Commonsense Augmented Question Answering

Chloe Eggleston
University of Massachusetts Amherst
ceggleston@umass.edu

ABSTRACT

Commonsense knowledge encompasses the concepts and relations that are assumed by speakers of a language, and thus rarely stated explicitly. For example, one can read a paper, but not read a sandwich; this is rarely stated by any speaker, but intuitively understood by any listener. From a computational lens however, we cannot directly access or reference these shared societal assumptions about language and concepts; this need has led to the development of knowledge graphs for commonsense reasoning; specifically, the CommonSense Knowledge Graph (CSKG), ATOMIC^{10X}, and ATOMIC$^{20}_{20}$. These knowledge graphs have shown promising results in tasks spanning Information Retrieval and Natural Language Processing, specifically natural language inference and machine reading comprehension (MRC) in particular. In this paper, I plan to examine the utility of these by applying them to various question answering benchmarks.

1 PROBLEM DESCRIPTION

I intend to examine the utility and applicability of commonsense knowledge, specifically in the form of knowledge graphs (KG), for question answering tasks. This will be assessed by through fine-tuning transformer models on KG-generated outputs, which will be evaluated on zero-shot question answering for the datasets.

Commonsense reasoning has been of interest in Information Retrieval [10, 16] and Natural Language Processing [1, 12], along with research initiatives from the Allen Institute for AI and DARPA. However, I have not found a study examining how well they perform on retrieval based question answering systems, as most papers evaluate them only on commonsense-oriented question answering benchmarks.

Commonsense approaches provide value to information retrieval and complement the teachings in COMPSCI 646 by offering means to generate zero shot models, which are helpful due to the large cost of training models on large document collections. They also are explicitly catered to question answering tasks, a topic included in the COMPSCI 646 curriculum.

2 MOTIVATION

The key motivation to study commonsense knowledge is that it allows for a generalizable pretraining method that has reasoning applicable to most question answering tasks. Moreover, the need for robust zero shot question answering models has grown over time, as fine-tuning ever and ever larger transformer models to various tasks are becoming computationally infeasible. The generation of large factual knowledge graphs also has impacts in a variety of IR and NLP settings, with examples including using them for named entity recognition (which is used as a task in NLP and a preprocessing technique in IR). Also, the ability to engrain domain-agnostic knowledge into a model can help for niche tasks with sparse data, as the underlying conceptual relations are likely partially covered by commonsense knowledge.

3 RELATED WORK

This paper will be attempting to expand upon the work of Ma et al. 2021 [12] and will be applying the resources consolidated into CSKG in Ilievski, Szekely, Zhang 2021 [8]. Specifically, I will attempt to recreate their knowledge graph generated fine-tuning approach that instilled commonsense knowledge in their models, and then attempt to use transfer learning to see how well it performs on both commonsense and non-commonsense focused question answering datasets (which is a deviation from the approach in both these papers). While the CSKG paper used only GPT-2 Large and RoBERTa Large for their resources, I also plan to examine transformer models designed in other ways for commonsense tasks such as COMET[4]. These models take a similar approaches on discrete existing knowledge graphs, with a focus on automatic completion of knowledge graph systems, and are not trained on an accumulated knowledge graph like CSKG.

4 RESEARCH QUESTIONS

This paper will investigate three fundamental questions concerning commonsense knowledge graphs and question answering:

(1) Does commonsense fine-tuning impact zero-shot machine reading comprehension performance?
(2) Does commonsense fine-tuning impact zero-shot generative question answering performance?
(3) Does commonsense fine-tuning impact factoid question answering more than non-factoid question answering?

5 DATA

This study will be focused on the following knowledge graphs (see Table 1): ATOMIC[14], ATOMIC^{10X}[20], CSKG[8], and CWWV (a partition of CSKG including Wikidata[19], ConceptNet[17], WordNet[5], and VisualGenome[9]).

