Commonsense Reasoning in Interpersonal Conflict

Ziyuan Cao

Department of Computer Science
Georgia Institute of Technology
4/17/2022

1 Introduction

People often made decision without active reasoning. For instance, if we want to leave our apartment, we leave through the door instead of the window without thinking about the consequence of stepping out of the window. In artificial intelligence research, that type of knowledge is called commonsense knowledge. Many prior research has focused on representing commonsense knowledge for computational use [20, 17]. Such representation can be utilized to endow language models the ability of commonsense reasoning. Recently, research has shown that large-scale pretrained models perform well on various NLP tasks [4, 14, 2]. Naturally, there has been research on probing the commonsense knowledge learned through pretrainig [3, 13]. In order to make NLP systems better fit into our social interaction, such systems need to gain more social commonsense knowledge [9]. Researchers proposed several benchmarks that require understanding of social situations and using social commonsense knowledge to make decisions or judgments to evaluate models' ability of social commonsense reasoning [18, 6, 11]. However, the social situations provided by those benchmarks are very simple. For instance, one situations in the benchmark SOCIAL CHEMISTRY is "asking my boyfriend to stop being friends with his ex" and one of the task is to give the social judgment to the situation. In those simple scenes, it is easier for the NLP models to capture the lexical correlation between the situations and the judgment so it is unclear if NLP models that perform well on those benchmark utilize social commonsense knowledge. To fill this gap, we propose to use the subreddit r/AmItheAsshole (r/AITA), a set of more complicated social situations, when conducting studies on social commonsense reasoning.

The subreddit r/AITA hosts more than 900,000 users' posts on their interpersonal conflicts with other people. As we show in Table 1, the posts are longer than the situations in [18, 6] and therefore contain richer background information. The comments usually contain users' judgments on if the poster is the one to blame in the interpersonal conflict. We test if morden NLP systems can predict the judgments of the subreddit users guiven the posts. To utilize existing social commonsense resources, we test whether intermediate finetuning can improve models' performance [16]. To get as much information from the text as possible, we utilize multi-task learning by finetuning BART [12] to jointly predict the judgment and the explanation given by the users.

Statistics	r/AITA	SocialChem [6]	SocialIQa [18]
Average	394	12	16
Median	382	11	15

Table 1: Number of words per situation

During processing the corpus, we identify that users' judgments are often not consistent regarding a given post. Additionally, across different post, the

divisiveness of users' judgments varies. Having recognized this variation across posts, we investigate how the divisiveness of judgments influences the difficulty in training models to predict the majority judgments. Even though users' judgments on some posts are consistent, we have to identify that the judgment of the users in r/AITA may not represent the most popular judgment in the society. In fact, the ethical relativism argues that people hold different moral rules due to differences of their cultures. We wish to investigate the moral bias of the trained models. We wish that this enable us to identify the moral bias shared by the community of r/AITA, if there is any consistent one. To achieve this goal, we propose to apply the adversarial attack technique on the judgment prediction model to generate a modified posts which only differs minimally from the original story but judged in the opposite way by the model. We call this pair of posts that differ only minimally in one text segment a minimal pair.

2 Related Work

2.1 Social Norm and Commonsense

Several previous works have focused on collecting a corpus to test the ability of NLP models of social commonsense reasoning and endow social norms to NLP models. In the setting of SocialIQa [18], the NLP system is given a social situation and a question and it needs to choose the correct answer to the question. The authors classify the questions into 6 types: wants, reactions, descriptions, motivations, needs, and effects. Answering the questions requires doing inferential reasoning of commonsense knowledge of those 6 categories. On the other hand, SocialChem [6] focuses on endowing social norms to NLP models. The dataset they proposed is composed of rule-of-thumbs (social norms), each accompanied with properties from 12 different dimension of people's judgments. The setting of their task is that, given a situation, the model will generate the relevant rule-of-thumbs and the properties of the rule-of-thumbs. They showed that state-of-the-art neural models can model the social norms well. Note a portion of SocialChem's situations comes from r/AITA. However, it only uses the titles but not the detailed descriptions of the posts. Scruples[15] also use r/AITA as their main corpus but the focus of Scruples is to predict the optimal performance of any classifier on r/AITA.

The problem of prior work, such as SocialIQa [18] and SocialChem [6], is that the social situations in those datasets are very simple and often composed of one sentence. The lack of the necessity of complicated reasoning make it unclear if the models that perform well on those datasets really perform social commonsense reasoning. We fill this gap by testing models on more complicated scenes.

