(COMET-)ATOMIC$^{20}_{20}$: On Symbolic and Neural Commonsense Knowledge Graphs

Jena D. Hwang[1*], **Chandra Bhagavatula**[1*], **Ronan Le Bras**[1], **Jeff Da**[1], **Keisuke Sakaguchi**[1], **Antoine Bosselut**[13] and **Yejin Choi**[12]

[1] Allen Institute for AI, WA, USA
[2] Paul G. Allen School of Computer Science & Engineering, WA, USA
[3] Stanford University, CA, USA
{jenah, chandrab, ronanl, jeffd, keisukes, antoineb, yejinc}@allenai.org

Abstract

Recent years have brought about a renewed interest in commonsense representation and reasoning in the field of natural language understanding. The development of new commonsense knowledge graphs (CSKG) has been central to these advances as their diverse facts can be used and referenced by machine learning models for tackling new and challenging tasks. At the same time, there remain questions about the quality and coverage of these resources due to the massive scale required to comprehensively encompass general commonsense knowledge.

In this work, we posit that manually constructed CSKGs will never achieve the coverage necessary to be applicable in all situations encountered by NLP agents. Therefore, we propose a new evaluation framework for testing the utility of KGs based on how effectively implicit knowledge representations can be learned from them.

With this new goal, we propose ATOMIC$^{20}_{20}$, a new CSKG of general-purpose commonsense knowledge containing knowledge that is not readily available in pretrained language models. We evaluate its properties in comparison with other leading CSKGs, performing the first large-scale pairwise study of commonsense knowledge resources. Next, we show that ATOMIC$^{20}_{20}$ is better suited for training *knowledge models* that can generate accurate, representative knowledge for new, unseen entities and events. Finally, through human evaluation, we show that the few-shot performance of GPT-3 (175B parameters), while impressive, remains ∼12 absolute points lower than a BART-based knowledge model trained on ATOMIC$^{20}_{20}$ despite using over 430x fewer parameters.

1 Introduction

Commonsense understanding and reasoning remain long-standing challenges in general artificial intelligence. However, large-scale language models have brought tremendous progress in the sub-field of natural language processing. Such large-scale language models (Radford et al. 2018; Devlin et al. 2019; Brown et al. 2020) trained on extreme-scale data have been shown to effectively adapt to diverse downstream tasks, achieving significant performance gains across natural language benchmarks (Wang et al. 2019). Interestingly, as these models have grown larger (and trained on larger amounts of data), their benchmark performance has continued to improve (Raffel et al. 2019) despite limited conceptual improvements, leaving open questions regarding the source of these remarkable generalization properties.

Recent work has hypothesized that many of these performance gains could be a result of language models being able to memorize facts in their parameters during training (Roberts, Raffel, and Shazeer 2020) that can be leveraged at evaluation time. As a result, a new paradigm of language models as knowledge bases has emerged (Petroni et al. 2019). In this setting, language models are prompted with natural language prefixes or questions, and they express knowledge through language generation. The initial success of this paradigm for representing commonsense knowledge (Davison, Feldman, and Rush 2019; Tamborrino et al. 2020)

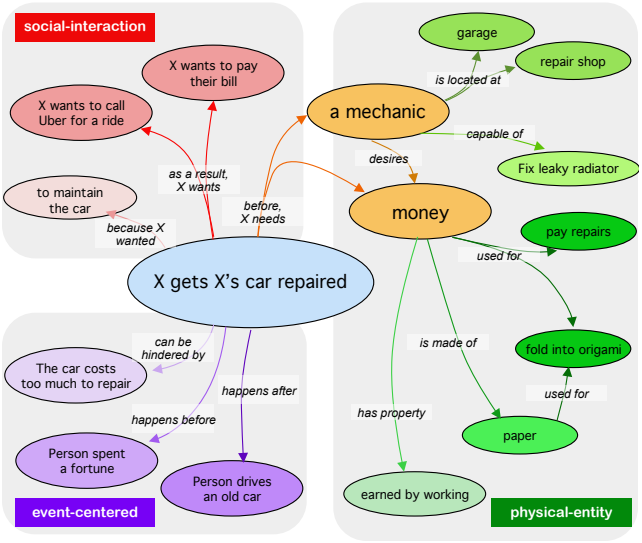

Figure 1: A tiny subset of ATOMIC$^{20}_{20}$, a large atlas of social and physical commonsense relations. Relations in the top-left quadrant reflects relations from ATOMIC.[1]

*The authors contributed equally to this work.

[1] ATOMIC$^{20}_{20}$ and ATOMIC-2020 can be used interchangeably, but for brevity we use ATOMIC$^{20}_{20}$ in this paper.

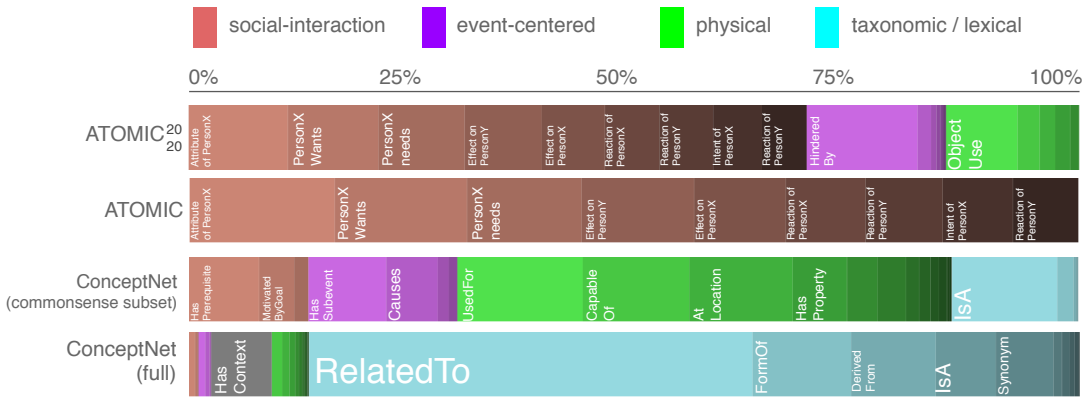

Figure 2: ATOMIC$^{20}_{20}$ tuple count distribution compared to ATOMIC (Sap et al. 2019) and CONCEPTNET, either its commonsense subset (Li et al. 2016) or the full set (Speer, Chin, and Havasi 2017).

has led to the optimistic claim that language models comprehensively encode commonsense knowledge, and remove the need for structured knowledge resources.

