CSUI at SemEval-2020 Task 4: Commonsense Validation and Explanation by Exploiting Contradiction

Kerenza Doxolodeo, Rahmad Mahendra
Faculty of Computer Science, Universitas Indonesia
Depok, West Java, Indonesia
kerenza.doxolodeo@ui.ac.id, rahmad.mahendra@cs.ui.ac.id

Abstract

This paper describes our submissions into the ComVe challenge, the SemEval 2020 Task 4. This evaluation task consists of three sub-tasks that test commonsense comprehension by identifying sentences that do not make sense and explain why they do not. In subtask A, we use Roberta to find which sentence does not make sense. In subtask B, besides using BERT, we also experiment with replacing the dataset with MNLI when selecting the best explanation from the provided options why the given sentence does not make sense. In subtask C, we utilize the MNLI model from subtask B to evaluate the explanation generated by Roberta and GPT-2 by exploiting the contradiction of the sentence and their explanation. Our system submission records a performance of 88.2%, 80.5%, and BLEU 5.5 for those three subtasks, respectively.

1 Introduction

Natural Language Understanding (NLU) has seen increasing attention in recent times. It requires not only the capability to deduce the semantics of the text but also commonsense knowledge of the world. One of the prominent tasks to test how well a machine recognizes and understands the human language is Natural Language Inference, in which given the pair of texts (a hypothesis and a premise), and the model deduces whether the premise entails or contradicts the hypothesis. Winograd Scheme Challenge (Morgenstern et al., 2016) is another task that provides a text and a question regarding the text with two possible answers, and the model must select the correct one. However, the question is designed that both answers are possible, grammatically wise, and the only way to answer is to have access to the implicit commonsense knowledge.

While the aforementioned tasks are the application of commonsense, SemEval 2020 Task 4: Commonsense Validation and Explanation (Wang et al., 2020) puts emphasis purely on the commonsense itself. In this paper, we present our submission system that relies on the Transformers-based model (Devlin et al., 2019). On the other hand, one of the biggest concerns of the ComVe subtask is its small size. We know that the explanatory NLU model is prone to over-fitting and may generate inconsistent explanation (Camburu et al., 2020). A potential solution is to augment the dataset by leveraging the existing commonsense knowledge.

2 Task Description

SemEval 2020 Task 4: Commonsense Validation and Explanation (ComVe) challenges three subtasks: identifying which member of a pair of a similar sentence does not make sense, selecting the best explanation why a sentence does not make sense, and generate an explanation from scratch why a sentence does not make sense. The dataset for ComVe task consists of 10K instances in train set and 1K instances in test set

Proceedings of the 14th International Workshop on Semantic Evaluation, pages 614–619
Barcelona, Spain (Online), December 12, 2020.

2.1 Subtask A: Validation

The model is required to select one of two sentences. Those two sentences only differ in a few words. However, one of them does not make sense. For instance, the texts ”*He put a turkey into the fridge*” and ”*He put an elephant into the fridge*”. The model should select the second one as the implausible sentence.

2.2 Subtask B: Explanation (Multiple Choice)

Given a sentence that does not make sense and three other sentences, the model should select among three which one is the best explanation why the first sentence against common-sense. For instance, ”*He put an elephant into the fridge*” is a sentence that does not make sense, and the three reasons: ”*An elephant is much bigger than a fridge*”, ”*elephants are usually gray while fridges are usually white*”, and ”*an elephant cannot eat the fridge*”. The first reason is the most precise explanation.

2.3 Subtask C: Explanation (Generation)

The model is challenged to generate an explanation of why the given sentence does not make sense. Evaluation is performed towards three samples of explanations created by human annotators with 4-gram BLEU. To prevent data leakage, using the result from subtasks A and B to assist the model in subtask C is not allowed.

3 Resources

Our submission system to ComVe challenge takes advantage of several resources, i.e. Transformers based mode, ConceptNet knowledge base, and Natural Language Inference dataset.

3.1 Transformers based model

Transformers is a language-translation model that removes the need for recurrent networks by simply using the attention network. Its breakthrough is the introduction of self-attention layers that produces attention-based embedding for each word instead of the entire text. Language models, such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and GPT-2 (Radford et al., 2019), have taken a subset of this model as part of their architecture. BERT and RoBERTa use the encoder while RoBERTa used the decoder. The difference between the two parts can be ignored safely for our understanding.

However, this difference warrants different methods of training. BERT and RoBERTa are pre-trained with Masked Language Modeling. We present them with sentences where some words have been blanked (masked), and the model is trained to predict this blanked word. Because the decoder is only able to output one word per time tick, GPT-2 is trained by receiving an incomplete sentence and asking to predict which word comes next.

