Can Language Models Learn Commonsense Truisms?

Anonymous ACL submission

Abstract

Language Models (LMs) pre-trained on extremely large corpora have achieved significantly better performances than previous models on Natural Language Understanding (NLU) benchmarks including commonsense reasoning tasks. Specifically, prior studies have shown that simply fine-tuned LMs can achieve performances on Winograd Schema Challenge (WSC) close to human-level. However, the crucial question of whether LMs can solve commonsense tasks due to the capability of reasoning or just shallow pattern matching has not been addressed. This work aims to extensively analyze the current LMs and probe their capacity to understand commonsense truisms regardless of how they are phrased. We found that current state-of-the-art LMs are just picking up statistical patterns in the training data and do not have the capacity to fully understand commonsesne truisms.

1 Introduction

Pre-trained Langauge representations like BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) have outperformed previous state-of-the-art models by a large margin on multiple NLU benchmarks, including those designed to test commonsense reasoning like CommonsenseQA (Talmor et al., 2019) and WSC (Levesque et al., 2012). Besides, several work has found that simple methods like fine-tuning these LMs can already yield a lot better results than the more sophisticated models (Trinh and Le, 2018; Kocijan et al., 2019). However, pre-trained LMs notoriously require an extremely large amount of training corpora and computing power. And the training objective of the most widely used LMs is simply predicting masked words in a sentence without the injection of any knowledge or logic rules (Liu et al., 2019).This naturally raises the question that whether LMs produce good results on commonsense reasoning datasets due to their ability to reason and use logic like humans, or they are dependent on some other clues.

The fact that LMs heavily depend on the training data likely makes LMs biased towards expressions that appear in the training corpora most frequently. However, the ability to reason about a specific piece of commonsense knowledge or truism (an undoubted or self-evident truth) should not be dependent on the way it is phrased. For example, consider the truism "Objects cannot be contained in containers smaller than themselves", the statements "A is larger than B, so *A cannot be contained in B*" and "A is larger than B, so *B can fit into A*" should both be correct and follow the truism despite the second half is phrased differently. If LMs can only make correct predictions if we mask "larger" in one of the sentences but not the other, then LMs are still lacking the ability to understand this commonsense truism and are just picking up statistical patterns in the training data. Besides, the reasoning ability should also be isolated from the entity information that is likely to be learned from training. For example, instead of testing statements like "A car is larger than a box, so a car cannot fit into a box", we will replace the real world entities with out-of-vocabulary (oov) ones like "uorijnat" and "okuzen" that are fictitious so that LMs cannot depend on the entity information to "cheat".

This work aims to answer the question of whether LMs can actually perform commonsense reasoning or they are just following patterns in the training corpora. We will first provide an extensive analysis using the state-of-the-art LM on the simple task of masked word prediction. In order to test if LMs can make the correct predictions regardless of how the truisms are expressed, we make small perturbations of the statement that preserve the semantics, like changing the word ordering or

using synonyms. We will provide a dataset with fictitious entities testing different types of commonsense truisms containing the perturbations. If LMs' performances are not consistent on the dataset, then we argue that LMs are not equipped with the ability to reason about commonsense. As a next step, we also aim to augment the LMs so that they can perform more robustly under perturbations.

In summary, the contributions of our work are three-folds: 1. We answer the crucial question of whether LMs can solve commonsense tasks due to the capability of reasoning or just shallow pattern matching has not been addressed. 2. By isolating reasoning from entity information and perturb commonsense truisms under the same semantics, we create a dataset that can be used by future work to test the ability for a model to understand commonsense truisms, which can be further transferred to different datasets. 3. We construct a dataset of commonsense truisms with perturbations to test the robustness of current models.

2 Related Work

2.1 Leveraging LMs for Commonsense Reasoning

Researchers have proposed models to leverage LMs to directly help solve commonsense benchmarks, specifically Winograd Schema Challenge (WSC). WSC proposes a coreference resolution task that requires commonsense reasoning. The datasets provides a sentence with a pronoun, and asks the machine to find the right candidate for the pronouns from two options. Trinh and Le (2018)'s method is very simple. They first substitute the pronoun in the original sentence with each of the candidate choices. The problem of coreference resolution then reduces to identifying which substitution results in a more probable sentence. They then use an LM to score the resulting two substitutions. They find that an ensemble of LMs trained on large text corpora ourperform previous methods using knowledge bases (KB) which are a lot more complicated.

