Commonsense Knowledge Reasoning and Generation with Pre-trained Language Models: A Survey

Prajjwal Bhargava and Vincent Ng
Human Language Technology Research Institute
University of Texas at Dallas
Richardson, TX 75083–0688
prajjwal.bhargava@utdallas.edu, vince@hlt.utdallas.edu

Abstract

While commonsense knowledge acquisition and reasoning has traditionally been a core research topic in the knowledge representation and reasoning community, recent years have seen a surge of interest in the natural language processing community in developing pre-trained models and testing their ability to address a variety of newly designed commonsense knowledge reasoning and generation tasks. This paper presents a survey of these tasks, discusses the strengths and weaknesses of state-of-the-art pre-trained models for commonsense reasoning and generation as revealed by these tasks, and reflects on future research directions.

Introduction

Commonsense knowledge is the information that is generally accepted by the majority of people concerning everyday life, encapsulating the practical knowledge about how the world works. Reasoning with commonsense knowledge is at the core of building natural language understanding models and, more broadly, AI systems that can reason about the world in the same way as humans do.

A vast amount of work on commonsense knowledge acquisition and reasoning has traditionally been conducted in the knowledge representation and reasoning community (see Zang et al. (2013) for a survey). For instance, there have been notable attempts to manually create large-scale commonsense knowledge bases (e.g., Cyc (Lenat 1995)) and automatically acquire commonsense knowledge from the Web (e.g., Open Mind Common Sense (Singh et al. 2002)). More recently, the Winograd Schema Challenge, a pronoun resolution task that requires the use of commonsense knowledge, was proposed as a practical alternative to the Turing Test (Levesque et al. 2012).

The advent of the neural natural language processing (NLP) era has revolutionized virtually all areas of NLP research. One of the major breakthroughs is arguably the development of *pre-trained* language models (PLMs). Specifically, researchers have discovered that neural models can be trained (via a process known as *pre-training*) on a large body of *unannotated* text to acquire general knowledge about language, including both linguistic and commonsense knowledge. This has sparked tremendous interest in the NLP community in examining what kind of commonsense knowledge PLMs possess and the extent to which such knowledge can be exploited to address commonsense knowledge reasoning and generation tasks in the last few years.

Our goal in this paper is to provide the general AI audience with a timely survey of the recent advances in the NLP community on commonsense knowledge reasoning and generation using PLMs. Specifically, the focus of this survey is (1) the kind of commonsense knowledge that PLMs possess and (2) the extent to which such knowledge can be exploited for recently designed commonsense knowledge reasoning and generation tasks. For an overview of what *linguistic* knowledge PLMs possess, we refer the reader to Rogers et al. (2020). For comprehensive surveys of the details and the inner workings of PLMs, we refer the reader to Qiu et al. (2020), Han et al. (2021), and Kalyan et al. (2021).

Pre-trained Language Models

In this section, we provide the reader with the background on PLMs needed to understand the rest of the paper.

For a long time, *supervised* learning has been the most successful learning paradigm in NLP. For instance, training a neural model to perform a classification task in a supervised manner primarily involves training an *encoder* to encode the input as a *task-specific representation* that would be useful for classifying a given sample. In contrast, for a text generation task (e.g., text summarization, machine translation), one would typically employ an encoder-decoder neural architecture, in which the encoder first encodes the input, and then the decoder generates the output sequence token by token based on both the encoded input and the tokens that have been generated so far. The performance of supervised models is often limited by the (typically small) amount of data they are trained on.

Pre-training offers a solution to the aforementioned data scarcity problem. The idea is to first train a model on one or more *self-supervised* learning tasks during a process known as *pre-training*, and the resulting model, in which the weights have already been initialized during pre-training, can be *fine-tuned* using the (potentially small amount of) task-specific data in a supervised fashion. Self-supervised learning tasks are NLP tasks for which the label associated with a training instance can be derived automatically from

the text itself. Consider, for instance, one of the most well-known self-supervised learning tasks, Masked Language Modeling (MLM) (Devlin et al. 2019). Given a sequence of word tokens in which a certain percentage of tokens is *masked* randomly, the goal of MLM is to predict the masked tokens. As can be imagined, a model for MLM can therefore be trained on instances where each one is composed of a partially masked sequence of word tokens and the associated "class" value is the masked tokens themselves. Because no human annotation is needed, a model can be pre-trained on a very large amount of labeled data that can be automatically generated. Various studies have shown that pre-training allows a model to learn universal language representations that encode both linguistic and commonsense knowledge, and a PLM, after being fine-tuned, can offer substantially improved performance on a wide variety of NLP tasks.

Existing PLMs differ primarily in terms of (1) what is being pre-trained (is it the encoder, the decoder, or both of them?); (2) the self-supervised learning tasks used; and (3) the network architecture. While early work has focused on pre-training the encoder (e.g., BERT (Devlin et al. 2019), RoBERTa (Liu et al. 2019), ELECTRA (Clark et al. 2019)) or the decoder (e.g., GPT-2 (Svyatkovskiy et al. 2019)), recent work has focused on jointly pre-training the encoder and the decoder (e.g., T5 (Raffel et al. 2020), BART (Lewis et al. 2020)). The most successful PLMs are based on the Transformer (Vaswani et al. 2017), a fully-connected self-attention model. Throughout this paper we will simply use the term PLMs to refer to Transformer-based PLMs.

