← Back to **the profile of Changzhi Sun** (/profile?id=~Changzhi_Sun1)

E-KAR: A Benchmark for Rationalizing Natural Language Analogical Reasoning **(/pdf?id=9kXOFRtrEj)**

Anonymous

17 Nov 2021 (modified: 14 Jan 2022) ACL ARR 2021 November Blind Submission Readers: Everyone

Abstract: The ability to recognize analogies is fundamental to human cognition. Existing benchmarks to test word analogy does not reveal the underneath process of analogical reasoning of neur models. Holding the belief that models capable of reasoning should be right for the right reasons, we propose a first-of-its-kind Explainable Knowledge-intensive Analogical Reasoning benchmark (KAR). Our benchmark consists of 1,665 problems sourced from the Civil Service Exams, which require intensive background knowledge to solve. Besides, we design a free-text explanation scheme t explain how an analogy is drawn, and manually annotate E-KAR with 8,325 knowledge-rich sentences of such explanations. Empirical results suggest that this benchmark is very challenging to som state-of-the-art models for both explanation generation and analogical question answering tasks, which invites further research in this area.

Revealed to Jiangjie Chen, Rui Xu, Ziquan Fu, Wei Shi, Zhongqiao Li, Xinbo Zhang, Changzhi Sun, Lei Li, Yanghua Xiao, Hao Zhou

16 Nov 2021 (modified: 16 Nov 2021) ACL ARR 2021 November Submission

Authors: *Jiangjie Chen (/profile?id=~Jiangjie_Chen1), Rui Xu (/profile?id=~Rui_Xu6), Ziquan Fu (/profile?id=~Ziquan_Fu1), Wei Shi (/profile?id=~Wei_Shi6), Zhongqiao Li (/profile?id=~Zhongqiao_Li1), Xinbo Zhang (/profile?id=~Xinbo_Zhang1), Changzhi Sun (/profile?id=~Changzhi_Sun1), Lei Li (/profile?id=~Lei_Li11), Yanghua Xiao (/profile?id=~Yanghua_Xia Hao Zhou (/profile?id=~Hao_Zhou5)*

TL;DR: We introduce a first-of-its-kind explainable analogical reasoning benchmark.

Data: zip (/attachment?id=xBphQ9ugrK&name=data)

Preprint: yes

Preferred Venue: ACL 2022

Consent: yes

Consent To Review: yes

Reply Type: all Author: everybody Visible To: all readers Hidden From: nobody **5 R**

[−] Supplementary Materials by Program Chairs

ACL ARR 2021 November Program Chairs

14 Jan 2022 ACL ARR 2021 November Paper2673 Supplementary Materials Readers: Program Chairs, Paper2673 Reviewers, Paper2673 Authors, Paper2673 Area Chairs

Data: zip (/attachment?id=lGCmsx9Pasz&name=data)

Note From EiCs: These are the confidential supplementary materials of the submission. If you see no entries in this comment, this means there haven't been submitted any.

[−] Meta Review of Paper2673 by Area Chair 4i8j

ACL ARR 2021 November Paper2673 Area Chair 4i8j

08 Jan 2022 (modified: 10 Jan 2022) ACL ARR 2021 November Paper2673 Meta Review Readers: Paper2673 Senior Area Chairs, Paper2673 Area Chairs, Paper2673 Authors, Paper2673 Reviewers, Program Chairs

Metareview:

- This paper proposes a new knowledge-intensive analogical reasoning benchmark.
    - The dataset is manually augmented with free-text explanations to rationalize how an analogy is drawn.
    - Explanations are collected carefully by abducting a suitable structure for the query and then mapping that onto each answer candidate.
    - All answers except the correct one should fail at certain structures.
    - The structure mapping process is then verbalized into NL to explain the analogical reasoning.
- Using this dataset, the authors proposed two tasks:
    - (i) multiple-choice analogical QA, and a more challenging
    - (ii) explanation/rationalization generation (for analogical reasoning task) task.
- The performance of several baselines (including PLMs) and insightful analysis are reported, warranting further research in this domain.