We take a more skeptical view of this capacity of language models – *Does scaling up language models actually endow them with commonsense knowledge?* While language models can successfully express certain types of knowledge, their best results are observed in narrowly specific conditions – we show (cf. §5) that they perform better when evaluated on knowledge bases that prioritize ontological relations and whose examples resemble language-like assertions (e.g., mango `IsA` fruit).[2] Consequently, the types of knowledge that can be directly accessed through the language model's interface remains limited.

However, prior work has also shown that training language models on knowledge graph tuples leads them to learn to express their implicit knowledge directly (Bosselut et al. 2019), allowing them to provide commonsense knowledge on-demand. These adapted *knowledge models* have exhibited promising results on commonsense benchmarks compared with methods that require linking entities to knowledge graphs (Shwartz et al. 2020; Liu et al. 2020). Inspired by these successes, we propose a dual use for commonsense knowledge bases going forward: as static graphs that can be linked to for discrete knowledge access, and as resources for adapting language models to hypothesize commonsense knowledge about un-annotated entities and events.

With this second purpose in mind, we propose evaluating commonsense knowledge resources based on the complementary information they can bring to pretrained language models. We construct ATOMIC$^{20}_{20}$, a new, high-quality knowledge graph with 1.33M commonsense knowledge tuples across 23 commonsense relations. We compare ATOMIC$^{20}_{20}$ with respect to its coverage and accuracy in competition with other highly used CSKGs, such as CONCEPTNET (Speer, Chin, and Havasi 2017). Our results show that ATOMIC$^{20}_{20}$ is able to cover more correct facts about more diverse types of commonsense knowledge than any existing, publicly-available commonsense knowledge resource. However, our results also indicate that there remains a large amount of exclusivity between these KGs, highlighting the challenge of creating resources that cover the scale and diversity of general commonsense knowledge.

Furthermore, we formalize the COMET framework of Bosselut et al. (2019) across different seed language models and training knowledge graphs, and evaluate the commonsense knowledge hypothesized by these adapted *knowledge models*. Our empirical study yields two promising conclusions. First, it confirms that KG-adapted language models learn to express knowledge more precisely than naive language models trained only on language. And second, we show that ATOMIC$^{20}_{20}$ as a transfer resource leads to COMET models that achieve the largest increase over their seed language model (across all seed LMs) for the commonsense knowledge types it covers, validating the importance of constructing knowledge resources with examples of knowledge not readily found in language models.

Key Contributions: In summary, we make three key contributions in this paper. We present ATOMIC$^{20}_{20}$—a new commonsense knowledge graph covering social, physical, and eventive aspects of everyday inferential knowledge (cf. §3). Next, we compare ATOMIC$^{20}_{20}$ with other prominent CSKBs head-to-head and show that our new *symbolic* knowledge graph is more accurate than any current CSKB (see Table 2) (cf. §4). Finally, we show that our new *neural* knowledge model COMET-ATOMIC$^{20}_{20}$ successfully transfers ATOMIC$^{20}_{20}$'s declarative knowledge to beat GPT-3, the largest pre-trained language model, in spite of using 400x fewer parameters (see Table 6) (cf. §5). This demonstrates the utility and importance of high-quality symbolic knowledge provided by ATOMIC$^{20}_{20}$ to generalize on commonsense information that LMs cannot expressively capture on their own (cf. §6).

[2] An observation supported by Brown et al. (2020)'s GPT-3 model, whose best few-shot performance on commonsense knowledge benchmarks comes on the PhysicalIQA (Bisk et al. 2020) and HellaSwag (Zellers et al. 2019) datasets.

2 Background

Commonsense Knowledge Graphs Large scale commonsense knowledge graphs are ubiquitous tools in natural language processing tasks as access to their facts allows models to learn to reason over commonsense knowledge to make predictions (Lin et al. 2019; Feng et al. 2020). In this work, we evaluate three existing knowledge graphs, CONCEPTNET, ATOMIC, and TRANSOMCS on their coverage and precision relative to our new resource ATOMIC$^{20}_{20}$.[3]

The **CONCEPTNET** (v5.7) knowledge graph (Speer, Chin, and Havasi 2017) consists of 36 relations focusing mostly on taxonomic and lexical knowledge (e.g., `RelatedTo`, `Synonym`, `IsA`) and physical commonsense knowledge (e.g., `MadeOf`, `PartOf`). CONCEPTNET (v5.7) contains 3.4M entity-relation tuples (in English) collected by crowdsourcing and merged with existing knowledge databases from DBPedia, WordNet, Wiktionary, and OpenCyc. Since the knowledge are derived from human efforts, the accuracy of CONCEPTNET (v5.7) knowledge is fairly high, though the quality does vary depending on the sources of knowledge and relation types. However, as highlighted in (Davis and Marcus 2015; Sap et al. 2019), and shown in Figure 2, the coverage of CONCEPTNET (v5.7) is limited to mostly taxonomic, lexical, and object-centric physical commonsense knowledge. In fact, out of 3.4M tuples, 90% of them correspond to taxonomic (e.g., `IsA`) or lexical (e.g., `Synonym`, `RelatedTo`) knowledge, making the commonsense portion of CONCEPTNET (v5.7) relatively small.

