Of Few-Shot Commonsense Knowledge Models

Anonymous ACL submission

Abstract

Providing natural language processing systems with commonsense knowledge is a critical challenge for achieving grounded language understanding. Recently, commonsense knowledge models (Bosselut et al., 2019) have emerged as a suitable approach for hypothesizing situation-relevant commonsense knowledge on-demand in natural language applications. However, these systems remain limited by the data investment required to train them — $O(10^5)$ manually annotated examples.

To address this limitation, we investigate training commonsense knowledge models in a few-shot setting with limited tuples per commonsense relation in the graph. We perform five separate studies on different dimensions of few-shot commonsense knowledge learning, providing a roadmap on best practices for training these systems efficiently. Importantly, we find that human quality ratings for knowledge produced from a few-shot trained system can achieve performance within 6% of knowledge produced from fully supervised systems.

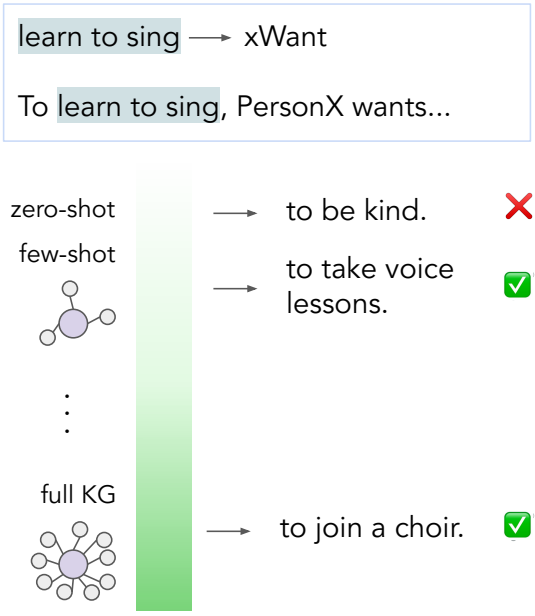

Figure 1: We show that commonsense knowledge models can be trained to effectively hypothesize commonsense knowledge in few-shot settings. The produced tuple quality approaches that of fully supervised systems.

1 Introduction

Language understanding systems that are grounded to world knowledge remain elusive. While large-scale language models have led to considerable advances on popular NLP benchmarks (Wang et al., 2019b,a), the desired interplay between knowledge and language representations is not reliably captured by current leading training schemes for NLP systems. Consequently, practitioners turn to commonsense knowledge graphs to ground language with explicit rules about the world.

However, commonsense knowledge graphs are limited in breadth and diversity, and can not be grown with high fidelity to the scale needed to model general purpose commonsense knowledge. Manual approaches to knowledge graph construction yield reliable and precise annotations of commonsense relationships, but are cost-prohibitive: human annotation is not cheap (Speer et al., 2017; Sap et al., 2019). Automatic, extraction-based methods achieve much greater scale (Zhang et al., 2020), but yield tuples of lower precision and questionable fidelity (Hwang et al., 2021). Reporting bias (Gordon and Van Durme, 2013) — the idea that obvious details go unstated in text (Grice et al., 1975) — limits the degree of useful commonsense knowledge that can be directly extracted from text.

More recently, commonsense knowledge models have emerged as a potential solution to this bottleneck (Bosselut et al., 2019). These models

learn to represent knowledge graphs implicitly and accessibly. They are pretrained on large text corpora (Radford et al., 2019; Lewis et al., 2020; Raffel et al., 2019) and then further fine-tuned on examples from a knowledge graph. This two stage process allows them to transfer implicit, but inaccessible, representations of knowledge learned from language (Petroni et al., 2019) to the task of hypothesizing declarative knowledge. Since their inception, they have become a popular mechanism for providing commonsense knowledge *on-demand* to downstream NLP systems (Chakrabarty et al., 2020; Ammanabrolu et al., 2020; Kearns et al., 2020; Majumder et al., 2020).

Despite their successes, commonsense knowledge models remain fundamentally restricted by the knowledge graph used to support knowledge transfer. They require large numbers of training examples (in the order of hundreds of thousands; Bosselut et al., 2019), and they can only learn the relations that make up the schema of the knowledge base. In cases where broader notions of common sense are required for situational understanding, these systems cannot generalize from the fixed relationships present in their seed KG. To reach broad applicability for NLP systems, knowledge models must be able to rapidly ingest new formulations of commonsense knowledge.