Capturing Commonsense Knowledge

How well do PLMs capture commonsense knowledge? Researchers have employed *probing* to answer this question. To probe a PLM for commonsense knowledge, most of the probing methods use a hand-crafted *template* to convert a relational fact from a knowledge base (KB), which is typically represented in the form of a triple $<s, r, o>$ where s is the SUBJECT, r is the RELATION, and o is the OBJECT, into a natural language sentence. One template needs to be defined for each relation. As illustrated in Figure 1, each triple having the KB relation "place of birth" would be translated to a sentence of the form "SUBJECT was born in OBJECT". Note that a template keeps the entities intact while approximately the RELATION to a set of hand-coded verbs/relations that can generalize on numerous entities (e.g., ATLOCATION may be translated to "is in"). One of the entities (i.e., the SUBJECT or the OBJECT) in the sentence will then be masked. The resulting *clozed* sentence can then provide an automated and flexible way to probe PLMs for stored knowledge. Specifically, if a PLM can fill in the blank in the given clozed sentence correctly (i.e., the answer is the same as the entity that appears in the triple originally used to generate the sentence), then the PLM is considered to possess the knowledge being probed. Note that we are not asking the PLM to derive *new* knowledge: only inference is performed by the PLM to check for stored knowledge.

Several key observations are being revealed via probing experiments. First, **PLMs are becoming a promising alternative to KBs.** BERT shows signs of capturing relational knowledge in a zero-shot setting reasonably well compared to supervised alternatives (Petroni et al. 2019). It can recall factual knowledge for some relations such as one-to-one, but it struggles to perform well on other relations such as N-to-M and N-to-one.

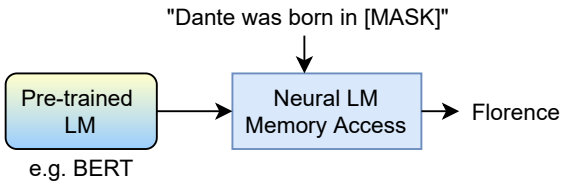

Figure 1: A common approach to probe PLMs for stored knowledge (Petroni et al. 2019).

Prompt	Model Predictions
A ___ has fur, is big, has claws, has teeth, is an animal, eats, is brown, and lives in woods.	**bear**, wolf, cat, ...

Table 1: Masked token predictions about stereotypical assumptions get refined as more properties are appended (Weir et al. 2020).

Second, **PLMs do not generalize well on unseen entities**. While BERT can predict the top 100 triples mined from Wikipedia fairly accurately, which suggests that BERT can generalize to specific data sources (Davison et al. 2019), PLMs do not generalize well on entities not encountered during pre-training due to their heavy reliance on memorization in the pre-training process (Logan et al. 2019).

Third, **PLMs can perform comparisons and categorizations of entities**. Specifically, they can compare physical objects along a particular attribute such as weight or size (e.g., a chair is *smaller than* a room) (Goel et al. 2019). When it comes to categorization, they work reasonably well on knowledge types that are ontological in nature, such as "mango isA fruit" (Hwang et al. 2021).

Finally, analyzing the top-k predictions made by PLMs on the association between an entity and its attributes, we see that **PLMs can learn stereotypical associations reasonably well** from large text corpora. As more properties are appended to provide contextual knowledge (see the example in Table 1), the performance of RoBERTa-L increases, with predictions going from being sensible to more acceptable as per human interpretation. Specifically, PLMs do better on functional (e.g., "eat fish") and encyclopedic (e.g., "are found in forests") knowledge than on visual-perceptual variants (e.g., "bears have fur"). Although this result is encouraging, when asked for widely acceptable properties about an entity, the ranked predictions provided by PLMs do not correlate strongly with those of humans (Weir et al. 2020).

Reasoning with Commonsense Knowledge

In this section, we examine how well PLMs perform commonsense reasoning by considering five types of commonsense reasoning tasks.

Linguistic Reasoning

Linguistic reasoning is concerned with understanding text for which the correct interpretation requires commonsense knowledge. A representative benchmark for linguistic reasoning is WINOGRANDE (Sakaguchi et al. 2020), which consists of Winograd schema-inspired problems that require linguistic, social or physical reasoning (Levesque et al. 2012). As an example, given the sentence "The plant took up too much room in the urn, because the ___ was large" and two answer candidates "plant" and "urn", the goal is to determine which candidate should be used to fill in the blank.

Several observations can be made based on the performance of PLMs on this and other linguistic reasoning tasks.

First, **BERT shows poor linguistic sensitivity** and becomes fragile on negated and misprimed sentences (Kassner and Schütze 2020; Ettinger 2020). For example, BERT fails to distinguish between the two sentences “Birds cannot [MASK]”) and “Birds can [MASK]” and tends to get distracted if they are prepended with "misprimes" such as “Talk? Birds can [MASK]”. The fact that its predictions do not change with such major changes indicates that BERT does not attend to the prominent cues in the desired manner.

Second, **PLMs perform poorly on numerical knowledge out-of-the-box** (Lin et al. 2020a; Wallace et al. 2019; Chen et al. 2019; Bhagavatula et al. 2020). For instance, given the sentence "Birds have [MASK] legs", BERT predicts "four" to be the answer, suggesting that pre-training does not facilitate the acquisition of numerical knowledge.

Third, **as the sentences in a given reasoning task require more turns of logical reasoning (i.e., the task becomes increasingly complex), BERT’s performance deteriorates** (Zhou et al. 2020b; Richardson and Sabharwal 2020; Huang et al. 2019). These are typically sentences with complex semantics such as riddles, where PLMs are required to understand figurative language (Lin et al. 2021a).

Several attempts have been made to improve the *robustness* of PLMs for linguistic reasoning tasks.

Semantic similarity. Niu et al. (2021) show that semantic similarity matching can be used to make PLMs robust against irrelevant factors such as word frequencies. Specifically, we can first use PLMs to generate plausible answers so that we can compute the similarity between each generated answer and each of the provided answer candidates, and then we can select the answer candidate that has the highest similarity score as the correct answer.

Attention maps. Klein and Nabi (2019) show that attention maps obtained from BERT can be used for coreference resolution in long sentences, suggesting their potential usefulness for pronoun disambiguation.

Numerical reasoning. To improve numerical reasoning, Geva et al. (2020) pre-train BERT on two numerical tasks, one involving predicting what comes after a sentence such as "3+4+5" and the other involving answering numerical questions (e.g., given a historical passage about Spain, answer the question of "How many Japanese families were in Spain?"), so that BERT is endowed with the ability to understand computations expressed in pseudo-natural language (text+digits).

