SOCIAL IQA: Commonsense Reasoning about Social Interactions

Maarten Sap[⋆] [◇♡] **Hannah Rashkin**[⋆] [◇♡] **Derek Chen**[♡] **Ronan Le Bras**[◇] **Yejin Choi**[◇♡]
[◇]Allen Institute for Artificial Intelligence, Seattle, WA, USA
[♡]Paul G. Allen School of Computer Science & Engineering, Seattle, WA, USA
{msap,hrashkin,dchen14,yejin}@cs.washington.edu
{ronanlb}@allenai.org

Abstract

We introduce SOCIAL IQA, the first large-scale benchmark for commonsense reasoning about social situations. SOCIAL IQA contains 38,000 multiple choice questions for probing *emotional* and *social* intelligence in a variety of everyday situations (e.g., Q: "*Jordan wanted to tell Tracy a secret, so Jordan leaned towards Tracy. Why did Jordan do this?*" A: "*Make sure no one else could hear*"). Through crowdsourcing, we collect commonsense questions along with correct and incorrect answers about social interactions, using a new framework that mitigates stylistic artifacts in incorrect answers by asking workers to provide the right answer to a different but related question. Empirical results show that our benchmark is challenging for existing question-answering models based on pretrained language models, compared to human performance (>20% gap). Notably, we further establish SOCIAL IQA as a resource for transfer learning of commonsense knowledge, achieving state-of-the-art performance on multiple commonsense reasoning tasks (Winograd Schemas, COPA).

1 Introduction

Social and emotional intelligence enables humans to reason about the mental states of others and their likely actions (Ganaie and Mudasir, 2015). For example, when someone spills food all over the floor, we can infer that they will likely want to clean up the mess, rather than taste the food off the floor or run around in the mess (Figure 1, middle). This example illustrates how Theory of Mind, i.e., the ability to reason about the implied emotions and behavior of others, enables humans to navigate social situations ranging from simple conversations with friends to complex negotiations in courtrooms (Apperly, 2010).

Figure 1: Three context-question-answers triples from SOCIAL IQA, along with the type of reasoning required to answer them. In the top example, humans can trivially infer that Tracy pressed upon Austin because there was no room in the elevator. Similarly, in the bottom example, commonsense tells us that people typically root for the hero, not the villain.

While humans trivially acquire and develop such social reasoning skills (Moore, 2013), this is still a challenge for machine learning models, in part due to the lack of large-scale resources to train and evaluate modern AI systems' social and emotional intelligence. Although recent advances in pretraining large language models have yielded promising improvements on several commonsense inference tasks, these models still struggle to reason about social situations, as shown in this and previous work (Davis and Marcus,

[⋆] Both authors contributed equally.

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 4463–4473,
Hong Kong, China, November 3–7, 2019.

2015; Nematzadeh et al., 2018; Talmor et al., 2019). This is partly due to language models being trained on written text corpora, where reporting bias of knowledge limits the scope of commonsense knowledge that can be learned (Gordon and Van Durme, 2013; Lucy and Gauthier, 2017).

In this work, we introduce Social Intelligence QA (SOCIAL IQA), the first large-scale resource to learn and measure *social* and *emotional* intelligence in computational models.[1] SOCIAL IQA contains 38k multiple choice questions regarding the pragmatic implications of everyday, social events (see Figure 1). To collect this data, we design a crowdsourcing framework to gather contexts and questions that explicitly address social commonsense reasoning. Additionally, by combining handwritten negative answers with adversarial question-switched answers (Section 3.3), we minimize annotation artifacts that can arise from crowdsourcing incorrect answers (Schwartz et al., 2017; Gururangan et al., 2018).

This dataset remains challenging for AI systems, with our best performing baseline reaching 64.5% (BERT-large), significantly lower than human performance. We further establish SOCIAL IQA as a resource that enables transfer learning for other commonsense challenges, through *sequential finetuning* of a pretrained language model on SOCIAL IQA before other tasks. Specifically, we use SOCIAL IQA to set a new state-of-the-art on three commonsense challenge datasets: COPA (Roemmele et al., 2011) (83.4%), the original Winograd (Levesque, 2011) (72.5%), and the extended Winograd dataset from Rahman and Ng (2012) (84.0%).

Our contributions are as follows: (1) We create SOCIAL IQA, the first large-scale QA dataset aimed at testing social and emotional intelligence, containing over 38k QA pairs. (2) We introduce question-switching, a technique to collect incorrect answers that minimizes stylistic artifacts due to annotator cognitive biases. (3) We establish baseline performance on our dataset, with BERT-large performing at 64.5%, well below human performance. (4) We achieve new state-of-the-art accuracies on COPA and Winograd through sequential finetuning on SOCIAL IQA, which implicitly endows models with social commonsense knowledge.