In response, we perform the first comprehensive study of the few-shot learning potential of commonsense knowledge models. We explore five different dimensions for knowledge model transfer in the few-shot setting: learning vs. context augmentation, model scale, input representation, example selection, and learning schedule. Along these axes, we provide novel insights into the few-shot learning behavior of knowledge models. Our empirical results showcase the following main takeaways:

- Few-shot learning is more flexible than few-shot adaptation (§4.1), allowing a 16x smaller model to exceed generalization performance of GPT-3 (Brown et al., 2020).
- Under similar experimental settings, commonsense knowledge models with more parameters generalize more effectively (§4.2).
- Expressive prompting that uses language descriptions of relations yields more efficient transfer — 10x-100x fewer examples (§4.3).
- When annotating examples for few-shot transfer, situation breadth (*i.e.*, diverse knowledge tuple heads) is more beneficial than situation depth (*i.e.*, multiple tails per head) (§5.1).
- Pretraining on other relations improves zero-shot relation induction, but benefits subside when as few as 30 examples of a new relation are available for transfer (§5.2).

2 Background

In this work, we investigate commonsense knowledge model performance under different model and training settings for few-shot learning. Below, we describe background concepts that are helpful for contextualizing these analyses.

2.1 Commonsense Knowledge Graphs

Commonsense knowledge graphs are structured, relational representations of commonsense knowledge. In our study, we use the ATOMIC$^{20}_{20}$ (Hwang et al., 2021) knowledge graph, 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 social and physical aspects of common human experiences. Across its 1.33M tuples, ATOMIC$^{20}_{20}$ captures information about 23 relationship types: 9 relations about social interactions, 7 physical-entity commonsense relations, and 7 event-centered commonsense relations. Example head entities and relations can be found in Table 2. The ATOMIC$^{20}_{20}$ knowledge graph is adversarially split into training, development, and test subsets such that no head entities in one set appear in any other. This property allows models trained on these resources to be evaluated on their capacity to generalize commonsense relationships to new entities.

2.2 Commonsense Knowledge Models

Commonsense knowledge models represent facts by learning to encode a commonsense knowledge graph (Bosselut et al., 2019). They are seeded with language models and are provided knowledge tuples as training data to learn to hypothesize knowledge relationships through language generation. After training on a large collection of tuples from a knowledge graph, they learn the structure and relationships of that knowledge graph. Furthermore, because they are seeded with pretrained language models, they learn to generalize the relationships to other entities about which the language model implicitly encodes knowledge (Petroni et al., 2019; Roberts et al., 2020). Consequently, they can be

Relation r	Prompt
ObjectUse	h is used for t
AtLocation	You are likely to find a h in a t
xIntent	Because of h, PersonX wanted t
xWant	After h, PersonX would want t.
xAttr	h is seen as t
isAfter	Something that happens after h is t
oWant	As a result of h, others would want t

Table 1: Examples of natural language prompts used to represent input for knowledge models. Prompts significantly speed up transfer in few-shot learning settings.

used to produce precise knowledge on-demand for any entity that can be expressed through language.

2.3 Few-Shot Learning vs. Augmentation

Recently, the term "few-shot" has taken two meanings: the classical definition of training on limited examples, and a new definition linked to few-shot context augmentation (Brown et al., 2020). In few-shot learning, language models are trained directly on a limited set of examples from the knowledge graph. In few-shot augmentation (or adaptation), models are given examples as a prepended augmentation of their context and can attend to these examples to recognize the structure of the task. While the results of few-shot adaptation have been impressive, classical few-shot learning has key advantages. Larger example sets will exceed maximum context windows for transformer language models, limiting the number of examples that can be used for few-shot adaptation. This bottleneck potentially lowers performance when additional training examples could be used for adaptation (§4.1).

3 Experimental Setup

To address the challenge of few-shot learning of knowledge models, we set up empirical studies to evaluate the effect of different modeling, training, and data considerations. In this section, we outline our final approach for training a few-shot learner. In the following sections (§4, §5), we describe studies that led to this system's design.

3.1 Input

The schema of Atomic$^{20}_{20}$ contains 23 relations across social and physical commonsense scenarios. Each link in the knowledge graph is composed of a {head *h*, relation *r*, tail *t*} triplet. The head and tail entities in the triplet are natural language words or phrases. Commonsense knowledge models are trained by providing the tokens of *h* and *r* as inputs to the model and learning to generate the tokens of *t*. In this work, rather than initializing *r* with a new token and random embedding (Bosselut et al., 2019), we automatically format the input tuples into natural language prompts to represent the relation (Feldman et al., 2019; Jiang et al., 2020). Table 3 shows examples of such prompts.

3.2 Few-Shot Data Generation

We source training tuples from the Atomic$^{20}_{20}$ training set (Hwang et al., 2021). When constructing few-shot training sets, we set a target number n of examples to randomly sample from the knowledge graph for each relation (*i.e.*, $n = 5 \implies 5$ examples $\times 23$ relations = 115 total training examples). Unless stated otherwise (§5.1), for each relation, n examples are selected by randomly selecting n head entities, and then selecting one tail entity (connected through that relation) for each head entity. This procedure ensures a diversity of head entities for training because each head can have a multiple connected tail entities in the graph.