3.2 ConceptNet

ConceptNet (Speer et al., 2017) is a crowd-sourced undirected knowledge graph that connects words and phrases to its assertion. Each word is represented as an edge, and they may be connected by vertex known as relations. The relations have labels that describe them, such as ”RelatedTo,” ”CapableOf.” The network has been converted as word embedding and achieved competitive results in SAT analogies.

3.3 NLI Datasets

Current existing Natural Language Inference datasets include SNLI (Bowman et al., 2015) and MultiNLI (Williams et al., 2018). In those dataset, the relation between premise and hypothesis sentences are labeled into one of three categories (entailment, contradictory, neutral). In our experiment, we use MultiNLI. Even though it is a bit smaller than SNLI in term of number of instances (400K instances of MLNI vs 550K instances of SNLI), MNLI dataset is designed to encompass more diverse topics.

4 System Description

4.1 Subtask A

We fine-tune a Roberta model to classify the sentence. Instead of having the model takes the pair of sentences as the input and selecting which one is against common sense (as in SNLI model), we train our model to tackle single sentence classification. Given a sentence, we compute how likely it does not make sense. Since the instances of subtask A are in form of the pair of sentences, we have independent prediction for each sentence in the pair. The result of final prediction is the sentence with highest posterior probability among two sentences in the input instance.

In addition, we attempt to use ConceptNet to construct an additional dataset. First, we extract NP phrases from the training dataset using a constituency tree parser (Kitaev et al., 2019) and search for their relations in ConceptNet. We decide not to include groups of relationships with less than 100 relationships to ensure the diversity of the generated explanations. There are only 4 types of relationship who has more than 100 relationships obtained : 734 ”RelatedFor” relationships, 454 ”AtLocation” relationships, 281 ”IsA” relationships, 141 ”CapableOf” pairs, 107 ”UsedFor” relationships.

Each relationship becomes an example of a sentence that make senses. To form its nonsense counterpart, we select another relationship by random, which does not share the same subject or object but the same relationship. The sentence that does not make sense is formed by the first relationship's subject and the second relationship's object.

For example, we have a relationship that mentions ”*Dog is capable of barking*” This relationship becomes an example of a logical sentence. To form corresponding nonsense sentence, we select randomly another relationship that does not have dog or barking as its subject or object, but shares ”*is capable of*” as its relation. Suppose that we have ”*Thief is capable of stealing*”. We combine the two of them by the first relationship's subject (”*dog*”) and the second relationship's object (”*stealing*”) to form the nonsense knowledge ”*Dog is capable of stealing*”.

4.2 Subtask B

We fine-tune the BERT model for sentence classification. However, we do not proceed the sentences and three options of explanations as the input at once, as we often see in multiple-choices QA model. Our model takes the sentence and a single explanation as the input at one time and predict the relation between them. Instead of feeding the sentence and the explanation as two separate sentences, which is often found in the entailment model, we combine the input into single complex sentence following the pattern ”[SENTENCE] does not make sense because [EXPLANATION].” Similar to model in subtask A, we select the explanation with the highest probability.

Furthermore, we explore the possibility of using the NLI model to replace the original dataset. We hypothesize that a sentence that does not make sense and its explanation should be contradictory. Therefore, a model can perform well in this task by finding which explanation contradicts the sentence the most. One potential benefit is, since the NLI dataset is significantly bigger than the task's dataset, we may have a more robust model. For this experiment, we use a BERT model that is fine-tuned with MNLI. The explanation that is predicted as contradiction with the highest probability, according to the softmax layer, is chosen for the task.

4.3 Subtask C

We experiment with the Encoder/Decoder that takes the sentence as the input and outputs the explanation why it does not make sense. However, the model is only able to produce incoherent sentences. We hypothesize that it occurs due to the model unfitting since the small size of the train set. Since the data is relatively small, we tackle subtask C with a model that does not use the train set.

We hypothesize that every explanation follows the pattern

”[NP Subject][Auxiliary verb + (optional too)][One or two words explanation].

For example, we can explain that ”*You can put giraffe inside a fridge*” is nonsense since [*giraffe*][*is too*][*big*]. We can explain that ”*fish dies inside water*” is nonsense since [*fish*][*has*][*fin*].

The first part generates all the possible [NP Subject][Modals]. We generate the possible subjects by parsing the constituency tree of the sentence using (Kitaev et al., 2019) and take all NP phrases. The auxiliary verb is fixed list that contains [can't / can / is / are / isn't / aren't / are too / is too]. Finally, to form incomplete sentences, we use every permutation of NP and auxiliary verb using the template

” (sentence) makes no sense because [NP] [auxiliary]”

We complete those sentences with an explanation. We use two models: Roberta and GPT-2. For Roberta, we use its capability of the Cloze task. Cloze task challenges model to fill in missing words. (Taylor, 1953). We pretend that this explanation is the missing word, and we ask the model to predict what these missing words are. For each sentence, we create two versions that the model has to fill: one that only has one missing word and followed by a full stop and one that has two missing words and followed by a full stop. The addition of full stop is crucial step to discourage Roberta from filling the missing word with a full stop.

