Do Children Texts Hold The Key To Commonsense Knowledge?

Julien Romero
Télécom SudParis
jromero@telecom-sudparis.eu

Simon Razniewski
Max Planck Institute for Informatics
srazniew@mpi-inf.mpg.de

Abstract

Compiling comprehensive repositories of commonsense knowledge is a long-standing problem in AI. Many concerns revolve around the issue of *reporting bias*, i.e., that frequency in text sources is not a good proxy for relevance or truth. This paper explores whether children's texts hold the key to commonsense knowledge compilation, based on the hypothesis that such content makes fewer assumptions on the reader's knowledge, and therefore spells out commonsense more explicitly. An analysis with several corpora shows that children's texts indeed contain much more, and more typical commonsense assertions. Moreover, experiments show that this advantage can be leveraged in popular language-model-based commonsense knowledge extraction settings, where task-unspecific fine-tuning on small amounts of children texts (childBERT) already yields significant improvements. This provides a refreshing perspective different from the common trend of deriving progress from ever larger models and corpora.

1 Introduction

Compiling commonsense knowledge (CSK) is a long-standing problem in AI (Lenat, 1995). Automated text-extraction-based approaches to CSK compilation, like Knext (Gordon et al., 2010), TupleKB (Dalvi Mishra et al., 2017), Quasimodo (Romero et al., 2019), COMET (Hwang et al., 2021) or Ascent (Nguyen et al., 2021) typically struggle with *reporting bias* (Gordon and Van Durme, 2013; Mehrabi et al., 2021), in particular an under-reporting of basic commonsense assertions. This is a crux of commonsense: If knowledge is assumed to be commonplace, such as that *rain is wet* or *cars have wheels*, there is little need to utter it explicitly. In contrast, statements that contradict commonsense are more frequently reported, leading to inappropriate images of the real world, e.g., that fires are more often cold than hot (e.g., 238 vs. 173 literal occurrences in the English Wikipedia).

Children's material may partially counter this bias: As children's knowledge is still growing, seemingly obvious assertions may still be frequently expressed explicitly in such material. Note that this is not a binary question of whether some knowledge is expressed or not, but more a ranking problem: Prominent CSK repositories often do not struggle to recall relevant statements (e.g., Ascent (Nguyen et al., 2021) contains 2800 assertions for "elephant"), but struggle to rank them properly. This is especially true for language-model based approaches of CSK compilation (Hwang et al., 2021; West et al., 2022), which by design can assign every token in the vocabulary a probability, but should do so in sensible order.

This paper investigates (i) whether children's texts are a promising source for CSK and (ii) whether small corpora can still boost knowledge extraction from large language models. Specifically, we analyze the density and typicality of CSK assertions in children's text corpora and show how fine-tuning existing language models on them can improve CSK compilation. Data and models, including a childBERT variant, can be found at https://www.mpi-inf.mpg.de/children-texts-for-commonsense.

2 Background

Prominent manual efforts towards CSK compilation include ConceptNet (Speer et al., 2017), Atomic (Sap et al., 2019), and the integrated CSKG (Ilievski et al., 2021). Prominent text extraction projects are Knext (Gordon et al., 2010), TupleKB (Dalvi Mishra et al., 2017), Quasimodo (Romero et al., 2019) and Ascent (Nguyen et al., 2022). Each carefully selects extraction corpora, like Wikipedia texts, user query logs, or targeted web search, to minimize extraction noise and maximize salience. Nonetheless, all struggle with extracting very basic CSK that is generally deemed

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 10954 - 10959
December 7-11, 2022 ©2022 Association for Computational Linguistics

too obvious to state explicitly. The utilized corpora are also small compared to what is typically used in language model pre-training. Therefore, pre-trained language models (PTLMs) have been employed directly for CSK extraction in a setting called prompting/probing (cf. the LAMA benchmark) (Petroni et al., 2019), where the BERT LM showed promising results in predicting ConceptNet assertions. They can also be employed with supervision, like in the COMET and the Atomic^{10x} system (Hwang et al., 2021; West et al., 2022). However, both PTLM-paradigms are grounded in frequencies observed in the original text corpora used for LM training, which are again subject to reporting bias.

3 Children Text Corpora

For understanding the nature of different text corpora, we rely on the Flesch Reading-ease score (FRE) (Flesch, 1979) that is based on the number of syllables, words, and sentences.

It generally ranges between 0 and 100, with 0-30 being considered difficult to read, 60-70 assumed standard, and above 80 easy.

