Commonsense Question Answering using Hybrid Models for Efficiency and Explainability

Master Thesis Proposal

Gabriel Breiner

TU Wien

gabriel.breiner@protonmail.com

April 14, 2022

Abstract

Enabling machines to apply common sense is on of the major hurdles in AI applicability. In the hopes of alleviating this problem, the field of question answering has been reinvigorated by a number of semantically challenging benchmarks over recent years. Many of them require knowledge that is not directly captured in the provided data and thus encourage aid of external knowledge sources. Most of the SOTA systems use large Knowledge Graphs in addition to their (mainly) word embedding based approaches. This Master Thesis attempts to contribute to the improvement of 2 of the major problems of these approaches: Firstly problems that come with utilization of large knowledge graphs (e.g. storage, computation and curation). Secondly the lack of explainability in the inference steps due to solely relying on representations that favor neural network architectures. We will attempt to do this by implementing a 'hybrid system' utilizing Graph Attention Networks and 4Lang Graphs, and focusing our evaluation on Explainability and Efficiency.

I. Introduction

Question Answering (QA) is a well established problem in NLP, with many available benchmarks. Many of these, especially the older ones, are akin to information retrieval (e.g sQuAD [32])and can be classified as "information-seeking" [38]. However common sense QA is an exception to this with benchmarks dating back to at least 2011 [36] and dicussion around the topic being ignited in 1959 by McCarthy [27]. With the ever increasing efforts to closing the semantic gap and advances in NLU in general [43], recently the focus of QA started to shift to questions requiring common sense again. This kind of knowledge is per definition not available in the training data of the given benchmark, but is rather implied to be common knowledge and as such need not be explicitly stated. Consider this scenario and follow-up question: "Jack needed some money, so he went and shook his piggy bank. He was disappointed when it made no sound. Did Jack find some money?" [28]

As human being, this question is easily answered, but imagine the reasoning that an intelligent agent would have to apply to arrive at our conclusion: Piggy Banks contain money in the form of coins, coins are made out of metal, metal produces a distinct sound. Since the piggy bank did not make a sound, it contains no coins and thus no money. Therefore Jack did not find any money.

Enabling intelligent agents to act on common sense knowledge would be a crucially important step in Natural Language Unterstanding and NLP as a whole. While many of the SOTA systems seem to be solving the aforementioned

tasks with high amounts of accuracy, the actual progress on the subject is not as evident: Most if not all of the top peformers of these benchmarks are Neural Network architectures that tune a huge number of parameters and are based on pre-trained language models like BERT [5], T5 [31] or GPT-* [30]. Models like these have been shown to learn from artifacts [10, 29], non-observable biases in the data that are irrelevant to the task and mostly introduced due to poor sample selection or other external biases (e.g. annotator bias [9]). Additionally most neural networks offer barely any interpretable feedback, which makes it not only hard to prove that they are not using artifacts to achieve spectacular results, but can also be seen as a major obstruction for the scientific process in general. Note that the interpretability issue is less prevalent in systems based on formal logic [3], which often rely on symbolic representations. While these systems tend to not be as common on the benchmark leaderboards as the aforementioned neural models, they are able to explain their decisions in a human-understandble way.
We propose a model that is able to tap into the strengths of systems based on logic and of pre-trained language models by combining them in a "Hybrid System" utilizing BERT* embeddings [5], Graph Attention Networks [45] and 4Lang Graphs [33]

II. Benchmarks

Most samples (questions) provided by benchmarks are triplets of question, answer and context, the latter being the "scenario" the question is asked about (e.g. a story, an article, a conversation, ...). QA tasks often employ a whole suite of different orthogonal characteristics, such a as their accepted form of answer (multiple-choice, extracted text spans, free language) or whether they require external knowledge (facts, that are not explicitly stated in question or context). A taxonomy of QA can be found in Rogers (2021) [38] and Storks (2019) [43]. It is also worth mentioning, that many of the current NLP tasks outside of QA can be parsed to be a QA task (e.g. "What is the summary of <context paragraph>" is a summarization task parsed to QA) [26]. This is why QA intersects with many other NLP tasks such as NLI, reading comprehension or information retrieval.
Definitions for common sense are generally very ambiguous, so defining common sense QA is just as hard. There seems to be an implication, that these are problems requiring a "rich understanding of the world" [4] - hinting at the need for data, that is either implicit or external in relation to a given benchmark. Most benchmarks chose to generate samples by pre-selecting tuples from knowledge graphs, then formulating them with the help of crowdworkers [44, 39]. Additionally samples can be created and selected adversarially, purposefully counteracting large parameter language models - with limited effectiveness [51]).
In order to capture common sense in many of its different domains (temporal, spacia, social, etc.) we opted to tackle a generalist dataset: **commonsenseQA** [44]. Additionally this dataset has been annotated with rationales (spans of explanations in the samples) by a third party [6], enabling us to quantifiably measure of explainability for our systems (see section V). Optionally we also have **QuAIL** [39] - another well annotated generalist commonsense QA dataset - which we will utilize in case we cannot answer any of the research questions sufficiently using our main benchmark.