Overall, based on the three reviews:

- The paper is well written and easy to understand.
- The task itself is very interesting, well-motivated, and offers a new difficult challenge to the NLP community.
- The experimental part and the analysis of the result are satisfactory.
- I believe that can be a good contribution to ACL/NAACL.

Summary Of Reasons To Publish:

- A knowledge-intensive analogical reasoning benchmark (best model scores an accuracy of 50%, the human baseline results in 78% accuracy on i)). Mastering the proposed challenge requires the incorporation of background knowledge and the ability to rationalize predictions via explanations. This dataset is can be helpful for the community and facilitates future research.
- The dataset additionally contains free-text human written rationales, which are collected carefully with multiple quality checking steps. These explanations make the dataset more unique of its kind. The quality of the dataset is also shown by significant improvement in the QA task (when augmented with rationales).
- Insightful analysis and experiments, showing directions for future works. The authors carry out a detailed error analysis, e.g., identifying models that struggle to generate negated facts.
- The paper is well structured and easy to follow.

Summary Of Suggested Revisions:

- The models evaluated on i) differ from the model evaluated on ii). For i) the authors test pre-trained methods Word2Vec, GloVe, FastText, BERT, RoBERTa adn fine-tuned methods BERT and RoBERTA but for ii) they evaluate the Seq2Seq models BART and T5. Therefore, the explanations are not justifying the model's predictions. I suggest the authors either justify this or revise their models.
- wA4q wants to see more clarity on "knowledge-intensity" (what aspect of the dataset entails this property?)
- There are some questions regarding the language of the language. It seems to me that you are only evaluating one language (Chinese?) even though the dataset contains both English and Chinese. Why not both? (is the data translated from Chinese to English?)
- Would be good discuss if there are any direct/indirect extrinsic applications for the proposed setup.

Overall Assessment: 4 = There are minor points that may be revised

Suggested Venues:
NAACL or ACL

[−] Official Review of Paper2673 by Reviewer XdfW

ACL ARR 2021 November Paper2673 Reviewer XdfW

28 Dec 2021 (modified: 28 Dec 2021) ACL ARR 2021 November Paper2673 Official Review Readers: Program Chairs, Paper2673 Senior Area Chairs, Paper2673 Area Chairs, Paper2673 Reviewers, Paper2673 Authors

Paper Summary:
The paper proposes a new benchmark for the task of analogical reasoning, framing it as two tasks over a new corpus. The tasks consist of a multiple-choice question answering and the generation of explanations related to the query and each candidate answer. In the experimental part, the benchmark is used to test a set of neural language models, whose performances are not satisfactory and highlight the difficulty of this benchmark.

The paper is well written and easy to understand. The background section provides all the necessary material to understand and contextualize the task. The task itself is very interesting, well-motivated, and offers a new difficult challenge to the NLP community. The experimental part and the analysis of the result are satisfactory.

The dataset creation process seems to have some weak points but it is still acceptable if some additional details are provided. It is not clear whether the authors plan to release this dataset publicly or whether it will remain private.

Summary Of Strengths:
The task is well designed and motivated. It provides a difficult challenge that sets a high bar for current models and that will be very useful to evaluate future NLP-based reasoning systems.

Summary Of Weaknesses:
Additional details regarding the creation of the dataset would be helpful to solve some doubts regarding its robustness. It is not stated whether the dataset will be publicly released.

Comments, Suggestions And Typos:

