Team Solomon at SemEval-2020 Task 4: Be Reasonable: Exploiting large-scale language models for commonsense reasoning

Vertika Srivastava* **Sudeep Kumar Sahoo*** **Yeon Hyang Kim**
Rohit R.R **Mayank Raj** **Ajay Jaiswal**

Samsung R&D Institute India, Bangalore
{v.srivastava, sudeep.sahoo, purine.kim,
rohit.r.r, mayank.raj, ajay.jaiswal}@samsung.com

Abstract

In this paper, we present our submission for SemEval 2020 Task 4 - Commonsense Validation and Explanation (ComVE). The objective of this task was to develop a system that can differentiate statements that make sense from the ones that don't. ComVE comprises of three subtasks to challenge and test a system's capability in understanding commonsense knowledge from various dimensions. Commonsense reasoning is a challenging task in the domain of natural language understanding and systems augmented with it can improve performance in various other tasks such as reading comprehension, and inferencing.

We have developed a system that leverages commonsense knowledge from pretrained language models trained on huge corpus such as RoBERTa, GPT2, etc. Our proposed system validates the reasonability of a given statement against the backdrop of commonsense knowledge acquired by these models and generates a logical reason to support its decision. Our system ranked 2nd in subtask C with a BLEU score of 19.3, which by far is the most challenging subtask as it required systems to generate the rationale behind the choice of an unreasonable statement. In subtask A and B, we achieved 96% and 94% accuracy respectively standing at 4th position in both the subtasks.

1 Introduction

In today's digital age, information is shared widely in textual format via e-mails, news articles, social media posts and messages, blogs, internet forums, etc. We are now surrounded by textual information more than ever, which demands a meaningful understanding of a text by machines. Machines augmented with commonsense reasoning will be a key step towards achieving this. 'Commonsense Knowledge' also referred to as the background knowledge, is the understanding of the everyday world and the art of drawing inferences by manipulating the knowledge gathered. Humans are rational and have acquired a sense of reasoning by combining facts and beliefs from their day to day life. To an average person reasoning the fact that a person can have a pet dog but not pet dinosaur comes naturally and is fairly straightforward. Commonsense knowledge is assumed to be known to all and people typically tend to omit this while communicating with others, which makes it more challenging.

Current Natural Language Understanding (NLU) systems assisted with semantic representations, statistical methods, and distributional representations have shown better performance than humans on many benchmarks but there is a growing concern that these systems scratch only the surface of the human level of understanding of the world and thus are too shallow. Natural language is complex in nature and NLU systems have tried to derive useful meaning by capturing the context i.e. neighboring words and sentences but these systems fail miserably when the context is restricted or omitted. Such cases call for systems to delve deeper into understanding the background knowledge enjoyed by all humans. For a commonsense deprived machine, understanding that the sentence *'he put books in his pencil box'* is against common sense as *'a book is much bigger than a pencil box'* is difficult in the absence of the knowledge about the size of books in comparison with the size of a pencil box. Since the existing systems

*Equal Contribution

Proceedings of the 14th International Workshop on Semantic Evaluation, pages 585–593
Barcelona, Spain (Online), December 12, 2020.

Subtask A: Validation	S0: He put a turkey into the fridge. **S1: He put an elephant into the fridge.**
Subtask B: Explanation (Multi-Choice)	**Statement:** He put an elephant into the fridge. **Option A: An elephant is much bigger than a fridge.** Option B: Elephants are usually white while fridges are usually white. Option C: An elephant cannot eat a fridge.
Subtask C: Explanation (Generation)	**Statement:** He put an elephant into the fridge. Referential Reasons: 1. An elephant is much bigger than a fridge. 2. A fridge is much smaller than an elephant. 3. Most of the fridges aren't large enough to contain an elephant.

Table 1: Examples from the ComVE dataset.

either do not possess this knowledge or are rather weak in reasoning beyond the data provided to them. Through this task, we aim to empower machines to acquire this knowledge and perform better in many natural language tasks like question answering, fake news detection, etc.

