DISCOSENSE: Commonsense Reasoning with Discourse Connectives

Prajjwal Bhargava and **Vincent Ng**
Human Language Technology Research Institute
University of Texas at Dallas
Richardson, TX 75083-0688
prajjwalin@protonmail.com, vince@hlt.utdallas.edu

Abstract

We present DISCOSENSE, a benchmark for commonsense reasoning via understanding a wide variety of discourse connectives. We generate compelling distractors in DISCOSENSE using Conditional Adversarial Filtering, an extension of Adversarial Filtering that employs conditional generation. We show that state-of-the-art pre-trained language models struggle to perform well on DISCOSENSE, which makes this dataset ideal for evaluating next-generation commonsense reasoning systems.

1 Introduction

Much of the recent work in commonsense reasoning has focused on evaluating a pre-trained language model's (LM) ability to predict the most plausible ending/option given a context. Even after devising bias reduction techniques (Zellers et al., 2019b; Bras et al., 2020) to mitigate the effects of annotation artifacts and make the task difficult, state-of-the-art LMs have managed to achieve or even surpass human performance on numerous commonsense downstream tasks (Zellers et al., 2019b; Sakaguchi et al., 2020; Bhagavatula et al., 2020). Nevertheless, these LMs are still very far from being able to perform commonsense reasoning as well as humans. Hence, the fact that they have begun to ace existing benchmarks implies that time is ripe to design a new challenging benchmark that can reliably target their limitations.

Motivated by this observation, we present DISCOSENSE, a benchmark for performing commonsense reasoning through understanding a wide variety of discourse connectives. Figure 1 shows an example taken from DISCOSENSE. As can be seen, an example is composed of a *context* (e.g., "Our waitress was very nice, but she kept on forgetting my stuff.") and a discourse connective (e.g., "For example"), and the goal is to choose the most plausible ending out of four options. If we *ignore* the discourse connective, then all four options may seem plausible because we do not know what the writer's intent is. Once we consider both the context and the discourse connective, then it is clear that only option b) is plausible. The reason is that "For example" signals an EXEMPLIFICATION relation between its arguments, and what follows the discourse connective is expected to be an example of the waitress keeping on forgetting the writer's stuff. Using commonsense knowledge, we know that (1) "my beer and fries" is an example of "my stuff", and (2) her taking forever to bring the writer stuff implies she kept on forgetting his/her stuff.

Our waitress was very nice, but she kept on forgetting my stuff. **For example**

a) When I ordered the garlic shrimp, she remembered to add my requested garlic butter.
b) **She took forever to bring me my beer and fries.**
c) When I told her I wanted to use the free breakfast that was available she was not pleased.
d) For some customers, this is fine.

Figure 1: Example on commonsense reasoning with discourse connectives. The correct (i.e., most plausible) option is boldfaced.

What if we replace "For example" with "However" in the example? Since "However" signals a CONTRAST relation, options a) and d) both seem viable. Specifically, option a) describes a situation in which she did not forget the writer's stuff. While option d), unlike option a), does not describe any example that signals a contrast, one may infer a contrast between option d) and the context: being forgetful is fine for some customers. Nevertheless, option a) is arguably *more plausible* than option d) and should be chosen. The reason is that for d) to be sensible, one needs to assume that her forgetting *the writer*'s stuff implies that she is in general forgetful. Without this assumption, it may be strange for other customers to have an opinion on her forgetting the writer's stuff. In general, the most plausible option is the option that makes the smallest number of assumptions, and/or is the most

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 10295–10310
December 7-11, 2022

coherent given the context and the discourse connective. Considering the commonsense knowledge *and* the reasoning involved, it should not be difficult to see that this task is challenging.

Our contributions are four-fold. First, we create DISCOSENSE, a new dataset aimed at testing LMs' commonsense reasoning capabilities through discourse connectives. Second, we employ a controlled text generation based adversarial filtering approach to generate compelling negatives. Third, we establish baseline results on DISCOSENSE with numerous state-of-the-art discriminator models and show that they struggle to perform well on DISCOSENSE, which makes our dataset an ideal benchmark for next-generation commonsense reasoning systems. Finally, we show the efficacy of using DISCOSENSE as a transfer learning resource through sequential fine-tuning of LMs on DISCOSENSE followed by HELLASWAG and achieve near state-of-the-art results on the HELLASWAG test set. To stimulate work on this task, we make our code and data publicly available.[1]

2 Related Work

In this section, we discuss related work, focusing our discussion on the differences between DISCOSENSE and existing commonsense reasoning benchmarks. In addition, we present an overview of Adversarial Filtering, which will facilitate the introduction of the Conditional Adversarial Filtering mechanism we propose in Section 3.

Commonsense reasoning benchmarks. SWAG (Zellers et al., 2018) and HELLASWAG (Zellers et al., 2019b) are arguably the most prominent commonsense reasoning benchmarks. In SWAG, given a partial description along with four candidate endings, the task is to predict the most plausible ending. The synthetic options (a.k.a. distractors) are generated through a process called Adversarial Filtering (AF) (see below). HELLASWAG is an extension of SWAG that seeks to eliminate artifacts in the generated endings. Unlike SWAG and HELLASWAG, DISCOSENSE requires that the discourse connective be taken into account in the reasoning process, thus increasing the number of inference steps and potentially the task complexity. In addition, while the examples in SWAG and HELLASWAG come primarily from ActivityNet (a benchmark focused on dense captioning of temporal events), DISCOSENSE features a more diverse set of examples coming from varied domains that may only be solved with rich background knowledge.

There are benchmarks that aim to test different kinds of commonsense reasoning abilities, although none of them focuses on reasoning over discourse connectives. SocialIQA (Sap et al., 2019), for instance, focuses on social and emotional commonsense reasoning. ABDUCTIVE NLI (Bhagavatula et al., 2020) focuses on abductive reasoning. WINOGRANDE (Sakaguchi et al., 2020) contains Winograd schema-inspired problems, which are essentially hard pronoun resolution problems requiring world knowledge. PIQA (Bisk et al., 2020) examines physical commonsense reasoning. MCTACO (Zhou et al., 2019) and TIMEDIAL (Qin et al., 2021) focus on temporal reasoning in comprehension and dialogue formats.

More closely related to DISCOSENSE are commonsense reasoning benchmarks that involve reasoning with a particular kind of relations. COPA (Choice of Plausible Alternatives) (Roemmele et al., 2011) focuses exclusively on reasoning with CAUSAL relations and involves choosing the more plausible ending out of two (rather than four) options. P-MCQA (Qasemi et al., 2021) focuses exclusively on reasoning with PRECONDITION relations: given a commonsense fact, select the precondition that make the fact possible (enabling) or impossible (disabling) out of four options. δ-NLI (Rudinger et al., 2020), which aims to evaluate *defensible inference*, focuses exclusively on reasoning with the STRENGTHEN/WEAKEN relations: given a premise-claim pair where the premise supports the claim, generate a sentence that either strengthens or weakens the support. WINOVENTI (Do and Pavlick, 2021), which is composed of Winograd-style schemas, focuses exclusively on reasoning with ENTAILMENT relations: given two sentences with an entailment relation, such as "Pete says the pear is delicious. The pear is ____", the goal is to fill in the blank with one of two choices (e.g., "edible", "inedible"). There are two key differences between these datasets and DISCOSENSE. First, rather than focusing on a particular type of relation, DISCOSENSE encompasses 37 discourse connectives signaling different discourse relation types. Second, DISCOSENSE involves reasoning with discourse *connectives*, which is more complicated than reasoning with discourse relations. Specifically, as some connectives are sense-ambiguous

[1]For our code and data, see https://github.com/prajjwal1/discosense/.