III. Problem Definition

i. Large Graphs

Current SOTA hybrid models use established knowledge graphs, such as ConceptNet [42] or ATOMIC [40] (see VI for descriptions of some of these models). These are fairly large graphs - ConceptNet for example "contains over 21 million edges and over 8 million nodes" [42]. Processing them can take substantial amounts of computation, especially if the model using it requires costly operations [50, 21] such as path-detection or finding an optimal sub-graph.

Graphs of these magnitudes can pose problems with storage (e.g. IOT, mobile devices, etc.) or might not be obtainable (e.g. low-data settings, uncommon languages, ...)

ii. Lack of Explainability

Many systems utilizing external knowledge in the form of graphs may use them to retrieve additional information about the question, answer or context without using the graphs as working representation, but rather parsing its entities back into vectorized embeddings only (see VI: Graph Informed Models). These systems neglect some potential in regards to explainability, since graph structures inherently provide explainability through observation of the paths used during inference. This lack of explainability passes for many SOTA systems, because leaderboards and challenges barely try to incentivize other metrics than competence (e.g. accuracy) [25, 37].

iii. Proposed Solution

In order to create more compact and dynamic graphs we propose the use of 4Lang graphs. These graphs have already shown to perform well on semantic similarity [34] and lexical entailment tasks [18]. Primarily, they have many advantageous attributes, that distinguish them from established graph knowledge bases such as ATOMIC [40] or ConceptNet [42] (see [33]) 4Lang is actually a formalism that enables semantic graph construction, rather than a single semantic graph. Building graphs at run-time might allow us to create possibilities for inference, that may not be captured by static knowledge graphs. This has the added benefit of allowing us to omit graph-pruning. Furthermore 4Lang enables creation of graphs of concepts rather than of concrete words, which enables it to capture creative and everyday language use [33]. Becase the 4Lang formalism employs no inherent means of inference, but puts a large burden on the reasoning framework utilizing it [33], we propose experimenting with neural reasoning modules (see V).

Using Graph Attention Networks yields two essential benefits: Firstly it allows us to join LM and Graph representations such as BERT* and 4Lang, giving us the opportunity to use 4Lang in combination with other SOTA systems. Secondly they allow us to preserve graph structures as working representations and employ a (neural) reasoning mechanism, that allows us to do explainable inference. Furthermore Graph Attention Networks and other Graph Neural Networks [17] have already been applied succesfully [21, 8, 50] on one of our chosen benchmarks (commonsenseQA) [44].

IV. Research Questions

1. Explainability
   (a) How well can the system describe its inference process?
   (b) Are the explanations comprehensive?
   (c) Are the explanations sufficient?
2. Efficiency
   (a) How much computational effort is needed during training and inference?
   (b) What is the influence of the training size on Explainability and Competence?
3. Comparison
   (a) How does the hybrid system compare to the other systems [1] in the notion of Explainability defined above.
   (b) How does the hybrid system compare to the other systems [1] in the notion of Efficiency defined above.

V. Methodology and Approach

In principal our approach to solving this thesis' problems is to follow the datascience life cycle as outlined by Blitzstein and Pfister [14]: (1) Identify the problem, (2) Get and explore

[1] see section V

the data, (3) Build, fit, evaluate a model, (4) Communicate results. Keep in mind, that these steps are part of a cycle and will be visited iteratively. For a description of the model see sections ii

i. Cycles

While these may be subject to change the initial plan is to iterate development of the following models in order. Since these models (disregarding the baseline models) are building on top of each other modularily, we are able to assess the effectiveness of a module by doing an ablation study.