[1] Available at https://tinyurl.com/socialiqa

SOCIAL IQA		
# QA tuples	train	33,410
	dev	1,954
	test	2,224
	total	**37,588**
Train statistics		
Average # tokens	context	14.04
	question	6.12
	answers (all)	3.60
	answers (correct)	3.65
	answers (incorrect)	3.58
Unique # tokens	context	15,764
	question	1,165
	answers (all)	12,285
	answers (correct)	7,386
	answers (incorrect)	10,514
Average freq. of answers	answers (correct)	1.37
	answers (incorrect)	1.47

Table 1: Data statistics for SOCIAL IQA.

2 Task description

SOCIAL IQA aims to measure the social and emotional intelligence of computational models through multiple choice question answering (QA). In our setup, models are confronted with a question explicitly pertaining to an observed *context*, where the correct answer can be found among three competing options.

By design, the questions require *inferential* reasoning about the social causes and effects of situations, in line with the type of intelligence required for an AI assistant to interact with human users (e.g., know to call for help when an elderly person falls; Pollack, 2005). As seen in Figure 1, correctly answering questions requires reasoning about motivations, emotional reactions, or likely preceding and following actions. Performing these inferences is what makes us experts at navigating social situations, and is closely related to Theory of Mind, i.e., the ability to reason about the beliefs, motivations, and needs of others (Baron-Cohen et al., 1985).[2] Endowing machines with this type of intelligence has been a longstanding but elusive goal of AI (Gunning, 2018).

[2] Theory of Mind is well developed in most neurotypical adults (Ganaie and Mudasir, 2015), but can be influenced by age, culture, or developmental disorders (Korkmaz, 2011).

ATOMIC

As a starting point for our task creation, we draw upon social commonsense knowledge from ATOMIC (Sap et al., 2019) to seed our contexts and question types. ATOMIC is a large knowledge graph that contains inferential knowledge about the causes and effects of 24k short events. Each triple in ATOMIC consists of an event phrase with person-centric variables, one of nine inference dimensions, and an inference object (e.g., "PersonX pays for PersonY's ___", "xAttrib", "generous"). The nine inference dimensions in ATOMIC cover causes of an event (e.g., "X needs money"), its effects on the agent (e.g., "X will get thanked") and its effect on other participants (e.g., "Y will want to see X again"); see Sap et al. (2019) for details.

Given this base, we generate natural language contexts that represent specific instantiations of the event phrases found in the knowledge graph. Furthermore, the questions created probe the commonsense reasoning required to navigate such contexts. Critically, since these contexts are based off of ATOMIC, they explore a diverse range of motivations and reactions, as well as likely preceding or following actions.

3 Dataset creation

SOCIAL IQA contains 37,588 multiple choice questions with three answer choices per question. Questions and answers are gathered through three phases of crowdsourcing aimed to collect the *context*, the *question*, and a set of *positive* and *negative* answers. We run crowdsourcing tasks on Amazon Mechanical Turk (MTurk) to create each of the three components, as described below.

3.1 Event Rewriting

In order to cover a variety of social situations, we use the base events from ATOMIC as prompts for context creation. As a pre-processing step, we run an MTurk task that asks workers to turn an ATOMIC event (e.g., "PersonX spills ___ all over the floor") into a sentence by adding names, fixing potential grammar errors, and filling in placeholders (e.g., "Alex spilled food all over the floor.").[3]

3.2 Context, Question, & Answer Creation

Next, we run a task where annotators create full context-question-answers triples. We automatically generate question templates covering the nine commonsense inference dimensions in ATOMIC.[4] Crowdsourcers are prompted with an event sentence and an inference question to turn into a more detailed context[5] (e.g. "Alex spilled food all over the floor and it made a huge mess.") and an edited version of the question if needed for improved specificity (e.g. "What will Alex want to do next?"). Workers are also asked to contribute two potential correct answers.

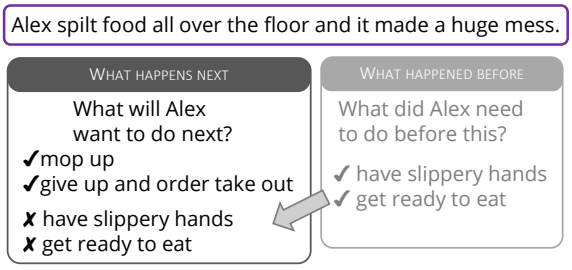

Figure 2: Question-Switching Answers (QSA) are collected as the correct answers to the wrong question that targets a different type of inference (here, reasoning about what happens before instead of after an event).

3.3 Negative Answers

In addition to correct answers, we collect four incorrect answer options, of which we filter out two. To create incorrect options that are adversarial for models but easy for humans, we use two different approaches to the collection process. These two methods are specifically designed to avoid different types of annotation artifacts, thus making it more difficult for models to rely on data biases. We integrate and filter answer options and validate final QA tuples with human rating tasks.