Dataset	Model	Human
SWAG (Zellers et al., 2018)	91.71	88
αNLI (Bhagavatula et al., 2020)	91.18	92.9
Hellaswag (Zellers et al., 2019b)	93.85	95.6
CosmosQA (Huang et al., 2019)	91.79	94
PIQA (Bisk et al., 2020)	90.13	94.9
SocialIQa (Sap et al., 2019)	83.15	88.1
MC-TACO (Zhou et al., 2019)	80.87	75.8
WinoGrande (Sakaguchi et al., 2020)	86.64	94
ProtoQA (Boratko et al., 2020)	54.15	74.03
VCR (Zellers et al., 2019a)	63.15	85

Table 1: Status of how competitive current commonsense reasoning benchmarks are for state-of-the-art pretrained language models.

(e.g., the connective "since" may serve as a temporal or causal connective (Pitler and Nenkova, 2009)), a LM will likely need to (implicitly) perform sense disambiguation in order to perform well on DISCOSENSE.

There are datasets and knowledge bases where the semantic/discourse/commonsense relations are explicitly annotated and which can provide data sources from which commonsense reasoning benchmarks can be derived. Examples include (1) the Penn Discourse TreeBank (Prasad et al., 2008), where two sentences or text segments are annotated with their discourse relation type, if any; (2) COREQUISITE (Qasemi et al., 2021), which is used to provide the commonsense facts and the human-generated preconditions in the P-MCQA dataset mentioned above; (3) SNLI (Bowman et al., 2015), where each premise-hypothesis pair is annotated as ENTAILMENT, CONTRADICTION, or NEUTRAL; (4) ATOMIC$^{20}_{20}$ (Hwang et al., 2021), which is a commonsense knowledge graph where the nodes correspond to propositions and the edges correspond to social/physical commonsense relations; and (5) SOCIAL-CHEM-101 (Forbes et al., 2020), which is a collection of statements about commonsense social judgments made given everyday situations.

One of the motivations behind the creation of DISCOSENSE is that state-of-the-art LMs have managed to achieve or even surpass human performance on various commonsense reasoning benchmarks. Table 1 shows the best accuracies achieved by existing LMs on 10 widely used commonsense reasoning benchmarks and the corresponding human performance levels. As can be seen, existing LMs have managed to achieve an accuracy of more than 80% on eight of these benchmarks.

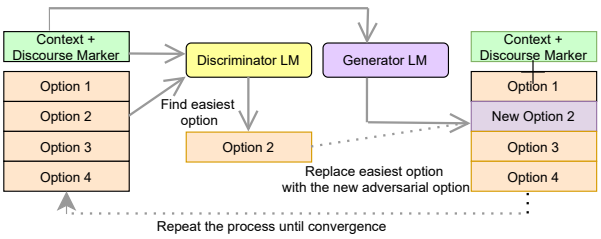

Figure 2: Components of Adversarial Filtering.

Adversarial filtering (AF). Originally proposed by Zellers et al. (2018), AF aims to create examples that would be difficult for models to solve, specifically by replacing the easy options in correctly-solved examples with difficult ones. As shown in Figure 2, AF has three components: data (i.e., examples with multiple options, one of which is correct), a discriminator LM (a classifier that is used to solve each example) and a generator LM (a model that generates new options for an example). In each AF iteration, the discriminator LM is trained on the training set and used to solve each example in the test set. If a test example is incorrectly solved (i.e., the discriminator LM chooses the wrong option), the example is deemed sufficiently difficult and no change is made to it. On the other hand, if a test example is correctly solved, then AF seeks to increase its difficulty by replacing the *easiest* option (i.e., the generated option that the discriminator LM classifies with the highest confidence) with a new option generated by the generator LM. Training a new discriminator LM in each AF iteration ensures that the dataset is not just adversarial for one LM but a class of LMs, as training different instances of the same type of LMs results in models that have differently learned linguistic representations. This process is repeated on all correctly classified examples in the test set until the performance on the test set converges.

3 DISCOSENSE

3.1 Task Description

DISCOSENSE aims to measure the commonsense inference abilities of computational models through the use of discourse connectives. The correct endings can be obtained after understanding the purpose of the given discourse connectives. Given a context $c = (s, d)$, which is composed of a contextual sentence s and a discourse connective d as well as a set of four options $O = \{o_1, o_2, o_3, o_4\}$, the task is to predict the most plausible ending $o_i \in O$.

Data Source	DISCOSENSE Train	DISCOSENSE Test
DISCOVERY Train	Bottom 7%	-
DISCOVERY Validation	-	100%
DISCOFUSE train	Top ∼54k w/ DC	-

Table 2: Data sources for DISCOSENSE and its composition before human verification. DC refers to those samples in DISCOFUSE that are concerned with the discourse connective phenomenon.

3.2 Dataset Creation

To assemble DISCOSENSE, we focus on source datasets that contain two sentences connected through a discourse connective. Specifically, we use two peer reviewed academic datasets, DISCOVERY (Sileo et al., 2019) and DISCOFUSE (Geva et al., 2019). In DISCOVERY, each sentence is composed of two sentences connected via a discourse connective for the purpose of learning joint sentence representations with discourse connectives. DISCOFUSE, on the other hand, is assembled for the task of sentence fusion (i.e., joining several independent sentences into a single coherent sentence). We only consider those examples where a discourse connective is needed for sentence fusion, and include in DISCOSENSE the fused sentences in the Wikipedia[2] split of DISCOFUSE. Since these datasets contain sentences from Common Crawl[3] and Wikipedia articles, DISCOSENSE is diverse in the topics it covers. Importantly, since by construction the discourse connective is crucial in solving the underlying tasks (i.e., sentence representation learning and sentence fusion), the crucial role played by the discourse connectives in these sentences makes them suitable for our use case. Details of how the DISCOVERY and DISCOFUSE sentences are used to create DISCOSENSE are shown in Tables 2 and 3.

3.3 Generating Options

Next, we describe how we generate challenging options for DISCOSENSE using an improved version of AF that we call Conditional Adversarial Filtering (CAF). CAF follows the AF procedure in Figure 2, only differing from AF in terms of (1) the generator LM (Section 3.3.1), (2) the discriminator LM (Section 3.3.2), and (3) how the generator LMs are used to generate options (Section 3.3.3).