ComVE comprises of three subtasks, subtask A is a validation task where a system has to choose the unreasonable statement between a pair of given statements. Under subtask B the participating system has to pick up the reason which explains the rationale behind the unreasonable statement selected in subtask A. Subtask C, which by far is the most challenging task and requires the system to generate the reason explaining the argument behind its choice in the first task. We have illustrated some samples from the dataset for all the three subtasks in Table 1.

Under this task, we have developed a system that leverages commonsense knowledge gained by pretrained language models from their huge training corpora. We have used models like BERT, RoBERTa, GPT2, etc. Our system seeks to utilize the language model's world knowledge and identify commonsense facts in the task-specific dataset with task-centric finetuning. The model with little finetuning and task-specific modifications such as transforming the input and adding a score comparator achieved significant gains on all the subtasks. Our system ranked 2nd[1] in subtask C (Explanation with Generation) with a BLEU score of 19.3. In subtask A and B, we achieved 96% and 94% accuracy respectively standing at 4th rank in both the subtasks.

This paper is organized as follows, in Section 2, we briefly review some of the popular works in the domain of commonsense. In Section 3, we describe the task and the dataset. Section 4 gives details of our system and individual setup for each subtask. In Section 5, we have dicussed the experiments and the results. We conclude the paper in Section 6.

2 Related work

In the recent past, there have been several lines of research focussing on commonsense reasoning. Multiple tasks and datasets have been proposed which tests machine's intelligence pivoting on commonsense, like The Winograd Schema Challenge (Levesque et al., 2012) which aims to resolve ambiguities arising out of pronouns, using commonsense. Mostafazdeh et al., (2016) released the ROCStories corpus and introduced a Story Cloze Test, in which a system is given a four-sentence 'context' and two alternative endings to the story, called the 'right ending' and the 'wrong ending'. This task challenges a system to understand the context and predict the correct ending. Another dataset, SWAG (Zellers et al., 2018) involves predicting the next scene given the current one to evaluate grounded commonsense inference. It's also a multiple-choice dataset with 4 possible continuations to the given description. Devlin et al., (2018) finetuned their $\text{BERT}_{\text{LARGE}}$ model on the SWAG dataset to beat the human performance and achieve

[1] Leaderboard link: `https://competitions.codalab.org/competitions/21080#results`

state-of-the-art results with 86.3% accuracy on the test set.

All of these works involve multiple-choice datasets, where one has to choose the right option without providing any justification as to why the system chose a particular option. This raises a concern in the system's capability in actually understanding the choice made. None of the above work inspects a direct understanding of commonsense by demanding a logical reason for the choice. Wang et al,. (2019) released a dataset, which requires a system to choose an unreasonable statement from a given pair and also predict the right reason behind its choice. They also utilized state-of-the-art language models (LM) like BERT for Sen-Making and Explanation task which are similar to subtask A and B respectively of ComVE. However, they have reported a decline in performance on finetuning BERT. They achieved the best result with finetuned ELMO in the Sen-Making task and with pretrained BERT in the Explanation task.

Rajani et al., (2019) developed a Commonsense Auto-Generated Explanation (CAGE) framework for commonsenseQA task (Talmor et al., 2018). CAGE in its main approach of 'Reasoning' finetunes a language model conditioned on the question and all the plausible answer choices. The language model utilized Common Sense Explanations (CoS-E) as a referential reason while training. CoS-E is a manually constructed dataset comprising of reasons given by users who in turn were provided the question, all answer choices, and the correct label. LM trained via this approach was used to augment a classifier to predict the answer for the multiple-choice question posed initially. They leveraged CoS-E to assist in predicting the right answer. We have taken inspiration from their work in our subtask C to generate a logical reason as to why the unreasonable statement is against commonsense.

Subtask A

Dataset	Label-0	Label-1	Total
Train	4979	5021	10000
Dev	518	479	997

Subtask B

Dataset	A	B	C	Total
Train	3195	3362	3443	10000
Dev	344	327	326	997

Table 2: Dataset decription for Subtasks A and B.