The **ATOMIC** (Sap et al. 2019) knowledge graph consists of 880K of tuples across 9 relations that cover social commonsense knowledge (e.g, X gets X's car repaired `xIntent` to maintain the car), including dynamic aspects of events such as causes and effects, `if-then` conditional statements, and mental states. The ATOMIC dataset is collected and validated completely through crowdsourcing.

The **TRANSOMCS** (Zhang et al. 2020a) knowledge graph consists of 18.48M tuples that were automatically converted from syntactic parses of sentences from various web sources including Wikipedia, Yelp, and Reddit. The set of relations used for the mapping is copied from CONCEPTNET. Although TRANSOMCS is much larger than other commonsense knowledge graphs, the precision of the extracted knowledge is significantly lower compared to other resources (cf. §4), and performs poorly as an adaptation resource relative to other KGs (cf. §5).

For this work we have selected three large scale CSKGs that retain a closed class of relational types that are comparable to one another. Other commonsense KBs in existence such as Quasimodo (Romero et al. 2019) provide a wider variety of fine-grained relations.

Language Models as Knowledge Bases Recent work hypothesizes that pretrained language models represent commonsense knowledge implicitly (Petroni et al. 2019; Roberts, Raffel, and Shazeer 2020). However, the results motivating these observations are often limited to narrowly scoped subsets of commonsense knowledge that primarily include taxonomic knowledge (e.g., mango `IsA` fruit) and that are often found explicitly stated in text. However, commonsense facts are often implied (Gordon and Van Durme 2013), and as will be seen in our studies (cf. §4), state of the art neural models struggle to express implicit commonsense knowledge that involves complex relationships.

To overcome this limitation, Bosselut et al. (2019) take the best of both worlds between commonsense knowledge graphs and pretrained language models. The commonsense transformer, or COMET, adapts pretrained neural language models by training on example tuples from commonsense knowledge graphs. It takes a head/source phrase and a relation (e.g., take a nap `Causes`) and generates the tail/target phrase (e.g., have energy). Bosselut et al. (2019) show that COMET trained on the CONCEPTNET and ATOMIC knowledge graphs is able to adapt to generate novel (and valid) commonsense knowledge tuples.

Importantly, these neural *knowledge models* can produce commonsense knowledge on-demand for any head entity that can be expressed through language. This flexibility allows them to be used out-of-the-box, and they have been applied to new, previously unexplored tasks, such as sarcastic comment generation (Chakrabarty et al. 2020), therapy chatbots (Kearns et al. 2020), and automated story plot generation (Ammanabrolu et al. 2020). These contributions show that progress on knowledge models opens up new downstream applications that were challenging to model before.

3 ATOMIC$^{20}_{20}$

We present ATOMIC$^{20}_{20}$, a commonsense knowledge graph with 1.33M everyday inferential knowledge tuples about entities and events. ATOMIC$^{20}_{20}$ represents a large-scale commonsense repository of textual descriptions that encode both the social and the physical aspects of common human everyday experiences, collected with the aim of being complementary to commonsense knowledge encoded in current language models. ATOMIC$^{20}_{20}$ introduces 23 commonsense relations types. They can be broadly classified into three categorical types: 9 commonsense relations of social-interaction, 7 physical-entity commonsense relations, and 7 event-centered commonsense relations concerning situations surrounding a given event of interest. The full inventory of ATOMIC$^{20}_{20}$ relations is listed in Table 1.

In terms of physical and event-centered commonsense, by far, the two largest new relations in ATOMIC$^{20}_{20}$ are `ObjectUse` and `HinderedBy`. For `ObjectUse`, we focused on *affordances* of everyday objects such as "popcorn bucket" that may be used for "holding popocorn" or "storing things". For `HinderedBy`, we explore the notion that many events in real world can be defeasible (Lascarides and Asher 1991) by collecting hindrances to goals that may be useful for tasks such as counterfactual reasoning. For example X's desires to adopt a cat may be hindered by finding out that X is allergic to cats, which would necessitate X to adjust future actions accordingly (say, opt for hypoallergenic options like tortoises).

[3] We were unable to include Cyc (Lenat 1995) in our study due to the discontinuation of its research license and the cost of the commercial license (over $1M). CONCEPTNET includes a subset of Cyc – OpenCyc.

	Head	Relation	Tail	Size
PHYSICAL-ENTITY	bread	ObjectUse	make french toast	165,590
		AtLocation*	basket; pantry	20,221
		MadeUpOf	dough; wheat	3,345
		HasProperty*	cooked; nice to eat	5,617
	baker	CapableOf*	coat cake with icing	7,968
		Desires*	quality ingredients	2,737
		Not Desires*	bad yeast	2,838
EVENT-CENTERED	X runs out of steam	IsAfter	X exercises in the gym	22,453
		HasSubEvent	become tired	12,845
		IsBefore	X hits the showers	23,208
		HinderedBy	drinks too much coffee	106,658
		Causes	takes a break	376
		xReason	did not eat breakfast	334
	X watches ___ anyway	isFilledBy	the game; the TV	33,266
SOCIAL-INTERACTION	X runs out of steam	xNeed	do something tiring	128,955
		xAttr	old; lazy; lethargic	148,194
		xEffect	drinks some water	115,124
		xReact	tired	81,397
		xWant	to get some energy	135,360
	X votes for Y	xIntent	to give support	72,677
		oEffect	receives praise	80,166
		oReact	grateful; confident	67,236
		oWant	thank X; celebrate	94,548

Table 1: Relations in ATOMIC$^{20}_{20}$ along with illustrative examples and their respective size. Relations that reflect semantically identical categories to CONCEPTNET is marked with an asterisk (*).