- **BOW** (baseline)
  A bag of words model using logistic regression (word probabilities taken as explanations)
- **BERT* + LIME** (baseline)
  A neural model purely based on a BERT* embedding (RoBERTa [22], ALBERT [19], etc.) - explained with LIME [35]
- **QA-GNN** (baseline)
  Using the unmodified implementation at (https://github.com/michiyasunaga/qagnn).
- **BERT*less** (target v1)
  Utilizing Graph Attention Networks [45] initialized without BERT* on 4Lang Graphs. [33]
- **Minimal** (target v2)
  Extends *BERT*less* by intializing the working Graph with BERT*
- **Extended** (target v3)
  Extends *Minimal* by making use of 4Lang's expansion mechanism [33, 18]
- **Full** (target v4)
  Extends *Extended* by enriching GAT's Message Passing with more information (e.g. edge and node types - similarly to QA-GNN [50])

ii. Model Description

The goal of any of the described models is to assess the plausiblity of a given answer a of a set of possible answers $A = \{a_1, a2, ...a_n\}$ to being the correct answer to the question Q given the context C. We will call the tuple of Q, C and a_1 the question-answer tuple $t_1 = (Q, C, a_1)$.
Firstly for each tuple t_a we will establish our working graph which is a 4Lang Graph $G_a = (N_a, E_a)$, where N_a is the set of nodes and E_a is the set of edges between them. Depending on the origin of the nodes we will define the node attributes $A = \{a_q, a_c, a_a\}$ to indicate if the node appeared in question, context, answer or any combination of them.
de Embedding e_n. While these representations will be updated with more diverse information (we will call them h_n from then on), we will intialize them using TransE [21]. Additionally we will append a special node z to N_a that is representative of Q as a whole and connected to every other $n \in N_a$. [50] We will initialize this node using BERT* which allows the model to distribute LM information in the next stage.
With Node Representations established we have two essential benefits of using Graph Attention Networks: Firstly we can use Message Passing to update a given nodes representation h_n by updating it with information (the message) from M_n, its neighborhood: $h_n = f(\sum_{m \in M_n}(\alpha_{nm}\mu_{nm})) + h_n$ where f is a Multilayer Perceptron and μ is the message consisting of node-to-node-relevant information (node-type, edge types, etc.) Secondly with the attention weights α, we can learn how strongly nodes are exchanging messages and thus create weighted edges for G_a that can later serve as interpretative artifacts of the inference process.
To assess the plausibility of tuple t_a, we will asses the plausibility of G_a as a whole which is represented by feeding a concatenation of selected properties through a MLP: $p_a = MLP([e_z, h_z, pooling(N_a)])$.
Lastly we apply a Softmaxlayer to choose the most plausible of the Graph representations, optimizing the whole model using cross entropy loss.

iii. Evaluation

As mentioned previously we will evaluate our system on 2 dimensions: Efficiency and Ex-

plainability - viewing all of the corresponding metrics in the context of an ablation study.
Competence will be the "official" measure of evaluating any of the given benchmarks and is measured in accuracy. We will display this value for our models, but we will not optimize their development solely on this metric.
Efficiency can be understood as competence in relation to another metric. In our case we will calculate Floating Point Operations (FPO) to quantify the computational workload of a given model during training and inference. This metric finds increasing adoption (e.g. in Green AI [41]).
Explainability of the system will be judged on two levels: on a qualitative and a functionally grounded level [7] with respect to two aspects of explainability: Comprehensiveness and Sufficiency.
Qualitative: We will study a random set of infered samples and their inference artifacts (graph and weights).
Functionally Grounded: We compute quantifiable metrics from human gold-label rationales and explanations provided by ERASER [6].
Comprehensiveness: We will observe if the calculated weights portray plausible patterns of reasoning to answer the question [7].
Sufficiency: We want to know if the working graph itself provides the right entities and connections to enable satisfying classification [7].
Note that in order to make the BERT* model comparable in terms of Explainability, we will use LIME [35] to create post-hoc explanations.