3.3 Training

Once a training set of examples is produced, the model is trained on this subset to minimize the negative log-likelihood of the tokens of the tail entity for each tuple. We use a constant learning rate of 0.001, a mini-batch size of 4, and train the model for 3 epochs. Unless stated otherwise (§4.2), we use T5-11B (Raffel et al., 2019) as a seed language model for all experiments. We checkpoint the model after each few-shot training epoch and select the best one based on training loss (a development set is available, but using it for early stopping would violate the few-shot setting).

3.4 Evaluation

We evaluate the knowledge hypothesized by the trained few-shot models using human and automatic evaluations. For the human evaluation (Accept % in Table 3), we use the procedure described in Hwang et al. (2021). We ask annotators to label the quality of the {*given head, given relation, generated tail*} tuple using a 4-point Likert scale. Our scale corresponds to the following assessments of plausibility about the generated tuple: {*always/often true* $(+2)$, *sometimes/likely true* $(+1)$, *false/untrue* (-1), *nonsensical* (-2)}. We collect 3 annotations per relation, convert each annotation to an acceptability label (*i.e.*, $\{+1, +2\} \rightarrow$

Head h	Relation r	Generated Tail t (COMET)	Generated Tail t (GPT-3)
nail	ATLOCATION	construction site ✓	wall ✓
state highway	OBJECTUSE	statewide transportation ✓	drive a car ✗
video camera	OBJECTUSE	video recording ✓	record ✓
PersonX takes it to the vet	HINDEREDBY	PersonX doesn't have money to pay the vet ✓	PersonX gets a new pet ✗
PersonX makes PersonY very sick	HINDEREDBY	PersonX isn't close to PersonY. ✓	PersonY is not sick ✓
PersonX finds another job	ISAFTER	PersonX leaves job ✓	PersonX gets a new job ✗
PersonX falls ill	ISBEFORE	PersonX feels sick ✓	to be happy with personx ✗
PersonX gets a call for an interview	XATTR	qualified ✓	hopeful ✓
PersonX wants to learn how to swim	XATTR	PersonX isn't confident in the water ✓	to be able to swim ✗
PersonX falls ill	XEFFECT	is hospitalized ✓	they are under the weather ✓
PersonX sprays by a skunk	XEFFECT	PersonX will be sick ✓	their smell ✓
PersonX works really hard	XINTENT	to be appreciated ✓	to be rewarded ✓
PersonX misses class	XNEED	to have a valid excuse ✓	to be in class ✗
PersonX notices a strange smell	XWANT	to investigate the smell ✓	to get rid of it ✓
PersonX wants to learn how to sing	XWANT	to take voice lessons ✓	to learn how to sing ✗

Table 2: Examples of few-shot ($n = 5$) generations produced by COMET (T5) and GPT-3. We find that COMET (T5) is able to produce diverse and novel tail hypotheses despite learning from only a few examples.

✓, $\{-1, -2\} \rightarrow$ ✗) and use the majority label as the acceptability judgment. Evaluation agreement is measured using Fleiss's $\kappa = 0.49$ for the acceptability judgment (*i.e.*, moderate agreement between raters). Due to the cost of human annotation, and the number of experiments pursued, we use automated metrics, such as BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), ROUGE-L (Lin, 2004), CIDEr (Vedantam et al., 2015)) to report performance in most experiments.

4 How do knowledge models learn?

In this section, we survey modeling and training considerations for few-shot commonsense knowledge models (e.g. training using the COMET framework). First, we explore the general few-shot learning capability of large-scale language models. Then, we investigate the effect of model size on few-shot learning potential of language models. Finally, we explore the importance of natural language prompts for representing relations over symbolic representations.

4.1 Can commonsense knowledge models be trained in few-shots?

We find that the knowledge models trained in few-shots can learn to produce high-quality commonsense knowledge tuples (Table 3).

Motivation Most work on commonsense knowledge modeling is in fully supervised settings (Bosselut et al., 2019; Hwang et al., 2021), where models are trained on knowledge graphs with hundreds of thousands of examples. This scale requires expensive crowdsourcing, (Table 3), hindering efficient learning of new commonsense relationships. Pretrained language models offer a promising solution. They have shown that they can represent commonsense knowledge, and be queried for it in a zero-shot manner by converting knowledge graph relations to natural language prompts (Feldman et al., 2019). However, zero-shot induction of commonsense relationships from only language is not reliable (Jiang et al., 2020). For knowledge models to achieve broad applicability, they must efficiently learn new relationships from few examples.