Table 1: Commonsense Knowledge Graphs

KG	Triples	Relations
ATOMIC	732,723	9
ATOMIC$^{20}_{20}$	1,331,113	23
ATOMIC^{10X}	6,456,300	7
CSKG	6,001,531	58
CWWV	660,844	14

Relations: Unique types of connections in the Knowledge Graph,
Triples: Knowledge Graph Triplets in the form <node, relation, node>

I will attempt to evaluate the commonsense-enabled zero-shot question answering system based on the following types of question answering datasets (see Table 2):

- Machine Reading Comprehension question answering (CosmosQA [7], aNLI [2], CommonSenseQA 2 [18], and SocialIQA [15])
- Generative question answering (ProtoQA [3])
- Retrieval-oriented question answering (WikiQA [21] and ANTIQUE [6])

Table 2: Question Answering Datasets

Dataset	Task	Answers	Questions	Passages
aNLI	MRC	mc	1,532	n/a
SocialIQA	MRC	mc	1,954	n/a
CSQA 2	MRC	yes/no	2,540	n/a
CosmosQA	MRC	mc	2,985	n/a
ProtoQA	gen	entity	52	n/a
WikiQA	Retr.	passage	296	2,733
ANTIQUE	Retr.	passage	195	403,492

Answers: passage for passage retrieval, mc for multiple choice, yes/no for binary answering, entity for generating an entity not featured in text
Questions, Documents: ANTIQUE (test set), All others (dev set)

6 METHODOLOGY

6.1 Overview

Each of the question answering tasks represented by the datasets in Table 2 will be approached by one of these pretrained transformer models:

- GPT2_{XL} (pretrained model provided with GPT2[13])
- $\text{COMET}^{\text{DIS}}_{\text{TIL}}$ (GPT2_{XL} fine-tuned on ATOMIC^{10X})
- GPT2_{L} (pretrained model provided with GPT2[13])
- GPT2_{L}[AT,CWWV,CSKG] (GPT2_{L} fine-tuned on each)
- $\text{COMET}^{\text{DIS}}_{\text{TIL}}$ (GPT2_{XL} fine-tuned on ATOMIC^{2020})
- $\text{RoBERTa}_{\text{L}}$ (pretrained model provided with RoBERTa[11])
- $\text{RoBERTa}_{\text{L}}$[AT,CWWV,CSKG] ($\text{RoBERTa}_{\text{L}}$ fine-tuned on each)

Each of these models (excluding the FT ones) are available from their corresponding papers in the Huggingface binary format, allowing for less time to be spent worrying about implementation. The MRC tasks will be approached using Masked Language Modeling or Causal Language Modeling scores using an utility function from the CSKG paper, while the retrieval tasks will be approached using BM25Plus on the collection, and then have the top 25 passages from that re-ranked using the same models and scoring system. The generative tasks will be approached using a greedy search with temperature 0.175, length penalty 0.6, max length 20, and a sample of 300 candidates. These candidates will be fed back into the scoring methodology in the previous tasks and re-ranked accordingly.

Note that due to limited resources and time, the full capabilities of the generative and retrieval tasks will not be reached in the zero-shot setting. The goal here is to demonstrate the impact of the commonsense knowledge graph fine tuning on these models, so for example the limited scale of the reranking for the retrieval tasks may not be in line for a larger re-ranker or a full end-to-end system, yet it will still be uniform across the models.

6.2 Evaluation

I will be using the standard evaluation metrics associate with each of the QA tasks:

- WikiQA — MAP, MRR
- ANTIQUE — MAP, MRR, nDCG@{1,3,10}
- aNLI — Accuracy
- CSQA 2 — Accuracy
- ProtoQA — Max Ans.@{1,3,5,10}, Max Inc@{1,3,5}
- CosmosQA — Accuracy
- SocialIQA — Accuracy

MAP, MRR, and nDCG are standard information retrieval metrics, while accuracy is a common natural language processing metric. The Max Ans@k and Max Inc@k are custom metrics for ProtoQA that refer to the number of maximum correct and incorrect answers given k chances to answer correctly.

The entire evaluation scheme will consist of comparing these metrics on $\text{RoBERTa}_{\text{L}}$ and GPT2_{L} to their commonsense fine-tuned equivalents in the zero-shot setting.