3 Task and Dataset Description

The ComVE was formulated as a three-stage problem, where different subtasks assess a system's understanding of commonsense from a disparate perspective. The first subtask aims to empower the system to differentiate an unreasonable statement from a reasonable one and is proposed as a 'Validation' task. The next subtask, 'Explanation with Multiple-Choice' assess the system's capability to choose the right reason behind its choice for a particular statement to be unreasonable in the first subtask. The system's rationality is further tested by subtask C: 'Explanation with Generation', which expects the system to generate the reason explaining the rationality behind the system's choice of the irrational statement in subtask A. Table 1 shows an example of the dataset.

Subtask A is a two-class (or binary) classification problem, where a system has to choose from two natural language statements with similar wordings which one makes sense and which one doesn't. There are 10,000 sentence pairs in the training data, with each instance being labeled as either 0 or 1 depending on whether sentence 0 is unreasonable or sentence 1. The sentence pairs have been designed in a way that it is fairly easy for a human to pick the right statement but cannot be easily detected by commonsense deprived systems. Subtask B is a multi-class classification problem where a participating system has to pick the key reason from three options justifying why a given statement does not make sense. The training dataset for this subtask had 10,000 unreasonable sentences accompanied by three reasons for each. Dataset also had three noisy samples where just two options were provided. Subtask C is a text generation task where the objective is to generate a reason why the given statement is against common sense. The training dataset consists of 10000 unreasonable sentences and three referential reasons for each of them. We have provided the class-wise data distribution for the first two tasks in Table 2.

Dev data released for the evaluation phase had 997 samples for each subtask and hidden test data on which systems were finally evaluated had 1000 samples without labels. During the final evaluation, subtasks were kicked off sequentially with subtask A being opened first, followed by subtask C, and at the end, subtask B was started. This ensured that the there is no information leakage between the subtasks. More information on the tasks and the dataset can be found in Wang et al., (2020).

4 System Description

Our systems leverage commonsense knowledge from pretrained language models via transfer learning, thus we first briefly discuss the language models used in our system's core in Section 4.1. Subsequently, we explain the details of our models for each subtask. We have developed systems separately for each subtask which can be combined in the desired manner for an end to end commonsense pipeline.

4.1 Overview of Pretrained Language Models

BERT: Bidirectional Encoder Representations for Transformers, (Devlin et al., 2018) is a pretrained deep bidirectional transformer model producing context representations. It was trained on masked language modeling and the next sentence prediction objectives. BERT representations can be fine-tuned to many downstream NLP tasks by adding just one additional output layer for the target task, or it can be used as a feature for task-specific architectures. Using a fine-tuning setting, BERT has advanced state-of-the-art performances on a wide range of NLP tasks. We used pretrained BERT$_{\text{base-uncased}}$ with 110M parameters in our experiments.

ALBERT: Ian et al., (2019) introduced A Lite BERT (ALBERT) for learning language representations. It has two parameter reduction techniques that help it to increase the training speed and reduce memory consumption thus overcoming previous memory limitations of BERT. The authors have introduced the concept of parameter sharing across layers to prevent the growth in trainable parameters as the network's depth increases. They introduced a self-supervised loss for sentence order prediction in place of ineffective next sentence prediction of BERT. ALBERT$_{\text{large}}$ with 18M parameters was used in our experiments.

RoBERTa: Robustly Optimized BERT pre-training Approach (RoBERTa) (Liu et al., 2019) is an adaptation of BERT architecture trained with larger batches on 160 GB data from various domains. The paper mentioned that BERT was significantly undertrained and has the potential to outperform other transformer-based models with the right amount of data and design choices. RoBERTa was trained by dynamically modifying language masking while the next sentence prediction loss used in BERT was dropped. Other improvising techniques like larger input text sequences, byte pair encoding are used in training which seemingly improved the model performance in downstream tasks. It achieved state-of-the-art results in 4 of the 9 GLUE benchmark tasks during the time of publishing. For our experiments, we will be using RoBERTa$_{\text{large}}$ which has 355M hyperparameters.

GPT-2: Generative Pretrained Transformer 2 (Radford et al., 2019) is a large transformer-based language model trained on a dataset of 8 million web pages. GPT-2 is trained with a simple objective of predicting the next token, given all of the previous tokens within some text. This model shares the same architecture as GPT, with more than 10X the parameters and trained on more than 10X the amount of data. It displays a broad set of capabilities, including the ability to generate conditional and unconditional text samples of unprecedented quality. For downstream tasks involving text generation, the model performs better than all the other transformer-based language models. We will be using GPT-2$_{\text{large}}$ model with 762M parameters for our experiments.

