premise} *so/because* {correct_answer} *Explanation:* {explanation}" if including the QA context, and "*Rate this explanation:* {explanation}" otherwise. [5] In both cases, the expected

output is the rating as a decimal number. We followed the original Balanced COPA split and used the explanations for the original development questions, setting aside 5% for validation, as training data, and explanations for the test questions as test data.

The results are shown in Table 3. Each value is the average Pearson coefficient over three runs. Overall, correlation with human ratings increased with the size of the model and was generally higher when providing the QA context. However, even the best performing setting only reached a moderate correlation of 0.58. This shows the potential of future explanation scoring systems trained with human-scored explanation data, with still much room for improvement.

5. Discussion

Outlook: COPA-SSE as a Resource for Commonsense Reasoners. We created this dataset with several uses in mind: it can serve as training data for (textual) explanation generation models, or as representations of "ideal" subgraphs to use as gold data for graph-based reasoners or to compare with existing KGs. COPA-SSE's textual explanations can be used to improve language model (LM)-based systems such as Commonsense Auto-Generated Explanations (CAGE) (Rajani et al., 2019), the system CoS-E was first intended for, which uses a LM to generate explanations as an intermediate step during training and inference. Its triple-like format can in turn be useful for improving graph-based reasoning systems such as QA-GNN (Yasunaga et al., 2021). While still outperformed by current state-of-the-art systems, graph-based systems have the benefit of being more interpretable than purely LM-based systems due to having an accessible internal

---

[4] For example, *"sun"* and *"under the sun"* are connected ($s = 0.76$), *"shadow"* and *"shadows"* are merged ($s = 0.93$).

[5] COPA asks for the cause or the effect of a premise. An example input is as follows: "Rate this explanation: *My body cast a shadow over the grass.* because *The sun was rising.* Explanation: *Sunrise causes casted shadows.*" The gold rating for this explanation is 3.6 (out of 5).

Model	Size	Expl. only	Expl.+QA context
T5-Small	60M	0.322	0.190
T5-Base	220M	0.476	0.504
T5-Large	770M	0.535	0.556
T5-3B	3B	0.530	0.569
T5-11B	11B	0.515	**0.576**

Table 3: Pearson's correlation between human ratings of COPA-SSE explanations and ratings outputted by fine-tuned T5 implementations of various sizes (num. parameters) given the same explanations. "Expl.+QA context" are the results when the models received both the explanation and the QA context as the input, and "Expl. only" when the input only included explanations. All values are averages over three runs.

reasoning structure. At the time of writing, the top-performing graph-based system is QA-GNN (Yasunaga et al., 2021), a system combining LMs and GNNs to extract and weigh relevant knowledge from a KG, then perform reasoning over the extracted subgraph. Our aggregated graph-form explanations (§3.5) can be considered as idealized versions of relevant subgraphs, thus offering gold examples for improving the extraction and relevance scoring steps in such a system. Even though the explanations are in a similar format, their degree of freedom made it possible to collect new information that might not have been present in Concept-Net. We intend to further explore this direction with the primary goal of steering graph-based systems into being more interpretable.

6. Conclusion

We introduced a new crowdsourced dataset of explanations for Balanced COPA in a triple-based format intended for advancing graph-based QA systems and clear comparison with existing commonsense KGs. The dataset provides relevant and minimal information needed to bridge the question and answer. Our dataset includes explanations in text form and raw triple form as written by crowdworkers, and post-processed versions with similar nodes being merged or connected. This dataset can serve to improve explanation generation in text-based or graph-based approaches.

7. Acknowledgements

8. Bibliographical References

Bowman, S. R. and Dahl, G. (2021). What will it take to fix benchmarking in natural language understanding? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4843–4855, Online, June. Association for Computational Linguistics.

Daniel, F., Kucherbaev, P., Cappiello, C., Benatallah, B., and Allahbakhsh, M. (2018). Quality control in crowdsourcing. *ACM Computing Surveys*, 51(1):1–40, Apr.

Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S., and Smith, N. A. (2018). Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana, June. Association for Computational Linguistics.

Hwang, J. D., Bhagavatula, C., Bras, R. L., Da, J., Sakaguchi, K., Bosselut, A., and Choi, Y. (2021). COMET-ATOMIC 2020: On symbolic and neural commonsense knowledge graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38.

Rajani, N. F., McCann, B., Xiong, C., and Socher, R. (2019). Explain yourself! Leveraging language models for commonsense reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4932–4942.

Reimers, N. and Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11.

Roberts, A., Raffel, C., Lee, K., Matena, M., Shazeer, N., Liu, P. J., Narang, S., Li, W., and Zhou, Y. (2019). Exploring the limits of transfer learning with a unified text-to-text transformer. Technical report, Google.

Rogers, A., Gardner, M., and Augenstein, I. (2021). QA dataset explosion: A taxonomy of nlp resources for question answering and reading comprehension. *arXiv e-prints*, pages arXiv–2107.

Saha, S., Yadav, P., Bauer, L., and Bansal, M. (2021). ExplaGraphs: An explanation graph generation task for structured commonsense reasoning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7716–7740, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguis-

tics.

Yasunaga, M., Ren, H., Bosselut, A., Liang, P., and Leskovec, J. (2021). QA-GNN: Reasoning with language models and knowledge graphs for question answering. In *North American Chapter of the Association for Computational Linguistics (NAACL)*.

9. Language Resource References

Kavumba, P., Inoue, N., Heinzerling, B., Singh, K., Reisert, P., and Inui, K. (2019). When choosing plausible alternatives, Clever Hans can be clever. In *Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing*, pages 33–42, Hong Kong, China, November. Association for Computational Linguistics.

Roemmele, M., Bejan, C. A., and Gordon, A. S. (2011). Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *2011 AAAI Spring Symposium Series*.

Speer, R., Chin, J., and Havasi, C. (2017). ConceptNet 5.5: An open multilingual graph of general knowledge.