VI. State of the Art

Progress in common sense question answering can be seperated in essentially 3 areas: resources, approaches and benchmarks [43]. Since we touched on resources and benchmarks already in chapters II and V we will discuss SOTA approaches to commonsense reasoning in regards to the relevant categories of approaches from the perspective of this thesis.
Graph Informed Models use KGs to expand the question context, essentially appending additional terms or concept definitions. These will be transformed into BERT* representations ([5, 22, 19]) and processed by fully neural models [49, 46]. These models achieve high results on the leaderboards.
Graph Embedding Models use KGs to create Embeddings from graphs [1, 20, 17, 42] for concepts or words, but abandon graph structures as their working representations. These representations are either used directly in Neural Models or joined with BERT*. [48, 47, 52]
Hybrid Models - our target model. They operate similar to Graph Embedding Models by applying GNNs ([20, 17, 45], but keep graph structures as working representations in order to operate on path-level or gain interpretable feedback. [24, 21, 8, 50].
Purely **Formal Logic** based systems are rare for the benchmarks mentioned in this paper. Some of them have been applied for other NLI tasks such as Lexical Entailment (e.g. a 4Lang approach [18]). But even they ofteny rely on some form of combination of logic and PTLM. [15, 11, 12]. A more theoretical survey on logic-based commonsense can be found in Davis (2017) [3].
Transfer Learning approaches are not inherently related to our target model, but should be mentioned due to their increasing success on benchmarks [2, 13]. These models train on large amounts of NLI problems that have been rephrased to QA problems and the Common Crawl [31, 16, 23].

VII. Planned Contributions

- Exploring the use of **4Lang Graphs** in QA benchmarks.
- Appyling **Graph Attention Networks with 4Lang Graphs**.
- Exchange large top-down Knowledge Graphs with smaller **bottom-up Semantic Graphs** for hybrid systems.
- Creating an **explainable** system using neural SOTA **LM representations** (BERT*).
- Promoting more diverse means of Evaluation by prioritizing **Explainability and Efficiency**.

References

[1] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Durán, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, page 2787–2795, Red Hook, NY, USA, 2013. Curran Associates Inc.

[2] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge, 2018.

[3] Ernest Davis. Logical formalizations of commonsense reasoning: A survey. *J. Artif. Intell. Res.*, 59:651–723, 2017.

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

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[6] Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. ERASER: A benchmark to evaluate rationalized NLP models. *CoRR*, abs/1911.03429, 2019.

[7] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning, 2017.

[8] Yanlin Feng, Xinyue Chen, Bill Yuchen Lin, Peifeng Wang, Jun Yan, and Xiang Ren. Scalable multi-hop relational reasoning for knowledge-aware question answering. *CoRR*, abs/2005.00646, 2020.

[9] Mor Geva, Yoav Goldberg, and Jonathan Berant. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. *CoRR*, abs/1908.07898, 2019.

[10] Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. Annotation artifacts in natural language inference data, 2018.

[11] Izumi Haruta, Koji Mineshima, and Daisuke Bekki. Combining event semantics and degree semantics for natural language inference. *CoRR*, abs/2011.00961, 2020.

[12] Hai Hu, Qi Chen, Kyle Richardson, Atreyee Mukherjee, Lawrence S. Moss, and Sandra Kuebler. MonaLog: a lightweight system for natural language inference based on monotonicity. In *Proceedings of the Society for Computation in Linguistics 2020*, pages 334–344, New York, New York, January 2020. Association for Computational Linguistics.

[13] Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Cosmos qa: Machine reading comprehension with contextual commonsense reasoning, 2019.

[14] Hanspeter Pfister Joe Blitzstein. Harvard data science course. CS109 Lecture, 2013.

[15] Aikaterini-Lida Kalouli, Richard Crouch, and Valeria de Paiva. Hy-NLI: a hybrid system for natural language inference. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5235–5249, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

[16] Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. Unifiedqa: Crossing format boundaries with a single qa system, 2020.

[17] Thomas N. Kipf and Max Welling. Semi-supervised classification with

graph convolutional networks. *CoRR*, abs/1609.02907, 2016.

[18] Adam Kovacs, Kinga Gemes, Andras Kornai, and Gabor Recski. Explainable lexical entailment with semantic graphs. *Natural Language Engineering*, pages 1–24, 2022.

[19] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations, 2020.

[20] Yujia Li, Daniel Tarlow, Marc Brockschmidt, and Richard Zemel. Gated graph sequence neural networks, 2017.

[21] Bill Yuchen Lin, Xinyue Chen, Jamin Chen, and Xiang Ren. Kagnet: Knowledge-aware graph networks for commonsense reasoning, 2019.

[22] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.