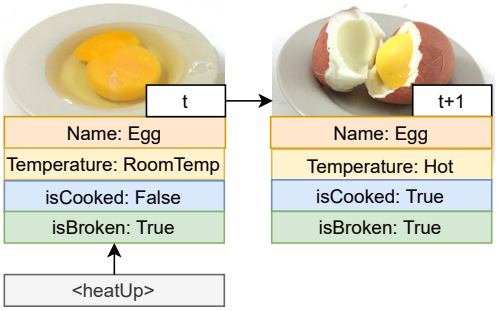

Figure 2: Simulating what might happen next in order to enable PLMs to encode language "form" and "meaning" (Zellers et al. 2021).

Reasoning about Physical World

Physical commonsense reasoning involves understanding concepts based on the physical properties of objects, including the *affordances* of objects (i.e., the actions applicable to them) and how they can be manipulated. A representative benchmark for physical commonsense reasoning is PIQA (Bisk et al. 2020). Given a sentence such as "When boiling butter, when it’s ready, you can ___ ", the goal is to fill in the blank with one of two answer options, such as "Pour it onto a plate" and "Pour it onto a jar".

Perception and interaction are among the key components behind how humans learn to reason about the physical world. However, static input representations which current neural models are fed are inadequate since they cannot compensate for the information humans derive from being in a dynamic physical world. So, a key question posed by PIQA is whether PLMs can reason over physical commonsense questions without interacting with the physical world.

Several observations can be made about the performance of PLMs on physical commonsense reasoning questions. First, **PLMs predominantly learn property associations that are explicitly mentioned in text**, achieving higher accuracies on entities that have simple affordances (e.g., "spoon") than on entities that have a long tail of affordances (e.g., "water"). Second, **PLMs struggle to understand fundamental relations** (e.g., "before/after", "top/bottom") and **find it hard to reason when common objects are used in unconventional ways** (e.g., a glue stick is used as a paperweight). Finally, although neural representations are dexterous at capturing the affordances ("boats can be driven") and properties ("boats require fuel") of objects, **PLMs struggle to understand the connection between affordances and properties** (Forbes et al. 2019; Zhao et al. 2020).

In light of the weakness associated with the lack of interaction with the physical world, Zellers et al. (2021) explore the benefits of providing PLMs access to world dynamics. World dynamics include information that one would obtain after interacting with objects. As an example, consider Figure 2. If an action `heatUp` is applied to a pan, the model will learn that the temperature of an egg has risen to become `hot` and is now in a `cooked` state. Predicting object states after an action has been taken drastically improves a PLM’s ability to make correct inferences about object states.

Abductive Reasoning

Abductive reasoning involves finding the most likely explanation for a set of incomplete observations. There are at least two representative benchmarks for abductive reasoning, COSMOSQA (Huang et al. 2019) and HELLASWAG (Zellers et al. 2019b). COSMOSQA is a commonsense comprehension task where, given a context, the goal is to choose the answer to the question based on the context from four answer candidates. This benchmark contains questions that require abductive reasoning, such as "what might I continue to do after the situation described in the context?" HELLASWAG is a benchmark in which the goal is to choose the best plausible ending of a given context out of four options.

Several observations can be made. First, Huang et al. (2019) attribute the errors made by their model on COSMOSQA to two reasons. First, **PLMs struggle on examples where the context becomes intricate enough to require cross-sentence interpretation and reasoning**. In such examples, PLMs are required to understand the important parts of the passage and jointly attend to the identified parts. In addition, **PLMs do not understand what situations are inconsistent with human commonsense**. For example, they may choose "leaving a baby alone at home is not safe" over "she would try to find a babysitter" when asked the question "what would happen if she does not find a daycare". Interestingly, when one of the answer options is "None of the above", PLMs often struggle to choose this option since the words in this option do not provide enough signal for why this option might be correct.

Second, **PLMs struggle with selecting the most plausible ending given a context for** HELLASWAG. Since a given context can have multiple correct endings, determining which one would be the most plausible requires prior reasoning of what humans relate to the most with their commonsense knowledge. Zellers et al. (2018, 2019b) show that when more surface cues are eliminated, PLMs are less likely to be able to predict the most plausible ending even though it might be trivial for humans to do so.

Social Reasoning

Social reasoning involves modeling the mental states of others and their likely actions to the extent that reasoning can be performed over their behaviors and emotions. A representative benchmark for social reasoning is SOCIALIQA (Sap et al. 2019b), which evaluates commonsense reasoning based on social situations and interactions. Consider the following example taken from SOCIALIQA, in which the correct answer is boldfaced:

"Context": "Tracy had accidentally pressed upon Austin in the small elevator and it was awkward."
"Question" : "Why did Tracy do this?"
"Choice A": "get very close to Austin";
"Choice B": "squeeze into the elevator";
"Choice C": "get flirty with Austin"

Several observations can be made. First, BERT finds examples of effects ("what will happen to X?") and motivation ("why did X do that to Y?") easier than those that involve understanding descriptions ("how would you describe X?") (Sap et al. 2019b). Second, it performs better on examples where the answer exhibits cues about emotions than those involving spatial commonsense (Bhagavatula et al. 2020).

Multimodal Reasoning

Textual representations are restricted to what can be expressed through natural language and therefore are unable to represent the multi-modal information that humans could have access to or infer by being in a dynamic world, such as a constant stream of images and a sequence of interactions in the physical world. Vision naturally becomes a next step towards enabling learning through joint interaction (Baldwin 1995). However, merely using raw visual images along with their descriptions is by no means sufficient to provide grounded understanding (Marasović et al. 2020). For example, inferring the intentions of the entities in images can only be well dealt with if we have some prior information (either behavioral or temporal) to rely on to make justifiable inferences. This has led the community to look into approaches that can help provide a tighter integration of linguistic and visual modalities.