2.2 Adversarial Attack

There has been a plenty of work on generating adversarial examples for text input, following the work of adversarial attack in computer vision. Different from image inputs, the difficulties of generating adversarial examples for text are two-fold: text inputs are composed of discrete tokens and small arbitrary perturbations can be noticeable and may produce ungrammatical sentences or sentences with strange meanings. [1] tackled these difficulties by practicing word replacement where they constrain the candidate word by word embedding similarity and fitness of the candidate in the context. To search for a valid replacement strategy, they use the genetics algorithm. In another work, BAE (BERT-based adversarial examples), the authors followed a similar paradigm but used more advanced methods [7]. They used BERT-MLM to search for potential replacement/insertion that fit the context well and used a Universal Sentence Encoder to control the semantic similarity. However, those two methods only work

3 Data Preprocessing

The subreddit, r/AITA, hosts posts of users describing their interpersonal conflicts with others. A post is composed of a title and one or more paragraphs of selftext (description). For each post, there are comments of other reddit users. As per the guidelines of r/AITA, the comments should indicate the judgment of the commenters. We collected the posts and comments on the r/AITA up to February 2022 using the Pushshift API [1]. During preprocessing, we remove the deleted posts (with empty selftext) and their accompanying comments. Following [15], we use regular expressions to extract the judgment provided by the comments. We use a binary scheme for the judgments where we assign one of the two labels: (1) the poster is the one to blame and (2) the other person is the one to blame. We remove comments with ambiguous judgments and comments with judgments that do not fit in the binary scheme. We consider two features for each post: the number of comments and the proportion of the majority judgment. For the standard dataset that will be used for our most experiments, we only keep posts with more than 10 comments and with the proportion of the majority judgment greater than 90%.

4 Method

4.1 Automatic Classification of Judgments

One of our goals is to investigate the performance of pretrained language models on predicting the majority judgment of the community for a given post on an interpersonal conflict. To achieve this, we model this commonsense reasoning task as a binary classification task. Given the title and descriptive selftext

[1] `https://reddit-api.readthedocs.io/en/latest/`

of a post, we wish to predict whether the majority judgment rules that the poster is the one to blame or that that the other person is the one to blame. Formally, given $(T_i, S_i, y_i)_{i=1}^N$, the collection of tuples of titles, selftext, and majority judgments, the classification problem can be represented as minimizing the cross-entropy loss:

$$L = \sum_{i=1}^{N} y_i \log\left(p(y_i|T_i, S_i)\right) + (1 - y_i) \log(1 - p(y_i|T_i, S_i))).$$

We considered to use different subsets of the corpus for training and testing the classifiers to investigate the impacts of two features of a chosen subset on the performance of the classifiers. The two features are (1) the number of comments and (2) the proportion of the majority judgment. We hypothesize that the classifiers will perform better on the subsets of corpus with larger number of comments and with higher proportion of the majority judgment.

4.2 Intermediate Task Transfer Learning

Many corpora were constructed for endowing social commonsense knowledge to NLP models and for testing the social commonsense reasoning ability of NLP models [6, 18, 5]. To utilize social commonsense knowledge embedded in those corpora, we use the intermediate task finetuning. Prior studies have shown that this technique, finetuning a pretrained language model on relevant datasets before finetuning it on the target task, can improve the model's performance on the target task [16, 18]. We investigate whether this technique can improve the performance of our models on r/AITA. We first finetune the pretrained language model on SocialChem [6] and SocialIQa [18] before finetuning the model on the target task, r/AITA. We test the validation performance of the models on r/AITA when fine-tuning on those two corpus respectively.

4.3 Multi-task Learning

For each judgment, the user give his/her explanation alongside. To utilize as much information from the AITA corpus as possible, we propose to incorporate the textual explanations given by the users into the finetuning process of the classifiers. We achieve this by using the interpolation of multiple objective functions,

$$L = \alpha L_{\text{generation}} + (1 - \alpha) L_{\text{judgment}},$$

as the objective function to jointly train the model on predicting the judgments and on predicting the textual explanation of the judgments. We choose BART [12] as the base architecture and add a classification head on top of the hidden state of $\langle \text{eos} \rangle$. Prior studies have shown positive results of learning to classify and generate jointly [10].

4.4 Generating Minimal Pairs

We recognize that the social judgment given by the community of r/AITA may contain its unique biases. The classifiers trained on r/AITA then inevitably inherit the biases contained in the corpus generated by r/AITA. However, the biases in r/AITA may not be the same as the biases in other community. It is then necessary to develop a method to systematically investigate and compare the biases in different communities. We tackle this problem by a variation of adversarial attack methods.