Experiments We train COMET models as described in §3. In this experiment, we set the number of training examples per relation n=5 (the number of adaptation examples often selected for GPT-3). For our few-shot augmentation baselines (*e.g.*, GPT-{2,3}), we prepend the training examples (for the same relation) to the start of the input sequence. The baseline model conditions on these additional examples to generate the tokens of the predicted tail. For the fully supervised setting, we train on all tuples in ATOMIC$^{20}_{20}$. We also report the scores

Methodology	Model	BLEU-2	METEOR	ROUGE-L	CIDEr	Accept %	Cost / r
zero-shot	GPT-2 XL	2.8	8.2	9.8	4.7	36.6	$0
few-shot ($n = 5$)	GPT-2 XL (augmentation)	5.7	10.2	13.8	6.6	38.3	$0.50
	GPT-3 (augmentation)	15.3	18.2	25.5	17.5	73.0	
	COMET (T5) (learning)	21.9	19.5	25.7	19.2	78.6	
fully supervised	COMET (GPT-2 XL)	24.8	29.2	48.5	65.3	72.5	$4,347
	COMET (BART)	28.6	33.0	**49.5**	**65.8**	84.5	
	COMET (T5)	**28.6**	**33.5**	47.1	59.7	**84.6**	

Table 3: Comparison between various methods of training knowledge models. Few-shot ($n = 5$) knowledge models transfer well in both the learning (COMET) and augmentation (GPT-3) settings, suggesting that many of the beneficial effects knowledge modeling can be achieved from limited example sets. This result is promising for learning knowledge models across large relation sets, as the total costs of annotating few-shot example sets is much than large-scale crowdsourcing ($0.10 per tuple).

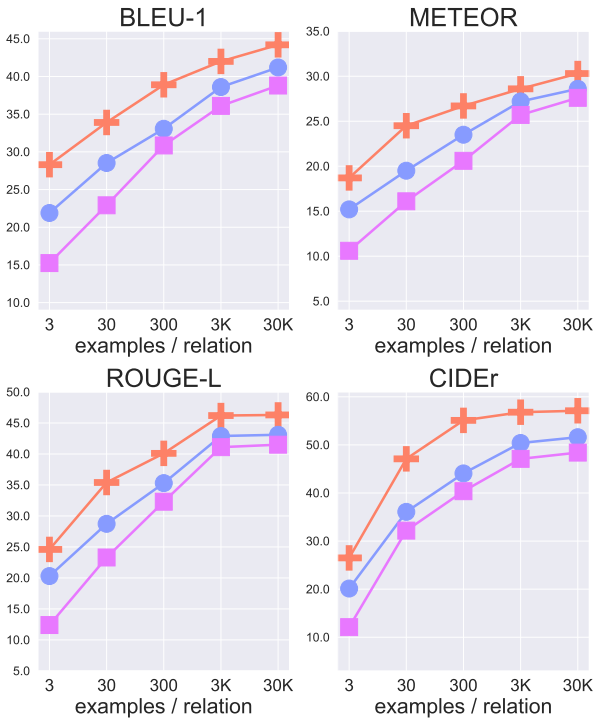

Figure 2: Effect of model size for commmonsense knowledge modeling. Small (60M;), Large (770M;), 11B (). The difference between model sizes is greatest in the few-shot settings.

for fully supervised COMET models from Hwang et al. (2021) seeded with GPT-2 XL (Radford et al., 2019) and BART (Lewis et al., 2020).

Findings We find that both the few-shot adaptation and few-shot learning settings are able to produce high quality tuples for large models. Using only n=5 tuples per relation, both GPT-3 and COMET (T5) outperform the fully-supervised COMET (GPT-2 XL) model. We also see a significant improvement from the few-shot COMET (T5) model over GPT-3, indicating that few-shot learning may be a richer adaptation strategy than few-shot augmentation. A model with 16× fewer parameters is able to transfer more successfully to the task. Finally, we observe that zero-shot knowledge elicitation from language (GPT-2 XL) is not a viable way of hypothesizing language tuples. In Table 2, we show examples of generations comparing the few-shot settings across different relations.

4.2 How does model size affect knowledge model learning?

We find that larger models generalize better in few-shot settings.

Method We train a few-shot COMET (T5) model across different pretrained language model sizes. We implement the COMET model with T5-Small (∼60M parameters), T5-Large (∼770M parameters), and T5-11B (∼11B parameters), and record their performance across different values of n for the few-shot training sets. Training hyperparameters remain the same between model sizes.

Findings In Figure 2, we observe that the performance difference between model sizes is largest in few-shot settings, but gradually decreases as more examples are available for training. However, the larger seed language models provide a consistent improvement regardless of the number of examples available for training.

4.3 How do prompts influence knowledge model learning?

Using natural language prompts to express relations accelerates learning commonsense knowledge models in few-shot settings (Figure 4).

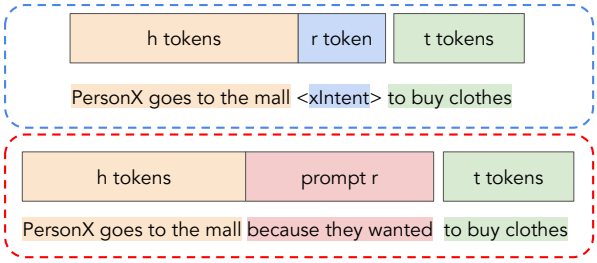

Figure 3: Illustration of knowledge model prompting. With prompting, the commonsense relation is converted using a natural language template.

