SWAGex at SemEval-2020 Task 4: Commonsense Explanation as Next Event Prediction

Wiem Ben Rim
School of Computing
Tokyo Institute of Technology
Tokyo, Japan
wiem.benrim@nlp.c.titech.ac.jp

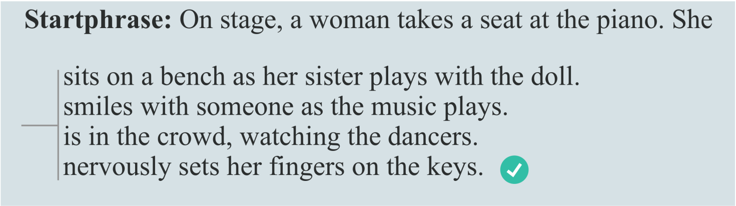
Naoaki Okazaki
School of Computing
Tokyo Institute of Technology
Tokyo, Japan
okazaki@c.titech.ac.jp

Abstract

We describe the system submitted by the SWAGex team to the SemEval-2020 Commonsense Validation and Explanation Task. We used multiple methods on the pre-trained language model BERT (Devlin et al., 2018) for tasks that require the system to recognize sentences against commonsense and justify the reasoning behind this decision. Our best performing model is BERT trained on SWAG fine-tuned for the task. We investigate the ability to transfer commonsense knowledge from SWAG to SemEval-2020 by training a model for the Explanation task with Next Event Prediction data.

1 Introduction

Commonsense reasoning has long been a challenge in the field of Natural Language Processing (NLP). Recently, the emergence of large pre-trained language models achieving state-of-the-art results in several NLP tasks, such as BERT and GPT (Devlin et al., 2018; Radford et al., 2019) has also influenced commonsense tasks to regain popularity. Natural Language Inference (NLI) is one of the most popular tasks in commonsense reasoning research. It consists of determining if a hypothesis is correct, given a premise. Situations With Adversarial Generation (SWAG) task is closely related to NLI (Zellers et al., 2018). It consists of determining the most appropriate ending given a start phase as a context. The start phrases and endings in SWAG dataset are video captions. We show an example in Figure 1. The task is to find the ending that describes a possible (future) world that can follow the given sentence even when it is not strictly entailed. Making such inference necessitates a rich understanding of everyday physical situations (Zellers et al., 2018).



Startphrase: On stage, a woman takes a seat at the piano. She

sits on a bench as her sister plays with the doll.
smiles with someone as the music plays.
is in the crowd, watching the dancers.
nervously sets her fingers on the keys. ✓

Figure 1: SWAG dataset example. The startphrase offers one sentence as context, and a beginning part of the second sentence. The task is to predict its appropriate ending.

In this paper, we explore the ability of SWAG to infer commonsense explanation to tackle Semeval 2020 Task 4: **Com**monsesense **V**alidation and **E**xplanation (**ComVE**). To this end, we followed the directions given by the organizers of the task (Wang et al., 2020) and attempted a contribution to their two main tasks.

The first main task is concerned with the commonsense validation of a sentence, which is also called *Sen-making*. It consists of providing two sentences that have similar wordings, one being against commonsense