Handwritten Incorrect Answers (HIA) The first method involves eliciting handwritten incorrect answers that require reasoning about the context. These answers are handwritten to be similar to the correct answers in terms of topic, length, and style but are subtly incorrect. Two of these answers are collected during the same MTurk task as the original context, questions, and correct answers. We will refer to these negative responses as handwritten incorrect answers (HIA).

Question-Switching Answers (QSA) We collect a second set of negative (incorrect) answer

[3] This task paid $0.35 per event.

[4] We do not generate templates if the ATOMIC dimension is annotated as "none."

[5] Workers were asked to contribute a context 7-25 words longer than the event sentence.

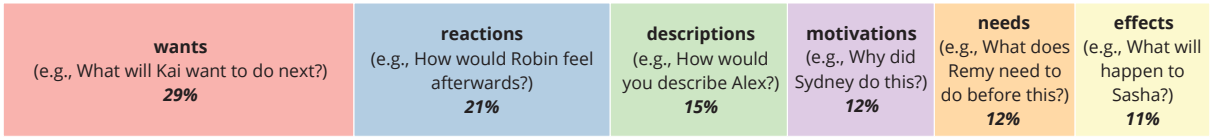

Figure 3: SOCIAL IQA contains several question types which cover different types of inferential reasoning. Question types are derived from ATOMIC inference dimensions.

candidates by switching the questions asked about the context, as shown in Figure 2. We do this to avoid cognitive biases and annotation artifacts in the answer candidates, such as those caused by writing incorrect answers or negations (Schwartz et al., 2017; Gururangan et al., 2018). In this crowdsourcing task, we provide the same context as the original question, as well as a question automatically generated from a different but similar ATOMIC dimension,[6] and ask workers to write two correct answers. We refer to these negative responses as question-switching answers (QSA).

By including answers to a different question about the same context, we ensure that these adversarial responses have the stylistic qualities of correct answers and strongly relate to the context topic, while still being incorrect, making it difficult for models to simply perform pattern-matching. To verify this, we compare valence, arousal, and dominance (VAD) levels across answer types, computed using the VAD lexicon by Mohammad (2018). Figure 4 shows effect sizes (Cohen's d) of the differences in VAD means, where the magnitude of effect size indicates how different the answer types are stylistically. Indeed, QSA and correct answers differ substantially less than HIA answers ($|d| \leq .1$).[7]

3.4 QA Tuple Creation

As the final step of the pipeline, we aggregate the data into three-way multiple choice questions. For each created context-question pair contributed by crowdsourced workers, we select a random correct answer and the incorrect answers that are least entailed by the correct one, following inspiration from Zellers et al. (2019a).

For the training data, we validate our QA tuples through a multiple-choice crowdsourcing task where three workers are asked to select the right answer to the question provided.[8] In order to ensure even higher quality, we validate the dev and test data a second time with five workers. Our final dataset contains questions for which the correct answer was determined by human majority voting, discarding cases without a majority vote. We also apply a lightweight form of adversarial filtering to make the task more challenging by using a deep stylistic classifier to remove easier examples on the dev and test sets (Sakaguchi et al., 2019).[9]

To obtain human performance, we run a separate task asking three new workers to select the correct answer on a random subset of 900 dev and 900 test examples. Human performance on these subsets is 87% and 84%, respectively.

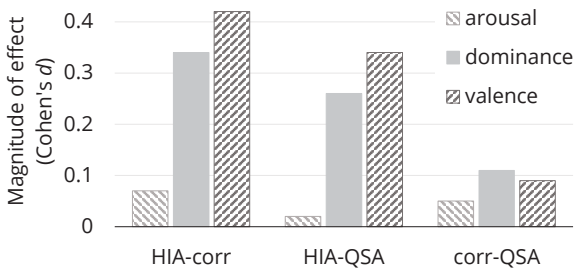

Figure 4: Magnitude of effect sizes (Cohen's d) when comparing average dominance, arousal and valence values of different answer types where larger $|d|$ indicates more stylistic difference. For valence (sentiment polarity) and dominance, the effect sizes comparing QSA and correct answers are much smaller, indicating that these are more similar tonally. Notably, all three answer types have comparable levels of arousal (intensity).

3.5 Data Statistics

To keep contexts separate across train/dev/test sets, we assign SOCIAL IQA contexts to the same partition as the ATOMIC event the context was based on. Shown in Table 1 (top), this yields a

[6] Using the following three groupings of ATOMIC dimensions: {xWant, oWant, xNeed, xIntent}, {xReact oReact, xAttr}, and {xEffect, oEffect}.