[2] https://en.wikipedia.org/
[3] https://commoncrawl.org/

Data	Generator LM
DISCOVERY Train	last 93%
DISCOVERY Test	100%

Table 3: Data used to train the generator LMs in Conditional Adversarial Filtering.

3.3.1 Conditional Generator LM

Pre-training does not explicitly teach how important a particular token or text span is in contributing to the semantics of a sentence. Hence, to be able to generate sentences that are coherent with not only the context but also the discourse connective, we propose to use Controllable Text Generation, which aims to provide a more granular control over how generation happens to match a particular attribute. In the context of Transformer-based LMs, there are two lines of research on controllable text generation. One examines how to steer generation by fine-tuning an extra set of parameters while keeping the base (unconditionally trained) model fixed (Dathathri et al., 2020; Qin et al., 2020; Zhang et al., 2020; Krause et al., 2020), while the other involves conditionally training a generative model on a control variable to generate text w.r.t. a prompt prefix. We adopt the latter approach, extending CTRL (Keskar et al., 2019) to explicitly steer generation w.r.t. discourse relations by using discourse connectives as control codes, as described below.

Training. The input to CTRL is as follows:

$$\underline{\text{input}}\text{: } [d] + [\text{context}] - \underline{\text{label}}\text{: } [\text{ending}]$$

where d is a discourse connective. Specifically, each input context for CTRL is prepended with a connective, and the training task for CTRL is to learn the conditional distribution $p(e|d, \text{context})$ over possible endings e. The predicted ending is then compared with the human generated ending to compute loss. Since the original CTRL model is pre-trained with control codes suitable for open-ended text generation, we fine-tune CTRL on the portion of DISCOVERY shown in Table 3 using all the 174 connectives present in the selected splits. Comparing Tables 2 and 3, we can see that the data the generator LM is fine-tuned on is not part of DISCOSENSE. Doing so ensures that the endings generated by the generator LM are different from the ground truth (i.e., the human written endings).

Decoding. We use Nucleus sampling (Holtzman et al., 2020) for generating options for the training set with the value of p set to 0.7, which means the

weights of the tail of the probability distribution are ignored (i.e., tokens with a cumulative probability mass of less than 0.3 are left out). Additionally, we use a length penalty of 0.8 to restrict the length of the generations to match the average length of the ground truth to avoid the induction of length bias.

Efficacy of conditional generation. Recall that we propose the use of conditional generation, specifically the use of discourse connectives as control codes, in our generator LM because of our hypothesis that the resulting LM would generate options that are more compliant with the purpose of the discourse connective. To test this hypothesis, we compare the *text generation* capability of CTRL with that of GPT2-XL, a model that is trained unconditionally and has nearly the same number of parameters (1.6B) as CTRL, under the *same* evaluation setting. Specifically, both LMs are fine-tuned on the same data (see Table 3) using the same machine (a 2x Quadro RTX 8000 with a batch size of 24). The only difference between them lies in the format of the training examples: in CTRL the discourse connective is used as the control code and therefore precedes the context, whereas in GPT2-XL, the discourse connective follows the context.

The two LMs are then independently applied to generate exactly one option for each example in the DISCOVERY validation set. CTRL achieves a much lower perplexity than GPT2-XL (2.39 vs. 2.53), which suggests that conditional training improves the quality of the generated sentences.

3.3.2 Discriminator LM

We use ROBERTA-LARGE (Liu et al., 2019) as the discriminator LM, which takes the context, the discourse connective, and the four endings as input and predicts the most plausible ending. This LM is trained on the randomly shuffled training split of DISCOSENSE and applied to the DISCOSENSE test set to get the confidence scores associated with its predictions.

3.3.3 Generating Options

Next, we describe how we generate options for the examples in DISCOSENSE. Recall that each example contains one of 174 discourse connectives. Rather than generating options for examples that contain any of these 174 connectives, we select 37 discourse connectives and generate options only for examples that contain one of them. The connectives that are discarded are primarily those that impose few constraints on the endings to be generated given the context according to preliminary experiments. For instance, the connective "and" is discarded because numerous endings are equally plausible. Similarly for connectives that signal a temporal relation (e.g., "before", "after"): they also tend to allow numerous equally plausible endings, as can be seen in examples such as "John went to eat lunch after [ending]". The 37 connectives that we end up choosing are shown in Table 4. These connectives are less likely to yield options that look equally plausible to human annotators and which are indicative of different kinds of discourse relations, such as EXEMPLIFICATION (e.g., "for instance"), CONCESSION (e.g., "although"), COMPARISON (e.g., "in contrast"), and CAUSAL (e.g., "as a result"). 94k examples in DISCOSENSE contain one of the 37 connectives.

although	in other words	particularly
as a result	in particular	rather
by contrast	in short	similarly
because of this	in sum	specifically
because of that	interestingly	subsequently
but	instead	thereafter
consequently	likewise	thereby
conversely	nevertheless	therefore
for example	nonetheless	though
for instance	on the contrary	thus
hence	on the other hand	yet
however	otherwise	
in contrast	overall	

Table 4: Discourse connectives present in DISCOSENSE.

To generate the options for these 94k sentences, we begin by training 20 generator LMs on a randomly shuffled order of the generators' training data (see Table 3) and then inserting them into a circular queue. Although the underlying data is the same, random shuffling ensures that the learned representations of these 20 models are different. Since each example needs to have 3 synthetic options, we use the first 3 generator LMs from the circular queue to generate the initial options for each example. After that, we begin CAF. In each CAF iteration, we (1) train the discriminator LM (see Section 3.3.2) on the DISCOSENSE training set for 4 epochs and use it to filter out the options deemed as easiest by the discriminator LM; and (2) use the next generator LM in the circular queue to generate the options for the examples whose easiest option is removed by the discriminator LM. In other words, a different discriminator LM is used in each CAF iteration, and a generator LM in the

DiscoSense		
# Context Answer tuples	train	9299
	test	3757
	total	**13056**

Statistics		Train / Test
Average # tokens	context	22.08 / 22.51
	answers (all)	18.62 / 18.92
	answers (correct)	16.94 / 18.18
	answers (incorrect)	18.51 / 18.5
Unique # tokens	context	32577 / 16858
	answers (all)	43992 / 27406
	answers (correct)	26836 / 15078
	answers (incorrect)	41158 / 25900

Table 5: Data statistics for DISCOSENSE.

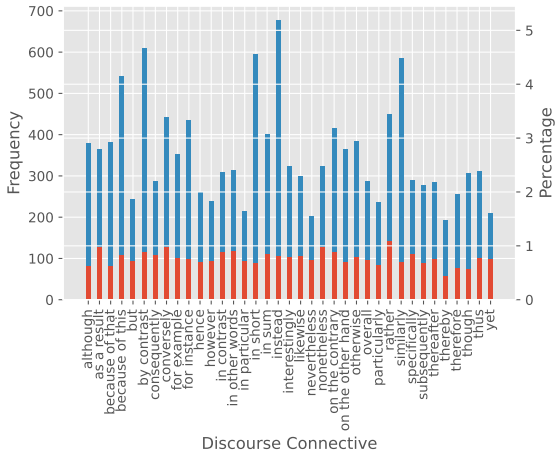

Figure 3: Distribution of examples over discourse connectives in DISCOSENSE.

circular queue is used once every 20 CAF iterations. CAF is run separately for the DISCOSENSE training and test sets. After running CAF for approximately 150 iterations, the average accuracy of a discriminator LM decreased from 86–90% to 34% on the DISCOSENSE test set.