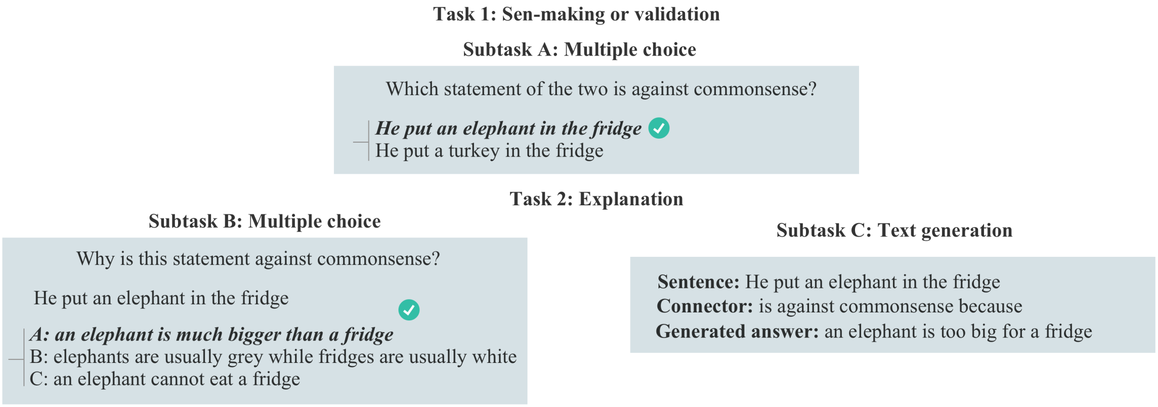
Proceedings of the 14th International Workshop on Semantic Evaluation, pages 422–429
Barcelona, Spain (Online), December 12, 2020.

and the other not. The sentences given do not require specialized knowledge and are trivial for humans, but remain challenging for machines. The second main task is Commonsense Explanation that aims at finding the key reason why a statement does not make sense. While the first task tests the system's performance in choosing which sentence is more likely to be true, the second one takes this a step further by providing us with the decision making process for commonsense reasoning.

We attempted several methods that we detail in later sections of the paper to tackle the tasks mentioned above. Our main contribution starts with the assumption that our Commonsense Explanation task is closely related to SWAG task. We tested this hypothesis with several experiments to confirm the usefulness of training a BERT model on SWAG, then testing it on the current dataset. We achieved nearly a 5.0% improvement using the above method, confirming our assumption that the same type of knowledge could be leveraged from Next Event Prediction in SWAG task to solve the ComVE task. For the Explanation task, our system currently ranks 14^{th} with an accuracy score of 84.6%. In post-evaluation, we improved our system and obtained an accuracy of 85.3%, also taking the 11^{th} place. We also demonstrated the method used to bridge the gap between the two datasets and explored the limitations of this approach with different layouts of data. Nevertheless, we encountered difficulties in the validation task, indicating that context sentences contribute highly to the efficacy of the model and questioning the world knowledge that can be encompassed in BERT. The extent of this is not explored in detail in this paper, but we consider it an exciting direction for future work. The pre-processing and model code we have used will be publicly available [1].

2 Background

For clarity, we follow the organizers in naming three subtasks. The *Sen-making* or Validation task is referred to as **subtask A**, and its aim is to find which sentence out of two given sentences is against commonsense. The Explanation task is to justify why a statement is against commonsense, and it is further divided into two subtasks. In **subtask B**, the system is given a sentence that is against commonsense *He put an elephant in the fridge*, along with three possible explanations. The system is expected to pick the explanation that is the most relevant and makes more sense; in this case, the correct answer would be *An elephant is much bigger than a fridge*. **Subtask C** is a text generation task that is concerned with answering the same question as to the previous subtask; given a statement that is against commonsense, the system should generate a corresponding answer. Figure 2 shows an example input for each subtask.



Task 1: Sen-making or validation

Subtask A: Multiple choice

Which statement of the two is against commonsense?

He put an elephant in the fridge ✓
He put a turkey in the fridge

Task 2: Explanation

Subtask B: Multiple choice

Why is this statement against commonsense?

He put an elephant in the fridge ✓

A: an elephant is much bigger than a fridge
B: elephants are usually grey while fridges are usually white
C: an elephant cannot eat a fridge

Subtask C: Text generation

Sentence: He put an elephant in the fridge
Connector: is against commonsense because
Generated answer: an elephant is too big for a fridge

Figure 2: SemEval-2020 task 4: Commonsense validation and explanation (ComVE).