We first present a formalism for adversarial attack on NLP models that is commonly used in previous studies [7, 1, 8]. Given a classifier $f : S \to C$, where S is the text input space and C is the class space, and an text input x with $f(x) = c$, the goal of adversarial attack is to find x_{adv} with $f(x_{adv}) \neq c$ and the difference between x_{adv} and x is not perceptible by humans. Existing approaches of adversarial attack search for variation of the original text by replacement, deletion, and insertion of tokens with imperceptibility constraints [7, 1, 8]. Usually, imperceptibility requires semantic similarity and syntactical soundness.

We then explain how adversarial attack can be applied to investigate the biases of a classifier. For illustration, consider that we want to investigate the biases of a classifier on the identity of the subject of the story. We then can use the adversarial examples generation algorithm to generate a pair of stories where only the subjects are different. If the classifier assigns different labels for the two stories, it suggests that the classifier use biases on the subject for prediction.

Suppose we want to investigate the biases on identity of people involved in the story. To pinpoint the identity involved in the story, we use a semantic role labeling model [2], based on [19], to extract all the spans containing ARG0 for some verbs. Based on the definition of semantic roles in PropBank, being an ARG0 is a good heuristic that a span contain the identity of an agent. Having those spans extracted, we follow the methods in [7]. We replace those spans with the mask token and collected the top-k predictions by RoBERTa-base for each mask. Then among those collected candidates for replacement, we search for one that changes the result of the classifier. Note that, considering the purpose of investigating biases, we don't need to force semantic similarity between the original text and the adversarial example.

Similarly, if we want to investigate the biases on other aspects of the situation (e.g., location), we can use other appropriate labeling models to extract the spans that contain the features in interest. Then we can apply the adversarial attack method that is constrained to modify those extracted spans.

[2] `https://demo.allennlp.org/semantic-role-labeling`

5 Experiments

We trained RoBERTa-base on different subsets of r/AITA with different number of comments and different proportion of the majority judgment. Each model is trained for 5 epochs with learning rate 1×10^{-5}. The results are in Table 2 and Table 3. In Table 2, the posts in all subsets have at least 10 comments and it shows the difference of models' performance when the constraints of proportion of major judgment are different. When creating those subsets, we make the size of each subset the same to avoid influence of the size of training data. Similarly, in Table 3, we control the proportion of the major judgment and vary the number of comments. Both of the two experiments show that the model perform worse when it is more difficult for humans to make the decision.

	Least major proportion	Accuracy	Precision	Recall
Least number of comments = 10	90%	0.86	0.86	0.87
	80%	0.82	0.83	0.82
	70%	0.79	0.82	0.79

Table 2: Performance of RoBERTa training and testing on different subsets when controlling the number of comments

	Least number of comments	Accuracy	Precision	Recall
Least major proportion = 90%	20	0.86	0.88	0.87
	10	0.83	0.84	0.83
	5	0.82	0.86	0.8

Table 3: Performance of RoBERTa training and testing on different subsets when controlling the agreement

To test how intermediate task transfer learning influence models' performance on r/AITA, we first finetuned the pretrained language models on SocialIQa and SocialChem. For the finetuning on SocialIQa, we follow the same task formulation as the author used: a multiple choice task [18]. For the finetuning on SocialChem, we formulate it as a classification task where the model predict the social judgment given the situation. For both SocialIQa and SocialChem, we finetuned the models for 3 epochs. We then trained the finetuned models on r/AITA for 5 epochs. The results of both BERT-base and RoBERTa-base is in Table 4. Note that, for both BERT-base and RoBERTa-base, finetuning on SocialChem produces comparable results as not finetuning. On the

	Accuracy	Precision	Recall
BERT-base	0.75	0.77	0.78
SocialIQa → BERT-base	0.73	0.74	0.78
SocialChem → BERT-base	0.75	0.77	0.78
RoBERTa-base	**0.79**	**0.82**	**0.79**
SocialIQa → RoBERTa-base	0.77	0.79	0.79
SocialChem → RoBERTa-base	0.78	0.81	0.78
BART-Joint-Generation-Classification	0.76	0.81	0.73

Table 4: Results of intermediate task transfer learning on r/AITA with **70%** agreement and **10** comments

other hand, finetuning on SocialIQa produces worse results. One reason for the better performance of finetuning on SocialChem compared with SocialIQa is that the domain of SocialChem is more similar to that of r/AITA where both of them are about social norm and commonsense. In general, it is worth investigating why practicing intermediate finetuning on both datasets does not produce better performance.