[23] Nicholas Lourie, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Unicorn on rainbow: A universal commonsense reasoning model on a new multitask benchmark. In *AAAI*, 2021.

[24] Shangwen Lv, Daya Guo, Jingjing Xu, Duyu Tang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, and Songlin Hu. Graph-based reasoning over heterogeneous external knowledge for commonsense question answering. *CoRR*, abs/1909.05311, 2019.

[25] Ana Marasović. Nlp's generalization problem, and how researchers are tackling it. *The Gradient*, 2018.

[26] Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. The natural language decathlon: Multitask learning as question answering, 2018.

[27] John McCarthy. Programs with common sense. In *Proceedings of the Teddington Conference on the Mechanization of Thought Processes*, pages 75–91, London, 1959. Her Majesty's Stationary Office.

[28] Marvin Minsky. Commonsense-based interfaces. *Commun. ACM*, 43(8):66–73, aug 2000.

[29] Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. Hypothesis only baselines in natural language inference, 2018.

[30] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[31] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer, 2020.

[32] Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[33] Gábor Recski. Building Concept Definitions from Explanatory Dictionaries. *International Journal of Lexicography*, 31(3):274–311, 05 2017.

[34] Gábor Recski and Judit Ács. Mathlingbudapest: Concept networks for semantic similarity. pages 543–547, 01 2015.

[35] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. *CoRR*, abs/1602.04938, 2016.

[36] Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S. Gordon. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *Logical Formalizations of Commonsense Reasoning, Papers from the 2011 AAAI Spring Symposium, Technical Report SS-11-06, Stanford, California, USA, March 21-23, 2011*. AAAI, 2011.

[37] Anna Rogers. How the transformers broke nlp leaderboards, Jun 2019.

[38] Anna Rogers, Matt Gardner, and Isabelle Augenstein. QA dataset explosion: A taxonomy of NLP resources for question answering and reading comprehension. *CoRR*, abs/2107.12708, 2021.

[39] Anna Rogers, Olga Kovaleva, Matthew Downey, and Anna Rumshisky. Getting closer to ai complete question answering: A set of prerequisite real tasks. *Proceedings of the AAAI Conference on Artificial Intelligence*, (05):8722–8731, Apr. 2020.

[40] Maarten Sap, Ronan LeBras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. Atomic: An atlas of machine commonsense for if-then reasoning, 2019.

[41] Roy Schwartz, Jesse Dodge, Noah A. Smith, and Oren Etzioni. Green AI. *CoRR*, abs/1907.10597, 2019.

[42] Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge, 2018.

[43] Shane Storks, Qiaozi Gao, and Joyce Y. Chai. Commonsense reasoning for natural language understanding: A survey of benchmarks, resources, and approaches. *CoRR*, abs/1904.01172, 2019.

[44] Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. Commonsenseqa: A question answering challenge targeting commonsense knowledge, 2019.

[45] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks, 2018.

[46] Liang Wang, Meng Sun, Wei Zhao, Kewei Shen, and Jingming Liu. Yuanfudao at semeval-2018 task 11: Three-way attention and relational knowledge for commonsense machine comprehension, 2018.

[47] Xiaoyan Wang, Pavan Kapanipathi, Ryan Musa, Mo Yu, Kartik Talamadupula, Ibrahim Abdelaziz, Maria Chang, Achille Fokoue, Bassem Makni, Nicholas Mattei, and Michael Witbrock. Improving natural language inference using external knowledge in the science questions domain. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):7208–7215, Jul. 2019.

[48] Weiwen Xu, Huihui Zhang, Deng Cai, and Wai Lam. Dynamic semantic graph construction and reasoning for explainable multihop science question answering, 2021.

[49] Yichong Xu, Chenguang Zhu, Ruochen Xu, Yang Liu, Michael Zeng, and Xuedong Huang. Fusing context into knowledge graph for commonsense reasoning. *CoRR*, abs/2012.04808, 2020.

[50] Michihiro Yasunaga, Hongyu Ren, Antoine Bosselut, Percy Liang, and Jure Leskovec. Qa-gnn: Reasoning with language models and knowledge graphs for question answering, 2021.

[51] Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. Swag: A large-scale adversarial dataset for grounded commonsense inference, 2018.

[52] Yuyu Zhang, Hanjun Dai, Kamil Toraman, and Le Song. Kg$\hat{2}$: Learning to reason science exam questions with contextual knowledge graph embeddings, 2018.