We investigate three children text corpora:

1. **Children Book Test (CBT)** The CBT dataset (Hill et al., 2016) contains 108 children books such as *Alice's Adventures in Wonderland* extracted from the Gutenberg Project. It targets children around 12-14 years old and is about 30 MB in total.
2. **C4-easy** C4 (Raffel et al., 2020) is a cleaned version of Common Crawl's web crawl corpus that was used to train the T5 language model. It is approximately 305 GB in size. We derive *C4-easy* by restricting the corpus to documents with an FRE greater than 80, retaining 40.827.011 documents, which are 11% of C4.
3. **InfantBooks** We newly introduce the InfantBooks dataset, composed of 496 books targeted at kids from 1-6 years. It is based on Ebooks from websites like freekidsbooks.org, monkeypen.com and kidsworldfun.com, which we collected, transcribed, and cleaned. The final dataset consists of 496 books with 2 MB of text.[1]

As a baseline, and to rule out that observed improvements stem only from general training on more data, we also compare with employing the whole C4 corpus. In Table 1, we compare the corpora according to average document length, vocabulary size, and readability. In Table 2, we make the same comparison with the number of distinct words, the number of frequent words (with a relative frequency greater than 0.01%), and the cumulative frequency of the top 1000 words.

Corpus	Avg. doc. len.	Vocab. size	Readability (FRE)
C4	*411 words*	*151k*	*60 (Standard)*
CBT	57k words	63k	62 (Standard)
C4-easy	317 words	106k	86 (Easy)
InfantBooks	659 words	18k	91 (Very Easy)

Table 1: Text corpora considered for pretraining/finetuning, sorted by FRE.

Corpus	Dist. Words	freq. words	Cumul. freq. top 1k
C4	8M	994	68%
CBT	5M	874	82%
C4-easy	8M	908	75%
InfantBooks	5M	1031	82%

Table 2: Text corpora statistics.

4 Analysis

CSK Density. Although CBT and InfantBooks are too small for comprehensive text extraction, it is informative to see how dense CSK assertions are stated in them, i.e., the relative frequencies of CSK assertions per text.

We used the CSLB (Devereux et al., 2014) dataset, a large crowdsourced set of basic CSK assertions, like *alligator: is scary / is long / is green*. We focused on the top 4,245 properties for 638 subjects stated at least five times. For each corpus, we computed the relative frequencies with which these statements appear (w/ lemmatization).

Table 3 shows the results. As one can see, InfantBooks has the highest relative density of CSK assertions, 3x as many as C4 per sentence, 5x more per word.

To further explore the relation of text simplicity and CSK density, we grouped C4 documents into buckets based on their FRE. For a sample of 10k documents per bucket, Figure 1 reports the per-word frequencies of CSK assertions, considering all spotted CSK assertions (blue) or only distinct ones (red). The results are shown in Figure 1. As one can see, CSK density increases significantly with easier readability, and only the most simple documents suffer from a lack of diversity (decrease in blue line).

[1]The dataset is available at https://www.mpi-inf.mpg.de/children-texts-for-commonsense.

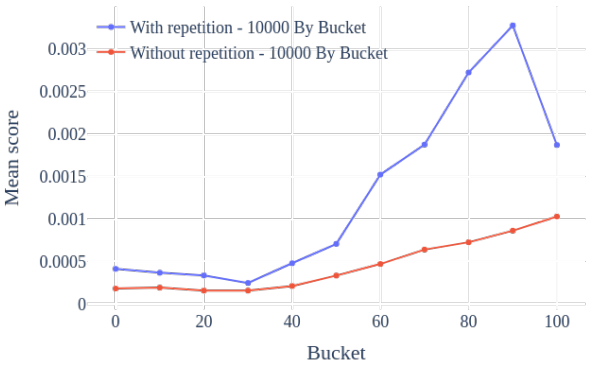

Figure 1: Relative CSK assertion frequency (per word) by C4 bucket of different readability (FRE).

Dataset	Mean freq./sent.	Mean freq./word
C4 (sample)	$9.99e^{-3}$	$4.86e^{-4}$
CBT	$1.03e^{-2}$	$4.94e^{-4}$
C4-easy (sample)	$1.10e^{-2}$	$7.65e^{-4}$
InfantBooks	$\mathbf{3.07e^{-2}}$	$\mathbf{2.56e^{-3}}$

Table 3: CSK density in different corpora.