Our main contribution targets subtask B. Nevertheless; we present our findings for the other subtasks. We trained BERT on SWAG, a dataset of 113k multiple-choice questions about a rich spectrum of grounded situations. We leveraged this knowledge to modify our system for the ComVE task. To implement this, we used BERT with a multiple-choice head on top, and trained it on a modified version of the SWAG

[1] `https://github.com/PhaelIshall/SemEval2020-Code`

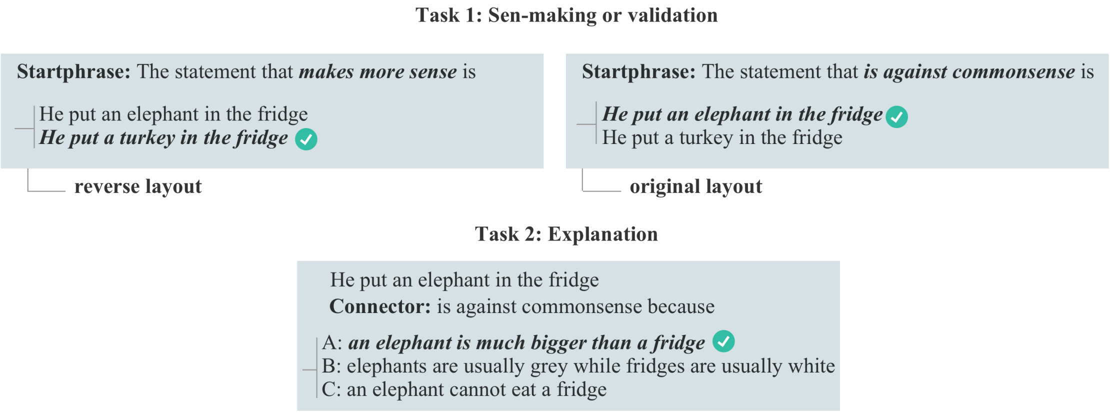
dataset; we changed the number of the given choices and the structure of the question for improved results. We then fine-tuned on our Explanation dataset, which consisted of 11k multiple-choice of sentences and possible explanations.

For subtask A, we gave our system a fixed question for all instances and two choices between a sentence against commonsense, and a sentence that agrees with commonsense. We used these results to verify our hypothesis that SWAG and ComVE datasets share the same type of commonsense knowledge. Through subtask A, we also investigated the extent that BERT encapsulates the world knowledge. For subtask C, we simply fine-tuned GPT-2, which performed reasonably well.

3 System overview

Our system is based on BERT with an initialized multiple-choice classification head on top, which is trained for each of the following subtasks separately. We prepared our data following the format seen in Figure 3. Figure 4 shows our model.

3.1 Subtask A: Commonsense Validation



Task 1: Sen-making or validation

Startphrase: The statement that ***makes more sense*** is

He put an elephant in the fridge
He put a turkey in the fridge ✓

reverse layout

Startphrase: The statement that ***is against commonsense*** is

He put an elephant in the fridge ✓
He put a turkey in the fridge

original layout

Task 2: Explanation

He put an elephant in the fridge
Connector: is against commonsense because

A: ***an elephant is much bigger than a fridge*** ✓
B: elephants are usually grey while fridges are usually white
C: an elephant cannot eat a fridge

Figure 3: Modified ComVE data as system input for subtask A and subtask B

The original SWAG dataset instances are mainly comprised of a start phrase split into two sentences, followed by four endings that represent the multiple choices, as shown in Figure 1. An example of a start phrase is *"On stage; a woman takes a seat at the piano. She"* followed by four candidate endings to the given sentence. Only one of the endings is plausible given the context; *"nervously sets her fingers on the keys."*, as opposed to *"sits on a bench as her sister plays with the doll"*, among others.

In subtask A, the original ComVE training data contains instances of the following format *(sentence1, sentence2, answer)* where the answer is one of the given input sentences that is against commonsense as shown in Figure 2. Thus, we modified the SWAG dataset in this direction by reducing the number of choices from four to two endings. For this purpose, we maintained the correct endings along with a randomly chosen distractor that represented an incorrect answer. The resulting dataset will be referred to as SWAG(A2) since it contains only two endings.