7 MACHINE READING COMPREHENSION

This section will demonstrate the performance of all the models across the four machine reading comprehension tasks.

Table 3: MRC Accuracies

Model	**aNLI**	**SocialIQA**	**CSQA 2**	**CosmosQA**
Random	0.5000	0.3333	0.5000	0.2500
GPT2_{XL}	0.5796	0.3961	0.5132	0.3075
$\text{COMET}^{\text{DIS}}_{\text{TIL}}$	0.5601	0.3895	0.5183	0.2848
GPT2_{L}	0.5691	0.4074	0.5150	0.3122
$\text{GPT2}_{\text{CSKG}}$	0.5914	0.4719	0.5012	0.3551
$\text{GPT2}_{\text{ATOM}}$	0.5933	0.4795	0.5161	0.3621
$\text{GPT2}_{\text{CWWV}}$	0.5777	0.4273	0.5138	0.3487
COMET^{20}_{20}	0.5307	0.3777	0.4805	0.2600
RoBERTa_{L}	0.5574	0.4575	0.5193	0.4573
$\text{RoBERTa}_{\text{CSKG}}$	0.7050	0.5420	0.5179	0.4516
$\text{RoBERTa}_{\text{ATOM}}$	0.7167	0.5430	0.5122	0.4372
$\text{RoBERTa}_{\text{CWWV}}$	0.6971	0.4811	0.4783	0.4633

7.1 aNLI

Abductive Natural Language Inference, also known as aNLI, is a task that tests a Question Answering system's ability to understand a narrative and hypothesize why the events are occuring after each other. It presents the model with a circumstance, such as "It was a gorgeous day outside", and a later outcome like "She asked her neighbor for a jump-start" and asks the model which of the underyling causes are more likely to lead to the second observed event, here giving "Mary decided to drive to the beach, but her car would not start due to a dead battery", and "It made a weird sound upon starting".

The task was implemented by arranging the first event in the narrative, then appending either hypothesis, then ending it with the second event in the narrative. These were compared with the aforementioned language modeling scores, and then the one with the least loss was selected as the result.

This task showed (like all others in the category) that the RoBERTa based models performed better than the GPT2 based models, likely due to masked language modelling of the base ROBERTa model. Moreover, in this task specifically, the difference in the baselines and the commonsense knowledge graph finetuned models were most pronounced. This is likely due to the event oriented concepts engrained in ATOMIC and FrameNet in particular, which likely covered significantly more word senses and argument structures than the baseline RoBERTa training alone.

7.2 SocialIQA

SocialIQA is a task that seeks to test a Question Answering system's ability to infer commonsense reasoning concerned with social interactions. It prompts the system with a social interaction, asks them a question about it, and gives them three possible conclusions. The focus on social reasoning was inspired by the ATOMIC knowledge graph, yet it was generated independently through crowdsourcing.

The task was implemented through appending together the context, the question, and each answer independently, then taking the selection with the minimal language modelling score loss like mentioned before. Similarly to the last section, the coverage of social reasoning inherent to ATOMIC, and CSKG which includes ATOMIC seems to have significant impact on performance.

7.3 CommonSense QA 2

CommonSense QA 2 is a task that seeks to test a Question Answering system's ability to recognize commonsense knowledge and distinguish it from falsehoods. It is structured as a list of statements, each being either true or false. It requires the model to answer yes or no to each of these statements. Most of the entities mentioned within it originate from ConceptNet, and the falsifications were generated through a gamified crowdsourcing experiment.

The task was implemented through appending the string "It is true that: ", or "It is not true that: " with the statement. The lack of any language to perform the language modelling score on was a significant challenge in this task, and as a result, these strings were used as a temporary replacement. However, possibly since they compose few function words, they did not provide the model much to distinguish each one from the other. As a result, this was the least performant of the machine reading comprehension tasks, with some models dropping even below random chance.