3.3.4 Other Implementation Details

For the models we use in CAF, we obtain the pre-trained weights and the implementations from Hugging Face Transformers (Wolf et al., 2019). These models are trained using the AdamW optimizer (Loshchilov and Hutter, 2019) with a learning rate of $2e^{-5}$. The training of each generator LM is performed on a 2x Quadro RTX 8000 with a batch size of 24 and typically lasts for 3 days. The training of a discriminator LM is performed on a RTX 3090 with a batch size of 16 and typically lasts for 5–6 hours.

3.4 Human Verification

Next, we perform human verification of the examples for which we have generated options. The verification proceeds in two steps. In Step 1, we ask three human verifiers to independently identify the correct option for each example, removing an example if at least one person fails to identify the correct option. We repeat this process until the number of examples that survive this verification reaches 13,056.[4] In Step 2, we ask three human verifiers not involved in Step 1 to independently identify the correct option for each of the 13,056 examples verified in Step 1. We compute for each verifier the accuracy of choosing the correct option and use the average accuracy as the human performance on DISCOSENSE. Appendix A contains the details on how the human verifiers are recruited and the annotation instructions we present to them.

3.5 Dataset Statistics

Statistics on DISCOSENSE are shown in Table 5, in which we report the average number of tokens in (1) the context, (2) the ground truth and (3) the generated endings. The number of unique tokens provides a rough characterization of the richness of the vocabulary. In addition, we report the distribution of the examples over the discourse connectives in DISCOSENSE in Figure 3.

4 Evaluation

4.1 Baseline Systems

Our baselines are composed of prominent LMs with different kinds of Transformer architectures. First, we consider models that are pre-trained in a BERT-like fashion and share architectural similarities, including the base and large variants of BERT (Devlin et al., 2019) and ROBERTA (Liu et al., 2019), as well as ALBERT-XXLARGE-V2 (Lan et al., 2020). As an extension, we select LONGFORMER BASE, which is pre-trained in the same manner as ROBERTA but has a sparse attention matrix.[5] From the autoregressive/decoder based networks, we experiment with XLNET LARGE (Yang et al., 2019), which maximizes the learning of bidirectional contexts and GPT2-XL. For

[4] This is the maximum number of examples we can verify given budgetary constraints.

[5] Some endings are longer than the others. The use of LONGFORMER allows us to see whether a sparse attention matrix can better exploit the length of an ending than other models.

Model	Accuracy / std
Random Guess	25.0
BERT-BASE (110M)	32.86 / 0.45
BERT-LARGE (336M)	34.25 / 1.04
ROBERTA-BASE (125M)	34.11 / 0.45
ROBERTA-LARGE (355M)	34 / 0.2
ALBERT-XXLARGE-V2 (223M)	50.91 / 1.44
LONGFORMER BASE (435M)	35.29 / 0.77
XLNET LARGE (340M)	36.71 / 0.77
FUNNEL-TRANSFORMER-XL (468M)	35.22 / 1.94
ELECTRA-LARGE	65.87 / 2.26
Human Performance	95.40 / 0.20

Table 6: Accuracies (best results obtained among 8 epochs when averaged over 5 runs with random seeds) of the LMs on the DISCOSENSE test set.

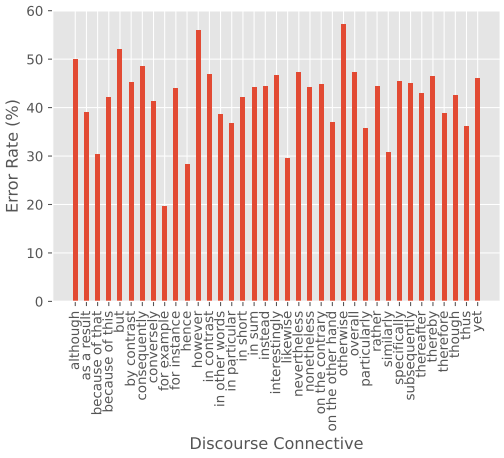

Figure 4: Error rate of ELECTRA on each discourse connective in the DISCOSENSE test set.

models trained with a different pre-training objective, we experiment with ELECTRA-LARGE (Clark et al., 2020) and FUNNEL-TRANSFORMER-XL (Dai et al., 2020), the latter of which is pre-trained in a similar manner as ELECTRA-LARGE.

We obtain the implementations of these LMs from Hugging Face Transformers. We fine-tune them on the DISCOSENSE training set using a 4-way cross-entropy loss in the same way as the discriminator LMs in CAF are trained (see Section 3.3.4) and evaluate them on the test set.

4.2 Results and Discussion

Results on the test set, which are expressed in terms of accuracy, are shown in Table 6. A few points deserve mention.

First, all baselines perform better than random guess (row 1). This implies that while CAF is used to remove easy options, there may still be artifacts in the data that could be exploited by the LMs.

Second, models sharing a similar pre-training objective as that of BERT, such as ROBERTA and LONGFORMER, are among the worst baselines. A similar trend is observed with XLNET. Although ALBERT has the Masked Token Prediction task in its pre-training objective, its architectural differences (i.e., larger hidden states and parameter sharing) and its Sentence Order Prediction objective seem to help it learn inter-sentence coherency properties better than its BERT counterparts.

Third, pre-training appears to play a predominant role in our task. While the BERT family of models are trained with the masked-LM objective, the pre-training objective of ELECTRA (the best baseline) is designed to determine if a token in a human-written sentence has been replaced by a generator. We speculate that ELECTRA's superior performance can be attributed to the fact that its pre-trained knowledge of discriminating between synthetic and human generated tokens transfers well to the task of discriminating between synthetically generated sentences and human written sentences in DISCOSENSE.[6] Nevertheless, the fact that it only achieves an accuracy of 65.87% is indicative of the challenges DISCOSENSE has for existing LMs. Note that this accuracy is much lower than those achieved by LMs on many commonsense reasoning benchmarks (see Table 1). These results suggest that DISCOSENSE is a challenging benchmark for state-of-the-art LMs.

Finally, we report human performance in the last row of Table 6. Details of how these numbers are obtained are discussed in Section 3.4. As can be seen, the accuracy achieved by the best baseline, ELECTRA, lags behind that of humans by nearly 30% points.

4.3 Quantitative Error Analysis

We perform a quantitative error analysis of our best-performing model, ELECTRA. Specifically, we compute for each discourse connective the percentage of examples in the DISCOSENSE test set that are misclassified by ELECTRA, with the goal of gaining a better understanding of the discourse connectives that are perceived as easy as well as those that are perceived as difficult as far as commonsense reasoning is concerned.