Kocijan et al. (2019) extend the previous work by fine-tuning BERT (Devlin et al., 2019) on Winograd-like datasets and get even better results. One of the training objectives of BERT is masked word prediction and they utilize this fact by masking the pronoun in WSC and ask BERT to predict the right word. To get more data for fine-tuning, they generate Winograd-like datasets from Wikipedia. Results show that they can improve upon the previous SOTA methods by around 8%.

These methods utilizing LMs are conceptually very simple, and they already yield better results on WSC. This shows that LMs, especially latest ones that are trained on huge corpora (RoBERTa already gets around 89%) (Liu et al., 2019).

2.2 Probing LMs of Exploitation of Statistical Cues

Some previous work has shown that LMs' performances on tasks like Nartural Language Inference (NLI), Argument Reasoning Comprehension (ARC), and identifying paraphrases are due to statistical cues from the dataset (Niven and Kao, 2019; McCoy et al., 2019; Zhang et al., 2019). Specifically, Niven and Kao (2019) probes BERT on the task of argument mining. They find statistical cues by defining *applicability, productivity, and coverage*, and they find not to be one of the most important cues for this task. They also create an adversarial set where they negate the statement and invert the label, and find BERTs performance drops significantly.

McCoy et al. (2019) focus on the task of NLI. They define three syntactic heuristics exploited by the models and construct a dataset named HANS based on them. For each heuristic, they generate five templates for examples that support the heuristic and five templates for examples that contradict it. And then they augment LMs with HANS and find that it helps to make the model more robust. Zhang et al. (2019) provides a challenging paraphrase identification dataset PAWS. Challenging sentence pairs are generated by controlled word swapping and back translation, followed by fluency and paraphrase judgments by human raters. They find that existing state-of-the-art models fail miserably on PAWS when trained on existing resources, but some perform well when given PAWS training examples.

2.3 Probing LMs of Commoonsense Knowledge

Very recently, there are several papers proposing methods to examine that whether pre-trained LMs have commonsense knowledge. Kwon et al. (2019) introduce a new knowledge probing test designed to analyze whether the LMs understand structured common sense knowledge as semantic triples in an external repository specifically ConceptNet (Liu and Singh, 2004). They generate sentences through predefined predicate patterns. For example, a pred-

icate pattern of the "Antonym" relation can be "s and o are opposite ." The predicate patterns of relations are collected from the Open Mind Common Sense (OMCS) dataset (Singh et al., 2002). Our approach is different from theirs since we are testing different types of commonsense truisms instead of relations from ConcepNet.

Zhou et al. (2019) study the commonsense ability of several LMs by testing them on seven challenging benchmarks, finding that language modeling and its variants are effective objectives for promoting models commonsense ability while bi-directional context and larger training set are bonuses. They additionally find that current models do poorly on tasks require more necessary inference steps. Finally, they test the robustness of models by making dual test cases, which are correlated so that the correct prediction of one sample should lead to correct prediction of the other. Interestingly, the models show confusion on these test cases, which suggests that they learn commonsense at the surface rather than the deep level. Our work distinguishes from this paper by considering our own benchmark of truisms and better defining the perturbation types instead of simply "adding, deleting, replacing, or swapping" words in the test instances.

3 Analysis of LMs for Commonsense Truisms

In this section, we show analysis of LMs' capabilities to reason about commonsense truisms. We will first describe our task to test the LMs and use an example truism to demonstrate the perturbations. Then we show how we construct our dataset of three different types of commonsense truisms with perturbations. Lastly, we will show the results.

3.1 Perturbed Masked Word Prediction

Masked word prediction is the most important training objective in pre-training LMs like BERT and RoBERTa. Given a sentence, random words in it are masked and the LMs have to predict the masked words. It is considered a bidirectional training of language understanding. Here, in order to test whether LMs can reason about commonsense truisms or just learning frequent text patterns in the training corpora, we perturb the original truism statement under the same semantics and compare the performances for all perturbations. As an illustrative example, we will focus on the truism "Objects cannot fit into containers smaller than themselves". And we mainly consider five types of perturbations: asymmetry (applied to premise or conclusion), negation, antonym, paraphrasing, and any combination of the above.