To test how jointly learning to generate and classify perform, we also train the BART model mentioned in Section 4.3 on the same subset as we have tested the intermediate finetuning methods on. The result is in Table 4. Note that it outperforms BERT-base but it does not reach the performance of RoBERTa-base. This shows that the model does benefit from the multi-task learning scheme.

To investigate the types of bias that the algorithm in Section 4.4 can find out, we apply the algorithm to 16 situations in the corpus. Here we describe the insight behind one adversarial example. The situation is between a dad and his stepchild where the comments judge that the dad is the one to blame. Here's one sentence from the situation:

> She's 19, almost 20, and I have three sons aged 18, 16 and 15.

When the algorithm changed it to

> She's 19, almost 20, and they have three sons aged 18, 16 and 15,

the model's prediction changes. One explanation is that this modification implies that this situation involves another parent and that changes the model's prediction.

6 Conclusion

We propose to use corpora with more complicated situations such as r/AITA to study models' performance on social commonsense reasoning. We collected

the posts and comments on r/AITA up to February 2022. We formulated the social commonsense reasoning task as a binary classification task and evaluated the performance of RoBERTa-base on different subsets of the corpus. The results show that the model perform worse on the situations that are harder for humans to reach an agreement. We tested whether intermediate finetuning and multi-task learning can improve the performance on the hard subset. The results show that intermediate finetuning does not help for both BERT-base and RoBERTa-base. Multi-task learning outperforms vanilla BERT-base but does not reach the performance of vanilla RoBERTa-base. Lastly, we develop an approach for investigating biases of models using adversarial attack. To make this approach more powerful, adversarial attack algorithms that can practice action beyond replacement needs to be developed. Then we can investigate biases beyond identity biases, such as biases on actions, where the practice of generation/decoding is needed during the attack.

References

[1] Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. Generating natural language adversarial examples. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2890–2896, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[2] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[3] Joe Davison, Joshua Feldman, and Alexander Rush. Commonsense knowledge mining from pretrained models. 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 1173–1178, Hong Kong, China, November 2019. Association for Computational Linguistics.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, June 2019. Association for Computational Linguistics.

[5] Denis Emelin, Ronan Le Bras, Jena D. Hwang, Maxwell Forbes, and Yejin Choi. Moral stories: Situated reasoning about norms, intents, actions, and their consequences. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 698–718, Online and Punta

Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[6] Maxwell Forbes, Jena D. Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 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, November 2020. Association for Computational Linguistics.

[7] Siddhant Garg and Goutham Ramakrishnan. BAE: BERT-based adversarial examples for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6174–6181, Online, November 2020. Association for Computational Linguistics.

[8] Chuan Guo, Alexandre Sablayrolles, Hervé Jégou, and Douwe Kiela. Gradient-based adversarial attacks against text transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5747–5757, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[9] Dirk Hovy and Diyi Yang. The importance of modeling social factors of language: Theory and practice. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 588–602, Online, June 2021. Association for Computational Linguistics.

[10] Tatsuya Ide and Daisuke Kawahara. Multi-task learning of generation and classification for emotion-aware dialogue response generation. *arXiv preprint arXiv:2105.11696*, 2021.

[11] Liwei Jiang, Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Maxwell Forbes, Jon Borchardt, Jenny Liang, Oren Etzioni, Maarten Sap, and Yejin Choi. Delphi: Towards machine ethics and norms. *arXiv preprint arXiv:2110.07574*, 2021.

[12] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online, July 2020. Association for Computational Linguistics.

[13] Bill Yuchen Lin, Seyeon Lee, Rahul Khanna, and Xiang Ren. Birds have four legs?! NumerSense: Probing Numerical Commonsense Knowledge of Pre-Trained Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6862–6868, Online, November 2020. Association for Computational Linguistics.

[14] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[15] Nicholas Lourie, Ronan Le Bras, and Yejin Choi. Scruples: A corpus of community ethical judgments on 32, 000 real-life anecdotes. *arXiv preprint arXiv:2008.09094*, 2020.

[16] Yada Pruksachatkun, Jason Phang, Haokun Liu, Phu Mon Htut, Xiaoyi Zhang, Richard Yuanzhe Pang, Clara Vania, Katharina Kann, and Samuel R. Bowman. Intermediate-task transfer learning with pretrained language models: When and why does it work? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5231–5247, Online, July 2020. Association for Computational Linguistics.

[17] Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3027–3035, 2019.

[18] Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 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, November 2019. Association for Computational Linguistics.

[19] Peng Shi and Jimmy Lin. Simple bert models for relation extraction and semantic role labeling. *arXiv preprint arXiv:1904.05255*, 2019.

[20] Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-first AAAI conference on artificial intelligence*, 2017.