Results are shown in Figure 4. As we can see,

[6]While FUNNEL TRANSFORMER employs the same pre-training strategy as ELECTRA, we speculate that the pooling mechanism it uses to compress hidden states offsets the benefits it receives from its pre-training strategy on this task.

the misclassification rates are highest for those discourse connectives that express contrast (e.g., "otherwise", "however", "but", "although"). A plausible explanation for this result is that it is often hard to anticipate what a human would have in mind if they are trying to indicate the opposite of what they mean to say. On the other hand, the model finds it easy to predict sentences where the discourse connective signals compliance and exemplification (e.g., "similarly", "likewise", "hence", "because of that", "for example").

4.4 Qualitative Error Analysis

To better understand the mistakes made by ELECTRA, we manually inspected 100 randomly selected examples that are misclassified and identified four major reasons why they are misclassified.

1. Less plausible endings. This category contributes to 21% of the errors where the model chooses a less plausible ending. Choosing a less plausible option could be associated with a partial understanding of the context or unwarranted assumptions. In Example 1 of Figure 5, the model makes the assumption that whatever is applicable to grass is also applicable to trees. However, the option it ends up picking is non-factual in nature because of the phrase "7000 years ago".

2. Abstract associations. 14% of the errors are made due to the formation of abstract associations between concepts. The model seems to rely on certain spans of context for classification rather than understand the semantics in its entirety. In Example 2 of Figure 5, the model seems to wrongly associate "energy dense nutrients" with "obesity" and fails to understand that the context is discussing the correlation between nutrient deficit diet and people belonging to lower income groups.

3. Complex Context Understanding. 23% of the examples are misclassified due to the fact that a deeper than usual reasoning is needed to understand the context. In Example 3 of Figure 5, we see that the context is about something weighing on a mind, indicating that the author may be faced with a pressing situation. The connective "but" indicates that while the situation being dealt with is problematic or stressful, the author would still pursue it, making option c) the most plausible. Here, the model fails to understand what it means to have something weighing on mind and what that can make a person do, in this case, "ask bigger questions".

1) Although the grasses were only a moment old, they appeared as if they were months old. **Likewise**

a) Similar phenomena occurred with the ancient trees around the earth 7,000 years ago.
b) The dinosaurs were not billions of years old.
c) Several seeds were found encased within stems that are several months old, but they seemed quite fresh and alive.
d) The trees, although only a day old when they sprouted forth, were nevertheless like trees years old as they were fully grown.

2) Low income people are less likely to consume a healthy diet than wealthier people, and energy dense nutrients poor diets are preferentially consumed by persons of lower socioeconomic status. **Consequently**

a) Nutrients associated with these diets may be potentially contributing to obesity and diabetes.
b) Metabolic syndrome is primarily related to obesity.
c) Their health is at greater risk from diet related illness.
d) A great number of persons suffering from obesity related diseases receive inadequate nutritional care.

3) It weighs on a mind, all this **but**

a) You have to live it if you want to know whats on it.
b) All that means in practice.
c) It does make me want to back up and ask even bigger questions.
d) In a kind of perverse way, I don't really feel sad.

Figure 5: Examples misclassified by ELECTRA (misclassified options in pink; ground truths in **green**).

4. Lack of understanding of the discourse connective. In many cases it is difficult to pinpoint the reason why an example is misclassified. Hence, if a misclassified example is not covered by any of the first three categories, we attribute the mistake to a lack of understanding of the discourse connective. This category contributes to 42% of the errors.

4.5 Role of Context and Discourse connective

To better understand the role played by the context and the discourse connective in a LM's reasoning process, we conduct two ablation experiments. In the first experiment, we remove the discourse connective, so only the context and the endings are available to the LMs. In the second experiment, we strip the context *and* the discourse connective, exposing only the endings to the LMs.

Results of these experiments are shown in the C+E column and the E column of Table 7 respectively. For comparison purposes, the results obtained by not removing anything are shown in the

Model	C+D+E	C+E	E
BERT-BASE	32.86 / 0.5	33.95 / 0.8	32.27 / 1.3
BERT-LARGE	34.25 / 1.0	33.30 / 0.6	30.29 / 0.9
ROBERTA-BASE	34.11/ 0.5	32.73 / 0.4	32.49 / 0.7
ROBERTA-LARGE	34.99 / 0.2	34.94 / 0.9	32.70 / 0.7
ALBERT-XXL	50.91/ 1.4	50.87 / 1.2	38.04 / 0.4
LONGFORMER	35.29 / 0.8	36.82 / 0.8	33.18 / 1.0
XLNET LARGE	36.71 / 0.8	36.68 / 1.1	31.87 / 0.4
FUNNEL-XL	35.22 / 1.9	34.62 / 4.9	30.76 / 4.2
ELECTRA-LARGE	65.87 / 2.3	56.75 / 0.8	43.33 / 2.2

Table 7: Accuracies (best results among 8 epochs when averaged over 5 runs with random seeds) on the DISCOSENSE test set with specific pieces of information from the input removed. D, C, and E refer to the discourse connective, the context and the endings respectively. The numbers following '/' are the standard deviations.

C+D+E column. As can be seen, when the discourse connective is removed, performance drops for all baselines except for BERT-BASE and LONGFORMER, and when both the discourse connective and the context are removed, performance drops for all baselines. In the case of the best baseline, ELECTRA, performance drops abruptly as information is withdrawn (C+E: 17.91% and E: 44.27%), thus highlighting its reliance on both pieces of information for its competitive performance. Overall, these results suggest that reasoning over both the context and the connective is necessary for this task. It is worth mentioning, though, that even when both the context and the connective are removed, all the LMs still manage to achieve an accuracy of more than 30%. Additional experiments are needed to determine the reason why they perform considerably better than random guess when only the endings are given. This will most likely involve an examination of whether there are systematic differences between the human-generated sentences and their automatically generated counterparts.

4.6 DISCOSENSE for Transfer Learning

Next, we look at DISCOSENSE from the perspective of a transfer learning source. Specifically, to understand whether fine-tuning a LM on DISCOSENSE can improve its performance on a related dataset, HELLASWAG, we perform sequential fine-tuning, where we fine-tune each baseline LM on the DISCOSENSE training set followed by the HELLASWAG training set (both for 4 epochs). Note that discourse connectives are removed from the input because HELLASWAG does not have them.

Results on the *validation* split of HELLASWAG are shown in Table 8. Specifically, the HS column shows the results of the baselines on HELLASWAG and the DS→HS shows the results of the baselines after sequential fine-tuning. As we can see, sequential fine-tuning yields performance improvements with almost all LMs. Notably, the improvement is more pronounced for ELECTRA (4.3%) than for ALBERT and the BERT-based models. One plausible reason is that ELECTRA does not struggle as much in understanding DISCOSENSE as the BERT-based models do, and as a result, it shows a bigger improvement, possibly benefitting from the diverse contextual nature of DISCOSENSE.

