SemEval-2020 Task 4: Commonsense Validation and Explanation

Mohammad Karimiabdolmaleki
University of Alberta
Edmonton, AB
karimiab@ualberta.ca

Ali Zamani
University of Alberta
Edmonton, AB
azamani1@ualberta.ca

1 Literature Review

With the rapid development of artificial intelligence and Machine Learning (ML) algorithms, computers have shown promising results in understanding the context and recording better performance than humans in some benchmarks (Wang et al., 2017). Commonsense validation is one of the Natural Language Understanding (NLU) tasks that has received increasing attention throughout recent years (Wang et al., 2020b).

Commonsense reasoning has been studied in different ways in natural language and has recently attracted a lot of attention (Wang et al., 2020a). The study (Pai, 2020), utilized pre-trained Language Models (LM) and fine tuned them for subtask A and B. The model scored a high accuracy of 95.9% for subtask A and 90.8% for subtask B which was close to human performance on the tasks. The paper (Mohammed and Abdullah, 2020), participated in subtask A and applied ensembling of four different state-of-the-art pre-trained models (BERT (Devlin et al., 2019), ALBERT (Lan et al., 2020), RoBERTa (Liu et al., 2019), and XLNET (Yang et al., 2019)) and achieved 96.2% accuracy for subtask A. Consequently, having a pipeline of fine-tuned state-of-the-art pre-trained models is a common approach that results in a superior performance.

The paper (Doxolodeo and Mahendra, 2020), utilized RoBERTa for subtask A to figure out which sentence does not make sense. In subtask B, they used BERT alongside replacing the dataset with Multi Natural Language Interface (MNLI) corpus (Neelakantan et al., 2017) to have a wider range of possible explanations. In subtask C, the MNLI corpus from subtask B was used to examine the explanation generated by RoBERTa and GPT-2 (Radford et al., 2019a) by extracting the contradiction between a sentence and their explanations. The study resulted in 88.2% and 80.5% accuracy for subtasks A and B.

The paper (Liu et al., 2020) uses an ALBERT-based model for subtask A, while for subtask B, it uses a multiple-choice model enhanced with a hint sentence mechanism to select the reason from given options about why a statement contradicts the commonsense. The achieved accuracy score of the tasks were 95.6% and 94.9% for subtask A and B, respectively. (Ben Rim and Okazaki, 2020) have also used BERT LM to identify whether a sentence contradicts the commonsense (validation task) and justify the main reason of this choice (explanation task). The model obtained an accuracy of 88.7% and 85.3% for subtask A and B, respectively. On the other hand, the work in (Collins et al., 2020) uses the pre-trained GPT Head Model from OpenAI (Radford and Narasimhan, 2018) to calculate perplexity scores based on the sequence of tokens in input sentences and achieves an accuracy of 75% for subtask A.

2 Methods

2.1 Preprocessing

In subtask A, each dataset record consists of two sentences with a label corresponding to the index of the against commonsense sentence (0 or 1). For our binary classification purposes, using ML models, first, we separate these two sentences and assign label 1 to the sentence that makes sense and 0 to the sentence that does not make sense. In fig. 1, the frequency of each class variable is shown. As expected, the dataset is balanced in terms of target value distribution since we divided each record into two records, one for class 1 and the other for class 0. The next step is applying several preprocessing techniques to the sentences. First of all, all of the sentences were converted to lower case, and all punctuations were removed. The second step is

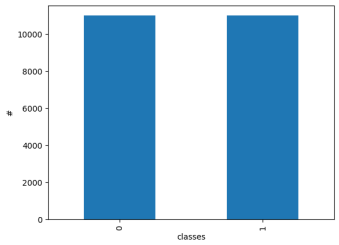

Figure 1: Target variable frequency for subtask A

removing stop words and stemming all the words to decrease the vocabulary size. Consequently, we have a clean dataset to apply any ML or deep learning algorithm to it. It is worth mentioning that we will use the same techniques to have a clean dataset for subtask B.

2.2 Models

2.2.1 Subtask A

Since commonsense validation and explanation was a competition task, it is not a wonder that most of the literature has used state-of-the-art pre-trained models. As described in the preprocessing section, subtask A is a binary classification task. Consequently, we are able to use ML and deep learning algorithms to identify whether a sentence makes sense or not. As a baseline, we implemented logistic regression, naive Bayes, and SVM in scikit-learn with Term Frequency-Inverse Document Frequency (TF-IDF) (Das and Chakraborty, 2018) input representations to have a taste of the data complexity. We have also used pre-trained LMs (GPT2 (Radford et al., 2019b), BERT) to estimate the probability of a sentence. The sentence is given as input to the pre-trained LM, and the output is the perplexity of that sequence. The sentence which obtains a lower perplexity is the sentence that makes sense. That being said, we will continue to improve our ML models by using more robust and powerful representations and will fine-tune the LMs based on (Ben Rim and Okazaki, 2020) work.