There are two well-known benchmarks for multimodal reasoning. VISUAL COMMONSENSE REASONING (VCR) (Zellers et al. 2019a) seeks to answer cognition-level questions from images. Concretely, given an image with a list of regions and a question, the goal is to choose the answer to the question out of a set of possible candidates and provide a rationale that can explain why the chosen answer is correct. An example can be found in Figure 3. VISUAL COMMONSENSE GRAPHS (Park et al. 2020) checks how well PLMs reason about the dynamic context from a static image and an event. Specifically, given an image and a textual description of an event at present, the task is to generate the rest of the visual commonsense graph that is connected to the event. For example, given an image of a man who is drowning in the river and a textual description of the event, the goal is to generate a commonsense graph with nodes such as "the man wanted to save himself from drowning", "the man is waiting for help", "the man senses his own death", and "the man needs to swim towards the river bank". Empirical results reveal that for both benchmarks, models that exploit both visual and textual information outperform those that only use textual information. This suggests that **visual features help make higher quality commonsense inferences**.

Temporal Reasoning

Time is an inherent aspect of events. Broadly, temporal reasoning involves two subtasks. *Temporal attribute prediction* involves understanding an event mentioned in text or dialogue through reasoning with its temporal dimensions such as the duration of the event, when the event typically happens, how long the event is going to be stationary, and how often it happens. *Temporal relation identification* involves understanding how an event is temporally related to other events mentioned in the same text or dialogue (e.g., did an event take place *before* or *after* another event?). Temporal reasoning is challenging because the timestamp associated with an event and the aforementioned temporal dimensions

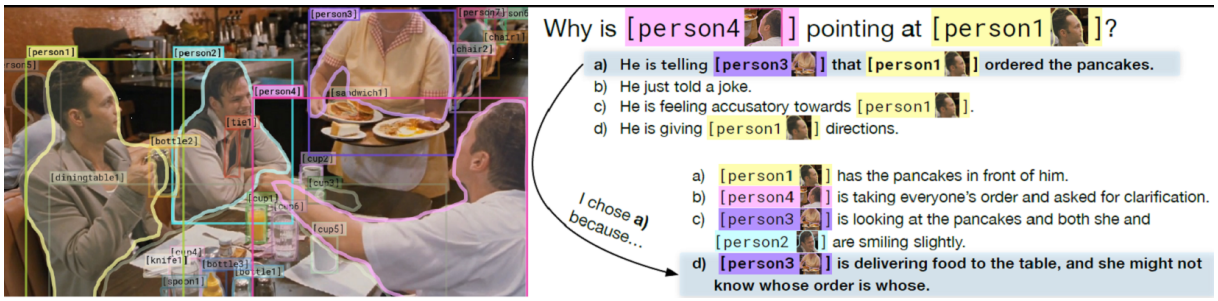

Figure 3: Learning to reason about dynamic context from a static image (Zellers et al. 2019a).

may not be mentioned explicitly and therefore need to be inferred.

Two commonly-used benchmarks have been developed for temporal reasoning. MC-TACO (Zhou et al. 2019) is a question-answering benchmark involving temporal commonsense comprehension. Here is an example:

> "Context": The massive ice sheet, called a glacier, caused the features on the land you see today
> "Question": When did the glacier start to impact the land's features ?
> "options": **a) centuries ago**; b) hours ago; c) 10 years ago; **d) tens of millions of years ago**

TIMEDIAL (Qin et al. 2021) involves temporal reasoning in dialogue. Here is an example:

> A: May we see the wine list please.
> B: Sure. Our special wine today is a 1989 Chardonnay.
> A: I'd like a bottle please.
> B: I'll need to see your ID please.
> A: Here you go.
> B: Sorry about the inconvenience, you look so young. I had to make sure you are over .
> **a) 21 years old**; b) 30 years old; c) 4 years old; **d) 18 years old**

Ideally, one can train *time-aware* PLMs to address these temporal reasoning tasks. An obstacle to the development of such PLMs concerns the lack of large-scale KBs that incorporate the notion of time into the facts that they encode over entities and events. For instance, the LOCATION relation (i.e., where a person lives) and the EMPLOYMENT relation (i.e., the company a person is affiliated with) are dependent on time, but existing KBs typically fail to encode the time period for which a given relation holds true. Such time-aware KBs should also encode temporal commonsense knowledge such as "if a student attends a university, s/he will likely graduate and work after a few years".

Given the lack of such KBs, time-aware PLMs can only be trained on the annotated training data provided by MC-TACO and TIMEDIAL. For instance, Zhou et al. (2020a) propose TACO-LM, a BERT-based PLM that is trained to be temporally aware by contextually estimating duration and time via (1) extracting the important events and their temporal information and then (2) asking the model to predict the masked tokens that talk about some temporal aspect. However, TACO-LM only provides marginal improvements over BERT w.r.t. duration, frequency, and when the event typically takes place. More recently, Qin et al. (2021) have shown that fine-tuned PLMs struggle to perform well on TIMEDIAL primarily because they largely fail to understand the context of the given dialogue and instead simply rely on shallow cues about the temporal patterns in the context.

Generating Commonsense Knowledge

Commonsense knowledge generation is a critical component in building commonsense knowledge resources. Broadly, we can divide commonsense knowledge generation tasks into two categories, as described below.

Knowledge Base Completion

A KB is a collection of relational facts, each of which is represented as a triple $<s, r, o>$, where s is the SUBJECT, r is the RELATION, and o is the OBJECT. KB completion is the task of automatically inferring missing facts by reasoning about the information already present in the KB.

To date, the most successful knowledge generation approach with PLMs is arguably Commonsense Transformer (COMET) (Bosselut et al. 2019). COMET can be used to generate o given s and r after being pre-trained on a knowledge base such as ConceptNet (Speer et al. 2017; Singh et al. 2002), which represents (mostly taxonomic) commonsense knowledge as a graph of concepts (words or phrases) connected by relations (edge types), or ATOMIC (Sap et al. 2019a), which is a large-scale KG consisting of textual descriptions of inferential knowledge (*if-then* relations).