The SWAG task, by definition, aims to find the most natural and plausible ending of a given start phrase. Therefore, for each data instance in SWAG, the selected ending corresponds to the ending that agrees the most with commonsense. This is unsurprising since the dataset is constructed from temporally adjacent video captions. However, since subtask A consists of selecting the instance that is **against** commonsense, arguably the opposite of SWAG, some data modifications were necessary. Hence, we experimented with two data layouts, as shown in Figure 3. The first is the *reversed* data layout; we conducted an experiment where the **correct answer is the sentence that agrees with commonsense**. Such a layout included an inversion of the original labels, as well as the addition of a start phrase. The input to our model became (sentence1, sentence2, *The sentence that makes more sense is*, answer) where the answer, in this case, represented the sentence that agrees with commonsense. We made this change based on the assumption that the model has been trained to recognize sentences that agree with commonsense because it was trained on SWAG. The second layout is the *original* data layout, **where the correct answer corresponds to the sentence that is against commonsense**, and where the data input followed the format (sentence1, sentence2, *The sentence that is against commonsense is*, answer).

In SWAG, the start phrases such as *"On stage; a woman takes a seat at the piano. She"* offers a context (or premise) on which the appropriate ending is conditioned. Whereas in our case, the start phrases are fixed and do not share context with the input sentences. Thus, we expected the task to be considerably more difficult and to perform worse than on the SWAG dataset. We consider any change in performance between the original and reverse layout of ComVE and SWAG as an insight to better understand BERT and commonsense reasoning.

3.2 Subtask B: Commonsense Explanation I

For subtask B, we modified the original SWAG dataset by reducing the number of endings to only three. We refer to the resulting dataset as SWAG(B3). We randomly assigned the excluded ending from the set of incorrect endings for each instance. This time, the start phrase was not fixed. However, it consisted of a sentence that is against commonsense such as *I put an elephant in the fridge* concatenated with a connector, namely *is against commonsense because* then followed by one explanation at a time. This is shown in Figures 4 and 5. We experimented with multiple connectors such as *is not possible. Because*, and the one mentioned above and found that they provided similar results.

Once trained on SWAG, the model was fine-tuned on ComVE data, which was also modified as specified in Figure 2. This task can be closely approximated to the SWAG Next Event Prediction task. We hypothesize that the same type of commonsense knowledge is used to tackle both tasks, and we verify our hypothesis through the previously mentioned setup. In order for the system to correctly answer the multiple-choice question in subtask A, it requires knowledge about the world, including time, positional, and size reasoning, among others. Concretely, in order for the system to justify why *He put an elephant in the fridge* is against commonsense, it would need to have some understanding of the respective sizes of an elephant and a standard refrigerator and deduce that the statement is incorrect. Moreover, it would need to make that connection between the reasoning and the incorrect sentence. The extent of such knowledge that BERT contains remains unclear in the research community (Davison et al., 2019).