For GPT-2, we provide the incomplete sentence and ask the model to complete it with one word and two words. Since the explanation of why the sentence does not make sense is something that makes sense, the information regarding the explanation may be found in the corpus The transformer who is trained on a corpus should have learned the fact and be able to finish the sentence successfully. For instance, if we have an incomplete sentence of ”Putting an elephant inside a fridge does not make senses because an elephant is too,” we hypothesize that it is reasonable to assume that our corpus will mention the fact that an elephant is big, so the model should complete the sentence with ”big.” Finally, we evaluate all these candidate explanations which have been completed. We use the SNLI model to find the best explanation. For the task, we strip the template from the sentence and submit the [NP][auxiliary verb][one or two words explanation] for evaluation.

4.4 System Configuration

We use ktrain's implementation of cased large BERT (Maiya, 2020) to train the vanilla model for subtask B submission. The models for other subtasks harness the implementation of Hugging Face's transformers (Wolf et al., 2019).

For subtask C, due to technical constraints, instead of using the big GPT-2 directly, we use the distilled version. A distilled model is a model whose architecture is smaller and trained to replicate the gradient of the bigger model (Sanh et al., 2019). This cause a slight drop with accuracy, but a model that is more manageable due to smaller size.

5 Result

In overall, performance of our model surpass the baseline (Wang et al., 2019), that are a fine-tuned ELMO with an accuracy of 74.1 % for subtask A and the BERT model with an accuracy of 45.6 %.

Subtask	Model	Performance	
A	Roberta	Accuracy	88.2 %
A	Roberta + ConceptNet	Accuracy	89.1 %
B	BERT	Accuracy	80.5 %
B	BERT + MNLI	Accuracy	64.7 %
C	Roberta	BLEU	5.5
C	GPT-2	BLEU	5.5

Table 1: Summary of Submission Model Evaluation

For subtask A, the Roberta model records an accuracy of 88.2% and finishes in 22nd out of 39 teams. Moreover, the Roberta model that uses ConceptNet to generate an additional dataset, which we submit after the evaluation period, shows a slight improvement regarding accuracy. The BERT model that we submit for subtask B reaches 80.5% and positions 16th place out of 27 teams. However, the model that is trained with only SNLI gets an accuracy of 64.7%. We see a significant accuracy drop from the SNLI model, but it should be high enough that it should be able to support the model for subtask C. Both GPT-2 and Roberta achieve the BLEU score of 5.5. GPT-2 was evaluated by human grader and marked of 0.73/3, although we discover that the sentences created by Roberta should be more coherent than Roberta. From its BLEU, the model finish in 12th place out of 17 teams. However, sorted by human evaluation, the model places in 14th place.

From the submission into the subtask A, we observe that in around 70% our inaccurate prediction involve changing the last word in the instance (shown in Table 2). We suspect that is suffered by the nature of attention. Since both sentences are similar, both sentences may emphasize identically, possibly with a low level of attention to the last words, and therefore the single word difference is missed.

Logical sentence	Nonsense sentence
crack addicts are addicted to crack	crack addicts are addicted to chocolate milk
the generator was able to power the house	the generator was able to power the continent
She went to the grocery store to get bananas.	She went to the grocery store to get an aneurysm.

Table 2: Sample of Misclassified Sentences

6 Summary

We observe that training Transformer model to take the single sentence as an input and classify the sentence within the pair independently can achieve a high performance in subtask A and subtask B. We have explored that the contradiction relationship between the sentence that does not make sense and the explanation may work on commonsense validation and explanation task.

References

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal, September. Association for Computational Linguistics.

Oana-Maria Camburu, Brendan Shillingford, Pasquale Minervini, Thomas Lukasiewicz, and Phil Blunsom. 2020. Make up your mind! adversarial generation of inconsistent natural language explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4157–4165, Online, July. Association for Computational Linguistics.

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

Nikita Kitaev, Steven Cao, and Dan Klein. 2019. Multilingual constituency parsing with self-attention and pre-training. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3499–3505, Florence, Italy, July. Association for Computational Linguistics.

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*.

Arun S. Maiya. 2020. ktrain: A low-code library for augmented machine learning.

Leora Morgenstern, Ernest Davis, and Charles L Ortiz. 2016. Planning, executing, and evaluating the winograd schema challenge. *AI Magazine*, 37(1):50–54.

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

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. In *NeurIPS EMC*$^{2}Workshop$.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Wilson L Taylor. 1953. “cloze procedure”: A new tool for measuring readability. *Journalism quarterly*, 30(4):415–433.

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. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4020–4026, Florence, Italy, July. Association for Computational Linguistics.

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.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

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.