# DISCOSENSE: Commonsense Reasoning with Discourse Connectives

## 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

This paper addresses the critical need for challenging benchmarks that can reliably target the limitations of current pre-trained language models (LMs) in commonsense reasoning. State-of-the-art LMs have achieved or even surpassed human performance on numerous commonsense downstream tasks. 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 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.

## 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 and HELLASWAG 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, for instance, focuses on social and emotional commonsense reasoning. ABDUCTIVE NLI focuses on abductive reasoning. WINOGRANDE contains Winograd schema-inspired problems, which are essentially hard pronoun resolution problems requiring world knowledge. PIQA examines physical commonsense reasoning. MCTACO and TIMEDIAL 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) focuses exclusively on reasoning with CAUSAL relations and involves choosing the more plausible ending out of two (rather than four) options. P-MCQA 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. NLI, 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, which is composed of Winogradstyle 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

| Dataset | Model | Human |
|---|---|---|
| SWAG | 91.71 | 88 |
| NLI | 91.18 | 92.9 |
| Hellaswag | 93.85 | 95.6 |
| CosmosQA | 91.79 | 94 |
| PIQA | 90.13 | 94.9 |
| SocialIQa | 83.15 | 88.1 |
| MC-TACO | 80.87 | 75.8 |
| WinoGrande | 86.64 | 94 |
| ProtoQA | 54.15 | 74.03 |
| VCR | 63.15 | 85 |

Table 1: Status of how competitive current common-sense reasoning benchmarks are for state-of-the-art pre-trained language models.

Figure 1: Components of Adversarial Filtering.

with discourse connectives, which is more complicated than reasoning with discourse relations. Specifically, as some connectives are sense-ambiguous

(e.g., the connective “since” may serve as a temporal or causal connective), 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, where two sentences or text segments are annotated with their discourse relation type, if any; (2) COREQUISITE, which is used to provide the commonsense facts and the human-generated preconditions in the P-MCQA dataset mentioned above; (3) SNLI, where each premise-hypothesis pair is annotated as ENTAILMENT, CONTRADICTION, or NEUTRAL; (4) ATOMIC20, 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, 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

Adversarial filtering (AF). Originally proposed by, AF aims to create examples that would be difficult for models to solve, specifically by replacing the easy options in correctlysolved 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.

| 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.

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

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

# 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 = o1, o2, o3, o4, the task is to predict the most plausible ending oi belongs to O.

## 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 and DISCOFUSE. 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 split of DISCOFUSE. Since these datasets contain sentences from Common Crawl 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).

### 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 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 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:

input: <d> <contexts> label: <endings>

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, 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 openended 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 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 GPT2XL, 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 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 gen-

erated 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
because of this
because of that
but
consequently
conversely
for example
for instance
hence
however
in contrast
in other words
in sum
interestingly
instead
likewise
nevertheless
nonetheless
on the contrary
on the other hand
otherwise
overall
particularly
specifically
subsequently
thereafter
thereby
therefore
though
thus
yet

Table 4: Discourse connectives present in DISCOSENSE.

| | DiscoSense | |
|---|---|---|
| | train | 9299 |
| Context Answer | test | 3757 |
| tuples | total | 13056 |
| | Statistics | Train / Test |
| | context | 22.08 / 22.51 |
| Average | answers (all) | 18.62 / 18.92 |
| | answers (correct) | 16.94 / 18.18 |
| tokens | answers (incorrect) | 18.51 / 18.5 |
| | context | 32577 / 16858 |
| Unique | answers (all) | 43992 / 27406 |
| tokens | answers (correct) | 26836 / 15078 |
| | answers (incorrect) | 41158 / 25900 |

Table 5: Data statistics for 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 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

### 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. These models are trained using the AdamW optimizer 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

| 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.

reaches 13,056. 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 and ROBERTA, as well as ALBERT-XXLARGE-V2. As an extension, we select LONGFORMER BASE, which is pre-trained in the same manner as ROBERTA but has a sparse attention matrix. From the autoregressive/decoder based networks, we experiment with XLNET LARGE, which maximizes the learning of bidirectional contexts and GPT2-XL. For

models trained with a different pre-training objective, we experiment with ELECTRA-LARGE and FUNNEL-TRANSFORMER-XL, 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 4way 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 pretrained 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. Nevertheless, the fact that it only achieves an accuracy of 65.87

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

### 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,

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.

Less plausible endings. This category contributes to 21 perentt 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".

Abstract associations. 14 percent 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.

Complex Context Understanding. 23

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.

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.

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).

make a person do, in this case, “ask bigger questions”.

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

### 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,