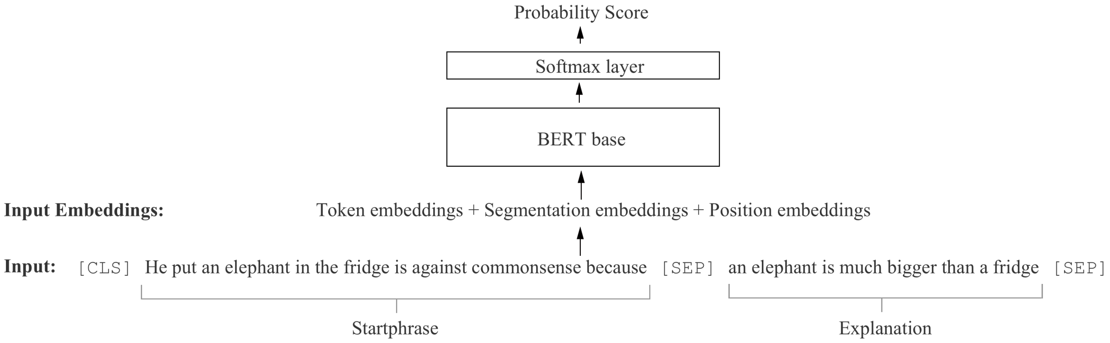
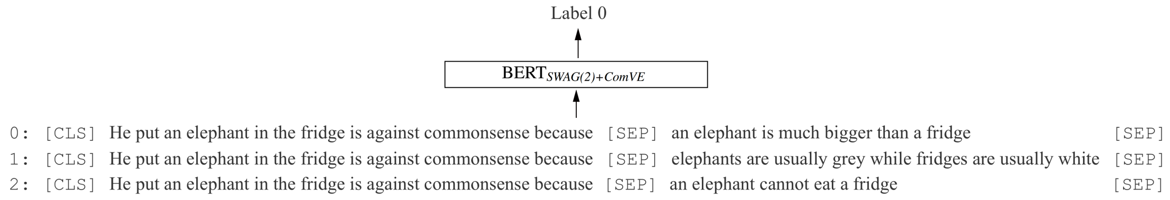

Probability Score
Softmax layer
BERT base
Input Embeddings:
Token embeddings + Segmentation embeddings + Position embeddings
Input:
[CLS]
He put an elephant in the fridge is against commonsense because
[SEP]
an elephant is much bigger than a fridge
[SEP]
Startphrase
Explanation

Figure 4: BERT model modified for Commonsense Explanation Task

3.3 Subtask C: Explanation Text Generation

For subtask C, we used the pre-trained GPT-2 model and fine-tuned it on the correct answers for subtask B. We limited the text generated to 20 words and process the resulting generated text by removing any extra words.



0: [CLS] He put an elephant in the fridge is against commonsense because [SEP] an elephant is much bigger than a fridge [SEP]
1: [CLS] He put an elephant in the fridge is against commonsense because [SEP] elephants are usually grey while fridges are usually white [SEP]
2: [CLS] He put an elephant in the fridge is against commonsense because [SEP] an elephant cannot eat a fridge [SEP]

Figure 5: Input to the model consisting of a concatenation of the sentence and each explanation. The model chooses the explanation with the highest internal probability score extracted from the softmax layer.

4 Experiments

4.1 Baselines

As a baseline, we fine-tuned BERT and GPT-2 model and used a metric that can be approximated to perplexity that indicates if a sequence is likely to occur (Wang et al., 2020). The underlying assumption is that sequences with lower approximated perplexity are more likely to agree with commonsense. Language models trained on large amounts of data give a higher probability to words that often occur in a given context and a lower probability to others that do not. An example of such an unlikely sequence of words is *He put an elephant in the fridge*.

For subtask A, the sentence with the lower perplexity was assumed to be the correct sentence. As for subtask B, for each data instance, we concatenated the sentence that is against commonsense along with the candidate explanations in this manner: *He put an elephant in the fridge is against commonsense because an elephant is too big for the fridge*. Since BERT is a bidirectional language model, the product of the conditional probabilities of each word given its context in a sentence can be interpreted as the perplexity. The context of a given word, is all other words in the sentence that appear before and after the current one. The conditional probability for a word given its context is equivalent to the probability of a masked word in a masked language model like BERT. For GPT-2, the perplexity was calculated as the exponential of the evaluation loss over the number of tokens. The full results are reported in Table 1. The organizers of the task provided baselines for the task in their paper; we report the value of Human Performance for comparison. (Wang et al., 2020).

4.2 Experimental setup

For training, we used the SWAG official data, with 73k instances for training and 20k for development. In the fine-tuning stage, we used the data that was provided by the task organizers with 10k training data and 1k development data. We tested on 1k instances by submitting to the official leaderboard and reporting the returned accuracy. SWAG and ComVE data were prepared as described in the System Description section. We used the open-source BERT transformers implementation by Huggingface(Wolf et al., 2019). We used BERT base uncased and trained on two GeForce RTX 2080 Ti GPUs for three epochs with a learning rate of 5×10^{-5}. We kept these parameters fixed throughout the experiments to focused on the effect of the changes in data setup on the overall result.