CSK Typicality. Human associations and CSK resources contain a mix of salient but atypical assertions (*lion attacks human*), and typical but basic assertions (*lion drinks water*) (Chalier et al., 2020). To evaluate whether children's texts can help extract more typical assertions, we use PTLM prompting. Prompting is a standard procedure for extracting knowledge from masked-LMs (Liu et al., 2022). Given a relation of interest, e.g., "LocatedAt", one uses a generic textual pattern like "<entity> can be found at [MASK]", to obtain suggestions from an LM.

We used the BERT-large model[2], and fine-tuned it on each of the four corpora, up to a maximum of 48 hours (using one NVIDIA Quadro RTX 8000). We used a learning rate of $1e^{-5}$.

We then used 1,600 CSK triples from Quasimodo (Romero et al., 2019), human-rated as *very typical* (score [1,2[), or *typical* (score [2,3[), or *plausible* (score [3,4[) as targets. For each of these triples, we masked the objects, and used the fine-tuned BERT model to predict token likelihoods for the mask.

In Table 4 we report the mean reciprocal rank (MRR) for each group. Pre-training on children corpora has a slight edge over the vanilla BERT model and a significant edge over pre-training on the general C4 corpus.

[2]https://huggingface.co/bert-large-uncased

BERT-finetuning	Very typical	Typical	Plausible
None	0.361	0.200	0.153
InfantBooks	**0.364**	0.198	0.152
CBT	**0.364**	**0.205**	0.148
C4-easy	0.326	0.204	**0.156**
C4	0.312	0.195	0.155

Table 4: MRR of typical statements.

5 CSK Generation

We next evaluate whether pre-training on children texts helps in two PTLM-based methods for CSK generation: (i) via prompting from a LM pretrained on those corpora (LAMA method) (Petroni et al., 2019), and (ii) via supervised learning, based on a LM fine-tuned on those corpora (COMET method) (Hwang et al., 2021).

Unsupervised Generation via PTLM Prompting. The LAMA probe is based on prompting-based (CSK) generation. Originally, it used ConceptNet (Speer et al., 2017) together with the Open Mind Common Sense (Singh et al., 2002) (OMCS), which was used to construct ConceptNet and contains 20k items. ConceptNet contains triplets, and each triplet is associated with a sentence in OMCS. When a triplet has an object composed of one token, then the object is masked in the original OMCS sentence. The problem with this approach is that OMCS sentences are unnatural template-based phrases, not suited for LM completion. For example, in the LAMA probe, we find that "One of the things you do when you are alive is [MASK]." (think), "Something that might happen while analyzing something is [MASK]." (education) or "Similarity between a trash container and a clock: both can be found in a [MASK]." (school).

We therefore decided to add new commonsense-related datasets, to better evaluate how much commonsense knowledge a language model contains. First, we used CSLB (Devereux et al., 2014) that contains 20k basic properties about given subjects. The sentences are generally simpler than in OMCS. Second, we exploit 2.4k human-generated CSK sentences produced as evaluation data in the Quasimodo project (Romero et al., 2019; Romero and Razniewski, 2020). Finally, we exploited sentence sources of top statements web-extracted for the same KB, based on QA forums and search engine query logs. We here either masked the object or the predicate for all the sentences. Table 5 shows a sample of probes for each dataset. We make public (link above) the new probes and a sample of

generations.

Dataset	Sample Probes
ConceptNet	Something that might happen while fiddling is [MASK] a string. **breaking** Geriatrics is a type of [MASK]. **nursing** A person doesn't want to be [MASK]. **crippled**
CSLB	Dolphin is an [MASK]. **animal** Mug hold [MASK]. **tea** Cat chase [MASK]. **strings**
Q'modo-eval	Salmons are [MASK]. **fish** Ducks [MASK] in water. **live** Barbers cut our [MASK]. **hair**
Quasimodo	Pencils are made of [MASK]. **graphite** Bumblebees collect [MASK]. **pollen** Rockets need fuel in [MASK]. **space**

Table 5: Sample prompts for LAMA-style evaluation.

We used the BERT-large model pretrained as is, or pre-trained on any of the children's text corpora, as in the previous section, to generate predictions for each of the masked probes. The resulting performances are reported in Table 6. We report MRR and Hits@k for the PTLM's predictions in each case. While there are no improvements on the idiosyncratic ConceptNet data, pre-training on InfantBooks performs consistently at the top in the three other settings. CBT also shows consistent gains, while the broader C4-easy helps little. Examples are shown in Table 7.

Supervised PTLM-based Generation. Beyond prompting, an established technique to obtain CSK assertions is supervised learning based on LMs. We adapted the COMET-ATOMIC-2020 system (Hwang et al., 2021), which in turn is based on the GPT-2 LM. Like the previous prompt-based extraction, the core component is a pre-trained LM, which we can fine-tune on different text corpora. Compared to the previous method, however, a relation-specific step of supervised learning is added, in which the model is trained to generate objects for given subject-relation pairs.