3.1.1 Types of Perturbation

The perturbations types mentioned above can all be considered as syntax changes that aim to modify how the truisms are phrased without changing the semantics or the commonsense knowledge tested in the truism. In the following part, we will first demonstrate each type of the perturbation, and then show that how we combine several types of perturbations together.

1. Original Truism: "A is larger than B, so A cannot fit into B." This is the original template of the truism we consider. It is in the form of "premise, conclusion", since we want to give the LMs enough context to make the inference.

2. Asymmetric Premise: "B is larger than A, so A can fit into B." We swap A and B only in the premise (second half), and replace "cannot" with "can" to follow the right logic. The intuition is that it is likely that in the training corpora of LMs, the order of the entities in a sentence is the same for the first half and the second half. Thus the asymmetry of entities created by swapping only the first half or second half will make it unfamiliar to the models that only remember textual patterns. Humans can understand

3. Asymmetric Conclusion: "A is larger than B, so B can fit into A." We swap A and B only in the conclusion (second half), and replace "cannot" with "can" to follow the right logic. The intuition of this perturbation is the same as the previous one, except that we change the ordering in the second half.

4. Negation: "A is not larger than B, so A can fit into B." For the negation type, we negate the premise and flip "cannot" to "can" in the conclusion accordingly. Some previous work examining deep neural models for exploitation of statistical cues (McCoy et al., 2019) has found that negation can confuse models greatly.

5. Antonym: "A is smaller than B, so A can fit into B." We change the adjective in the premise to its opposite and replace "cannot" with "can" to follow the right logic. This is to test that whether LMs understand the relations between words and the semantic difference the change brings.

6. Paraphrasing without Inversion: "A is larger than B, so A cannot be put into B." For paraphrasing, we consider two sub-cases, one that the logic key word like "can" needs not to be flipped and one that it needs to be flipped to make the logic correct and we call them "paraphrasing without inversion" or "paraphrasing with inversion", accordingly. In this specific example, the conclusion part (second half) is paraphrased from "fit into" to "be put into", but the "cannot" is kept same.

7. Paraphrasing with Inversion: "A is larger than B, so A can contain B." In this specific example, the conclusion part (second half) is paraphrased from "fit into" to "contain", and the "cannot" is flipped to "can", so we call this type as "with inversion"

8. Asymmetric Premise + Negation: "B is not larger than A, so A cannot fit into B." As mentioned earlier, we not only consider the single-type perturbations, but we also combine two or three single types together. Here we show an example combining "asymmetric premise" with "negation", where we first swap the order of the entities in the premise, and then negate the comparative adjective. Note that the logic needs not to be flipped. We will omit other combinations of two single types of perturbations.

9. Asymmetric conclusion + Antonym + Paraphrasing with Inversion: "A is smaller than A, so B can contain A." Here we show an example combining 3 single types: "asymmetric conclusion", "antonym", and "paraphrasing with inversion", where we first swap the order of the entities in the conclusion, swap "larger" to its opposite word "smaller", and then paraphrase "be put into" using "contain". We will omit other combinations of three single types of perturbations.

3.1.2 Scoring Metrics

In this part, we will show two settings for our experiments that test commonsense reasoning abilities of LMs.

Binary Setting We first evaluate by simply comparing the rankings of the masked word and the other candidate. For example, in the truism: "A is larger than B, so A cannot fit into B." We can mask "cannot" and treat "can" as the wrong answer. In the binary setting, we feed the masked sentence to the LMs and give the model score 1 if the right answer appears higher than the wrong answer, and 0 otherwise. This setting tests only the rankings of predicted words without taking into account of the scores, which makes the results not influenced by outlier words that have an extremely high score, but also omit some nuances.

Ratio Setting To address the nuances in the ranking scores of predicted words, we further propose a metric called ratio setting. Using the example above, we denote the predicted score for "cannot" as $score_{right}$ and that for "can" as $score_{wrong}$. Then we calculate our final score using: $(score_{right} - score_{wrong})/score_{right}$. The more positive the final score is, the better the performance according to this metric, and vice versa. This setting can take into account the score difference given by the models, but it can also make the results more easily influenced by outlier words that have an extremely high score or low one.