[7] Cohen's $|d| < .20$ is considered small (Sawilowsky, 2009). We find similarly small effect sizes using other sentiment/emotion lexicons.

[8] Agreement on this task was high (Cohen's κ=.70)

[9] We also tried filtering to remove examples from the training set but found it did not significantly change performance. We will release tags for the easier training examples with the full data.

total set of around 33k training, 2k dev, and 2k test tuples. We additionally include statistics on word counts and vocabulary of the training data. We report the averages of correct and incorrect answers in terms of: token length, number of unique tokens, and number of times a unique answer appears in the dataset. Note that due to our three-way multiple choice setup, there are twice as many incorrect answers which influences these statistics.

We also include a breakdown (Figure 3) across question types, which we derive from ATOMIC inference dimensions.[10] In general, questions relating to what someone will feel afterwards or what they will likely do next are more common in SOCIAL IQA. Conversely, questions pertaining to (potentially involuntary) effects of situations on people are less frequent.

Model	Accuracy (%) Dev	Test
Random baseline	33.3	33.3
GPT	63.3	63.0
BERT-base	63.3	63.1
BERT-large	**66.0**	**64.5**
w/o context	52.7	–
w/o question	52.1	–
w/o context, question	45.5	–
Human	86.9*	84.4*

Table 2: Experimental results. We additionally perform an ablation by removing contexts and questions, verifying that both are necessary for BERT-large's performance. Human evaluation results are obtained using 900 randomly sampled examples.

4 Methods

We establish baseline performance on SOCIAL IQA, using large pretrained language models based on the Transformer architecture (Vaswani et al., 2017). Namely, we finetune OpenAI-GPT (Radford et al., 2018) and BERT (Devlin et al., 2019), which have both shown remarkable improvements on a variety of tasks. OpenAI-GPT is a uni-directional language model trained on the BookCorpus (Zhu et al., 2015), whereas BERT is a bidirectional language model trained on both the BookCorpus and English Wikipedia. As per previous work, we finetune the language model representations but fully learn the classifier specific parameters described below.

Multiple choice classification To classify sequences using these language models, we follow the multiple-choice setup implementation by the respective authors, as described below. First, we concatenate the context, question, and answer, using the model specific separator tokens. For OpenAI-GPT, the format becomes `_start_ <context> <question> _delimiter_ <answer> _classify_`, where `_start_`, `_delimiter_`, and `_classify_` are special function tokens. For BERT, the format is similar, but the classifier token comes before the context.[11]

For each triple, we then compute a score l by passing the hidden representation from the classifier token $h_{CLS} \in \mathbb{R}^H$ through an MLP:

$$l = W_2 \tanh(W_1 h_{CLS} + b_1)$$

where $W_1 \in \mathbb{R}^{H \times H}$, $b_1 \in \mathbb{R}^H$ and $W_2 \in \mathbb{R}^{1 \times H}$. Finally, we normalize scores across all triples for a given context-question pair using a softmax layer. The model's predicted answer corresponds to the triple with the highest probability.

5 Experiments

5.1 Experimental Set-up

We train our models on the 33k SOCIAL IQA training instances, selecting hyperparameters based on the best performing model on our dev set, for which we then report test results. Specifically, we perform finetuning through a grid search over the hyper-parameter settings (with a learning rate in $\{1e{-}5, 2e{-}5, 3e{-}5\}$, a batch size in $\{3, 4, 8\}$, and a number of epochs in $\{3, 4, 10\}$) and report the maximum performance.

Models used in our experiments vary in sizes: OpenAI-GPT (117M parameters) has a hidden size H=768, BERT-base (110M params) and BERT-large (340M params) hidden sizes of H=768 and H=1024, respectively. We train using the HuggingFace PyTorch (Paszke et al., 2017) implementation.[12]

[10] We group agent and theme ATOMIC dimensions together (e.g., "xReact" and "oReact" become the "reactions" question type).

[11] BERT's format is `[CLS] <context> [UNUSED] <question> [SEP] <answer> [SEP]`

[12] https://github.com/huggingface/pytorch-pretrained-BERT

	Context	Question		Answer
(1)	Jesse was pet sitting for Addison, so Jesse came to Addison's house and walked their dog.	What does Jesse need to do before this?	✓ 🤖	(a) feed the dog (b) get a key from Addison (c) walk the dog
(2)	Kai handed back the computer to Will after using it to buy a product off Amazon.	What will Kai want to do next?	✓ 🤖	(a) wanted to save money on shipping (b) Wait for the package (c) Wait for the computer
(3)	Remy gave Skylar, the concierge, her account so that she could check into the hotel.	What will Remy want to do next?	🤖 ✓	(a) lose her credit card (b) arrive at a hotel (c) get the key from Skylar
(4)	Sydney woke up and was ready to start the day. They put on their clothes.	What will Sydney want to do next?	🤖 ✓	(a) go to bed (b) go to the pool (c) go to work
(5)	Kai grabbed Carson's tools for him because Carson could not get them.	How would Carson feel as a result?	🤖 ✓	(a) inconvenienced (b) grateful (c) angry
(6)	Although Aubrey was older and stronger, they lost to Alex in arm wrestling.	How would Alex feel as a result?	🤖 ✓	(a) they need to practice more (b) ashamed (c) boastful