4.2 Subtask A

For subtask A, we fine-tuned several task-specific pretrained BERT based classifiers, where input is a sentence and label is whether the sentence is unreasonable or not. Dataset for the classifiers was prepared by splitting a given input pair into two separate sentences and each of these was passed through the model to generate probability score for unreasonability. We have added a score comparison system on top of the model to predict the final label for the sentence pair by comparing the unreasonability score of both the sentences. Fine-tuning a BERT based model on our task consists of further training it on a task-specific dataset with masked language modeling (MLM) loss. We conducted our first experiments with a pretrained BERT model (BERT-CS). Systems with the same approach were implemented with

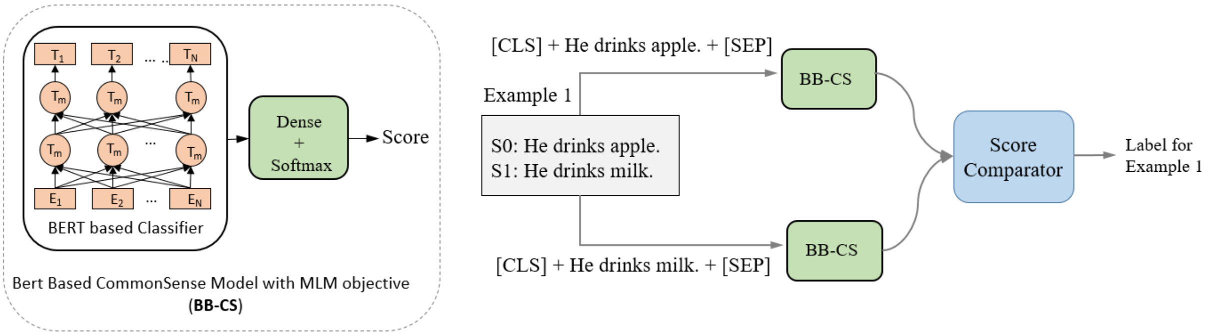

Figure 1: System for Subtask A

ALBERT (ALBERT-CS) and RoBERTa (RoBERTa-CS) architectures where RoBERTa based system seems most promising since it was trained on a large corpus and captures external knowledge convincingly.

4.3 Subtask B

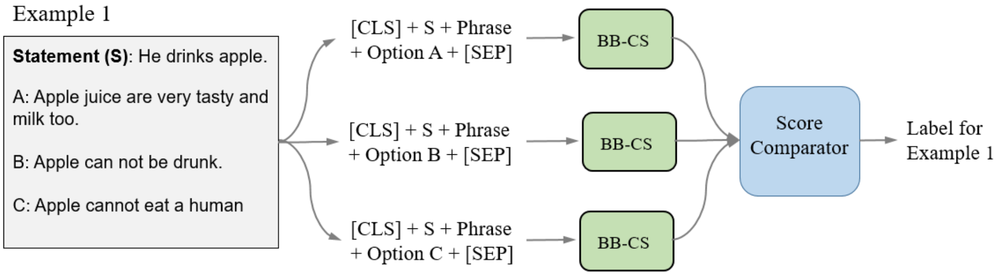
Example 1
Statement (S): He drinks apple.
A: Apple juice are very tasty and milk too.
B: Apple can not be drunk.
C: Apple cannot eat a human
[CLS] + S + Phrase + Option A + [SEP]
[CLS] + S + Phrase + Option B + [SEP]
[CLS] + S + Phrase + Option C + [SEP]
BB-CS
BB-CS
BB-CS
Score Comparator
Label for Example 1

Figure 2: System for Subtask B where, *Phrase* is a connecting negation phrase like, *"This does not makes sense because,"*, *S* is an unreasonable statement, and BB-CS is the same as described in Figure 1.

Subtask B is a multiple-choice task where a system has to identify the key reason to explain the irrationality of the given unreasonable statement (see Section 3). To achieve this, we have built a system that can understand the relation between a choice and the unreasonable statement, and also comprehend that the choice is justifying the logical reason behind it. To augment this further we have used *connecting negation phrases* between the choices and the unreasonable statement.