For subtask A and B, we reported accuracy as an evaluation measure as this is a multiple-choice task. For subtask C, we reported both BLUE and human evaluation scores. The latter is conducted by the organizers where three humans evaluate 100 random samples from the test set and give a score from 0 to 3; zero represents grammatically incorrect and incomprehensible answers, and 3 is an appropriate and robust explanation of why the given statement does not make sense.

4.3 Subtask A: Validation

We submitted our results for post-evaluation and received an accuracy of 88.7% on the test set, taking the 16^{th} place on the leaderboard.

	Model	Task A		Task B	
		dev	test	dev	test
Proposed Systems	Bert$_{\text{ComVE}}$	**89.2**	81.3	**81.0**	80.5
	Bert$_{\text{SWAG+ComVE}}$	88.1	**88.7**	80.8	**85.3**
	Bert$_{\text{SWAG+ComVE}}$ (no connector)	—	—	—	81.0
	Bert$_{\text{ComVE}}$ (no connector)	—	—	—	81.9
	Bert$_{\text{SWAG}}$	26.2	—	80.6	—
Baselines	GPT-2*	**73.5**	—	37.1	—
	Bert$_{\text{ComVE}}$*	70.1	—	35.4	—
	Bert base*	69.8	—	**45.6**	—
	Human performace	99.1		97.3	

Table 1: Accuracy of different methods on subtask A (original layout) and B. * Represent models where perplexity was used. Bert$_{\text{data}}$ is BERT fine-tuned on the given data. Bert$_{\text{SWAG}}$ refers to BERT trained on SWAG(A2) for subtask A and SWAG(B3) for subtask B. Values in bold represent the highest accuracies in the different categories: Proposed Systems on test set and development set as well as Baselines on development set.

In order to test our assumption that knowledge gained from training on SWAG task would be better leveraged by using a reversed layout, we ran experiments on both layouts explained in the System Description section. As we expected, the data layout had an impact on the performance of the models. The results are indicated in Table 2 where BERT$_{\text{SWAG(A2)}}$ obtained an accuracy of 86.7% on the development set of SWAG(A2). When tested on the reversed data layout of ComVE, the same model performed better than all baselines, obtaining 79.4% accuracy and confirming that training on SWAG can be useful for the validation task.

In contrast, the accuracy of BERT$_{\text{SWAG(A2)}}$ was 26.2% when tested on the original layout. The 53.2% difference in accuracy indicates that the start phrase *The sentence that is against commonsense is* in the original layout was discarded by the model which may be because it does not offer any additional context information. Instead, BERT proceeded to choose the sentence that agreed with commonsense, which was learned by training on the SWAG data. This vast difference in performance leaves open questions about the model's world knowledge since it failed to adjust to the inversion of the data. More concretely, when given the same sentences, the model succeeded in answering that *He put a turkey in the fridge* makes more sense than *He put an elephant in the fridge*. However, it failed to answer which sentence of the two is *against commonsense*. We did not analyze this topic further as it is outside of the scope of this paper and instead leave it for future work.

Data	Model	Reversed	Original
Development set	BERT$_{\textit{ComVE}}$	**89.2**	88.2
	BERT$_{\text{SWAG(A2)}}$	79.4	26.2
	BERT$_{\textit{SWAG(A2)+ComVE}}$	88.1	**88.6**
Test set	BERT$_{\text{SWAG(A2)+ComVE}}$	87.3	88.2

Table 2: Development and test accuracy of different models on subtask A with reversed and original layouts. Bert$_{\text{dataset}}$ is BERT fine-tuned on the given dataset

Our system BERT$_{\text{SWAG(2)+ComVE}}$ did not perform the best when tested on the development set of subtask A. However, its performance was comparable to BERT$_{\text{ComVE}}$ and surpassed it on the test set as shown in Table 1. It is possible that the test data contained more instances similar to SWAG that offered a familiar context than those in the development set. All proposed systems performed better than the baseline models GPT and BERT, which obtained an accuracy of 73.5% and 69.8% on the development set.