Table 3: Example CQA triples from the SOCIAL IQA dev set with BERT-large's predictions (🤖: BERT's prediction, ✓: true correct answer). The model predicts correctly in (1) and (2) and incorrectly in the other four examples shown here. Examples (3) and (4) illustrate the model choosing answers that might have happened before, or that might happen much later after the context, as opposed to right after the context situation. In Examples (5) and (6), the model chooses answers that may apply to people other than the ones being asked about.

5.2 Results

Our results (Table 2) show that SOCIAL IQA is still a challenging benchmark for existing computational models, compared to human performance. Our best performing model, BERT-large, outperforms other models by several points on the dev and test set. We additionally ablate our best model's representation by removing the context and question from the input, confirming that reasoning over both is necessary for this task.

Learning Curve To better understand the effect of dataset scale on model performance on our task, we simulate training situations with limited knowledge. We present the learning curve of BERT-large's performance on the dev set as it is trained on more training set examples (Figure 5). Although the model does significantly improve over a random baseline of 33% with only a few hundred examples, the performance only starts to converge after around 20k examples, providing evidence that large-scale benchmarks are required for this type of reasoning.

Error Analysis We include a breakdown of our best model's performance on various question types in Figure 6 and specific examples of errors in the last four rows of Table 3. Overall, questions related to pre-conditions of the context (people's motivations, actions needed before the context) are less challenging for the model. Conversely, the model seems to struggle more with questions relating to (potentially involuntary) effects, stative descriptions, and what people will want to do next.

Examples of errors in Table 3 further indicate that, instead of doing advanced reasoning about situations, models may only be learning lexical associations between the context, question, and answers, as hinted at by Marcus (2018) and Zellers et al. (2019b). This leads the model to select answers with incorrect timing (examples 3 and 4) or answers pertaining to the wrong participants (examples 5 and 6), despite being trained on large amounts of examples that specifically distinguish proper timing and participants. For instance, in (3) and (4), the model selects answers which

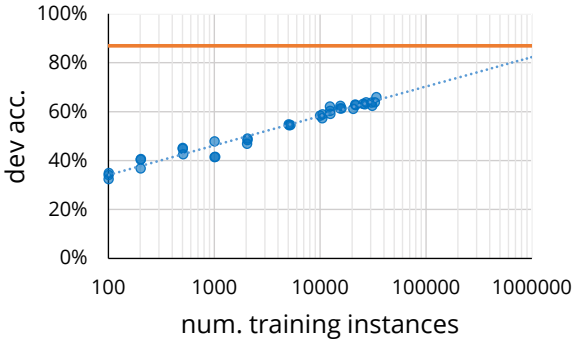

Figure 5: Dev accuracy when training BERT-large with various number of examples (multiple runs per training size), with human performance (86.9%) shown in orange. In order to reach >80%, the model would require nearly 1 million training examples.

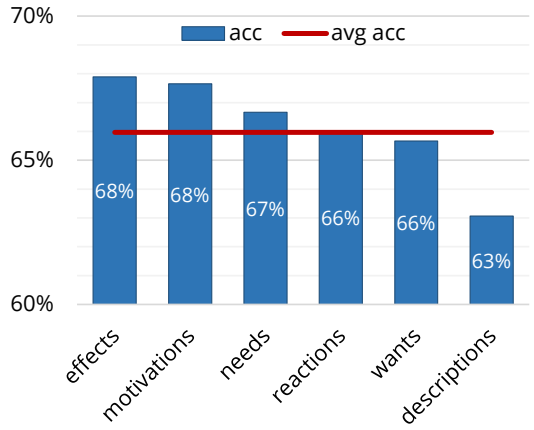

Figure 6: Average dev accuracy of BERT-large on different question types. While questions about effects and motivations are easier, the model still finds wants and descriptions more challenging.

are incorrectly timed with respect to the context and question (e.g., "arrive at a hotel" is something Remy likely did before checking in with the concierge, not afterwards). Additionally, the model often chooses answers related to a person other than the one asked about. In (6), after the arm wrestling, though it is likely that Aubrey will feel ashamed, the question relates to what Alex might feel–not Aubrey.

Overall, our results illustrate how reasoning about social situations still remains a challenge for these models, compared to humans who can trivially reason about the causes and effects for multiple participants. We expect that this task would benefit from models capable of more complex reasoning about entity state, or models that are more explicitly endowed with commonsense (e.g., from knowledge graphs like ATOMIC).