Constrained Commonsense Text Generation

Next, we examine studies on how PLMs can be used to generate commonsense text subject to a set of constraints.

Tasks There are three benchmarks commonly used to evaluate commonsense generation approaches.

COMMONGEN (Lin et al. 2020b): Given a concept set (e.g., `{dog, frisbee, catch, throw}`), the goal is to generate a coherent sentence describing an everyday event using all the provided concepts.

COMMONSENSE EXPLANATIONS (COS-E) (Rajani et al. 2019): Given that a model selects an answer (from a set of candidates) to a given question, the goal is to generate an explanation of why the selected answer is correct. The resulting explanation may help us understand the reasoning that a model relies on to arrive at the selected answer.

αNLG (Bhagavatula et al. 2020): Given two observations/events that happen in two different timesteps, the goal is to generate a valid hypothesis h of what happened between the observations/events.

Challenges These benchmarks reveal that PLMs, when used as commonsense knowledge generators, suffer from several shortcomings.

- **Poor coherency:** The generated sentences do not necessarily adhere to the human notion of commonsense. For instance, given the concept set {dog, catch, throw, frisbee}, GPT2 generates the sentence "A dog throws a frisbee at a football player". Although this sentence is grammatically correct, it suffers from poor coherency.
- **Insufficient concept coverage:** PLMs continue to produce sentences that fail to include all concepts from the provided concept set. In the previous example, the concept "catch" was not used to generate the output.
- **Limited reasoning capability:** It is not clear what kind of reasoning is used by PLMs to arrive at an answer. Though a solution to explanation generation could shed light on this question, some studies show that existing approaches generate either trivial or noisy explanations, providing little or no evidence of how a PLM arrives at the selected answer (Ji et al. 2020a). Other studies show that the reasoning used by PLMs is often not fully correct (McCoy et al. 2019; Shwartz and Choi 2020). Overall, the reasoning capability of PLMs is far from satisfactory.

Improving Coherency and Concept Coverage Several approaches have been proposed to address two of the challenges, poor coherency and insufficient concept coverage.

Using prototypes. Guu et al. (2018) propose a sentence generation mechanism that involves selecting a sentence from the training data (known as a *prototype*) and editing it into a form that satisfies a given set of constraints. Their hypothesis is that it is easier to edit a sentence that is grammatically correct and semantically coherent than to generate one from scratch. If provided with a concept set `{trailer, shirt, side, sit, road}` from COMMONGEN, a PLM may generate "A man sits on the side of a trailer and a shirt", whereas a prototype such as "Two guys in red shirts are sitting on chairs, by the side of the road, behind that open trailer" may be edited by the PLM to form "a man in a white shirt and black pants sits on the side of a trailer on the road", which has better coverage and coherency.

Using knowledge graphs (KGs). KGs play a crucial role in enabling PLMs to improve the semantic correctness (and thus coherency) of text as they can provide PLMs with information that may not be captured reliably by PLMs, such as entity representations and their dependency relations (i.e., how concepts are related). For instance, Li et al. (2021) extract concept-specific relations from a KG and inject them into a PLM to make the generated text more coherent.

Reasoning over multi-hop relational paths in KGs. The sparse connections between the nodes in KGs may make it hard for PLMs to learn rich relations from them. These rich relations, however, may be needed by PLMs to generate commonsense sentences with rich semantic structures in order to boost coherency. A solution to this problem is to perform *multi-hop reasoning*, which involves reasoning over multiple edges/relations in a KG. When performing multi-hop reasoning, models are required to attend to different parts of a given context to answer a question. Figure 4 shows an example of a task known as *story ending* generation, where the goal is to generate the end of a story given the story context. In this example, external commonsense knowledge in the form of relational paths can guide the generation of the two nodes `lava` and `substance` by providing background knowledge such as `(volcano, madeOf, lava)`. Capturing what `lava` and `substance` that appear in the story ending refer to in the story context is non-trivial for PLMs, especially when the story is long. To address this drawback, Ji et al. (2020b) perform reasoning over multi-hop relational paths as a way to extract structural and semantic knowledge from a KG.

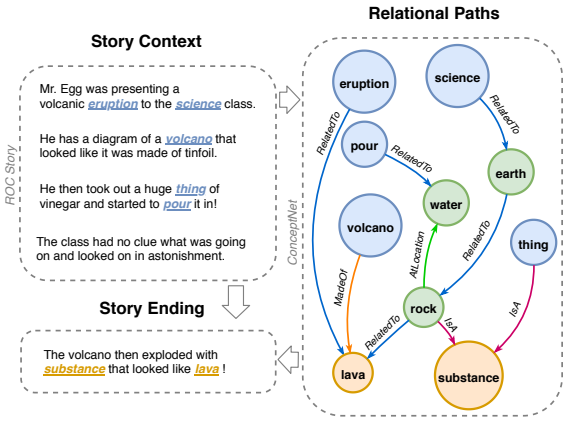

Figure 4: Using structural relational knowledge for multi-hop reasoning (Ji et al. 2020b). Blue nodes correspond to concepts in story, orange nodes correspond to those in story ending, and green nodes are intermediate concepts that connect their blue and orange counterparts.

Using iterative refinements. When provided with information about past and future events (as in αNLG), humans can easily reason about these events using contextual and prior knowledge. This kind of non-monotonic reasoning is crucial to improving generation coherency. However, non-monotonic reasoning is difficult for neural models since the generation process happens predominantly while conditioning on the left context (Welleck et al. 2019). To address this problem, Qin et al. (2020) propose a decoding approach that involves sampling from the combined output vector representations computed from both forward and backward propagation. In other words, iterative refinements are made on the generated text through alternating between forward and backward passes, yielding text with improved coherency.