We used ConceptNet for training and testing[3] and report precision and recall@k. The training and testing sets contain 186k and 23k triples.

The results are shown in Table 8. Due to the additional supervision on KB statements, differences are much smaller than in the previous setting. Nonetheless, we observe a consistent edge for fine-tuning on InfantBooks. To confirm this, we employed an additional human evaluation. For each subject-predicate pair in the test dataset, we computed the top-1 prediction for each model. For 300 examples where the predictions differed, we asked human annotators for their pairwise preference in terms of correctness. The results are shown in Table 9. InfantBooks again outperforms the vanilla BERT model and the more general C4-easy. Table 10 shows examples of generations.

We also performed an absolute evaluation of 600 pairs: For each pair, we asked a human evaluator to score the similarity between 1 (the worst) and 5 (the best). We obtained an average typicality of 2.57 for COMET-InfantBooks and 2.53 for COMET-C4. This is consistent with the results of the relative analysis - the gains in the KB-supervised settings are small, but consistently observable both in absolute and relative terms.

6 Conclusion

Our results positively confirm both starting questions: Dedicated children corpora contain more typical CSK assertions, and even small quality text corpora can boost the performance of large LMs for CSK extraction, especially concerning basic CSK assertions.

While the overall gains are still modest, these results affirm the role of data selection in commonsense knowledge compilation. Along with the exploitation of even stronger pre-trained language models, the quest for relevant text corpora thus continues.

7 Limitations

The following are notable limitations of our study, as well as of our approach.

Size of InfantBooks corpus A clear limitation of our newly presented InfantBooks corpus is its small size. Even if we were to ramp up the number of books (496 currently), text length in such books is inherently small, so this corpus will size-wise never compete with popular general web corpora. We alleviated this problem by a solution involving fine-tuning LMs trained on general corpora, but it remains that the presented corpus is not suitable for direct extraction tasks.

Flesch reading-ease vs. children content To further alleviate the size problem of the InfantBooks corpus, we also introduced the C4-easy corpus, based on filtering via the Flesch reading-ease

[3]Observe that in the previous unsupervised setting, the other datasets were needed because the *source sentences* underlying ConceptNet are unnatural. Its structured statements themselves are of good quality.

Corpus	ConceptNet			CSLB			Quasimodo-eval			Quasimodo		
	MRR	Hits@1	Hits@10	MRR	Hits@1	Hits@10	MRR	Hits@1	Hits@10	MRR	Hits@1	Hits@10
None	**0.270***	**0.188***	**0.429**	0.149	0.0911	0.259	0.433	0.327	0.633	0.212	0.138	0.357
InfantBooks	0.260	0.177	0.421	**0.174***	**0.106**†	**0.317***	**0.471**†	**0.374**†	**0.652**	**0.231***	**0.146**†	**0.404***
CBT	0.259	0.177	0.414	0.163	0.100	0.285	0.457	0.355	0.642	0.223	0.142	0.387
C4-easy	0.243	0.165	0.395	0.150	0.0877	0.274	0.390	0.288	0.586	0.200	0.123	0.356
C4	0.229	0.155	0.372	0.137	0.0777	0.251	0.457	0.356	0.637	0.182	0.111	0.325

Table 6: LAMA-style evaluation of CSK generation. (*: significantly better than all others with p-value$<$0.05, †: significantly better than having no corpus with p-value$<$0.05)

Masked Sentence	BERT-large	BERT-large finetuned on InfantBooks
Bears [MASK].	Club, ##kin, ##eed,##k, !, vs, ##ki, ##ville, Town, ##t	**bite**, **eat**, **walk**, sing, do, ##t, dance, say, **sleep**, kiss
Surgeons treat [MASK].	patients, them, children, wounds, animals, prisoners, **victims**, **people**, him, dogs	children, **pain**, wounds, **cold**, dogs, **burns**, birds, patients, **babies**, cats
Chefs know a lot about [MASK].	cooking, food, it, that, them, wine, this, you, fish, me	cooking, food, fish, wine, **baking**, **meat**, **recipes**, it, **dishes**, them
Researchers study [MASK].	it, them, this, children, women, him, animals, evolution, her, birds	animals, **art**, birds, **chemistry**, **insects**, it, **music**, **snakes**, **science**, them

Table 7: Sample Top Predictions (## means the letters are concatenated to the previous word). In bold good predictions that did not appear in the other method.