In the case of `ObjectUse`, we collected over 130K everyday object-use pairs by asking crowdworkers for necessary objects and their uses for each event in ATOMIC$^{20}_{20}$. For example, given "X eats popcorn" we elicited items such as "popcorn bucket" with their various expected uses. The number also reflects *atypical* usages gathered in a separate pass where workers were asked to provide creative or resourceful but *feasible* uses of the objects. Given "popcorn bucket", for instance, one might "wear it as a hat" for, say, a costume party. For `HinderedBy`, we crowdsourced over 100K tuples of hindrances to existing ATOMIC$^{20}_{20}$ events, asking the workers to provide situations or events that might pose as deterrence should the event be considered an achievable goal (see Appendix for further details). For social-interaction commonsense, we primarily incorporated tuples from ATOMIC, but also crowdsourced an additional 34K tuples using the same approach as Sap et al. (2019).

ATOMIC$^{20}_{20}$ also pulls commonsense tuples from the English subset of CONCEPTNET(v5.7) (latest version available; Speer, Chin, and Havasi 2017).[4] Of the 3.4M English tuples in CONCEPTNET(v5.7), a small subset of 172K tuples was selectively chosen to be integrated into ATOMIC$^{20}_{20}$ via elimination and crowdsourcing. This subset represents data carefully identified to reflect commonsense information dealing with qualitative human experiences. Among the eliminated data are tuples with edge weight ≤ 0.5, dictionary or etymologically based knowledge (e.g., synonyms/antonyms, inflections), lexical hyper/hyponymic lexical relationships such as `IsA` or `InstanceOf`, and relations based on lexical co-occurrence (e.g., `RelatedTo` or `LocatedNear`), which are easily recoverable from language models.[5] After selective removal of these relations and a post-processing step to ensure the removal of deterministic information such as geographic facts (e.g., "shenzhen" `AtLocation`"china"), tuples from each CONCEPTNET were examined for further splits or joins to align with the existing structure of ATOMIC$^{20}_{20}$. A random 10% tuples from each selected relations were then put through crowdsourced validity testing (akin to the process described later in §4). Tuples that were directly incorporated without further edits passed with an acceptance rate of 93% or higher. A subset of relations (i.e., `CapableOf`, `HasProperty`, `MotivatedByGoal`) were put through additional crowdsourcing to weed out tuples that were either invalid or found to hold prejudiced descriptions of human entities. In the end, only 5 relations (marked with an asterisk in Table 1) retain the CONCEPTNET's original meaning with a few relations that are cognates in ATOMIC$^{20}_{20}$ (more details in Appendix).

4 Symbolic Knowledge Graph Comparison

In this work, we compare our new ATOMIC$^{20}_{20}$ knowledge graph to three other prominent CSKGs: ATOMIC (Sap et al. 2019), CONCEPTNET[6] (Li et al. 2016), and TRANSOMCS (Zhang et al. 2020a). We measure the accuracy of tuples in each KG and compare the coverage of each CSKG w.r.t. other CSKGs head-to-head.

Accuracy Assessment

In order to assess the accuracy of the knowledge represented, 3K random instances were extracted from each of the knowledge graphs for a crowdsourced evaluation of the tuples.

Qualifying Crowdsource Workers. The evaluation was carried out through crowdsourcing on the Amazon Mechanical Turk platform. To ensure high-quality annotations, we qualified a pool of 173 workers through a paid qualification task that tested their ability to follow directions and provide reasonable answers to the qualification test. The qualification test contained 6 manually selected tuples from ATOMIC

[4] A CONCEPTNET(v5.7) fact is considered English if both the head and tail concepts are marked with '/en/' in the edge id.

[5] CONCEPTNET 5.7 defines weight as "the strength with which this edge expresses this assertion". A pilot crowdsource assessment step found any tuple with weight ≤ 0.5 unreliable w.r.t. its validity.

[6] Hereafter, as we focus on CSKGs, by ConceptNet, we refer to the commonsense subset, unless specified otherwise.

Knowledge Base	Accept	Reject	No Judgment
ATOMIC$^{20}_{20}$	**91.3**	**6.5**	2.2
ATOMIC	88.5	10.0	1.5
CONCEPTNET	88.6	7.5	3.9
TRANSOMCS	41.7	53.4	4.9

Table 2: Accuracy - Percentage (%) of tuples in the knowledge base evaluated by human crowdworkers as either always true or likely (Accept), farfetched/never or invalid (Reject), or unclear (No Judgment).

and CONCEPTNET, including both easy and tricky relations to annotate. A worker was qualified if they provided 100% acceptable answers. Workers providing 5 of 6 correct answers were also accepted only when they provided a reasonable written substantiation for their incorrect choice. Workers were paid an average of $15 per hour for their evaluations.

Human Evaluation Setup. Workers were presented with knowledge tuples in the form of (*head*, *relation*, *tail*) for annotation. To expedite the human assessment of the tuples, each *relation* (e.g., `xWant` or `AtLocation`) was translated into a human-friendly natural language form (e.g., "as a result, PersonX wants" and "located or found at/in/on", respectively; cf. Appendix). The workers were asked to rate the tuples along a 4-point Likert scale: *always/often* – the knowledge assertion presented is always or often true, *sometimes/likely* – it is sometimes or likely true, *farfetched/never* – it is false or farfetched at best, and *invalid* – it is invalid or makes no sense. Any tuples receiving the former two labels are ranked as **Accept** and latter two as **Reject**. The workers were also given a choice to opt out of assessment if the concepts were too unfamiliar for a fair evaluation (**No Judgment**). Each task (HIT) included 5 tuples of the same relation type, and each tuple was labeled by 3 workers. For the results, we take the majority vote among the 3 workers.

Results. ATOMIC$^{20}_{20}$ outperforms other KGs in crowdsourced accuracy as shown in Table 2.[7] ATOMIC ties with CONCEPTNET with reasonably high accuracy, while TRANSOMCS lags behind others with far lower accuracy. We provide a per-relation breakdown of accuracies in Table 3.