Model	HS	DS⟶HS
BERT-BASE-UNCASED	38.47	**40.38**
BERT-LARGE-UNCASED	44.36	42.54
ROBERTA-BASE	58.21	57.00
ROBERTA-LARGE	81.50	**82.34**
ALBERT-XXLARGE-V2	80.97	**81.47**
XLNET-LARGE	76.47	**76.56**
ELECTRA-LARGE	86.90	**91.50**
FUNNEL-TRANSFORMER-XL	86.88	**87.50**

Table 8: Results of sequential fine-tuning on the validation split of HELLASWAG.

Finally, we evaluate the sequentially fine-tuned ELECTRA-LARGE model on the HELLASWAG *test* split. The model achieves an accuracy of 90.76%, considerably outperforming its vanilla fine-tuning counterpart (85.75%) and only underperforming models that have 4x (e.g., He et al. (2021), Lourie et al. (2021)) and 32x more parameters (e.g., Lourie et al. (2021)) and are trained on 23x more data.

5 Conclusion

Motivated in part by the fact that existing pre-trained language models have surpassed human performance on numerous commonsense reasoning datasets, we introduced DISCOSENSE, a challenging benchmark that concerns commonsense reasoning with discourse connectives to determine the most plausible ending of a sentence. This task was made difficult by the synthesis of high quality complex examples, which was made possible through coupling highly competitive conditionally trained models for language generation with Adversarial Filtering. The best performing model on DISCOSENSE only achieved an accuracy of 65%, significantly lagging behind humans. This makes DISCOSENSE an ideal benchmark for next-generation commonsense reasoning systems.

Ethical Considerations

Following the guidelines in Mitchell et al. (2019), Bender and Friedman (2018), and Gebru et al. (2021), we believe that we have provided all the necessary information in our description of DISCOSENSE. In this section, we focus on ethical considerations.

Bias mitigation. While it may not be possible to eliminate all of the biases that exist in a dataset, we have certainly taken steps to mitigate biases in DISCOSENSE. Adversarial Filtering has been shown to be an effective de-biasing approach to remove annotation artifacts, and we have taken a step further to improve this approach through conditional text generation. In addition, to our knowledge, our work has used more capable generators and discriminators (adversarial filter) to synthesize text in comparison to other works (Zellers et al., 2018, 2019b; Bras et al., 2020).

Human annotator information. All annotators/verifiers were hired during Summer 2021 as student workers (20-25 hours/week) with full consent. All of them were undergraduate and graduate students aged around 20-24. The group comprised both male and female students with members belonging to different ethnicity, namely Asian, Caucasian, and Hispanic. All annotators were native English speakers. Additional details on the selection of annotators can be found in Appendix A.1. The annotators were compensated with a hourly rate of 10 US dollars.

Steps taken to protect annotators from harmful content. All annotators were provided with a thorough instructional training session in which they were instructed on how to annotate the data, how to go about the whole task, and what kind of examples to skip. Before we shared the data, we performed filtering of examples based on sensitive/offensive keywords. After the filtering process, we provided the annotators with a document that contains instructions on how to annotate and how to go about the whole task (see Appendix A.2). They were asked to follow their own pace (the amount of time they can spend per example). They were asked to attempt examples that were specifically related to commonsense reasoning tasks. Since the aforementioned keyword-based approach for filtering harmful content may not be able to identify all harmful/offensive documents, the annotators were provided with an opportunity to skip examples that they would consider offensive, sensitive or challenging enough to confuse them.

Is this dataset consistent with the terms of use and the intellectual property and privacy right of people? The most important term of use for this dataset is that it shall primarily be used for NLP research. The source text of this dataset was obtained from DISCOVERY and DISCOFUSE, both of which have been there for a long time. These datasets have been obtained from Common Crawl and Wikipedia data, which is public information. Therefore these data sources do not contain any information that is non-public. We agree with the authors of DISCOVERY and DISCOFUSE that these data sources do not seem to mis-represent any community nor can be used to identify a certain set of individual known outside public information. Through conditional text generation, which has been used to synthesize commonsense knowledge text with discourse connectives, we present text that is mostly suitable for commonsense reasoning tasks, making this work consistent with the terms of use. We believe that our work does not have use cases that would usually be considered out-of-bounds for NLP research.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? We have highlighted all the necessary information required by the user of this dataset to use it for their own use case. Each example has gone through multiple rounds of screening. We do not expect to see any risk being posed by the user of this dataset nor any financial harm associated with its use.

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? We will open-source all the models and data produced from this work immediately after publication. We plan to release it on a GitHub repository with the MIT license and also make it available on Hugging Face Datasets.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? We have provided all the essential information needed by a user to extend this work. They will have access to the data and the models which they can use for experimentation. We will continue to monitor the GitHub repository to resolve issues.

Limitations

Next, we discuss some limitations of our work.

Limitations of (Conditional) Adversarial Filtering. Recall that we seek to create a challenging commonsense reasoning benchmark by automatically removing annotation artifacts[7] and replacing easy options in examples via the design and use of Conditional Adversarial Filtering (CAF). Although CAF is an improved version of Adversarial Filtering (AF), which has been frequently used in the last few years in the construction of commonsense reasoning benchmarks, it is not without its limitations. Specifically, how well CAF can identify annotation artifacts and easy options and subsequently remove artifacts and replace easy options with difficult ones depends on how good the discriminator and the generators are. As discussed before, while the discriminator and generators we use in the creation of DISCOSENSE are stronger than those used in the creation of virtually all other commonsense reasoning benchmarks (e.g., SWAG and HELLASWAG), these discriminator and generators are still not perfect. In particular, the fact that the best-performing baseline, ELECTRA, achieves an accuracy that is substantially higher than random guess (i.e., 67% vs. 25%) is an indication that the discriminator and generators fail to remove all annotation artifacts and/or replace easy options with sufficiently difficult ones for state-of-the-art pre-trained language models. As pre-trained language models continue to improve, we do expect that the family of AF approaches will become more effective. Nevertheless, moving forward, researchers should think about whether there are alternative, non-AF-based approaches for creating challenging commonsense reasoning benchmarks that do not suffer from the limitations of AF-based approaches.

Coverage of discourse connectives. While we have taken measures to ensure that DISCOSENSE has a good coverage of discourse connectives, there are still many connectives that are not present in DISCOSENSE due to the ambiguity they give rise to. For example, having "and" does not make it clear what the next sentence should talk about given a context, meaning that it is likely for these connectives to have many endings that are equally plausible. Given budgetary constraints, we do not want to waste our human verification effort on identifying and filtering the potentially large number of examples that contain equally plausible endings as a result of these ambiguous discourse connectives. So, we have avoided their inclusion in DISCOSENSE thus far. However, to fairly evaluate how good a model is in reasoning with discourse connectives, we should augment DISCOSENSE with ambiguous discourse connectives in the future.

Types of reasoning. Although there is a high coverage in the types of commonsense reasoning DISCOSENSE aims to study (e.g., physical world reasoning, social commonsense reasoning, numerical reasoning, linguistic reasoning, temporal reasoning, abductive reasoning), there are other kinds of reasoning studied within the NLP literature that this benchmark does not aim to evaluate upon, such as multi-hop reasoning and symbolic reasoning. It is still not clear how adversarial approaches can be applied to make these kinds of reasoning difficult. We leave this component to a future work.