6 SOCIAL IQA for Transfer Learning

In addition to being the first large-scale benchmark for social commonsense, we also show that SOCIAL IQA can improve performance on downstream tasks that require commonsense, namely the Winograd Schema Challenge and the Choice of Plausible Alternatives task. We achieve state of the art performance on both tasks by sequentially finetuning on SOCIAL IQA before the task itself.

COPA The Choice of Plausible Alternatives task (COPA; Roemmele et al., 2011) is a two-way multiple choice task which aims to measure commonsense reasoning abilities of models. The dataset contains 1,000 questions (500 dev, 500 test) that ask about the causes and effects of a premise. This has been a challenging task for computational systems, partially due to the limited amount of training data available. As done previously (Goodwin et al., 2012; Luo et al., 2016), we finetune our models on the dev set, and report performance only on the test set.

Winograd Schema The Winograd Schema Challenge (WSC; Levesque, 2011) is a well-known commonsense knowledge challenge framed as a coreference resolution task. It contains a collection of 273 short sentences in which a pronoun must be resolved to one of two antecedents (e.g., in "The city councilmen refused the demonstrators a permit because *they* feared violence", *they* refers to the councilmen). Because of data scarcity in WSC, Rahman and Ng (2012) created 943 Winograd-style sentence pairs (1886 sentences in total), henceforth referred to as DPR, which has been shown to be slightly less challenging than WSC for computational models.

We evaluate on these two benchmarks. While the DPR dataset is split into train and test sets (Rahman and Ng, 2012), the WSC dataset contains a single (*test*) set of only 273 instances for evaluation purposes only. Therefore, we use the DPR dataset as training set when evaluating on the WSC dataset.

6.1 Sequential Finetuning

We first finetune BERT-large on SOCIAL IQA, which reaches 66% on our dev set (Table 2). We then finetune that model further on the task-specific datasets, considering the same set of hyperparameters as in §5.1. On each of the test sets,

Task	Model	Acc. (%) best	mean	std
COPA	Sasaki et al. (2017)	71.2	–	–
	BERT-large	80.8	75.0	3.0
	BERT-SOCIAL IQA	**83.4**	80.1	2.0
WSC	Kocijan et al. (2019)	72.5	–	–
	BERT-large	67.0	65.5	1.0
	BERT-SOCIAL IQA	**72.5**	69.6	1.7
DPR	Peng et al. (2015)	76.4	–	–
	BERT-large	79.4	71.2	3.8
	BERT-SOCIAL IQA	**84.0**	81.7	1.2

Table 4: Sequential finetuning of BERT-large on SOCIAL IQA before the task yields state of the art results (bolded) on COPA (Roemmele et al., 2011), Winograd Schema Challenge (Levesque, 2011) and DPR (Rahman and Ng, 2012). For comparison, we include previous published state of the art performance.

we report best, mean, and standard deviation of all models, and compare sequential finetuning results to a BERT-large baseline.

Results Shown in Table 4, sequential finetuning on SOCIAL IQA yields substantial improvements over the BERT-only baseline (between 2.6 and 5.5% max performance increases), as well as the general increase in performance stability (i.e., lower standard deviations). As hinted at by Phang et al. (2019), this suggests that BERT-large can benefit from both the large scale and the QA format of commonsense knowledge in SOCIAL IQA, which it struggles to learn from small benchmarks only. Notably, we find that sequentially finetuned BERT-SOCIAL IQA achieves state-of-the-art results on all three tasks, showing improvements of previous best performing models.[13]

Effect of scale and knowledge type To better understand these improvements in downstream task performance, we investigate the impact on COPA performance of sequential finetuning on less SOCIAL IQA training data (Figure 7), as well as the impact of the type of commonsense knowledge used in sequential finetuning. As expected, the downstream performance on COPA improves when using a model pretrained on more of SOCIAL IQA, indicating that the scale of the dataset

[13]Note that OpenAI-GPT was reported to achieve 78.6% on COPA, but that result was not published, nor discussed in the OpenAI-GPT white paper (Radford et al., 2018).

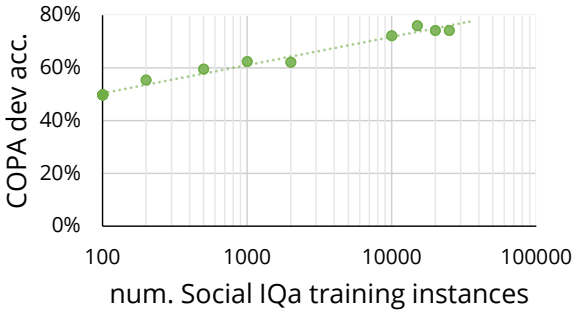

Figure 7: Effect of finetuning BERT-large on varying sizes of the SOCIAL IQA training set on the dev accuracy of COPA. As expected, the more SOCIAL IQA instances the model is finetuned on, the better the accuracy on COPA.

is one factor that helps in the fine-tuning. However, when using SWAG (a similarly sized dataset) instead of SOCIAL IQA for sequential finetuning, the downstream performance on COPA is lower (76.2%). This indicates that, in addition to its large scale, the social and emotional nature of the knowledge in SOCIAL IQA enables improvements on these downstream tasks.