Between ATOMIC$^{20}_{20}$ and ATOMIC, the variations in the assessed accuracies are not found to be statistically significant. Among the ATOMIC$^{20}_{20}$ and CONCEPTNET relations that represent *exact matches* (marked with * in Table 3), the differences are either not statistically significant or when they are, ATOMIC$^{20}_{20}$ improves upon the associated facts, reflecting that the preprocessing stages of CONCEPTNET integration were helpful in improving the quality of these relations (§3). Among *cognates* in ATOMIC$^{20}_{20}$ and CONCEPTNET relations, two sets of relations fare significantly worse in ATOMIC$^{20}_{20}$ than in CONCEPTNET. In the case of `ObjectUse`/`UsedFor`, this is likely due to the fact that ATOMIC$^{20}_{20}$'s `ObjectUse` includes atypical affordances (cf. §3). In an annotation setting where workers are asked to evaluate the truth or likelihood of an assertion rather than feasibility of use, a portion of the atypical usages are seen as 'farfetched' and thus, rejected. In the case of `MadeUpOf`/`MadeOf`, there may be some room for improvement for ATOMIC$^{20}_{20}$. Unlike the ATOMIC$^{20}_{20}$'s `HasSubEvent` label that successfully joins together CONCEPTNET's HAS(FIRST/LAST)SUBEVENT labels for an improved accuracy, ATOMIC$^{20}_{20}$'s `MadeUpOf` union of `MadeOf`, `PartOf`, and a subset of `HasA`, did not seem to have resulted in improved quality. The rest of the ATOMIC$^{20}_{20}$ cognates see a significantly higher or similar accuracy in comparison to CONCEPTNET.

[7] Overall inter-rater agreement measured by Fleiss' κ of 0.46 (moderate agreement; Fleiss 1971).

ATOMIC$^{20}_{20}$	ATOMIC	*Relation*	CN	T-OMCS
92.3		AtLocation*	89.4	34.3
93.9		CapableOf*	84.4	50.0
94.6		Causes	90.0	50.0
96.9		Desires*	96.3	48.2
93.9		HasProperty*	86.3	52.4
82.3		ObjUse/UsedFor	**96.3**	31.6
98.5		NotDesires*	96.3	
96.9		HasSubevent	88.1	57.7
		HasFirstSubevent	**93.8**	52.4
		HasLastSubevent	**95.6**	38.2
		HasPrerequisite	**94.4**	30.0
75.4		MadeUpOf/MadeOf	**88.1**	15.9
		PartOf	**71.9**	46.5
		HasA	**77.5**	43.5
96.9		HinderedBy		
96.2		isAfter		
95.4		isBefore		
96.2		isFilledBy		
		ReceiveAction	**84.4**	56.4
91.5	86.3	oEffect		
91.5	87.7	oReact		
88.5	**89.5**	oWant		
87.7	**91.0**	xAttr		
80.8	**87.2**	xEffect		
93.1	89.9	xIntent/MotivByGoal	84.4	27.1
87.7	85.1	xNeed		
90.8	**91.3**	xReact		
96.2		xReason		
82.3	88.4	xWant/CausesDesire	**90.0**	35.9

Table 3: KG accuracy values broken down by relation. Boxed cells indicate statistically significant difference from ATOMIC$^{20}_{20}$ values. Relational *cognates* have been grouped together and *exact matches* are asterisked (*) (cf. Table 1).

Coverage Assessment

We make a pairwise comparison between the CSKGs to assess their coverage with regards to the commonsense knowledge they contain. For a reliable head-to-head comparison, we map relations and tuples between various KGs.

Mapping Relations. Since ATOMIC$^{20}_{20}$ is built on existing ATOMIC relations, we primarily need to align relations between ATOMIC$^{20}_{20}$ and CONCEPTNET. We manually align them based on the definitions for the labels as supplied by

Source KB↓	Target KB→ ATOMIC	CN	T-OMCS	ATOMIC$^{20}_{20}$
ATOMIC	-	0.1	0.0	100.0
CONCEPTNET	0.3	-	5.5	45.6
TRANSOMCS	0.0	0.4	-	0.3
ATOMIC$^{20}_{20}$	60.2	9.3	1.4	-

Table 4: Coverage Precision - Average number of times (in %) a tuple in Source KB is found in Target KB.

Source KB↓	Target KB→ ATOMIC	CN	T-OMCS	ATOMIC$^{20}_{20}$
ATOMIC	-	0.3	0.0	60.1
CONCEPTNET	0.1	-	0.3	8.9
TRANSOMCS	0.0	7.6	-	1.3
ATOMIC$^{20}_{20}$	100.1†	47.8	0.4	-

Table 5: Coverage Recall - Average number of times (in %) a tuple in Target KB is found in Source KB. †This value is greater than 100 because multiple tuples in ATOMIC$^{20}_{20}$ can map to the same tuple in ATOMIC.

the two graphs, then the resulting alignment was verified by sampling at random approximately 20 instances per relation.

Mapping Tuples. In order to resolve syntactic differences in how the concepts are expressed in each of the KGs (e.g., ATOMIC's "PersonX eats breakfast" vs. CONCEPTNET's "eat breakfast"), we preprocess each of the head and tail concepts of each tuple in each KG in the following manner: (1) the concept is lowercased and stripped of extra spaces, punctuations, and stopwords; (2) any exact tuple duplicates within each KB removed, and (3) remaining content words are lemmatized according to their POS category. For ATOMIC and ATOMIC$^{20}_{20}$, an extra step is added to remove mentions of "PersonX", "PersonY" and "PersonZ" if occurring at the beginning of a string, and to replace with 'person' if they occur elsewhere (e.g, "PersonX greets PersonY").

Metrics. We use two metrics to evaluate the coverage of knowledge graphs. For each pair of CSKGs, we compute precision and recall with respect to a target KG. **Coverage precision** assesses the proportion of tuples in the source KG that are correct according to tuples in the target KG. **Coverage recall** reflects the proportion of tuples in the target KB that the tuples in the source KB successfully recalled.