3.2 Dataset Construction

Here we present how we construct commonsense truisms with perturbations mentioned before to test whether LMs can reason about commonsense or just learning textual patterns in the training corpora. We pose two constraints in the templates of the truisms and their perturbations. First, all truisms in our dataset are in the "premise, conclusion" style, since we want to give LMs enough context to reason about the fictitious entities that will be filled in the templates. Second, all templates involve comparisons between two objects or two people so that we can mask the comparative word like "larger" or "smaller" and evaluate the LMs based on which of these two opposite words is ranked higher.

We consider three types of commonsense knowledge: physical constraints, material properties, and social interactions. Truisms about physical constraints focus on testing the relationships between

certain attributes of objects and their implications in our physical world. The truism of "A is larger than B, so A cannot fit into B." mentioned before is of this type. Besides length, we also consider other attributes like hardness, heaviness, etc.. Truisms about material properties aim to test if LMs understand that each material has specific properties like "wool is softer than metal". An example of the truism is "A is made of wool and B is made of metal, so A is softer than B.". For truisms about social interactions, we want them to test commonsense knowledge about people in social lives. An example is "A won the competition and B lost, so A is happier than B.".

We ask humans to generate templates of truisms for each type with perturbation described in the previous subsection and we manually examine each truism to make sure it is correct and obvious for humans. For each category of commonsense knowledge, we collect around 20 truisms testing different knowledge in the category. Then for each truism, we have around 16 different perturbations. This results in 320 truisms for each type of knowledge and 960 in total. Lastly we randomly generate 100 fictitious words that range from 3 to 12 characters to fill in the templates at the locations of "A" and "B" , resulting in 96000 commonsense truism with entities.

3.3 Results

Here we show results of testing a state-of-the-art LM called RoBERTa on the task of perturbed masked word prediction using the commonsense truisms we construct. We present several different visualizations of our results to discuss our findings. We will present and discuss the results using the two settings that we introduced separately.

3.3.1 Binary Setting

Perturbations greatly affect the performances In this part we use a slightly different terminology to indicate the perturbation types. We use index from 1 to 8, where 1 to 4 are order changes (asymmetry of entities), and 5 to 8 are combinations of antonym perturbation with order change.

Figure 1 shows the experimental results on RoBERTa for the "A is larger than B, so A cannot fit into B" truism. We can see that RoBERTa performs perfectly with order change type 1 and 2, regardless of the adjective swap. However, for order change type 3 and 4, which is changing the order of only the second half of the truism, it can make zero right predictions, exactly opposite of the other syntax perturbation types. This demonstrates that the performance of the LMs for commonsense reasoning heavily depends on the way the truism is phrased. And in this case, the order change of the second half of the truism may be a lot less stated in the corpora, making the LMs confused when the truism is expressed this way.

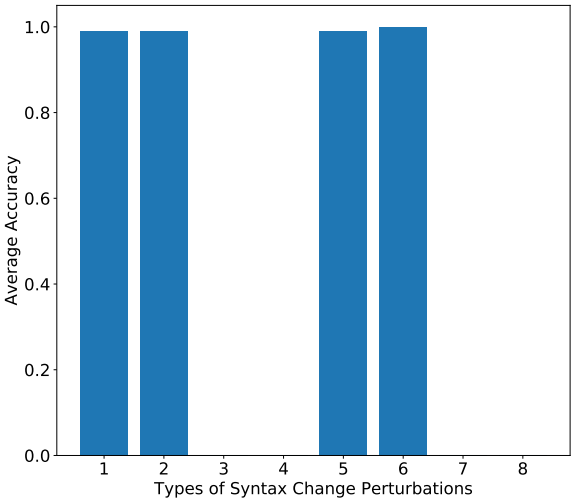

Figure 1: Results of average accuracy (y-axis) of the "A is larger than B, so A cannot fit into B" truism with different types of perturbations (x-axis) that match the index of the syntax change perturbations.

We also find that changing the lexicon influences the results a lot. As Figure 2 shows, the only change we make about the truism is using "contain" instead of "fit into" and replace the "cannot" with "can" to make the logic right. The results are totally reversed, where the order change types 1 and 2 make the performance of the LMs close to zero and types 3 and 4 will produce almost perfect predictions. This shows that the lexical used also affects RoBERTa's performance on commonsense reasoning a lot.

Inconsistency exists for all types of commonsense Since we have three different categories of commonsense truisms constructed, we also want to test if the inconsistency illustrated above also applies to other types of knowledge.