4.4 Subtask B: Multiple Choice Explanation

In the official evaluation, we obtained an accuracy of 84.6% and ranked 14^{th}. In the post-evaluation phase, we made some changes in the model checkpoint used for training, which improved the accuracy to 85.3% and ranked 11^{th}. In order to quantify the benefits of using SWAG data for our task, we ran experiments on 3 BERT models, respectively, fine-tuned on SWAG, ComVE, and both datasets. Our system $\text{Bert}_{\text{SWAG+ComVE}}$ showed nearly a 5.0% increase in accuracy compared to a model trained only on ComVE data without SWAG, as shown in Table 1. The performance boost confirms our hypothesis that the same type of knowledge could be leveraged from Next Event Prediction in SWAG and applied to the Explanation task. This is further confirmed by $\text{Bert}_{\text{SWAG}}$ model achieving comparable accuracy to $\text{Bert}_{\text{SWAG+ComVE}}$ and $\text{Bert}_{\text{ComVE}}$ on the development set.

To examine the impact of the connector in Figure 3 on the performance of the model, we ran an experiment where we removed *is against commonsense because of* from the input sequence. Thus, the new input consisted of the sentence that is against commonsense and its corresponding possible explanations. After training on SWAG(B3), fine-tuning on ComVE subtask B data without a connector achieved the same performance as not training on SWAG in both test and evaluation with an accuracy of 81%. This performance corresponds to a simple Multiple Choice BERT and indicates that SWAG offers no performance boost when no connector is used in the input.

All proposed models performed at least 50% better than the baselines on average. BERT baseline performed better than GPT baseline with 45.6% showing a clear decrease in performance once fine-tuned. The organizers of the ComVE task also found similar results, attributing the decrease in performance after fine-tuning BERT to making the model specific to the fine-tuning data and possibly corrupting the structure formed by Next Sentence Prediction (Wang et al., 2020).

4.5 Subtask C: Explanation text generation

For subtask C, we took the 11^{th} place with a BLUE value of 7.1%, along with a human evaluation score of 1.75 out of 3 and ranking 8^{th}. This score indicates that, on average, the sentences generated provided some relevant or simple reason but may contain grammatical errors and unnecessary or irrelevant parts. This can be used as a baseline for future work in the task. Table 3 shows three examples of the generated sentences. The sampled sentences are coherent. However, they are not necessarily relevant to the input given.

Fish can live in a land without water is against common sense because **No one can live without water.**
people are only able to be honest is against common sense because **They have to admit they are liars.**
If you want to dig a hole then you should use the spoon is against common sense because **A spoon is not for digging holes**

Table 3: Example outputs for subtask C generated by GPT-2, capped to 20 words a sentence and truncated after a full stop.

5 Conclusion

In this paper, we have explored the ability of BERT model to use the knowledge gained from SWAG dataset to solve Commonsense Validation and Explanation tasks. We found that training on SWAG without fine-tuning on ComVE data for subtasks A and B still performed better than other methods, proving that SWAG and ComVE Task 4 require some similar commonsense knowledge about the world. However, the difference in performance between the original and reserved layouts and a decrease in performance after fine-tuning on both datasets together raise questions about the extent of world knowledge contained in BERT model. We believe that injecting external knowledge into BERT, such as knowledge bases, could be an exciting addition to such systems and to be explored in future work.

Acknowledgements

This paper is based on results obtained from a project, JPNP18002, commissioned by the New Energy and Industrial Technology Development Organization (NEDO).

References

Joe Davison, Joshua Feldman, and Alexander Rush. 2019. 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.

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

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

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*, pages 4020–4026.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

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