Results. Tables 4 and 5 show a pairwise coverage precision and recall assessment among the CSKGs. ATOMIC$^{20}_{20}$ shows the widest coverage: ATOMIC$^{20}_{20}$ is able to recall all of ATOMIC (as expected) and just under half of CONCEPTNET. There is very little overlap between ATOMIC and CONCEPTNET, which is unsurprising as all of ATOMIC knowledge is focused on social behaviors CONCEPTNET does not cover while CONCEPTNET leans on physical commonsense which falls outside ATOMIC's scope. Overall, TRANSOMCS intersects very little with any of the other three KBs.

KG	Model	Accept	Reject	No Jdgm.
ATOMIC$^{20}_{20}$	GPT2-XL	36.6	62.5	0.9
	GPT-3	73.0	24.6	2.5
	COMET(GPT2-XL)	72.5	26.6	0.9
	COMET(BART)	**84.5**	**13.8**	1.7
ATOMIC	GPT2-XL	38.3	61.2	0.4
	COMET(GPT2-XL)	64.1	34.7	1.2
	COMET(BART)	**83.1**	**15.3**	1.6
CONCEPTNET	GPT2-XL	50.3	42.1	7.7
	COMET(GPT2-XL)	74.5	19.0	6.4
	COMET(BART)	**75.5**	**17.9**	6.6
TRANSOMCS	GPT2-XL	**28.7**	**53.5**	17.8
	COMET(GPT2-XL)	26.9	60.9	12.2
	COMET(BART)	23.8	65.9	10.3

Table 6: Human evaluation of generation accuracy (%). Each model uses greedy decoding to generate the *tail* of 5K randomly-sampled test prefixes (*head*, *relation*) from each knowledge graph. GPT2-XL, GPT-3 and BART have 1.5B, 175B and 440M parameters, respectively.

5 Neural Knowledge Graph Comparison

Language models are powerful tools for representing knowledge, but their ability to serve as generative knowledge bases is limited by the fact they are directly trained to represent the distribution of language. Previous work shows knowledge graphs can help language models better transfer as knowledge engines (Bosselut et al. 2019) by re-training them on examples of structured knowledge. As a result, a new purpose for knowledge graphs is to be useful in helping language models generalize to hypothesizing knowledge tuples.

Experimental Setup. To evaluate whether knowledge graphs can help language models effectively transfer to *knowledge models*, we train different pretrained language models on the knowledge graphs described in Section 4, which we describe below:

GPT2 (Radford et al. 2019) is a Transformer (Vaswani et al. 2017) based language model. In our experiments, we use the largest GPT2 model, GPT2-XL, that has 1.5B parameters. We fine-tune GPT2-XL on each of our CSKGs to predict the tail of a tuple (e.g., wheat) given the head (e.g., bread) and a relation (e.g., `MadeUpOf`). The hyperparameter settings used for training are described in more detail in Appendix. Additionally, we use GPT2-XL in a zero-shot setting as a baseline to measure the effect of transfer learning on knowledge graphs. For fair comparison, we convert each relation manually to an English language prompt expecting the tail of each tuple as output generated by the model.

BART (Lewis et al. 2020) is a Bidirectional and Autoregressive Transformer, an adaptation from BERT (Devlin et al. 2019) that is better suited for natural language generation (e.g., translation, summarization). Additional training details are provided in Appendix.

GPT-3 (Brown et al. 2020) is an autoregressive language model that has 175B (over 100X more parameters than

		Bleu-1	Bleu-2	Bleu-3	Bleu-4	METEOR	ROUGE-L	CIDEr	BERT Score
ATOMIC$^{20}_{20}$	GPT2-XL	0.101	0.028	0.010	0.003	0.082	0.098	0.047	0.395
	GPT-3	0.299	0.153	0.081	0.048	0.182	0.255	0.175	0.540
	COMET(GPT2-XL)	0.407	0.248	0.171	0.124	0.292	0.485	0.653	0.638
	COMET(BART)	**0.469**	**0.286**	**0.189**	**0.130**	**0.330**	**0.495**	**0.658**	**0.639**
ATOMIC	GPT2-XL	0.083	0.029	0.011	0.005	0.081	0.087	0.045	0.386
	COMET(GPT2-XL)	0.419	0.296	**0.228**	**0.189**	0.292	0.517	0.733	0.634
	COMET(BART)	**0.515**	**0.324**	0.220	0.159	**0.347**	**0.546**	**0.740**	**0.646**
CONCEPTNET	GPT2-XL	0.044	0.012	0.004	0.002	0.064	0.057	0.050	0.389
	COMET(GPT2-XL)	0.155	**0.119**	**0.095**	**0.078**	**0.134**	**0.193**	**0.425**	**0.552**
	COMET(BART)	**0.172**	0.111	0.072	0.049	0.130	0.184	0.368	0.535
TRANSOMCS	GPT2-XL	0.028	0.001	0.000	0.000	0.093	0.053	0.013	0.351
	COMET(GPT2-XL)	0.301	0.000	0.000	0.000	0.180	0.302	0.254	0.677
	COMET(BART)	**0.351**	**0.170**	**0.003**	0.000	**0.198**	**0.352**	**0.297**	**0.678**

Table 7: Automated metrics for the quality of the *tail* generations of the GPT2-XL language model and the knowledge models COMET(GPT2-XL) and COMET(BART). Each approach uses greedy decoding for sampled 5k test prefixes for each KG. The 5k prefixes correspond to the ones for the human eval. Similar results are obtained on the full test sets (cf. Appendix).

GPT2-XL) parameters and is trained on a corpus of web text. We use the GPT-3 API to *prime* the language model to generate the tail for a given prefix – *(head, relation)* pair. Thus, GPT-3 is evaluated in a few-shot setting. Additional details of our implementation are provided in Appendix.