1. Additional reference regarding explainable NLP Datasets: "Detecting and explaining unfairness in consumer contracts through memory networks" (Ruggeri et al 2021)
2. Some aspects of the creation of the dataset are unclear and the authors must address them. First of all, will the author release the dataset or will it remain private? Are the guidelines used to train the annotators publicly available? Having a single person responsible for the check at the end of the first round may introduce biases. A better practice would be to have more than one checker for each problem, at least on a subset of the corpus, to measure the agreement between them and, in case of need, adjust the guidelines. It is not clear how many problems are examined during the second round and the agreement between the authors is not reported. It is not clear what is meant by "accuracy" during the annotation stages.
3. Additional metrics that may be used to evaluate text generation: METEOR (http://dx.doi.org/10.3115/v1/W14-3348 (http://dx.doi.org/10.3115/v1/W14-3348)), SIM(ile) (http://dx.doi.org/10.18653/v1/P19-1427 (http://dx.doi.org/10.18653/v1/P19-1427)).
4. Why have the authors decided to use the colon symbol rather than a more original and less common symbol? Since the colon has usually a different meaning in natural language, do they think it may have an impact?
5. How much are these problems language-dependent? Meaning, if these problems were perfectly translated into another language, would they remain valid? What about the R4 category? Additional comments about these aspects would be beneficial for future works, cross-lingual transfers, and multi-lingual settings.
6. In Table 3, it is not clear whether the line with +epsilon refers to the human performance when the gold explanation is available or to the roberta performance when the golden explanation is available? In any case, both of these two settings would be interesting to know, so I suggest, if it is possible, to include them in the comparison if it is possible.
7. The explanation that must be generated for the query, the correct answer, and the incorrect answers could be slightly different. Indeed, if I am not making a mistake, the explanation for the incorrect answer must highlight the differences w.r.t. the query, while the explanation for the answer must highlight the similarity. It would be interesting to analyze these three categories separately and see whether if there are differences in the models' performances.

Overall Assessment: 4 = Strong: This paper is of significant interest (for broad or narrow sub-communities), and warrants acceptance in a top-tier *ACL venue if space allows.

Confidence: 3 = Pretty sure, but there's a chance I missed something. Although I have a good feel for this area in general, I did not carefully check the paper's details, e.g., the math or experimental design.

Best Paper: No

Replicability: 2 = They would be hard pressed to reproduce the results: The contribution depends on data that are simply not available outside the author's institution or consortium and/or not enough details are provided.

Datasets: 5 = Enabling: The newly released datasets should affect other people's choice of research or development projects to undertake.

Software: 1 = No usable software released.

Author Identity Guess: 3 = From the contents of the submission itself, I know/can guess at least one author's name.

[−] Official Review of Paper2673 by Reviewer wA4q

ACL ARR 2021 November Paper2673 Reviewer wA4q

28 Dec 2021 (modified: 28 Dec 2021) ACL ARR 2021 November Paper2673 Official Review Readers: Program Chairs, Paper2673 Senior Area Chairs, Paper2673 Area Chairs, Paper2673 Reviewers, Paper2673 Authors

Paper Summary:

The authors propose a new benchmark for analogy testing in neural models that reveals the underlying process of analogical reasoning: "Explainable Knowledge-intensive Analogical Reasoning" (E-KAR). Mastering the proposed challenge requires the incorporation of background knowledge and the ability to rationalize predictions via explanations. The dataset consists of 1,665 problems sourced from Chinese Civil Service Exams. Building on psycho-linguistic literature that suggests that analogical reasoning follows structure-mapping (a suitable structure of a source query has to be mapped onto a target), the authors developed a free-text explanation scheme. Following this schema, the authors annotated the dataset with 8,325 free-text explanations via crowed sourcing. The benchmark consists of two tasks: i) Analogical QA where analogy testing is formulated as multiple-choice QA: Given a query tuple following some structure, the model has to select from a list of candidate tuples most analogous to the query. ii) Explanation generation where the model has to justify the choice in i) by verbalizing the identified structure and justifying the choice of candidates.

Summary Of Strengths:

- The authors contribute a challenging dataset (best model scores an accuracy of 50%, the human baseline results in 78% accuracy on i)).
- The dataset was crowed-sources with quality control and a human baseline is reported.
- Prior work only tests two term analogy, 35% of the proposed dataset test three term analogy.
- The authors also evaluate static methods which is a valuable baseline to have.
- The authors carry out a detailed error analysis, e.g., identifying models struggle to generate negated facts.
- The authors acknowledge the issue in automatically evaluating generated explanations by carrying out a small scale manual evaluation of explanations.
- The paper is well structured and easy to follow.