Figure 3 shows the average of the largest difference in accuracy across perturbations for the three types of commonsense knowledge we considered. We can see that all of them are close to one, indicating that on average, RoBERTa performs very inconsistently under perturbations for all the commonsense truisms.

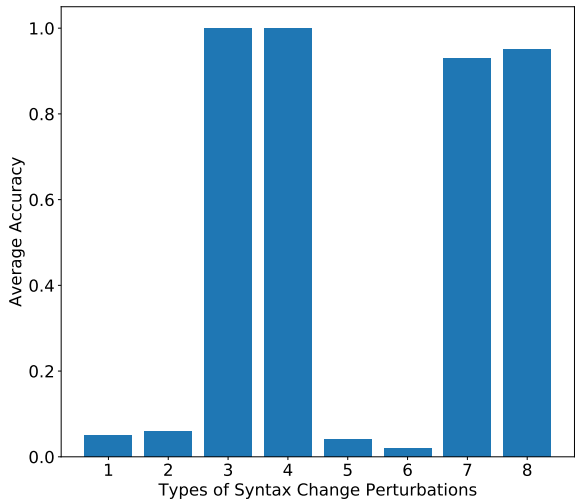

Figure 2: Results of average accuracy (y-axis) of the "A is larger than B, so A can contain B" truism with different types of perturbations (x-axis) that match the index of the syntax change perturbations. The only difference from Figure 1 is that we change the lexicon used, but the meaning is preserved.

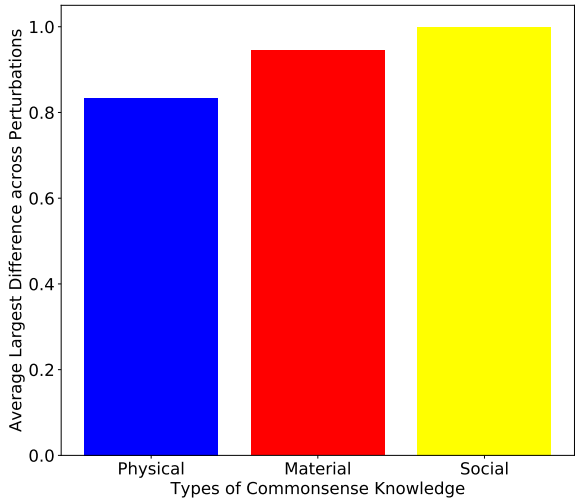

Figure 3: Results of average largest difference across perturbations (y-axis) of the constructed truism with different types of commonsense knowledge (x-axis). We can see that all average differences are close to 1, which means that the performances of RoBERTa are extremely inconsistent under perturbations.

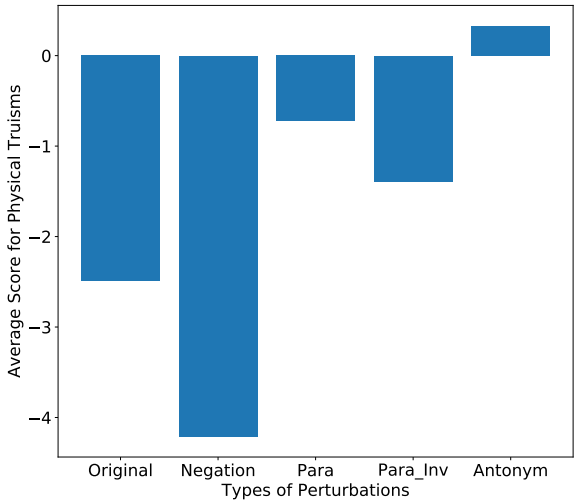

Figure 4: Results of average ratio score (y-axis) of the physical truisms with different types of perturbations (x-axis).

3.3.2 Ratio Setting

In this part, we will present and discuss our experimental results using the ratio score metric. We use the same truisms we collected and RoBERTa LM.

The average performance is very poor As mentioned earlier, the ratio scoring metric, although takes into account of more nuanced information, can be influenced by outliers very easily. For example, if in some perturbed truisms, the ratio is a large negative number due to that the score of the wrong answer is a lot higher than that of the right answer, then this instance will lower the average ratio score for truisms by a large margin. However, this still shows that LMs fail on some cases or some truisms miserably since it makes a confident wrong choice and this metric penalizes that.