Evaluation Setup. To assess language-to-knowledge transfer capabilities, we evaluate how language models generalize to new, unseen entities, concepts, or events. We split each knowledge graph into training, validation, and test sets such that the *heads* of the knowledge tuples do not overlap between these sets. This adversarial split forces the language models to generalize the relationships they learn from training on the knowledge graphs to the entities learned during language pretraining. Also, to avoid overpopulating the validation and test sets with generic *heads* (e.g., "I", "You", "He", "We", and "They" collectively account for over 2.2M tuple heads in TRANSOMCS), we enforce that the head of any knowledge tuple in the *dev* and *test* sets is involved in at most 500 tuples. Finally, we remove low-quality tuples from TRANSOMCS by imposing a confidence score of ≥ 0.5.

We score the tuples generated by these knowledge models using common evaluation metrics for text generation: BLEU (Papineni et al. 2002), ROUGE (Lin 2004), CIDEr (Vedantam, Lawrence Zitnick, and Parikh 2015), and BERT Score (Zhang et al. 2020b). For a subset of 5000 generated tuples from the test set of each knowledge graph, we also run the same human evaluation described in Section 4.

Results. We present our main results in Tables 6 and 7. First, we note the large divide between the zero-shot GPT2-XL model that produces commonsense knowledge without any fine-tuning and the two COMET models across the ATOMIC$^{20}_{20}$, ATOMIC, and CONCEPTNET knowledge graphs (Table 6). This large gap indicates that language models can benefit from learning facts from commonsense knowledge graphs. They do not have the means to precisely express this knowledge directly from just pretraining on language. This observation is supported by the gaps between these models in the automatic evaluations (Table 7), as well. Additionally, human evaluation of GPT-3 (Table 6) shows a $\sim$12 point deficit compared to the performance of COMET(BART), in spite of GPT-3 (175B) having over $\sim$430 times more parameters than COMET(BART) (406M). Similarly, we see a large gap in performance across all automated metrics in Table 7. The performance gap indicates that high-quality declarative knowledge is valuable even after the advent of extreme scale language models.

In addition to this main result, two particularly interesting observations emerge. First, we note that the gap between the zero-shot model and COMET is larger on the ATOMIC$^{20}_{20}$ and ATOMIC knowledge graphs, than on CONCEPTNET, supporting the reflection that ATOMIC$^{20}_{20}$ supports categories of knowledge that are more difficult to learn from pretraining. Second, the results on the human evaluation show that COMET models trained on TRANSOMCS are not able to generalize knowledge to new entities, implying that language models benefit more from accurate knowledge examples, which TRANSOMCS lacks (cf. §4).

6 Discussion

Do pretrained language models already encode commonsense knowledge? Our conclusions on this subject are mixed and hinge on the ambiguous meaning of what it means to *encode* knowledge. Despite the conclusions of prior work (Petroni et al. 2019; Roberts, Raffel, and Shazeer 2020; Tamborrino et al. 2020), our results in Table 6 are clear that language models fail to express large varieties of knowledge when prompted for it in a zero-shot manner. When converted to COMET models by training on a knowledge graph, their performance at hypothesizing knowledge tuples skyrockets – 47.9% absolute difference between COMET(BART) and GPT2-XL on ATOMIC$^{20}_{20}$.

However, the evaluation tuples are adversarially selected to not include head entities that were in the training set. The model must generalize its learned representations of relations to entities it has not observed these relationships for

during fine-tuning, meaning the representation of these entities is solely formulated from learning language. As a result, language models may still *encode* this knowledge in their parameters, even if they are not capable of *expressing* it directly. With this framing in mind, the COMET training paradigm proposed by Bosselut et al. (2019) can perhaps be viewed less as a means of learning *knowledge* from KGs, and more as a method of learning an *interface* for language models to hypothesize encoded knowledge through language generation. We look forward to future work in this space that attempts to disentangle these two ideas.

What considerations should be made when designing commonsense knowledge resources? Based on our results in Section 5, we outline desiderata for the design and development of future commonsense knowledge graphs. Because certain types of knowledge are already encoded and expressible by pretrained language models, CSKG designers should focus on collecting examples and categories of knowledge that are less likely to be known by language models. For example, of the 378 test tuples evaluated by the GPT2-XL zero-shot model that contained the `HinderedBy` relation, only 1.3% were deemed plausible by human raters – jumping to 85% plausibility for COMET(BART) – pointing to an advantage in constructing ATOMIC$^{20}_{20}$ with this relationship in mind (see Appendix for per-relation accuracy).

Second, commonsense knowledge resources should be designed with the goal of accuracy and relationship coverage. Because language models exhibit powerful adaptation (Brown et al. 2020), they can generalize many commonsense relationships as long they have examples on which to train. Consequently, we should construct commonsense resources that encapsulate larger numbers of relations so the knowledge in pretrained language models can be grounded to a variety of relationships. However, language models also benefit from learning from precise examples. Being able to train on a large collection of examples from TRANSOMCS (see Appendix) did not allow COMET models to generalize to unseen entities as these examples were not of sufficient quality (See Table 2). Resources should be carefully validated for the quality of their facts, an example set by Speer, Chin, and Havasi (2017) and Sap et al. (2019).

7 Conclusion

In this work, we formalize a use for commonsense knowledge graphs as transfer learning tools for pretrained language models. With this new purpose, we hypothesize that commonsense knowledge graphs should be designed to contain knowledge that is not already expressible by language models without difficulty (e.g., not taxonomic and lexical knowledge). Consequently, we propose ATOMIC$^{20}_{20}$, a novel commonsense knowledge graph containing tuples whose relations are specifically selected to be challenging for pretrained language models to express. Our empirical studies demonstrate that ATOMIC$^{20}_{20}$ contains high-accuracy knowledge tuples across multiple novel relations not found in existing CSKGs or expressible by LMs. Furthermore, we show that ATOMIC$^{20}_{20}$ can be effectively used as a training set for adapting language models as *knowledge models* to generate high quality tuples on-demand.