Summary Of Weaknesses:

- The models evaluated on i) differ from the model evaluated on ii). For i) the authors test pre-trained methods Word2Vec, GloVe, FastText, BERT, RoBERTa adn fine-tuned methods BERT and RoBERTA but for ii) they evaluate the Seq2Seq models BART and T5. Therefore, the explanations are not justifying the model's predictions. The authors do not give any justification for this mismatch.
- It is not clear to me what part of the dataset construction process insures knowledge-intensity.
- The language of the dataset is unclear to me: The attached dataset consists of English and Chinese samples. In the Methods section you list only English versions of pre-trained language models and in the appendix you refer to the Chinese versions.The examples given in the paper are mixed English and Chinese, e.g., Table 5. It seems to me that you are only evaluating one language (Chinese?) why not both? And how is the dataset translated?

Comments, Suggestions And Typos:

- Line 003: "does" --> do
- You might want to cite related work about negated facts: https://arxiv.org/abs/1907.13528 (https://arxiv.org/abs/1907.13528), https://arxiv.org/abs/2006.10413 (https://arxiv.org/abs/2006.10413), https://arxiv.org/abs/1911.03343 (https://arxiv.org/abs/1911.03343)
- Line 417: Is the fine-tuning also done for the variation using background knowledge?
- Why are you not evaluating seq2seq models on the task of analogy QA? I would strongly advocate for adding this in case of acceptance as this offers the ability to test the same model on analogy QA and explanation generation and see whether the quality of explanations correlates with correctness.
- Line 578: Is it a total of 3 students only?
- Why is the attached data only including 50 samples and not the whole dataset? It would have been reassuring to have a look at the full dataset.

Overall Assessment: 3.5

Confidence: 3 = Pretty sure, but there's a chance I missed something. Although I have a good feel for this area in general, I did not carefully check the paper's details, e.g., the math or experimental design.

Best Paper: No

Replicability: 5 = They could easily reproduce the results.

Datasets: 4 = Useful: I would recommend the new datasets to other researchers or developers for their ongoing work.

Software: 3 = Potentially useful: Someone might find the new software useful for their work.

Author Identity Guess: 1 = I do not have even an educated guess about author identity.

[−]

Official Review of Paper2673 by Reviewer LG6Q

ACL ARR 2021 November Paper2673 Reviewer LG6Q

24 Dec 2021 (modified: 28 Dec 2021) ACL ARR 2021 November Paper2673 Official Review Readers: Program Chairs, Paper2673 Senior Area Chairs, Paper2673 Area Chairs, Paper2673 Reviewers, Paper2673 Authors

Paper Summary:

This paper proposes a new knowledge-intensive analogical reasoning benchmark (an instance of word analogy recognition). The dataset is manually augmented with free-text explanations to rationalize how an analogy is drawn. Explanations are collected carefully by abducting a suitable structure for the query and then mapping that onto each answer candidate. All answers except the correct one should fail at certain structures. The structure mapping process is then verbalized into NL to explain the analogical reasoning. Using this dataset, the authors proposed two tasks: (i) multiple choice analogical QA, and a more challenging (ii) explanation generation (for analogical reasoning task) task. The performance of several baselines (including PLMs) and insightful analysis are reported, warranting further research in this domain.

Summary Of Strengths:

- A knowledge intensive analogical reasoning benchmark which is more challenging than traditional word analogy recognition setting. This dataset is definitely helpful for the community and facilitates future research.
- The dataset additionally contains free-text human written rationales, which is collected carefully with multiple quality checking steps. These explanations make the dataset more unique of its kind. The quality of the dataset is also shown by significant improvement in the QA task (when augmented with rationales).
- Insightful analysis and experiments, showing directions for future works.

Summary Of Weaknesses:

- Lack of extrinsic evaluation for explanation generation task (see comments)
- Simple and naive way of retrieving and incorporating knowledge, which made the collected knowledge corpus pointless.