7.4 CosmosQA

The last of the four machine reading comprehension tasks is CosmosQA, a question answering task concerned with contextual reading comprehension. It prompts the system with a paragraph of context and a question concerning the paragraph. The system has to select which one of four answers best answers the question in context.

The implementation for this task involved the same methods of appending the context, question, and candidate answer together, and minimizing over our scoring function loss. The point of interest that distinguishes this task from the others, however, is the significantly longer context passage for this task, which likely gave more information for the baseline models and commonsense models alike. Along with this reasoning, the performance on this task was rather uniform, or sometimes worse than the baselines, which likely is differences between pretraining on synthesized examples from a knowledge graph rather than human generated natural language.

8 GENERATIVE QUESTION ANSWERING

This section will demonstrate the performance of the generative models across the generative commonsense reasoning task - ProtoQA.

8.1 ProtoQA

Table 4: ProtoQA Evaluation Metrics

Model	A_1	A_3	A_5	A_{10}	I_1	I_3	I_5
GPT2$_{XL}$	14.54	19.88	21.04	22.96	15.33	20.38	21.54
COMET$^{DIS}_{TIL}$	3.77	9.67	12.94	14.25	3.77	9.81	13.62
GPT2$_{L}$	13.75	19.35	20.60	23.94	14.23	20.23	22.15
GPT2$_{CSKG}$	16.25	18.54	19.27	20.73	17.67	19.75	20.65
GPT2$_{ATOM}$	6.75	15.10	17.13	20.37	7.79	16.98	19.31
GPT2$_{CWWV}$	8.35	10.88	11.88	14.00	9.98	12.15	12.73
COMET$^{20}_{20}$	9.29	13.02	14.29	17.32	9.5	15.33	16.87

ProtoQA is a task that seeks to test a Natural Language Generation system's ability to understand the understanding a scenario and present a prototypical object that embodies the question being asked. This is structured as the gameshow Family Feud, in which a prompt like "Name a cause you are likely to donate to" is responded to by any free text entity. This text is tokenized, stopword filtered, and scored using the lemmatization and hyponyming in WordNet. It includes two core evaluation metrics: answers@k (max points awarded with k attempts), and incorrect answers@k (max points awarded until the system gives k incorrect attempts). These will be refrred to as A_k and I_k respectively in this paper.

As mentioned in Metholodgy, this task's performance was limited by the requirement to rank the samples generated by our generative language models, and because of this, many higher scoring answers were not included in the results since they were outside of the top 10, or after 10 incorrect answers. Further experimenting with RoBERTa based re-ranking system for the generated text is likely bound to fix this discrepancy. Regardless, the advantage of the causal language modelling on natural language is likely the reasoning why fine-tuning on that model with the commonsense synthetic data was ill-fitted for this task.

9 RETRIEVAL ORIENTED QUESTION ANSWERING

This section will demonstrate the performance of the RoBERTa based models across the two retrieval oriented question answering tasks.

9.1 WikiQA

WikiQA is an open domain retrieval oriented question answering task generated from Wikipedia pages. It is composed of prompts for questions with answers in passages from documents on Wikipedia. Likewise being generated from an encyclopedic resource, it is primarily oriented towards factoid question answering.

Table 5: WikiQA Evaluation Metrics

Model	MRR_{Psg}	MAP_{Psg}	MRR_{Doc}	MAP_{Doc}
Hier. BM25+	0.5693	0.5613	0.9056	0.9255
RoBERTa$_L$	0.3930	0.3842	0.7490	0.6152
RoBERTa$_L$[CSKG]	0.5284	0.5208	0.8825	0.7509
RoBERTa$_L$[ATOM]	0.4973	0.4887	0.8221	0.7125
RoBERTa$_L$[CWWV]	0.3966	0.3899	0.7348	0.5912

It was approached using a BM25+ document ranker, whose passages were individually ranked using the same ranked on themselves as a smaller collection. This, which will be referred to as Hierarchical BM25+, was later re-ranked with the RoBERTa based models based on the top 25 elements, with the remaining elements remaining static.