Acknowledgements

We would like to thank the anonymous reviewers for their valuable feedback. This research was supported in part by NSF (IIS-1524371), the National Science Foundation Graduate Research Fellowship under Grant No. DGE 1256082, DARPA CwC through ARO (W911NF15-1- 0543), DARPA MCS program through NIWC Pacific (N66001-19-2-4031), and the Allen Institute for AI. Computations on beaker.org were supported in part by credits from Google Cloud. TPU machines for conducting experiments were provided by Google.

References

Ammanabrolu, P.; Cheung, W.; Broniec, W.; and Riedl, M. 2020. Automated Storytelling via Causal, Commonsense Plot Ordering. *ArXiv* abs/2009.00829.

Bisk, Y.; Zellers, R.; Le Bras, R.; Gao, J.; and Choi, Y. 2020. PIQA: Reasoning about Physical Commonsense in Natural Language. In *AAAI*.

Bosselut, A.; Rashkin, H.; Sap, M.; Malaviya, C.; Çelikyilmaz, A.; and Choi, Y. 2019. COMET: Commonsense Transformers for Automatic Knowledge Graph Construction. In *ACL*.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krüger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. *ArXiv* abs/2005.14165.

Chakrabarty, T.; Ghosh, D.; Muresan, S.; and Peng, N. 2020. R^3: Reverse, Retrieve, and Rank for Sarcasm Generation with Commonsense Knowledge. In *ACL*.

Davis, E.; and Marcus, G. 2015. Commonsense Reasoning and Commonsense Knowledge in Artificial Intelligence. *Commun. ACM* 58(9): 92–103.

Davison, J.; Feldman, J.; and Rush, A. 2019. Commonsense Knowledge Mining from Pretrained Models. In *EMNLP-IJCNLP*, 1173–1178. Hong Kong, China.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*.

Feng, Y.; Chen, X.; Lin, B. Y.; Wang, P.; Yan, J.; and Ren, X. 2020. Scalable Multi-Hop Relational Reasoning for Knowledge-Aware Question Answering. *ArXiv* abs/2005.00646.

Fleiss, J. L. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin* 76(5): 378.

Gordon, J.; and Van Durme, B. 2013. Reporting bias and knowledge acquisition. In *AKBC '13*. ACM.

Kearns, W. R.; Kaura, N.; Divina, M.; Vo, C. V.; Si, D.; Ward, T. M.; and Yuwen, W. 2020. A Wizard-of-Oz Interface and Persona-based Methodology for Collecting Health Counseling Dialog. *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* .

Lascarides, A.; and Asher, N. 1991. Discourse relations and defeasible knowledge. In *ACL*.

Lenat, D. B. 1995. CYC: A large-scale investment in knowledge infrastructure. *Communications of the ACM* .

Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *ACL*.

Li, X.; Taheri, A.; Tu, L.; and Gimpel, K. 2016. Commonsense Knowledge Base Completion. In *ACL*.

Lin, B. Y.; Chen, X.; Chen, J.; and Ren, X. 2019. KagNet: Knowledge-Aware Graph Networks for Commonsense Reasoning. In *EMNLP/IJCNLP*.

Lin, C.-Y. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*.

Liu, Y.; Yang, T.; You, Z.; Fan, W.; and Yu, P. S. 2020. Commonsense Evidence Generation and Injection in Reading Comprehension. In *SIGDIAL*.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. BLEU: a method for automatic evaluation of machine translation. In *ACL*.

Petroni, F.; Rocktäschel, T.; Lewis, P.; Bakhtin, A.; Wu, Y.; Miller, A. H.; and Riedel, S. 2019. Language Models as Knowledge Bases? In *EMNLP*.

Radford, A.; Narasimhan, K.; Salimans, T.; and Sutskever, I. 2018. Improving language understanding by generative pre-training URL https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1(8): 9.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2019. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *ArXiv* abs/1910.10683.

Roberts, A.; Raffel, C.; and Shazeer, N. 2020. How Much Knowledge Can You Pack Into the Parameters of a Language Model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 5418–5426.

Romero, J.; Razniewski, S.; Pal, K.; Z. Pan, J.; Sakhadeo, A.; and Weikum, G. 2019. Commonsense properties from query logs and question answering forums. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 1411–1420. URL https://arxiv.org/pdf/1905.10989.pdf.

Sap, M.; Le Bras, R.; Allaway, E.; Bhagavatula, C.; Lourie, N.; Rashkin, H.; Roof, B.; Smith, N. A.; and Choi, Y. 2019. ATOMIC: An Atlas of Machine Commonsense for If-Then Reasoning. In *AAAI*.

Shwartz, V.; West, P.; Bras, R. L.; Bhagavatula, C.; and Choi, Y. 2020. Unsupervised Commonsense Question Answering with Self-Talk. *ArXiv* abs/2004.05483.

Speer, R.; Chin, J.; and Havasi, C. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *AAAI*.

Tamborrino, A.; Pellicanò, N.; Pannier, B.; Voitot, P.; and Naudin, L. 2020. Pre-training Is (Almost) All You Need: An Application to Commonsense Reasoning. In *ACL*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*.

Vedantam, R.; Lawrence Zitnick, C.; and Parikh, D. 2015. Cider: Consensus-based image description evaluation. In *CVPR*, 4566–4575.

Wang, A.; Pruksachatkun, Y.; Nangia, N.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; and Bowman, S. R. 2019. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. *ArXiv* abs/1905.00537.

Zellers, R.; Holtzman, A.; Bisk, Y.; Farhadi, A.; and Choi, Y. 2019. HellaSwag: Can a Machine Really Finish Your Sentence? In *ACL*.

Zhang, H.; Khashabi, D.; Song, Y.; and Roth, D. 2020a. TransOMCS: From Linguistic Graphs to Commonsense Knowledge. In *IJCAI*.

Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K.; and Artzi, Y. 2020b. BERTScore: Evaluating Text Generation with BERT. *ArXiv* abs/1904.09675.