2.2.2 Subtask B

Given a sentence that does not make sense and three possible explanations, this task is about choosing the best explanation that fits the incorrect sentence. Since the semantic complexity of this task is high, we do not believe in vanilla ML algorithms. Thus, we will adopt pre-trained language models (GPT, BERT, Roberta) and other state-of-the-art LMs for the explanation task.

3 Evaluation Protocol

Evaluation metrics are dependent on the method of choice for solving a problem. Since the dataset is balanced in terms of class variable frequency, we will use accuracy as the primary metric, but we will also report precision, recall, and F1-Score. Moreover, we will use K-fold cross-validation for subtask A to ensure that the model is robust, can generalize well, and does not overfit the training data. Also, for language modeling tasks in both of the subtasks, we will use probability and perplexity and will use accuracy to assess the performance of the models on the test set.

4 Results

This section will report the accuracy, precision, recall, and F1-Score associated with logistic regression, naive Bayes, and SVM. As shown in Table 1 & Table 2, the ML models could not perform better than chance, while both pre-trained LMs performance were noticeable.

	accuracy	precision	recall	f1-score
Logistic Regression	0.37	0.37	0.37	0.37
Naive Bayes	0.38	0.38	0.38	0.38
SVM	0.39	0.38	0.39	0.37

Table 1: Performance of ML models on subtask A

	accuracy
GPT-2	0.68
BERT	0.55

Table 2: Performance of pre-trained LMs on subtask A

5 Work Plan

The project roadmap and timeline is planned as follows:

Task	Due
ML models optimization	Nov. 7
LM implementation for subtask B	Nov. 20
LMs fine-tuning	Nov. 27
Error analysis and interpretation	Dec. 1
Final Report	Dec. 7

Table 3: Project Timeline

6 Repository URL

https://github.com/UOFA-INTRO-NLP-F21/f2021-proj-zamaniali1995/

References

Wiem Ben Rim and Naoaki Okazaki. 2020. SWAGex at SemEval-2020 task 4: Commonsense explanation as next event prediction. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 422–429, Barcelona (online). International Committee for Computational Linguistics.

Kris Collins, Max Grathwohl, and Heba Ahmed. 2020. Mxgra at SemEval-2020 task 4: Common sense making with next token prediction. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 569–573, Barcelona (online). International Committee for Computational Linguistics.

Bijoyan Das and Sarit Chakraborty. 2018. An improved text sentiment classification model using TF-IDF and next word negation. *CoRR*, abs/1806.06407.

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

Kerenza Doxolodeo and Rahmad Mahendra. 2020. UI at SemEval-2020 task 4: Commonsense validation and explanation by exploiting contradiction. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 614–619, Barcelona (online). International Committee for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations.

Shilei Liu, Yu Guo, Bochao Li, and Feiliang Ren. 2020. Lmve at semeval-2020 task 4: Commonsense validation and explanation using pretraining language model.

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.

Roweida Mohammed and Malak Abdullah. 2020. TeamJUST at SemEval-2020 task 4: Commonsense validation and explanation using ensembling techniques. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 594–600, Barcelona (online). International Committee for Computational Linguistics.

Arvind Neelakantan, Quoc V. Le, Martin Abadi, Andrew McCallum, and Dario Amodei. 2017. Learning a natural language interface with neural programmer.

Liu Pai. 2020. QiaoNing at SemEval-2020 task 4: Commonsense validation and explanation system based on ensemble of language model. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 415–421, Barcelona (online). International Committee for Computational Linguistics.

Alec Radford and Karthik Narasimhan. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019a. Language Models are Unsupervised Multitask Learners.

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

Cunxiang Wang, Shuailong Liang, Yili Jin, Yilong Wang, Xiaodan Zhu, and Y. Zhang. 2020a. Semeval-2020 task 4: Commonsense validation and explanation. In *SEMEVAL*.

Cunxiang Wang, Shuailong Liang, Yue Zhang, Xiaonan Li, and Tian Gao. 2020b. Does it make sense? and why? a pilot study for sense making and explanation.

Wenhui Wang, Nan Yang, Furu Wei, Baobao Chang, and Ming Zhou. 2017. Gated self-matching networks for reading comprehension and question answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 189–198.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.