Motivation Recent work has shows that prompts can help models elicit knowledge from pre-trained language representations (Feldman et al., 2019; Shin et al., 2020). However, eliciting knowledge through zero-shot prompting has drawbacks. First, because the language model is not explicitly trained for this purpose, the prompt may not yield salient knowledge in practice. Second, the output is sensitive to subtle variations in the construction of the language prompt (Jiang et al., 2020). Here, we explore whether prompts can accelerate few-shot learning in a stable manner.

Method We explore two different settings for modeling relations in knowledge models (Figure 3). In the first, we initialize natural language prompts for each relation following the procedure defined in Section 3.1. Example prompts for each relation can be found in Table 1. In the second, we follow Bosselut et al. (2019) and initialize a unique token for each relation, which is appended to the tokens of the head entity and maps to a unique learnable embedding for each relation.

Findings In Figure 4, we see that knowledge models can efficiently learn from fewer examples when relations are represented using natural language prompts. Prompts are especially important in the more restrictive few-shot settings where there is little signal to learn a relation embedding from scratch. Once approximately 3000 examples per relation are available for training, the performance between the models trained with and without prompts is similar, suggesting a point where prompts no longer help. Interestingly, we find a steeper slope for the prompt model when jumping from 3 to 30 examples per relation, implying that the model requires a minimum number of examples to map an understanding of the relation to the prompt words during few-shot training.

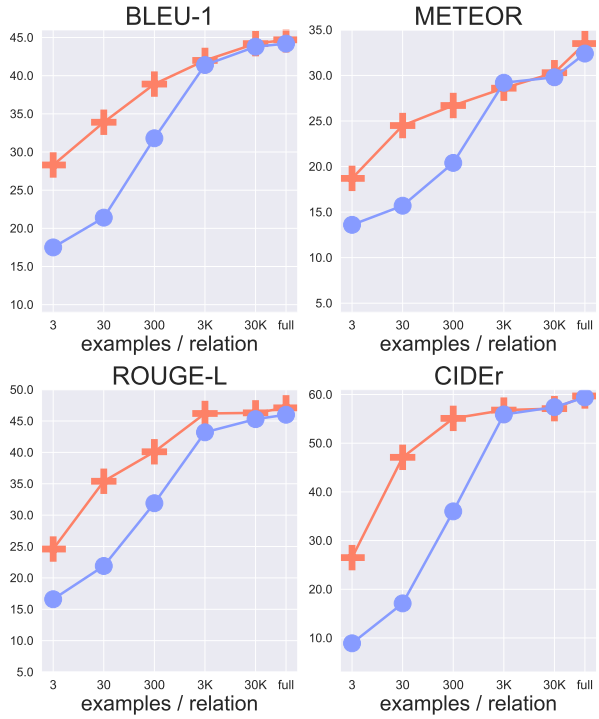

Figure 4: Comparison of training using natural language prompting (+) versus previous methods utilizing a relation embedding (●). We show that priming methods improves the data efficiency curve of knowledge model training.

5 How should we annotate new knowledge relations?

In this section, we design two studies to explore the effect of training set construction on few-shot learning of commonsense knowledge models, guiding future annotation efforts in commonsense knowledge graphs. First, we compare the tradeoff of example breadth and example depth by evaluating the learning improvement from training on different head entities for the same relation (*i.e.*, breadth) or different tail entities for the same head entity (*i.e.*, depth). Second, we explore whether pretraining on other relations in the graph can help few-shot transfer for a specific target relation, simulating a situation where we may want to learn a new commonsense relation in an online manner.

5.1 Heads or Tails: Example Breadth vs. Example Depth

Training knowledge models with diverse heads, rather than many tails per head leads to better few-shot learning performance (Figure 6).

Method In this setting, we compare two strategies for sampling examples for few-shot learning

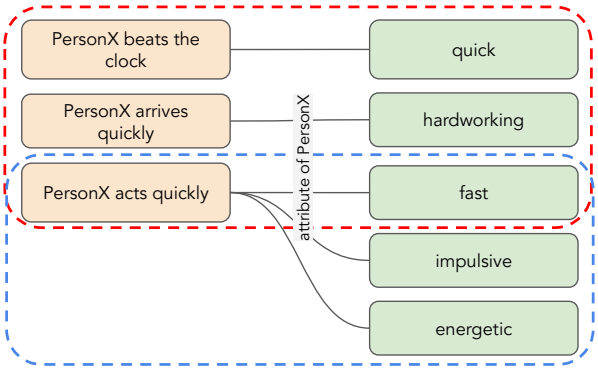

Figure 5: Illustration of sampling $n = 3$ tuples from a knowledge graph for the same relation by prioritizing diverse head entities (✚) or multiple tail entities per head entity (●).