Concluding Remarks

Despite recent progresses on using PLMs to address commonsense knowledge reasoning and generation tasks, many of these tasks are far from being solved. We conclude our discussion with key challenges in this area of research.

Dataset	Model	Human	Dataset	Model	Human
HellaSwag	93.85	95.6	WinoGrande	86.64	94.0
CosmosQA	91.79	94.0	SocialIQA	83.15	88.1
PIQA	90.13	94.9	VCR	63.15	85.0

Table 2: Results of state-of-the-art models and human baselines on widely-used commonsense reasoning benchmarks.

Improving benchmarks. While state-of-the-art models have achieved near-human performance on many of the benchmarks mentioned in this paper (see Table 2), the reasoning tasks underlying these benchmarks are still far from being solved. Consequently, it is not clear what the performance gains on a particular benchmark mean. Bender and Koller (2020) point out that acing a benchmark has led us to over-estimate the capability of PLMs, which in turn has given rise to misleading definitions of "understanding". It is therefore important to re-think what is being learned by PLMs and how benchmarks can be made more representative such that performance gains on them translate to meaningful progress towards the bigger goal of "understanding".

Reducing biases. Biases in benchmarks such as predictable question structures, annotation artifacts, and lexical overlap provide easy shortcuts for PLMs to arrive at correct answers without involving reasoning. To mitigate biases, researchers have used *adversarial filtering* wherein easily solvable options are replaced iteratively by new ones until the discriminator misclassifies it (Zellers et al. 2018; McCoy et al. 2019; Bras et al. 2020). To robustify data, several workarounds have been proposed that revolve around reducing lexical overlap, creating complex reasoning questions that require additional context, and employing adversarial approaches with newer models (Gardner et al. 2019). Bias reduction in benchmarks remains an active research area.

Exploring new components of commonsense knowledge. These are numerous components of commonsense knowledge that are partially understood and not covered by the present research. One primary reason for this is that we do not have a comprehensive understanding of how humans learn. A concrete example can be derived from Kahneman's (2011) cognitive system of intuition. There is no clear way of representing a human's mental and emotional states that can be readily used by our algorithms. Modeling multiple mental states with natural language is a highly non-trivial process (Sap et al. 2020).

Addressing the reporting bias. Much commonsense knowledge is assumed rather than mentioned explicitly in text (Grice 1975; Van Durme 2010; Gordon and Van Durme 2013). This results in what is known as the "reporting bias", which, when combined with the scarcity of training data for many NLP tasks, makes it hard for PLMs to receive appropriate signal about a particular concept (Zhao et al. 2020). This could lead to over-generalization of associations and amplification of biases (Shwartz and Choi 2020). How to address the reporting bias remains an open question.

Improving existing KGs. While many approaches rely on KGs to obtain rich contextual knowledge, existing KGs have at least two key weaknesses that can potentially limit their usefulness for commonsense reasoning. The first is *sparsity*: many concepts and relations are missing in existing KGs (Li et al. 2016). This sparsity problem in turn limits the amount of knowledge that can be extracted from KGs for commonsense reasoning (Zhao et al. 2020). The second is *non-contextualization*: finding the nodes that are most relevant to a query is difficult, particularly by propagation-based algorithms, because many of them are non-contextual in nature (Fadnis et al. 2019). To address the non-contextualization problem, Malaviya et al. (2020) have attempted to use the structural and semantic connections of the nodes in a KG to obtain contextual information. The resulting contextual information can then be explicitly encoded in a KG by creating additional nodes, which alleviates the sparsity problem. How to densify KGs and contextualize their nodes is an ongoing research topic (Wang et al. 2020).

Harnessing commonsense knowledge from different modalities. There are many instances wherein the visual modality is required along with text to make sense of a particular situation (Park et al. 2020). While humans learn commonsense knowledge predominantly through perception and interaction with the physical world, neural models are primarily trained on text data. Harnessing commonsense knowledge from different modalities can potentially take these models to the next level of performance.

Towards multilinguality. An important but underexplored direction is multi-lingual commonsense reasoning and generation. Studies have shown that the performance of cross-lingual PLMs is poor when evaluated on non-English commonsense reasoning benchmarks (Lin et al. 2021b). These models perform poorly when evaluated on a test set that was translated to English, leading to staggering transfer reasoning capabilities to other languages and restricting the research scope to only certain languages (Ponti et al. 2020).

References

Baldwin, D. A. 1995. Understanding the Link between Joint Attention and Language. *Joint Attention: Its Origins and Role in Development*, 131–158.

Bender, E. M.; and Koller, A. 2020. Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. In *ACL*.

Bhagavatula, C.; Bras, R. L.; Malaviya, C.; Sakaguchi, K.; Holtzman, A.; Rashkin, H.; Downey, D.; Yih, W.-T.; and Choi, Y. 2020. Abductive Commonsense Reasoning. In *ICLR*.

Bisk, Y.; Zellers, R.; Bras, R. L.; 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*.

Bras, R. L.; Swayamdipta, S.; Bhagavatula, C.; Zellers, R.; Peters, M. E.; Sabharwal, A.; and Choi, Y. 2020. Adversarial Filters of Dataset Biases. In *ICML*.

Chen, M.; D'Arcy, M.; Liu, A.; Fernandez, J.; and Downey, D. 2019. CODAH: An Adversarially-Authored Question Answering Dataset for Common Sense. In *3rd Workshop on Evaluating Vector Space Representations for NLP*.

Clark, K.; Luong, M.-T.; Le, Q. V.; and Manning, C. D. 2019. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. In *ICLR*.

Davison, J.; Feldman, J.; and Rush, A. 2019. Commonsense Knowledge Mining from Pretrained Models. In *EMNLP-IJCNLP*.

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

Ettinger, A. 2020. What BERT Is Not: Lessons from a New Suite of Psycholinguistic Diagnostics for Language Models. *TACL*, 8: 34–48.