Comments, Suggestions And Typos:

1. The language of the dataset is never stated in the paper explicitly. Is it a chinese dataset?
2. The author should better describe how their proposed knowledge-intensive analogy reasoning task is different from other traditional analogy recognition. Why is this benchmark the first of its kind? What aspects of it are different? This information should be clear from the introduction of the paper.
3. Line 207-208: Target domain and target problem can be confusing to readers. Maybe replacing one of them with another term?
4. Line 264-266: Not a big issue but “knowing tide is caused by solar/lunar gravity” is not an instance of commonsense knowledge but more of factual world knowledge.
5. Line 341 & Line 348 [suggestion]: In Section 4.3 you first introduced EG (the reader expects to get more details of EG first), but then you started

describing analogical QA as Task 1. It's more fluent/natural to have a consistent order.

6. Line 397: It's not clear whether the author is taking the difference between the sum of term pairs in query/candidate?
7. Line 408: do A,B,C, D refer to candidate answers? An example would be helpful for the prompt used in fine-tuning base models.
8. Line 415: Why choose the first sentence? the author should have explanations for the design choices.
9. Lots of details are missing in the description of baselines. For example, the prompt for the EG task is not clear. The author should have used an example of input to show how the models are trained/fine-tuned! Is the source structure part of the model output? or only the free-form text?
10. In the EG task, are the models aware of the gold label? In other words is it a post-hoc or ad-hoc rationalization?
11. Did the author try a variant where the EG model (possibly a separate model) is used to generate an explanation for the Query and then use that to ground a second model to generate E for each A candidate? This way the second model has access to the abductive source structure (As the author mentioned in section 4.3) to help it better generate reasonable E for candidates.
12. Line 443: It would be more natural to first answer Q3. This could help in designing methods that are better capable of analogical reasoning by mimicking the human process in solving the task.
13. Line 499: Doesn't this observation suggest that the author should have used a more smart way of retrieving relevant knowledge (or multiple of them)? e.g can the model first recognize the relationship between terms and then retrieve relevant knowledge accordingly? Otherwise the knowledge corpus wouldn't be useful? If there is any point in collecting such a knowledge corpus the author should have shown that through experiments.
14. Line 518: It's not evenly bad when R2 is the worst.
15. Line 521 (fig. 3b): btw, are these results normalized? As R2 happens to be the most common relation in the test set as well.
16. Table 4: why not including T5-large? Line 557: It’s very surprising that the models are not picking up these spurious correlations on negated clauses as you did not do anything to prevent it. Did the author report “90% gold explanations contain negated fact” using the same approach for removing them in the ablations? In other words Does the 90% only include explanations with negated clauses and not other cues such as contradiction discourse markers like but, whereas, etc. ?
17. Line 561: Where are the results of error analysis for EG tasks reported? What are the criterias for manual inspections? These info should be either stated in the paper or supplementary material.
18. It would have been interesting to have an extrinsic evaluation of generated explanations, by using them in the QA task (just as you did with gold explanations). This will show the potentials/quality of generated explanations. * Assuming that the EG model did not have access to gold label (which is btw not clear from the text and should be clarified)
19. Table 6: Observing these results, it's interesting to explore if QA model shows the same trend in getting help from generated E. compared to gold and knowledge (which is already included in Table 3)
20. Line 598: this means that although more than 54% of retrieved knowledge were found useful by humans, it hurt the QA model (as shown in Table 3). Does that suggest the issue is on how to incorporate these additional knowledge into QA?

Overall Assessment: 3.5

Confidence: 4 = Quite sure. I tried to check the important points carefully. It's unlikely, though conceivable, that I missed something that should affect my ratings.

Best Paper: No

Replicability: 3 = They could reproduce the results with some difficulty. The settings of parameters are underspecified or subjectively determined, and/or the training/evaluation data are not widely available.

Datasets: 4 = Useful: I would recommend the new datasets to other researchers or developers for their ongoing work.

Software: 1 = No usable software released.

Author Identity Guess: 1 = I do not have even an educated guess about author identity.