Figure 4 shows the average ratio score for several single-type perturbations on our physical truisms. We can see that except for "antonym" type, all others have a negative average score. This shows that on average, the ratio score indicates that LMs perform very poorly on collected truisms with perturbations. The antonym being the only positive may be due to that LMs can understand the semantics of some adjectives but not their opposite, possibly because of the frequency in the training corpora.

Similarly, Figure 5 shows that the average performance when applying the asymmetry perturbations also is very poor for physical truisms. We can also see that if we change the order of entities in the premise, the influence is higher. Interestingly, the

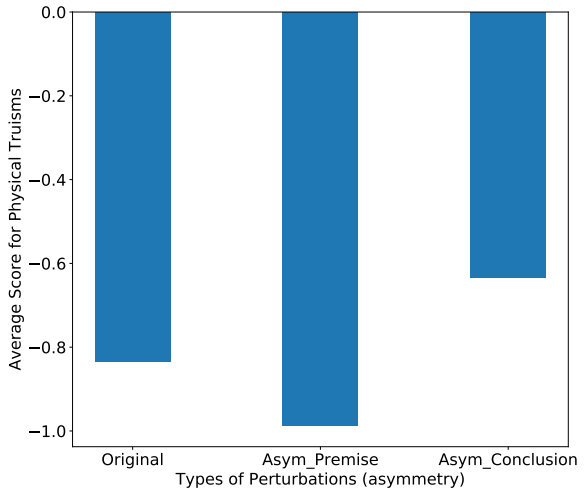

Figure 5: Results of average ratio score (y-axis) of the physical truisms with different types of perturbations (x-axis).

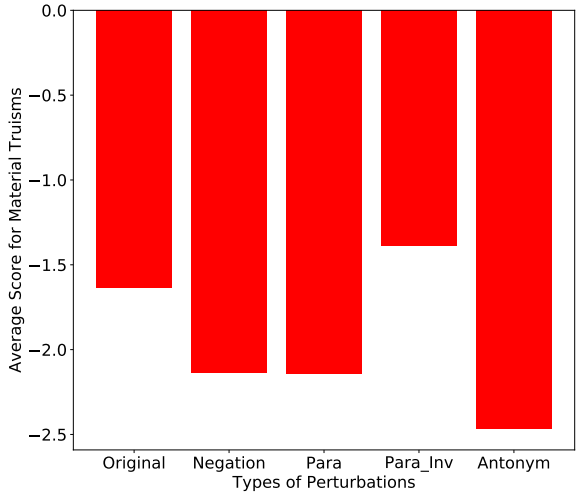

Figure 6: Results of average ratio score (y-axis) of the material truisms with different types of perturbations (x-axis).

original has a worse performance than when we apply asymmetry on conclusion for physical truisms.

LMs perform poorly on different types of commonsense We also present experimental results on material and social truisms.

Figures 6 and 7 show the results of single perturbation types as well as asymmetry perturbations for material truisms in our dataset. Again, we can see that all of them are well below zero, meaning that on average the poor performances of LMs on truisms with perturbations also exist for material-related commonsense truisms.

Finally, we also present results for social truisms shown in Figures 8 and 9, which strengthen

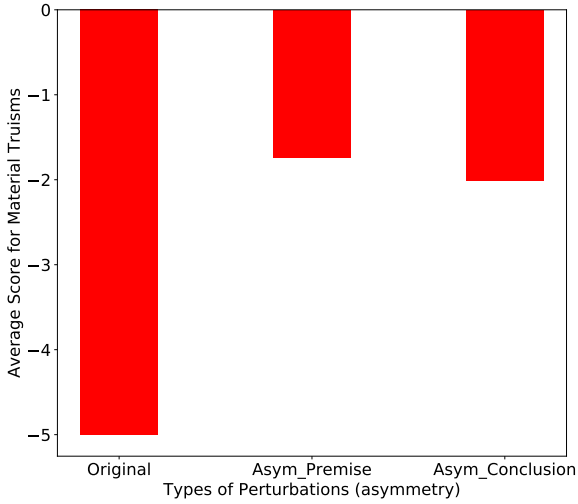

Figure 7: Results of average ratio score (y-axis) of the material truisms with different types of perturbations (x-axis).