(Figure 5). In the first, we sample diverse head entities (*i.e.*, unique seed instances), and limit any head entity to be sampled only once in the few-shot training set. In the second strategy, we sample multiple tail entities per head entity. Here, the model can learn from a richer set of commonsense relationships for each head entity, but at the expense of learning from fewer head entities overall. We sample head entities that are linked to at least 5 tail entities through the same relation in ATOMIC$^{20}_{20}$, and collect all tails for those sampled heads.

Findings In Figure 6, we see a large empirical gain when training on diverse heads rather than training on more tails per head.[1] While this difference is most notable for small n (*i.e.*, more extreme few-shot settings), a gap persists across all values of n. Consequently, when annotating commonsense knowledge graphs for few-shot transfer, we advise annotating diverse head entities for each relation, rather than annotating multiple inferences for the same head entities (*e.g.*, training with contrast sets; Gardner et al., 2020).

5.2 Does knowledge of other relations help to learn a *new* relation?

Models benefit from pretraining on the other relations in ATOMIC$^{20}_{20}$ in ultra few-shot settings. As the number of examples for a relation increases, pretraining no longer helps (Figure 8).

Method In this study, we select a subset of 6 relations from ATOMIC$^{20}_{20}$ as the few-shot set. Then, we pretrain the knowledge model on all examples

[1]We do not endorse extrapolating these findings to other common formulations of the "heads or tails?" question.

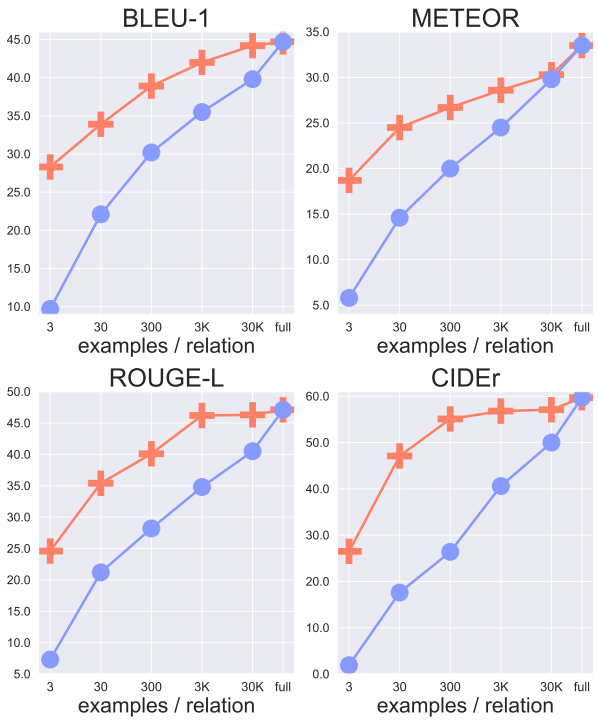

Figure 6: Comparison of training on examples with diverse head entities (✚) or multiple tail entities per head (●). Our results suggest that example breadth is more important for training knowledge models than example depth.

of the remaining relations in ATOMIC$^{20}_{20}$. As before, we use natural language prompts to represent relations, as we want knowledge about relations to transfer between them, which symbolic relation representations would hinder. We then perform few-shot training on the set-aside relations using n examples from each relation. Here, we cap our study at a maximum $n = 300$, as we see no improvement from pretraining with larger n. We compare these results to a baseline trained on n examples from the set-aside relations with no pretraining.

Findings We find that when more than 30 examples of a particular relation are available (*i.e.*, annotated), the benefit of pretraining on other relations evaporates. This result is surprising given the diversity of head entities available among the pretraining tuples. We would expect the model to see greater benefit from seeing many commonsense relationships during pretraining. Consequently, we conclude that annotating even a few examples of a new relation is the most fruitful way to improve a knowledge model's understanding of that relation.

However, it appears that knowledge models do

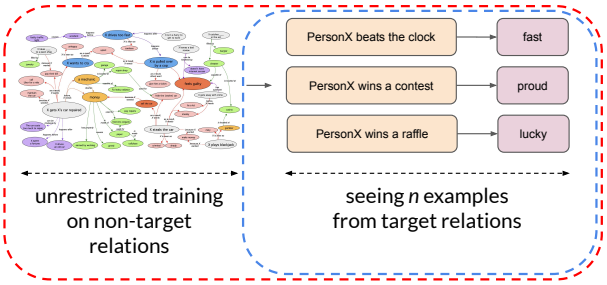

Figure 7: We investigate two training schedules: 1) Pre-training on the seed relations, followed by training on n examples of the target relations (✚); 2) Training only on n examples of the target relations (●).

manage to transfer some information between relations, with zero-shot relation induction improving drastically when the knowlege model is trained on other relations in the graph.