[7] An example of an annotation artifact would be that examples in a commonsense reasoning task can be solved by a pre-trained language model using unintended artifacts that exist in the data such as lexical overlap/similarity.

Acknowledgments

We thank the three anonymous reviewers for their insightful comments on an earlier draft of the paper. This work was supported in part by NSF Grant IIS-1528037. Any opinions, findings, conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views or official policies, either expressed or implied, of the NSF.

References

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen tau Yih, and Yejin Choi. 2020. Abductive commonsense reasoning. In *The 8th International Conference on Learning Representations*, Addis Ababa, Ethiopia.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020. PIQA: Reasoning about physical commonsense in natural language. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, pages 7432–7439, New York City, NY. AAAI Press.

Michael Boratko, Xiang Li, Tim O'Gorman, Rajarshi Das, Dan Le, and Andrew McCallum. 2020. Pro-

toQA: A question answering dataset for prototypical common-sense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1122–1136, Online. Association for Computational Linguistics.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew E. Peters, Ashish Sabharwal, and Yejin Choi. 2020. Adversarial filters of dataset biases. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 1078–1088. PMLR.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *The 8th International Conference on Learning Representations*, Addis Ababa, Ethiopia.

K. Bretonnel Cohen, Karën Fort, Gilles Adda, Sophia Zhou, and Dimeji Farri. 2016. Ethical issues in corpus linguistics and annotation: Pay per hit does not affect effective hourly rate for linguistic resource development on Amazon Mechanical Turk. In *Proceedings of LREC Workshop on ETHics In Corpus Collection, Annotation Application (ETHI-CA2 2016)*, pages 8–12.

Zihang Dai, Guokun Lai, Yiming Yang, and Quoc Le. 2020. Funnel-transformer: Filtering out sequential redundancy for efficient language processing. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020*, pages 4271–4282, Online. Curran Associates, Inc.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and play language models: A simple approach to controlled text generation. In *The 8th International Conference on Learning Representations*, Addis Ababa, Ethiopia.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Nam Do and Ellie Pavlick. 2021. Are rotten apples edible? challenging commonsense inference ability with exceptions. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2061–2073, Online. Association for Computational Linguistics.

Maxwell Forbes, Jena D. Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 2020. Social chemistry 101: Learning to reason about social and moral norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 653–670, Online. Association for Computational Linguistics.

Karën Fort, Gilles Adda, and K. Bretonnel Cohen. 2011. Last words: Amazon Mechanical Turk: Gold mine or coal mine? *Computational Linguistics*, 37(2):413–420.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.

Mor Geva, Eric Malmi, Idan Szpektor, and Jonathan Berant. 2019. DiscoFuse: A large-scale dataset for discourse-based sentence fusion. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3443–3455, Minneapolis, Minnesota. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: Decoding-enhanced bert with disentangled attention. *Computing Research Repository, arXiv:2006.03654*.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *The 8th International Conference on Learning Representations*, Addis Ababa, Ethiopia.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos QA: Machine reading comprehension with contextual commonsense reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2391–2401, Hong Kong, China. Association for Computational Linguistics.

Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2021. COMET-ATOMIC 2020: On symbolic and neural commonsense knowledge graphs. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, pages 6384–6392, Online. AAAI Press.

Nitish Shirish Keskar, Bryan McCann, Lav Varshney, Caiming Xiong, and Richard Socher. 2019. CTRL: A conditional transformer language model for controllable generation. *Computing Research Repository, arXiv:1909.05858*.

Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq Joty, Richard Socher, and Nazneen Fatema Rajani. 2020. GeDi: Generative discriminator guided sequence generation. *Computing Research Repository, arXiv:2009.06367*, abs/2009.06367.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. In *The 8th International Conference on Learning Representations*, Addis Ababa, Ethiopia.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *Computing Research Repository, arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *The 7th International Conference on Learning Representations*, New Orleans, Louisiana.

Nicholas Lourie, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. UNICORN on RAINBOW: A universal commonsense reasoning model on a new multitask benchmark. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence*, Online. AAAI Press.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*'19)*, pages 220–229, New York, NY. Association for Computing Machinery.

Emily Pitler and Ani Nenkova. 2009. Using syntax to disambiguate explicit discourse connectives in text. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 13–16, Suntec, Singapore. Association for Computational Linguistics.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Ehsan Qasemi, Filip Ilievski, Muhao Chen, and Pedro A. Szekely. 2021. CoreQuisite: Circumstantial preconditions of common sense knowledge. *Computing Research Repository, arXiv:2104.08712*.

Lianhui Qin, Aditya Gupta, Shyam Upadhyay, Luheng He, Yejin Choi, and Manaal Faruqui. 2021. TIMEDIAL: Temporal commonsense reasoning in dialog. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7066–7076, Online. Association for Computational Linguistics.

Lianhui Qin, Vered Shwartz, Peter West, Chandra Bhagavatula, Jena D. Hwang, Ronan Le Bras, Antoine Bosselut, and Yejin Choi. 2020. Back to the future: Unsupervised backprop-based decoding for counterfactual and abductive commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 794–805, Online. Association for Computational Linguistics.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew Gordon. 2011. Choice of Plausible Alternatives: An evaluation of commonsense causal reasoning. In *Proceedings of the AAAI Spring Symposium on Logical Formalizations of Commonsense Reasoning*. AAAI Press.

Rachel Rudinger, Vered Shwartz, Jena D. Hwang, Chandra Bhagavatula, Maxwell Forbes, Ronan Le Bras, Noah A. Smith, and Yejin Choi. 2020. Thinking like a skeptic: Defeasible inference in natural language. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4661–4675, Online. Association for Computational Linguistics.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Winogrande: An adversarial winograd schema challenge at scale. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, pages 8732–8740, New York City, NY. AAAI Press.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social IQa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4463–4473, Hong Kong, China. Association for Computational Linguistics.

Damien Sileo, Tim Van De Cruys, Camille Pradel, and Philippe Muller. 2019. Mining discourse markers for unsupervised sentence representation learning. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3477–3486, Minneapolis, Minnesota. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame,

Quentin Lhoest, and Alexander M. Rush. 2019. HuggingFace's Transformers: State-of-the-art natural language processing. *Computing Research Repository, arXiv:1910.03771*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R. Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019*, Vancouver, Canada. Curran Associates, Inc.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019a. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2019)*, pages 6720–6731, Long Beach, CA.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 93–104, Brussels, Belgium. Association for Computational Linguistics.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019b. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Jeffrey O. Zhang, Alexander Sax, Amir Zamir, Leonidas Guibas, and Jitendra Malik. 2020. Side-tuning: A baseline for network adaptation via additive side networks. In *Proceedings of the 16th European Conference on Computer Vision (ECCV 2020)*, pages 698–714, Glasgow, UK.