The impact of the re-ranking on the results is surprisingly bad, although it can be inferred that the reason why factoid question answering would be less in need of commonsense fine-tuning would be that it already includes the finely grained factoid nature that the commonsense KG imparts on the model in its fine-tuning.

9.2 ANTIQUE

In constract to WikiQA, ANTIQUE is a non-factoid oriented question answering task, mostly concerned with questions including complex answers, such as "why" and "how" questions. Since it was already produced with passages in a TREC format rather than joined into documents, the retrieval was done on a passage scale. The included levels of relevance from 1 to 4 allowed for nDCG to become an applicable benchmark as well.

The implementation similarly followed the BM25+ reranking methods mentioned above, however, unlike the previous section, there were noticeable advantages to certain commonsense models, specifically the ones that include ATOMIC in their source knowledge graph. The lower values for nDCG, however, imply that the reranking system was imperfect, especially since it had no means of distinguishing from levels 1 and 2 of relevance to 3 and 4, since it was zero-shot and thus never saw any training data. The claims that commonsense knowledge graphs can help with the missing factoid nature of the domain here is possible, albeit the MRR and MAP metrics were not significant enough to make any conclusion, moreover the fact that only 25 passages were re-reranked leads to more ambiguity.

Table 6: ANTIQUE Evaluation Metrics

Model	MRR	MAP	nDCG @ (1,3,10)		
BM25+	0.4172	0.1392	0.3790	0.4244	0.4705
RoBERTa$_L$	0.3801	0.1176	0.2763	0.2876	0.3128
RoBERTa$_L$[CSKG]	0.4661	0.1430	0.3521	0.3781	0.4141
RoBERTa$_L$[ATOM]	0.4551	0.1405	0.3472	0.3720	0.3936
RoBERTa$_L$[CWWV]	0.4040	0.1241	0.2989	0.3099	0.3256

10 CONCLUSION

Commonsense knowledge graphs have some limited utility in the zero-shot domain with tasks concerning with social inference and commonsense reasoning, yet in this study did not seem to extend towards commonsense prompted natural language generation (ProtoQA) nor factoid retrieval oriented question answering (WikiQA). There were noticeable differences between Factoid IR QA and Non-Factoid QA when it came to retrieval QA for example, with the commonsense re-ranking transformer models negatively impacting the BM25 oriented system.

Overall, the burgeoning interest in Commonsense reasoning and knowledge graphs lends it to domains such as Information Retrival, as rather than reasoning with noisy approximations such as implicit feedback, models could hopefully rely on factual knowledge graphs on which there was less noise and more firmly held reasons for inference. Moreover, if scaled properly, the assumptions encoded in these knowledge graphs would be transferrable and domain agnostic.

The inspiration for developing these and adapting them for Information Retrieval and Question Answering is a topic of current concern, especially with the community growing at Allen Institute for AI and the University of Soutern California in particular. This project in part was inspired by Filip Ieveski's presentation at the USC Information Sciences Institute REU that I attended, and he deserves an acknowledgement.

10.1 Extra credit

The scale of this project, with the large number of models and question answering tasks which each required researching and debugging could merit extra credit for this project.

REFERENCES

[1] Lisa Bauer, Yicheng Wang, and Mohit Bansal. 2018. Commonsense for Generative Multi-Hop Question Answering Tasks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 4220–4230. DOI: http://dx.doi.org/10.18653/v1/D18-1454

[2] Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen tau Yih, and Yejin Choi. 2020. Abductive Commonsense Reasoning. In *International Conference on Learning Representations*. https://openreview.net/forum?id=Byg1v1HKDB

[3] Michael Boratko, Xiang Li, Tim O'Gorman, Rajarshi Das, Dan Le, and Andrew McCallum. 2020. ProtoQA: A Question Answering Dataset for Prototypical Common-Sense Reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, 1122–1136. DOI: http://dx.doi.org/10.18653/v1/2020.emnlp-main.85

[4] Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: Commonsense Transformers for Automatic Knowledge Graph Construction. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, Anna Korhonen, David R. Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, 4762–4779. DOI: http://dx.doi.org/10.18653/v1/p19-1470

[5] Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database.* Bradford Books.