7 Related Work

Commonsense Benchmarks: Commonsense benchmark creation has been well-studied by previous work. Notably, the WinoGrad Schema Challenge (WSC; Levesque, 2011) and the Choice Of Plausible Alternatives dataset (COPA; Roemmele et al., 2011) are expert-curated collections of commonsense QA pairs that are trivial for humans to solve. Whereas WSC requires physical and social commonsense knowledge to solve, COPA targets the knowledge of causes and effects surrounding social situations. While both benchmarks are of high-quality and created by experts, their small scale (150 and 1,000 examples, respectively) poses a challenge for modern modelling techniques, which require many training instances.

More recently, Talmor et al. (2019) introduce CommonsenseQA, containing 12k multiple-choice questions. Crowdsourced using ConceptNet (Speer and Havasi, 2012), these questions mostly probe knowledge related to factual and physical commonsense (e.g., "Where would I not want a fox?"). In contrast, SOCIAL IQA explicitly separates contexts from questions, and focuses on the types of commonsense inferences humans perform when navigating social situations.

Commonsense Knowledge Bases: In addition to large-scale benchmarks, there is a wealth of work aimed at creating commonsense knowledge repositories (Speer and Havasi, 2012; Sap et al., 2019; Zhang et al., 2017; Lenat, 1995; Espinosa and Lieberman, 2005; Gordon and Hobbs, 2017) that can be used as resources in downstream reasoning tasks. While SOCIAL IQA is formatted as a natural language QA benchmark, rather than a taxonomic knowledge base, it also can be used as a resource for external tasks, as we have demonstrated experimentally.

Constrained or Adversarial Data Collection: Various work has investigated ways to circumvent annotation artifacts that result from crowdsourcing. Sharma et al. (2018) extend the Story Cloze data by severely restricting the incorrect story ending generation task, reducing the sentiment and negation artifacts. Rajpurkar et al. (2018) create an adversarial version of the extractive question-answering challenge, SQuAD (Rajpurkar et al., 2016), by creating 50k unanswerable questions. Instead of using human-generated incorrect answers, Zellers et al. (2018, 2019b) use adversarial filtering of machine generated incorrect answers to minimize surface patterns. Our dataset also aims to reduce annotation artifacts by using a multi-stage annotation pipeline in which we collect negative responses from multiple methods including a unique adversarial question-switching technique.

8 Conclusion

We present SOCIAL IQA, the first large-scale benchmark for social commonsense. Consisting of 38k multiple-choice questions, SOCIAL IQA covers various types of inference about people's actions being described in situational contexts. We design a crowdsourcing framework for collecting QA pairs that reduces stylistic artifacts of negative answers through an adversarial question-switching method. Despite human performance of close to 90%, computational approaches based on large pretrained language models only achieve accuracies up to 65%, suggesting that these social inferences are still a challenge for AI systems. In addition to providing a new benchmark, we demonstrate how transfer learning from SOCIAL IQA to other commonsense challenges can yield significant improvements, achieving new state-of-the-art performance on both COPA and Winograd Schema Challenge datasets.

Acknowledgments

We thank Chandra Bhagavatula, Hannaneh Hajishirzi, and other members of the UW NLP and AI2 community for helpful discussions and feedback throughout this project. We also thank the anonymous reviewers for their insightful comments and suggestions. This research was supported in part by NSF (IIS-1524371, IIS-1714566), DARPA under the CwC program through the ARO (W911NF-15-1-0543), DARPA under the MCS program through NIWC Pacific (N66001-19-2-4031), Samsung Research, and the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE-1256082.

References

Ian Apperly. 2010. *Mindreaders: the cognitive basis of" theory of mind"*. Psychology Press.

Simon Baron-Cohen, Alan M Leslie, and Uta Frith. 1985. Does the Autistic Child have a "Theory of Mind"? *Cognition*, 21(1):37–46.

Ernest Davis and Gary Marcus. 2015. Commonsense reasoning and commonsense knowledge in artificial intelligence. *Commun. ACM*, 58:92–103.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*.

José H. Espinosa and Henry Lieberman. 2005. Eventnet: Inferring temporal relations between commonsense events. In *MICAI*.