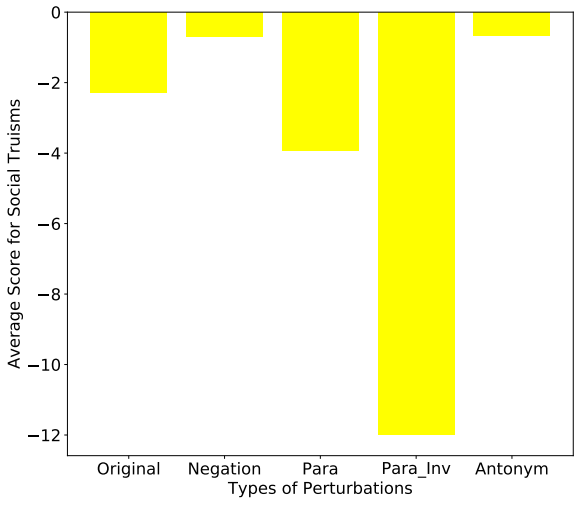

Figure 8: Results of average ratio score (y-axis) of the socail truisms with different types of perturbations (x-axis).

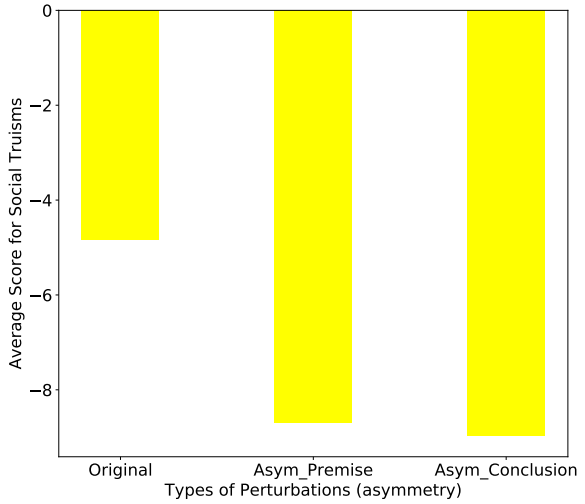

Figure 9: Results of average ratio score (y-axis) of the social truisms with different types of perturbations (x-axis).

our observation and shows that LMs can hardly understand truisms with perturbations, and are not robust.

4 Conclusion

In summary, we have conducted detailed survey on using LMs for commonsense reasoning and probing LMs for exploitation of statistical cues and commonsense knowledge, constructed a dataset of commonsense truisms with well-defined types of perturbations, provided analysis based on experimental results using our dataset and RoBERTa as the state-of-the-art LM. We find that in the binary setting, perturbations greatly affect the performances of the LM and the inconsistency exists for different types of truisms (physical, material, and social). Under the ratio setting, we show that the average performance of RoBERTa using our score metric is very poor regardless of the domain of commonsense knowledge, indicating that the there is still much improvement of LMs to understand commonsense.

References

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

Vid Kocijan, Ana-Maria Cretu, Oana-Maria Camburu, Yordan Yordanov, and Thomas Lukasiewicz. 2019. A surprisingly robust trick for winograd schema challenge. *arXiv preprint arXiv:1905.06290*.

Sunjae Kwon, Cheongwoong Kang, Jiyeon Han, and Jaesik Choi. 2019. Why do masked neural language models still need common sense knowledge? *arXiv preprint arXiv:1911.03024*.

Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*.

Hugo Liu and Push Singh. 2004. Conceptneta practical commonsense reasoning tool-kit. *BT technology journal*, 22(4):211–226.

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. *arXiv preprint arXiv:1907.11692*.

R Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference.

Timothy Niven and Hung-Yu Kao. 2019. Probing neural network comprehension of natural language arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4658–4664.

Push Singh, Thomas Lin, Erik T Mueller, Grace Lim, Travell Perkins, and Wan Li Zhu. 2002. Open mind common sense: Knowledge acquisition from the general public. In *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*, pages 1223–1237. Springer.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. 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 4149–4158.

Trieu H Trinh and Quoc V Le. 2018. A simple method for commonsense reasoning. *arXiv preprint arXiv:1806.02847*.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. Paws: Paraphrase adversaries from word scrambling. 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 1298–1308.

Xuhui Zhou, Yue Zhang, Leyang Cui, and Dandan Huang. 2019. Evaluating commonsense in pre-trained language models. *arXiv preprint arXiv:1911.11931*.