[6] Helia Hashemi, Mohammad Aliannejadi, Hamed Zamani, and W. Bruce Croft. 2019. ANTIQUE: A Non-Factoid Question Answering Benchmark. (2019). arXiv:cs.IR/1905.08957

[7] Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. 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)*. Association for Computational Linguistics, Hong Kong, China, 2391–2401. DOI: http://dx.doi.org/10.18653/v1/D19-1243

[8] Filip Ilievski, Pedro Szekely, and Bin Zhang. 2021. CSKG: The CommonSense Knowledge Graph. *Extended Semantic Web Conference (ESWC)* (2021).

[9] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S.

Bernstein, and Li Fei-Fei. 2016. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *International Journal of Computer Vision* 123 (2016), 32–73.

[10] Jingping Liu, Yanghua Xiao, Ao Wang, Liang He, and Bin Shao. 2020. CapableOf Reasoning: A Step Towards Commonsense Oracle. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20)*. Association for Computing Machinery, New York, NY, USA, 1797–1800. DOI:http://dx.doi.org/10.1145/3397271.3401251

[11] Zhuang Liu, Wayne Lin, Ya Shi, and Jun Zhao. 2021. A Robustly Optimized BERT Pre-training Approach with Post-training. In *Chinese Computational Linguistics - 20th China National Conference, CCL 2021, Hohhot, China, August 13-15, 2021, Proceedings (Lecture Notes in Computer Science)*, Sheng Li, Maosong Sun, Yang Liu, Hua Wu, Kang Liu, Wanxiang Che, Shizhu He, and Gaoqi Rao (Eds.), Vol. 12869. Springer, 471–484. DOI:http://dx.doi.org/10.1007/978-3-030-84186-7_31

[12] Kaixin Ma, Filip Ilievski, Jonathan Francis, Yonatan Bisk, Eric Nyberg, and Alessandro Oltramari. 2021. Knowledge-driven Data Construction for Zero-shot Evaluation in Commonsense Question Answering. *Proceedings of the AAAI Conference on Artificial Intelligence* 35, 15 (May 2021), 13507–13515. https://ojs.aaai.org/index.php/AAAI/article/view/17593

[13] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. (2019).

[14] Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019. ATOMIC: An Atlas of Machine Commonsense for If-Then Reasoning. *Proceedings of the AAAI Conference on Artificial Intelligence* 33, 01 (Jul. 2019), 3027–3035. DOI:http://dx.doi.org/10.1609/aaai.v33i01.33013027

[15] Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. 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)*. Association for Computational Linguistics, Hong Kong, China, 4463–4473. DOI:http://dx.doi.org/10.18653/v1/D19-1454

[16] A.B. Siddique, Fuad Jamour, Luxun Xu, and Vagelis Hristidis. 2021. Generalized Zero-Shot Intent Detection via Commonsense Knowledge. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21)*. Association for Computing Machinery, New York, NY, USA, 1925–1929. DOI:http://dx.doi.org/10.1145/3404835.3462985

[17] Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI'17)*. AAAI Press, 4444–4451.

[18] Alon Talmor, Ori Yoran, Ronan Le Bras, Chandrasekhar Bhagavatula, Yoav Goldberg, Yejin Choi, and Jonathan Berant. 2021. CommonsenseQA 2.0: Exposing the Limits of AI through Gamification.

[19] Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: A Free Collaborative Knowledgebase. *Commun. ACM* 57, 10 (Sept. 2014), 78–85. DOI:http://dx.doi.org/10.1145/2629489

[20] Peter West, Chandra Bhagavatula, Jack Hessel, Jena D. Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. 2021. Symbolic Knowledge Distillation: from General Language Models to Commonsense Models. (2021). arXiv:cs.CL/2110.07178

[21] Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. WikiQA: A Challenge Dataset for Open-Domain Question Answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portugal, 2013–2018. DOI:http://dx.doi.org/10.18653/v1/D15-1237