We formulated the given problem as a three-way binary classification task for each option. The dataset was transformed by constructing three input sequences per choice from the original sample. Each input sequence is a concatenation of the given unreasonable statement, connecting negation phrase, and one of the possible reasons. Connecting negation phrases like, *"This does not makes sense because,"* or *"No,"*, helps in constraining the model to learn a choice that explains the unreasonability of the statement. The system (see Figure 2) was developed by adding a task-specific layer on top of the pretrained models and finetuning them on task-specific data with MLM objective. We trained it on the modified input sequence as a binary classification problem with a softmax layer to produce a probability score for the sequence. An additional score comparator was used to merge the three binary classifiers. The score comparator analyzes scores from all three classifiers and arrives at the final prediction based on the maximum scoring sequence. BERT was used as a pretrained language model to develop BERT-Single and similarly, RoBERTa-Single was constructed with the RoBERTa language model, but our best performing system RoBERTa-Ens is an ensemble of 4 RoBERTa-Single based models with slight differences in their training.

4.4 Subtask C

As described in Section 3, the dataset consists of 10,000 unreasonable statements with three referential reasons each. In this subtask, we fine-tuned GPT-2 large model on the given training data and evaluated

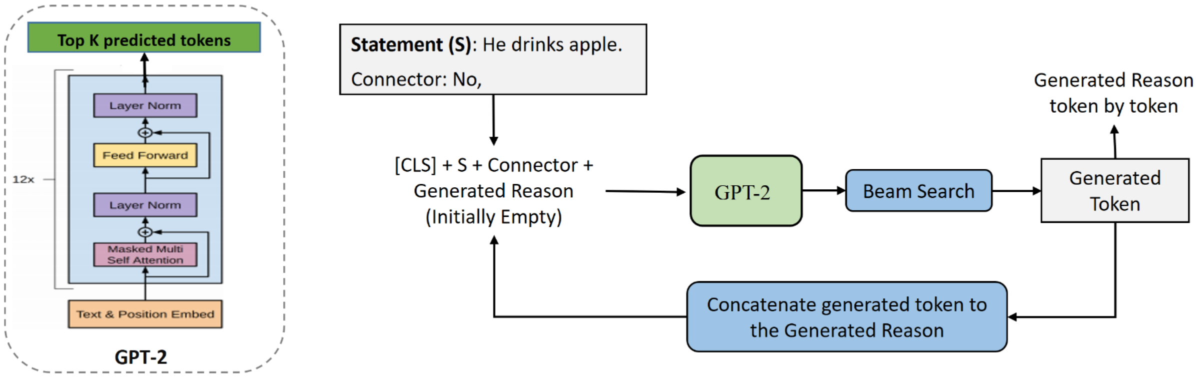
Top K predicted tokens
Layer Norm
Feed Forward
Layer Norm
Masked Multi Self Attention
12x
Text & Position Embed
GPT-2
Statement (S): He drinks apple.
Connector: No,
[CLS] + S + Connector + Generated Reason (Initially Empty)
GPT-2
Beam Search
Generated Token
Generated Reason token by token
Concatenate generated token to the Generated Reason

Figure 3: System for Subtask C

the system with the BLEU score. GPT-2 model has great capabilities to learn from raw text without the need for explicit labeling. Hence, this property of GPT-2 has been exploited here instead of using an encoder-decoder architecture to encode the unreasonable sentence and generate reason.

Each training row in the dataset has been converted to three separate samples based on the three referential reasons leading to 30,000 total input dataset size. The input to the system is fed sample wise instead of the original text chunk based training. Each sample is passed as *Unreasonable Statement* + *[No,]*+ *Referential Reason*. The model was fine-tuned on the cross-entropy loss to predict the next token at each step. We have used the beam search algorithm to generate the final output sequence instead of a greedy approach. The system outputs the probability score for each token in vocabulary to be the next element in the sequence. Following beam search, the top k sequences are separately appended with the input to generate the next token and this process is repeated till we reach the < |endoftext| > token. The generated reason is converted to lower case to avoid non-uniformity in the sentence structure leading to better match with the referential reasons and thus increasing the system's BLEU score.