MY Ganaie and Hafiz Mudasir. 2015. A Study of Social Intelligence & Academic Achievement of College Students of District Srinagar, J&K, India. *Journal of American Science*, 11(3):23–27.

Travis Goodwin, Bryan Rink, Kirk Roberts, and Sanda M Harabagiu. 2012. UTDHLT: Copacetic system for choosing plausible alternatives. In *NAACL workshop on SemEval*, pages 461–466. Association for Computational Linguistics.

Andrew S Gordon and Jerry R Hobbs. 2017. *A Formal Theory of Commonsense Psychology: How People Think People Think*. Cambridge University Press.

Jonathan Gordon and Benjamin Van Durme. 2013. Reporting bias and knowledge acquisition. In *Proceedings of the 2013 Workshop on Automated Knowledge Base Construction*, AKBC '13, pages 25–30, New York, NY, USA. ACM.

David Gunning. 2018. Machine common sense concept paper.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *NAACL-HLT*.

Vid Kocijan, Ana-Maria Cretu, Oana-Maria Camburu, Yordan Yordanov, and Thomas Lukasiewicz. 2019. A surprisingly robust trick for the winograd schema challenge. In *ACL*.

Baris Korkmaz. 2011. Theory of mind and neurodevelopmental disorders of childhood. *Pediatr Res*, 69(5 Pt 2):101R–8R.

Douglas B Lenat. 1995. Cyc: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38(11):33–38.

Hector J. Levesque. 2011. The winograd schema challenge. In *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*.

Li Lucy and Jon Gauthier. 2017. Are distributional representations ready for the real world? evaluating word vectors for grounded perceptual meaning. In *RoboNLP@ACL*.

Zhiyi Luo, Yuchen Sha, Kenny Q Zhu, Seung-won Hwang, and Zhongyuan Wang. 2016. Commonsense causal reasoning between short texts. In *Fifteenth International Conference on the Principles of Knowledge Representation and Reasoning*.

Gary Marcus. 2018. Deep learning: A critical appraisal. *CoRR*, abs/1801.00631.

Saif Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 174–184.

Chris Moore. 2013. *The development of commonsense psychology*. Psychology Press.

Aida Nematzadeh, Kaylee Burns, Erin Grant, Alison Gopnik, and Thomas L. Griffiths. 2018. Evaluating theory of mind in question answering. In *EMNLP*.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch. In *NIPS-W*.

Haoruo Peng, Daniel Khashabi, and Dan Roth. 2015. Solving hard coreference problems. In *HLT-NAACL*.

Jason Phang, Thibault Févry, and Samuel R. Bowman. 2019. Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks. *CoRR*, abs/1811.01088.

Martha E. Pollack. 2005. Intelligent technology for an aging population: The use of ai to assist elders with cognitive impairment. *AI Magazine*, 26:9–24.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative Pre-Training.

Altaf Rahman and Vincent Ng. 2012. Resolving complex cases of definite pronouns: The winograd schema challenge. In *EMNLP*, EMNLP-CoNLL '12, pages 777–789, Stroudsburg, PA, USA. Association for Computational Linguistics.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. *arXiv preprint arXiv:1806.03822*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy S. Liang. 2016. Squad: 100, 000+ questions for machine comprehension of text. In *EMNLP*.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S. Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Winogrande: An adversarial winograd schema challenge at scale. *ArXiv*, abs/1907.10641.

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. In *AAAI*.

Shota Sasaki, Sho Takase, Naoya Inoue, Naoaki Okazaki, and Kentaro Inui. 2017. Handling multiword expressions in causality estimation. In *IWCS*.

Shlomo S. Sawilowsky. 2009. New effect size rules of thumb. *Journal of Modern Applied Statistical Methods*, 8(2):597–599.

Roy Schwartz, Maarten Sap, Ioannis Konstas, Li Zilles, Yejin Choi, and Noah A Smith. 2017. The effect of different writing tasks on linguistic style: A case study of the ROC story cloze task. In *CoNLL*.

Rishi Kant Sharma, James Allen, Omid Bakhshandeh, and Nasrin Mostafazadeh. 2018. Tackling the story ending biases in the story cloze test. In *ACL*.

Robyn Speer and Catherine Havasi. 2012. Representing general relational knowledge in conceptnet 5. In *LREC*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *NAACL*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019a. From recognition to cognition: Visual commonsense reasoning. In *CVPR*.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In *EMNLP*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019b. Hellaswag: Can a machine really finish your sentence? In *ACL*.

Sheng Zhang, Rachel Rudinger, Kevin Duh, and Benjamin Van Durme. 2017. Ordinal common-sense inference. *Transactions of the Association of Computational Linguistics*, 5(1):379–395.

Yukun Zhu, Ryan Kiros, Richard S. Zemel, Ruslan R. Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 19–27.