Fadnis, K. P.; Talamadupula, K.; Kapanipathi, P.; Ishfaq, H.; Roukos, S.; and Fokoue, A. 2019. Path-Based Contextualization of Knowledge Graphs for Textual Entailment. *CoRR*, abs/1911.02085.

Forbes, M.; Holtzman, A.; and Choi, Y. 2019. Do Neural Language Representations Learn Physical Commonsense? In *CogSci*.

Gardner, M.; Berant, J.; Hajishirzi, H.; Talmor, A.; and Min, S. 2019. On Making Reading Comprehension More Comprehensive. In *2nd Workshop on Machine Reading for QA*.

Geva, M.; Gupta, A.; and Berant, J. 2020. Injecting Numerical Reasoning Skills into Language Models. In *ACL*.

Goel, P.; Feng, S.; and Boyd-Graber, J. 2019. How Pre-trained Word Representations Capture Commonsense Physical Comparisons. In *First Workshop on Commonsense Inference in NLP*.

Gordon, J.; and Van Durme, B. 2013. Reporting Bias and Knowledge Acquisition. In *AKBC*.

Grice, H. P. 1975. Logic and Conversation. In Ezcurdia, M.; and Stainton, R. J., eds., *The Semantics-Pragmatics Boundary in Philosophy*, 47.

Guu, K.; Hashimoto, T. B.; Oren, Y.; and Liang, P. 2018. Generating Sentences by Editing Prototypes. *TACL*, 6: 437–450.

Han, X.; Zhang, Z.; Ding, N.; Gu, Y.; Liu, X.; Huo, Y.; Qiu, J.; Zhang, L.; Han, W.; Huang, M.; Jin, Q.; Lan, Y.; Liu, Y.; Liu, Z.; Lu, Z.; Qiu, X.; Song, R.; Tang, J.; Wen, J.; Yuan, J.; Zhao, W. X.; and Zhu, J. 2021. Pre-Trained Models: Past, Present and Future. *CoRR*, abs/2106.07139.

Huang, L.; Le Bras, R.; Bhagavatula, C.; and Choi, Y. 2019. Cosmos QA: Machine Reading Comprehension with Contextual Commonsense Reasoning. In *EMNLP-IJCNLP*.

Hwang, J. D.; Bhagavatula, C.; Bras, R. L.; Da, J.; Sakaguchi, K.; Bosselut, A.; and Choi, Y. 2021. COMET-ATOMIC 2020: On Symbolic and Neural Commonsense Knowledge Graphs. In *AAAI*.

Ji, H.; Ke, P.; Huang, S.; Wei, F.; and Huang, M. 2020a. Generating Commonsense Explanation by Extracting Bridge Concepts from Reasoning Paths. In *AACL-IJCNLP*.

Ji, H.; Ke, P.; Huang, S.; Wei, F.; Zhu, X.; and Huang, M. 2020b. Language Generation with Multi-Hop Reasoning on Commonsense Knowledge Graph. In *EMNLP*.

Kahneman, D. 2011. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, LLC.

Kalyan, K. S.; Rajasekharan, A.; and Sangeetha, S. 2021. AMMUS : A Survey of Transformer-based Pretrained Models in Natural Language Processing. *CoRR*, abs/2108.05542.

Kassner, N.; and Schütze, H. 2020. Negated and Misprimed Probes for Pretrained Language Models: Birds Can Talk, But Cannot Fly. In *ACL*.

Klein, T.; and Nabi, M. 2019. Attention Is (not) All You Need for Commonsense Reasoning. In *ACL*.

Lenat, D. B. 1995. CYC: A Large-Scale Investment in Knowledge Infrastructure. *CACM*, 38(11): 33–38.

Levesque, H. J.; Davis, E.; and Morgenstern, L. 2012. The Winograd Schema Challenge. In *KR&R*.

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*.

Li, Y.; Goel, P.; Rajendra, V. K.; Singh, H. S.; Francis, J.; Ma, K.; Nyberg, E.; and Oltramari, A. 2021. Lexically-constrained Text Generation through Commonsense Knowledge Extraction and Injection. In *Workshop on CSKG*.

Lin, B. Y.; Lee, S.; Khanna, R.; and Ren, X. 2020a. Birds have four legs?! NumerSense: Probing Numerical Commonsense Knowledge of Pre-Trained Language Models. In *EMNLP*.

Lin, B. Y.; Wu, Z.; Yang, Y.; Lee, D.; and Ren, X. 2021a. RiddleSense: Reasoning about Riddle Questions Featuring Linguistic Creativity and Commonsense Knowledge. In *Findings of the ACL: ACL-IJCNLP 2021*.

Lin, B. Y.; Zhou, W.; Shen, M.; Zhou, P.; Bhagavatula, C.; Choi, Y.; and Ren, X. 2020b. CommonGen: A Constrained Text Generation Challenge for Generative Commonsense Reasoning. In *Findings of the ACL: EMNLP 2020*.

Lin, C.; Ouyang, Z.; Zhuang, J.; Chen, J.; Li, H.; and Wu, R. 2021b. Improving Code Summarization with Block-wise Abstract Syntax Tree Splitting. In *ICPC*.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR*, abs/1907.11692.

Logan, R.; Liu, N. F.; Peters, M. E.; Gardner, M.; and Singh, S. 2019. Barack's Wife Hillary: Using Knowledge Graphs for Fact-Aware Language Modeling. In *ACL*.

Malaviya, C.; Bhagavatula, C.; Bosselut, A.; and Choi, Y. 2020. Commonsense Knowledge Base Completion with Structural and Semantic Context. In *AAAI*.

Marasović, A.; Bhagavatula, C.; Park, J. s.; Le Bras, R.; Smith, N. A.; and Choi, Y. 2020. Natural Language Rationales with Full-Stack Visual Reasoning: From Pixels to

Semantic Frames to Commonsense Graphs. In *Findings of the ACL: EMNLP 2020*.

McCoy, T.; Pavlick, E.; and Linzen, T. 2019. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. In *ACL*.