5 Experiments and Evaluation

5.1 Subtask A

Table 3 shows the performance of the systems experimented for subtask A. From the results, we can observe that the best results are obtained using RoBERTa-CS with an impressive accuracy of 96% on the official test data. For RoBERTa-CS, we finetuned RoBERTa$_{\text{large}}$ pretrained model with a learning rate of 1e-5, dropout probability of 0.15, and a batch size of 32 for 7 epochs. BERT-CS and ALBERT-CS have achieved 87.7% and 81.1% accuracy respectively on the test data. Significant improvement of 8.3% by RoBERTa-CS over BERT-CS can be attributed to rigorous training and wider training dataset covering domains like news, stories, and Reddit of the pretrained RoBERTa. On samples where the model has to understand the irrationality with respect to time, duration or season, etc. (see Table 3), we found that RoBERTa-CS easily outperforms the others.

RoBERTa-CS was constructed with multiple task-specific components such as splitting our dataset sample and feeding them individually to the model and the score comparator at the top. To verify the necessity of these, we developed the RoBERTa-pairwise model. In the RoBERTa-pairwise, dataset format is kept intact and the input is fed as a sentence pair ([CLS] + Sentence 0 + [SEP] + Sentence 1 + [SEP]) with a softmax layer on top of the existing RoBERTa architecture and score comparator was completely shunned. It obtained 93.7% accuracy, recording a decline of 2.3% from RoBERTa-CS. Usually, BERT based architectures are trained on pairwise tasks by combining inputs with a separator, which tends to capture relations like: entailment, similarity, sentence order, etc. between the text pairs. Thus, when we fine-tuned a system based on the pairwise input as in RoBERTa-pairwise, some noise might have got added to the loss which has deteriorated the performance by struggling to capture a similar pairwise relation which is absent in the sample. On the other hand, an individual way of feeding input to the system captures the degree of reasonability of the statement and a simple comparison of the probability scores gives us the statement that is relatively more unreasonable.

We conducted an additional experiment by extracting final layer embeddings from the pretrained RoBERTa model and applied logistic regression on those as an input (RoBERTa-LR). This approach obtained an accuracy of 88.2% which is 7.8% less as compared to the best RoBERTa model. The results are in-line with the expectations as an end-to-end trained model learns details of the provided data while just applying a classification layer separately can't tune the embeddings to capture reasonability in a sentence. Yet, it has a minor gain of 0.5% over BERT-CS, which can be associated with better sentence representation being learned by RoBERTa's larger training corpus.

Model	Acc.
RoBERTa-CS	96.0
RoBERTa-pairwise	93.7
RoBERTa + LR	88.2
BERT-CS	87.7
ALBERT-CS	81.1

Samples	Predicted Label
S0: owls sleep at night S1: owls sleep at day	RoBERTa-CS: S0 BERT-CS: S1 ALBERT-CS: S1
S0: December is the 13th month of a year S1: December is the 12th month of a year	RoBERTa-CS: S0 BERT-CS: S1 ALBERT-CS: S1

Table 3: Result for Subtask A on the Test dataset (accuracy is in percentage). The second table shows a comparative analysis of the different models on some dataset samples.

5.2 Subtask B

RoBERTa-Ens is an ensemble of 4 RoBERTa-Single models trained on different connecting phrases, such as, *"No,"*, *", it is not true because"*, and *"This does not makes sense because,"*. One of the RoBERTa models for ensembling was trained on a new input sequence (*"Unreasonable Sentence:"* + Unreasonable statement + *"Reason:"* + One of the Reason Choice) which used a phrase in English to inject the task-specific information into the model. The ensemble is done by taking the average of probability from the 4 models for each given option and taking the option with the maximum score. The models are fine-tuned for 5 to 6 epochs with a learning rate of 1e-5, dropout of 0.1, batch size of 64, and 250 as the warmup steps for learning rate. It achieved an accuracy of 94% on the official test data as shown in Table 4. Among the RoBERTa-Single models, the system with a simple "No," connector performs the best with a 93.1% accuracy. Similar to subtask A, BERT-Single model performs subpar as compared to RoBERTa-Single with a drop of 10.6% accuracy.