Ben Zhou, Daniel Khashabi, Qiang Ning, and Dan Roth. 2019. "going on a vacation" takes longer than "going for a walk": A study of temporal commonsense understanding. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3363–3369, Hong Kong, China. Association for Computational Linguistics.

A Crowdsourcing Details

We performed crowdsourcing to determine if humans are able to identify the ground truth in each example present in the training and test sets of DISCOSENSE.

A.1 Selection of Annotators

We selected undergraduate and graduate student workers primarily on the basis of relevant background knowledge and their current skill set. Since the dataset features intricate English sentences, we shortlisted students who are native English speakers. We presented all shortlisted student workers with a one-hour training session followed by an evaluation test and ranked them based on their performance on the test. We then hired the top-performing students as our annotators and asked them to answer each example in DISCOSENSE. We closely monitored the progress and the performance of each annotator and provided timely feedback through virtual meetings on how his/her way of performing annotation can be improved by understanding the difficulties encountered by him/her and rectifying any problems if they have been doing the annotation in an incorrect manner. All of the annotators were made aware of how the annotated data would be used and its implications. The student workers who did not perform the annotation work in a satisfactory manner were replaced by the next best performing student on the waitlist. All student workers were paid $10/hr (Fort et al., 2011; Cohen et al., 2016).

A.2 Instructions Provided to Annotators

Below are the instructions we presented to the annotators during our one-hour tutorial. Each of them received a copy of these instructions plus numerous examples (one of them is shown below) at the end of the tutorial and was given an opportunity to ask clarification questions about the instructions and the annotation process in general after a closer examination of these instructions and examples.

Motivation. Academic research is an exploration activity to solve problems that have not been completely solved before. By this nature, each academic research work must sit at the frontier of the field and present novelties that have not been addressed by prior works. Machine Learning (specifically Deep Learning) has seen advancements in numerous domains such as personal assistants, machine translation, etc. What you may have seen is a particular Machine Learning algorithm being deployed in finished end products, but what we do not observe are the research challenges that were overcome behind the scenes to get there. What you will be involved in is one of the research tasks that is far from being solved. You will be contributing towards a meaningful task that has the potential to make scientific advancements. We are dealing with Commonsense Reasoning with regard to

Natural Language Processing. Specifically, we are addressing the problem of how machines can learn to reason in the manner humans do with textual data. This task often requires humans to make use of the information they have acquired about our world (e.g., how the physical world works), and reason about what they read and how it complies with what they already know. The reasoning works on multiple levels: firstly we understand the information we read, make sure we understood it properly, and then reason through various means about it to ensure that it makes sense with what we know. The following quote sums it up nicely.

> The brain is an abduction machine, continuously trying to prove abductively that the observables in its environment constitute a coherent situation.
>
> — Jerry Hobbs, ACL 2013 Lifetime Achievement Award winner

Current Deep Learning models are very good at picking unintended signals to arrive at the correct answer, but this is not what we intend them to do. For instance, if "not" is present within a sentence, then the model might be biased towards predicting negation even though that might not be the case. We require them to pick the right answer through a reasoning process that is as close to the human reasoning process as possible. To address this task, we need to build a dataset that aims at providing "correct" signals to our models, and hopefully the models can learn to reason reasonably well once they are trained on this dataset.

High level description of the task. The task description is straightforward.

> Given a context and a discourse connective, predict which ending is the best to the best of your knowledge/capability.

While reading the contexts, you should understand what the discourse connective is supposed to convey. A discourse connective is a word/phrase used to connect two sentences and reveal their relationship. Consider the following examples.

1. I am feeling hungry, **as a result**, I cooked lunch.
2. **Although** I liked reading the book, there were some major flaws.

> I love that he's able to use wired as a venue for launching future bestsellers **though**
>
> a) I think the wired article is a bit too long.
> b) I don't think its a bad thing for him to do so.
> c) I do agree with some of the other reviews that wired is not a very well written book.
> **d) Honestly, I might have preferred the podcast of his presentation on the topic.**

Figure 6: Example taken from the DISCOSENSE training set. The correct answer is **boldfaced**.

In these examples, the green-colored text is the context and the discourse connective is boldfaced. Notice how the text in orange is framed in accordance with the discourse connective. For instance, if "as a result" is present, then the ending will most likely be about the consequence of what is described in the context; in contrast, if "although" is the discourse connective, then the ending needs to take a contrasting standpoint. Hence, the discourse connective decides what the ending needs to talk about. We have provided a list of the discourse connectives you can expect in the "Role of discourse connectives" section. You are required to be completely familiar with the role each connective is supposed to play.

Figure 6 shows an example you might expect in the dataset. Any ending that violates what we know about how the physical world works or challenges our notions about anything in particular needs to be discarded. You have to choose the most plausible ending, which is the ending that is the most feasible amongst four endings. In cases where two or more endings seem equally feasible, the following criteria should be used.

- Any ending that seems to be indisputably correct can be regarded as the best.
- The best ending will not be inconsistent with the context and the discourse connective.
- If all three options seem incorrect and implausible and the remaining option makes the most sense relatively, then it should be chosen since the correct ending is the best ending by default.
- Two endings can make sense, but the ending that is likely to be more sensible is what you should consider as the best. For example:

- Context: A man is singing into a microphone.
- Ending 1: A man performs a song.
- Ending 2: A man is performing on stage.

Ending 2 is not incorrect but ending 1 makes more logical sense to us as humans. In such cases, mark what seems more logical to you.

Role of discourse connectives. Please make sure that you understand the role of each connective.

- although: in spite of the fact that; even though
- as a result: because of something
- because of this: for the reason that
- because of that: for the reason that
- but: used for joining two ideas or statements when the second one is different from the first one
- by contrast: used to express difference with something
- consequently: as a result
- conversely: introducing a statement or idea which reverses one that has just been made or referred to
- for example: used to introduce something
- for instance: as an example
- hence: as a consequence; for this reason
- however: used to introduce a statement that contrasts with or seems to contradict something that has been said previously
- in contrast: used to express difference with something
- in other words: to put it another way
- in particular: especially; specifically
- in short: to sum up; briefly
- in sum: to sum up; in summary
- instead: as an alternative or substitute
- interestingly: in a way that arouses curiosity or interest
- likewise: in the same way
- nevertheless: in spite of that; notwithstanding; all the same
- nonetheless: in spite of that; nevertheless
- on the contrary: conversely; used to intensify a denial of what has just been implied or stated by suggesting that the opposite is the case
- on the other hand: used to introduce a contrasting point of view, fact, or situation
- otherwise: in circumstances different from those present or considered; or else
- overall: taking everything into account
- particularly: especially
- rather: used to indicate one's preference in a particular matter; preferably
- similarly: in a similar way
- specifically: in a way that is exact and clear; precisely
- subsequently: after a particular thing has happened; afterward
- thereafter: after that time
- thereby: by that means; as a result of that
- therefore: for that reason; consequently
- though: despite the fact that; although
- thus: as a result or consequence of this; therefore
- yet: so far; up until the present or a specified or implied time; by now or then