6 Related Work

Commonsense Knowledge Graphs Our work uses ATOMIC$^{20}_{20}$ as the transfer commonsense knowledge graph, but other works have designed CSKGs as well. Part of ATOMIC$^{20}_{20}$ is built off ATOMIC, a crowdsourced graph with 9 social commonsense relations (Sap et al., 2019). ConceptNet (Speer et al., 2017), was also partially manually constructed from crowdsourced statements of commonsense knowledge. Recently, Zhang et al. (2020) built off the ConceptNet schema by constructing a knowledge graph using automatically converted syntactic parses from sentences. More recent works have explored adding context to knowledge graph head entities to provide a richer learning space for commonsense knowledge models (Mostafazadeh et al., 2020).

Commonsense Knowledge Models Our work uses commonsense knowledge models, first proposed by Bosselut et al. (2019), to learn from commonsense knowledge graphs. Hwang et al. (2021) also trained commonsense models on ATOMIC$^{20}_{20}$, but focused on fully-supervised learning. Other works have developed commonsense knowledge models that are grounded to visual scenes (Park et al., 2020; Da et al., 2020), requiring multimodal commonsense inference generation. Recent works extend commonsense knowledge models beyond generating single-hop inferences and generate multi-branch (Bosselut et al., 2021) and multi-hop (Wang et al., 2020b) inferential structures. Commonsense knowledge base completion is also a closely related task to commonsense inference generation (Li et al., 2016; Saito et al., 2018). Recent works on this task combine language and graph structure representations for improved generalization (Malaviya et al., 2020; Wang et al., 2020a).

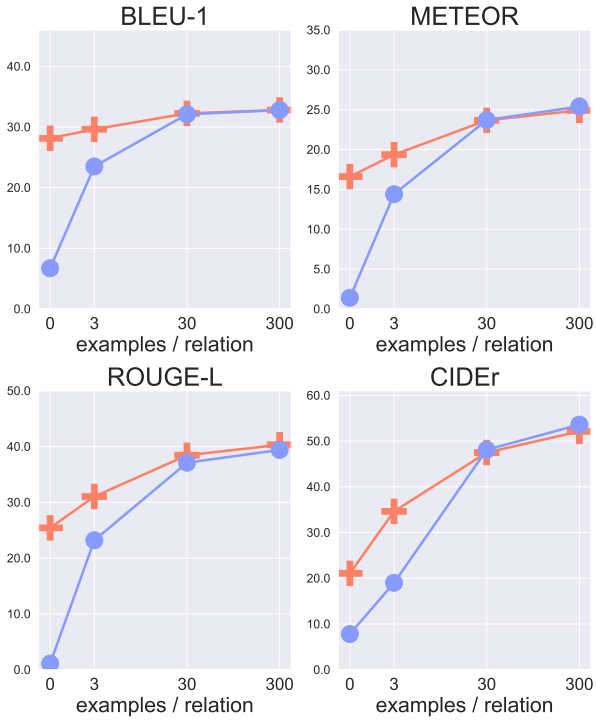

Figure 8: Comparison of few-shot training only on target relations (✚) compared to pre-training on seed relations in the ATOMIC$^{20}_{20}$ training set, and then few-shot training on the target relations (●).

7 Conclusion

In this work, we investigate five different dimensions of few-shot adaptation for knowledge model learning: adaptation strategy (*i.e.*, learning vs. augmentation), model size, relation prompting, few-shot training set selection, and knowledge graph pretraining. Our studies yield a roadmap for efficient few-shot learning of knowledge models. We use these insights to train a few-shot knowledge model that exceeds the performance of GPT-3 on commonsense knowledge hypothesization, and comes within 6% of the performance of a fully-supervised model trained on all of ATOMIC$^{20}_{20}$.

References

Prithviraj Ammanabrolu, W. Cheung, William Broniec, and M. Riedl. 2020. Automated storytelling via causal, commonsense plot ordering. *ArXiv*, abs/2009.00829.

S. Banerjee and A. Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *IEEvaluation@ACL*.

Antoine Bosselut, Ronan Le Bras, and Yejin Choi. 2021. Dynamic neuro-symbolic knowledge graph construction for zero-shot commonsense question answering. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence (AAAI)*.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, A. Çelikyilmaz, and Yejin Choi. 2019. Comet: Commonsense transformers for automatic knowledge graph construction. In *ACL*.

T. Brown, B. Mann, Nick Ryder, Melanie Subbiah, J. Kaplan, P. Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, G. Krüger, Tom Henighan, R. Child, Aditya Ramesh, D. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, E. Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, J. Clark, Christopher Berner, Sam McCandlish, A. Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *ArXiv*, abs/2005.14165.

Tuhin Chakrabarty, Debanjan Ghosh, Smaranda Muresan, and Nanyun Peng. 2020. R^3: Reverse, retrieve, and rank for sarcasm generation with commonsense knowledge. In *ACL*.

Jeff Da, M. Forbes, Rowan Zellers, Anthony Zheng, Jena D. Hwang, Antoine Bosselut, and Yejin Choi. 2020. Edited media understanding: Reasoning about implications of manipulated images. *ArXiv*, abs/2012.04726.