Corpus	P@5	P@10	R@5	R@10
None	3.76%	2.51%	15.4%	20.0%
C4-easy	3.78%	2.48%	15.5%	19.9%
C4	3.73%	2.47%	15.3%	19.7%
CBT	3.74%	2.48%	15.4%	19.9%
InfantBooks	**3.81%**	**2.52%**	**15.6%**	**20.1%**

Table 8: Supervised COMET-style generation of CSK.

InfantBooks Best	BERT Best	Same
13%	10%	77%
InfantBooks Best	**C4 Easy Best**	**Same**
22%	15%	63%

Table 9: Human preference of CSK generations.

Input (SP)	BERT-large	BERT-large fine-tuned on InfantBooks
bill, CapableOf	pay bill	amount nothing
standing up, HasSubevent	getting dizzy	falling down
hair, AtLocation	cabinet	hair salon
creating idea, Causes	solution	new idea
chew food, MotivatedByGoal	tastes good	eat

Table 10: Sample COMET Predictions

score (FRE). However, good readability is more of a necessary than a sufficient condition for inferring that content is intended for children, in other words, if there were not the results on InfantBooks, the positive results on C4-easy in isolation only proved a related hypothesis ("easier texts are the key to commonsense"), not the original one ("children texts are the key").

Dataset long-term availability InfantBooks contains copyrighted content. While we have checked national legislation, and believe that sharing the material for research purposes is permitted, should legal complaints occur, they would pose a risk to our ability to share this material long-term.

Acknowledgment

We thank Gerhard Weikum and Tuan-Phong Nguyen for discussions in earlier stages of this work, and the anonymous reviewers for their helpful comments.

References

Yohan Chalier, Simon Razniewski, and Gerhard Weikum. 2020. Joint reasoning for multi-faceted commonsense knowledge. In *AKBC*.

Bhavana Dalvi Mishra, Niket Tandon, and Peter Clark. 2017. Domain-targeted, high precision knowledge extraction. *TACL*.

Barry J Devereux, Lorraine K Tyler, Jeroen Geertzen, and Billi Randall. 2014. The centre for speech, language and the brain (CSLB) concept property norms. *Behavior research methods*, 46(4).

Rudolf Flesch. 1979. How to write plain English.

Jonathan Gordon, Benjamin Van Durme, and Lenhart K. Schubert. 2010. Learning from the web: Extracting general world knowledge from noisy text. *AAAI Workshops*.

Jonathan Gordon and Benjamin Van Durme. 2013. Reporting bias and knowledge acquisition. In *AKBC*.

Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. 2016. The goldilocks principle: Reading

children's books with explicit memory representations. In *ICLR*.

Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2021. (Comet-) Atomic 2020: On symbolic and neural commonsense knowledge graphs. In *AAAI*.

Filip Ilievski, Pedro Szekely, and Bin Zhang. 2021. Cskg: The commonsense knowledge graph. In *ESWC*.

Douglas B. Lenat. 1995. Cyc: A large-scale investment in knowledge infrastructure. *Commun. ACM*.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2022. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM CSUR*.

Ninareh Mehrabi, Pei Zhou, Fred Morstatter, Jay Pujara, Xiang Ren, and Aram Galstyan. 2021. Lawyers are dishonest? quantifying representational harms in commonsense knowledge resources. In *EMNLP*.

Tuan-Phong Nguyen, Simon Razniewski, Julien Romero, and Gerhard Weikum. 2022. Refined commonsense knowledge from large-scale web contents. In *TKDE*.

Tuan-Phong Nguyen, Simon Razniewski, and Gerhard Weikum. 2021. Advanced semantics for commonsense knowledge extraction. In *WWW*.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *EMNLP*.

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

Julien Romero and Simon Razniewski. 2020. Inside Quasimodo: Exploring construction and usage of commonsense knowledge. In *CIKM*.

Julien Romero, Simon Razniewski, Koninika Pal, Jeff Z. Pan, Archit Sakhadeo, and Gerhard Weikum. 2019. Commonsense properties from query logs and question answering forums. In *CIKM*.

Maarten Sap, Ronan LeBras, 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. *AAAI*.

Push Singh, Thomas Lin, Erik T Mueller, Grace Lim, Travell Perkins, and Wan Li Zhu. 2002. Open mind common sense: Knowledge acquisition from the general public. In *OTM Confederated International Conferences*.

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

Peter West, Chandra Bhagavatula, Jack Hessel, Jena D Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. 2022. Symbolic knowledge distillation: from general language models to commonsense models. *NAACL*.