Niu, Y.; Huang, F.; Liang, J.; Chen, W.; Zhu, X.; and Huang, M. 2021. A Semantic-based Method for Unsupervised Commonsense Question Answering. In *ACL-IJCNLP*.

Park, J. S.; Bhagavatula, C.; Mottaghi, R.; Farhadi, A.; and Choi, Y. 2020. VisualCOMET: Reasoning about the Dynamic Context of a Still Image. In *ECCV*.

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

Ponti, E. M.; Glavaš, G.; Majewska, O.; Liu, Q.; Vulić, I.; and Korhonen, A. 2020. XCOPA: A Multilingual Dataset for Causal Commonsense Reasoning. In *EMNLP*.

Qin, L.; Gupta, A.; Upadhyay, S.; He, L.; Choi, Y.; and Faruqui, M. 2021. TIMEDIAL: Temporal Commonsense Reasoning in Dialog. In *ACL-IJCNLP*.

Qin, L.; Shwartz, V.; West, P.; Bhagavatula, C.; Hwang, J. D.; Le Bras, R.; Bosselut, A.; and Choi, Y. 2020. Back to the Future: Unsupervised Backprop-based Decoding for Counterfactual and Abductive Commonsense Reasoning. In *EMNLP*.

Qiu, X.; Sun, T.; Xu, Y.; Shao, Y.; Dai, N.; and Huang, X. 2020. Pre-trained Models for Natural Language Processing: A Survey. *Science China Technological Sciences*, 63.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *JMLR*, 21(140): 1–67.

Rajani, N. F.; McCann, B.; Xiong, C.; and Socher, R. 2019. Explain Yourself! Leveraging Language Models for Commonsense Reasoning. In *ACL*.

Richardson, K.; and Sabharwal, A. 2020. What Does My QA Model Know? Devising Controlled Probes Using Expert Knowledge. *TACL*, 8: 572–588.

Rogers, A.; Kovaleva, O.; and Rumshisky, A. 2020. A Primer in BERTology: What We Know About How BERT Works. *TACL*, 8: 842–866.

Sakaguchi, K.; Bras, R. L.; Bhagavatula, C.; and Choi, Y. 2020. WINOGRANDE: An Adversarial Winograd Schema Challenge at Scale. In *AAAI*.

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

Sap, M.; Rashkin, H.; Chen, D.; Bras, R. L.; and Choi, Y. 2019b. Social IQA: Commonsense Reasoning about Social Interactions. In *EMNLP*.

Sap, M.; Shwartz, V.; Bosselut, A.; Choi, Y.; and Roth, D. 2020. Commonsense Reasoning for Natural Language Processing. In *ACL*.

Shwartz, V.; and Choi, Y. 2020. Do Neural Language Models Overcome Reporting Bias? In *COLING*.

Singh, P.; Lin, T.; Mueller, E. T.; Lim, G.; PerkinsWan, T.; and Zhu, L. 2002. Open Mind Common Sense: Knowledge Acquisition from the General Public. In *On the Move to Meaningful Internet Systems (OTM 2002)*.

Speer, R.; Chin, J.; and Havasi, C. 2017. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. In *AAAI*.

Svyatkovskiy, A.; Zhao, Y.; Fu, S.; and Sundaresan, N. 2019. Pythia: AI-assisted Code Completion System. In *KDD*.

Van Durme, B. 2010. *Extracting Implicit Knowledge from Text*. Ph.D. thesis, University of Rochester, Rochester, NY.

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 *NIPS*.

Wallace, E.; Wang, Y.; Li, S.; Singh, S.; and Gardner, M. 2019. Do NLP Models Know Numbers? Probing Numeracy in Embeddings. In *EMNLP-IJCNLP*.

Wang, P.; Peng, N.; Ilievski, F.; Szekely, P.; and Ren, X. 2020. Connecting the Dots: A Knowledgeable Path Generator for Commonsense Question Answering. In *Findings of the ACL: EMNLP 2020*.

Weir, N.; Poliak, A.; and Durme, B. V. 2020. Probing Neural Language Models for Human Tacit Assumptions. In *CogSci*.

Welleck, S.; Brantley, K.; Daumé, H.; and Cho, K. 2019. Non-Monotonic Sequential Text Generation. In *ICML*.

Zang, L.-J.; Cao, C.; Cao, Y.-N.; Wu, Y.-M.; and Cao, C.-G. 2013. A Survey of Commonsense Knowledge Acquisition. *Journal of Computer Science and Technology*, 28: 689–719.

Zellers, R.; Bisk, Y.; Farhadi, A.; and Choi, Y. 2019a. From Recognition to Cognition: Visual Commonsense Reasoning. In *CVPR*.

Zellers, R.; Bisk, Y.; Schwartz, R.; and Choi, Y. 2018. SWAG: A Large-Scale Adversarial Dataset for Grounded Commonsense Inference. In *EMNLP*.

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

Zellers, R.; Holtzman, A.; Peters, M.; Mottaghi, R.; Kembhavi, A.; Farhadi, A.; and Choi, Y. 2021. PIGLeT: Language Grounding Through Neuro-Symbolic Interaction in a 3D World. In *ACL*.

Zhao, Z.; Papalexakis, E.; and Ma, X. 2020. Learning Physical Common Sense as Knowledge Graph Completion via BERT Data Augmentation and Constrained Tucker Factorization. In *EMNLP*.

Zhou, B.; Khashabi, D.; Ning, Q.; and Roth, D. 2019. "Going on a vacation" takes longer than "Going for a walk": A Study of Temporal Commonsense Understanding. In *EMNLP*.

Zhou, B.; Ning, Q.; Khashabi, D.; and Roth, D. 2020a. Temporal Common Sense Acquisition with Minimal Supervision. In *ACL*.

Zhou, X.; Zhang, Y.; Cui, L.; and Huang, D. 2020b. Evaluating Commonsense in Pre-Trained Language Models. In *AAAI*.