J. Feldman, Joe Davison, and Alexander M. Rush. 2019. Commonsense knowledge mining from pretrained models. *ArXiv*, abs/1909.00505.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. 2020. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1307–1323, Online. Association for Computational Linguistics.

Jonathan Gordon and Benjamin Van Durme. 2013. Reporting bias and knowledge acquisition. In *Proceedings of the 2013 workshop on Automated knowledge base construction*, pages 25–30. ACM.

H Paul Grice, Peter Cole, Jerry Morgan, et al. 1975. Logic and conversation. *1975*, pages 41–58.

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. *AAAI*.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics (TACL)*.

William R. Kearns, Neha Kaura, Myra Divina, Cuong Viet Vo, Dong Si, Teresa M. Ward, and Weichao Yuwen. 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*.

M. Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, A. Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *ArXiv*, abs/1910.13461.

Xiang Li, Aynaz Taheri, Lifu Tu, and Kevin Gimpel. 2016. Commonsense knowledge base completion. In *ACL*, volume 1, pages 1445–1455.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *ACL 2004*.

Bodhisattwa Prasad Majumder, Harsh Jhamtani, Taylor Berg-Kirkpatrick, and Julian McAuley. 2020. Like hiking? you probably enjoy nature: Persona-grounded dialog with commonsense expansions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9194–9206, Online. Association for Computational Linguistics.

Chaitanya Malaviya, Chandra Bhagavatula, Antoine Bosselut, and Yejin Choi. 2020. Commonsense knowledge base completion with structural and semantic context.

Nasrin Mostafazadeh, Aditya Kalyanpur, Lori Moon, David Buchanan, Lauren Berkowitz, Or Biran, and Jennifer Chu-Carroll. 2020. GLUCOSE: GeneraLized and COntextualized story explanations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4569–4586, Online. Association for Computational Linguistics.

Kishore Papineni, S. Roukos, T. Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*.

J. Park, Chandra Bhagavatula, R. Mottaghi, Ali Farhadi, and Yejin Choi. 2020. Visualcomet: Reasoning about the dynamic context of a still image. In *ECCV*.

F. Petroni, Tim Rocktäschel, Patrick Lewis, A. Bakhtin, Y. Wu, Alexander H. Miller, and S. Riedel. 2019. Language models as knowledge bases? *ArXiv*, abs/1909.01066.

A. Radford, Jeffrey Wu, R. Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, W. Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *ArXiv*, abs/1910.10683.

Adam Roberts, Colin Raffel, and Noam Shazeer. 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)*, pages 5418–5426, Online. Association for Computational Linguistics.

Itsumi Saito, Kyosuke Nishida, Hisako Asano, and Junji Tomita. 2018. Commonsense knowledge base completion and generation. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 141–150.

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. *ArXiv*, abs/1811.00146.

Taylor Shin, Yasaman Razeghi, IV RobertLLogan, Eric Wallace, and S. Singh. 2020. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *ArXiv*, abs/2010.15980.

Robyn Speer, J. Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. *ArXiv*, abs/1612.03975.

Ramakrishna Vedantam, C. L. Zitnick, and D. Parikh. 2015. Cider: Consensus-based image description evaluation. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019a. Superglue: A stickier benchmark for general-purpose language understanding systems. In *NeurIPS*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019b. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *ICLR*.

Bin Wang, Guangtao Wang, J. Huang, Jiaxuan You, J. Leskovec, and C. J. Kuo. 2020a. Inductive learning on commonsense knowledge graph completion. *ArXiv*, abs/2009.09263.

Peifeng Wang, Nanyun Peng, Filip Ilievski, Pedro Szekely, and Xiang Ren. 2020b. Connecting the dots: A knowledgeable path generator for commonsense question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4129–4140, Online. Association for Computational Linguistics.

Hongming Zhang, Daniel Khashabi, Y. Song, and D. Roth. 2020. Transomcs: From linguistic graphs to commonsense knowledge. *ArXiv*, abs/2005.00206.

8 Appendix

8.1 Accuracy in zero-shot MLM setting

	BLEU-1	BLEU-2
T5 - Zero-shot	0.067	0.017

While little work has explored few-shot knowledge completion, recent works have investigated performance of zero-shot knowledge graphs (Petroni et al., 2019; Feldman et al., 2019). Thus, we investigate the ability of T5 to complete commonsense knowledge in a zero-shot setting. Different from the few-shot and supervised approaches, we do not use teacher forcing, but rather use prompts to leverage the masking objective of the language model pretraining. In addition to mask prediction, we try a couple variants. Since the mask only predicts several tokens at a time, for relations with longer responses (e.g. ATOMIC relations), we allow the model to predict up to 7 mask tokens in succession, or until the model predicts an empty string for the mask. We suggest that this is still only a workaround, and masked models are poor predictors of longer length tail entities.