Subtask B

Model	Accuracy (%)
RoBERTa-Ens	94.0
RoBERTa-Single + "No,"	93.1
BERT-Single	82.5

Subtask C

Model	BLEU Score
GPT-2 + BMS	19.34
GPT-2	16.92

Table 4: Results on the test dataset for Subtask B and C have been presented in the first and second table respectively.

5.3 Subtask C

In subtask C, GPT-2 model with beam search algorithm (GPT-2+BMS) was trained for a single epoch with all referential reasons of a statement line by line (instead of a block of text) with a batch size of 64 and a maximum sequence length of 128. The unreasonable sentence and one of the reasons are combined with a negating phrase, "No," while feeding to the system. The above-mentioned system in conjunction with k equal to 3 for beam search achieves a BLEU score of 19.34 and 2nd rank on the official leaderboard.

Replacing greedy search in place of beam search for reason generation decreases the BLEU score by 2.42 (see Table 4).

The GPT-2+BMS system can generate coherent reasons for the majority of the samples, yet in some cases, the system merely negated the input sentence to generate the reason as illustrated in Table 5. Training the model for larger epochs generates more complex and precise reasons at the cost of decreasing the BLEU score. This can be caused by the complexity of natural language where a reason can be explained in many ways and the ComVE dataset appeared to be containing very simple sentences. Furthermore, the BLEU score represents the precision of word order prediction and it doesn't take sentence structure, meaning, or recall into consideration which also justifies why the BLEU score dropped when the system generated coherent but complex reasons. In such cases, evaluating on BLEU along with the ROGUE score which captures recall of the generated sentence would have made more sense.

	Unreasonable Statement	**Generated Reason**
Good Examples	A soldier shot with a guitar.	A guitar is not a weapon.
	It is easy to see the stars on a clear day.	Stars appear in the night sky.
	She went to the grocery store to get an aneurysm.	An aneurysm is a medical condition.
	She eats a pillow after her workout.	Pillows are not edible.
Bad Examples	London goes in this bus.	A bus does not have wheels
	A television plays audio only.	A television does not play audio.

Table 5: In the above table we have shown some unreasonable statements from the dataset along with the reason generated by our best model (GPT-2+BMS).

6 Conclusion and Future Work

In this work, we have described our system for SemEval-2020 Task 4 on common sense validation and explanation. The proposed system leverages the background knowledge captured by large-scale transformer-based language models. This paper also discusses various ways of input manipulation in the architecture to improve the system's performance on the downstream tasks. Our official submission obtained an accuracy of 96% in subtask A and 94% in subtask B, both securing 4th position on the leaderboard. Our system also generates the rationale given an unreasonable sentence with a 19.34 BLEU score standing at 2nd rank on the leaderboard for subtask C. In future, we would like to train a joint model by combining systems for subtask A and C. This will provide an additional reasonable statement against the given unreasonable sentence to the system which can improve the reason generation capability.

Due to the limitation of language models in capturing external knowledge and their training being restricted by the dataset, the systems lagged behind humans in generating coherent reasoning for an unreasonable sentence. However, commonsense augmented systems can be incorporated in chatbots to create a more sensible conversation with a user. Also, it can help in detecting satirical articles in online news websites to combat the rising fake news problem.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

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

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*.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and evaluation framework for deeper understanding of commonsense stories. *arXiv preprint arXiv:1604.01696*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. *arXiv preprint arXiv:1906.02361*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2018. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *arXiv preprint arXiv:1811.00937*.

Cunxiang Wang, Shuailong Liang, Yue Zhang, Xiaonan Li, and Tian Gao. 2019. Does it make sense? and why? a pilot study for sense making and explanation. *arXiv preprint arXiv:1906.00363*.

Cunxiang Wang, Shuailong Liang, Yili Jin, Yilong Wang, Xiaodan Zhu, and Yue Zhang. 2020. SemEval-2020 task 4: Commonsense validation and explanation. In *Proceedings of The 14th International Workshop on Semantic Evaluation*. Association for Computational Linguistics.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. Swag: A large-scale adversarial dataset for grounded commonsense inference. *arXiv preprint arXiv:1808.05326*.