Commonsense Validation and Explanation in English Sentences

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

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

1 Introduction & Related Work

1.1 Introduction

Reasoning skills are a set of abilities related to people's critical thinking. While humans have these skills to some extent and level, computers have been showing difficulty in reasoning and understanding the intuition of a concept (e.g., text, images, videos, etc.) for several years. Nowadays, with the rapid development of artificial intelligence and machine learning algorithms, computers have shown promising results at understanding the context and even 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). Needless to say, commonsense validation has numerous applications in several Natural Language Processing (NLP) tasks (e.g., sentiment analysis, sarcasm detection, question answering, chatbots, etc.), which adds to its importance. The study is divided into three subtasks: 1) To determine which of two natural language statements with similar wordings make sense and which does not. 2) To choose the most important reason for a particular statement's incoherence from three possibilities. 3) To generate the reason that why a particular sentence contradicts the commonsense, with the help of pre-defined reasons. This project aims to employ machine learning, probabilistic language model, and deep neural networks to tackle commonsense validation and explanation based on the literature.

1.2 Related Work

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) employed an ensemble model and scored a high accuracy of 95.9% for subtask A and 90.8% for subtask B. Furthermore, the paper conducted a thorough examination of the subtask A data in which it discovered that the data can be classified into three categories, based on the sentence structure. 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. 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). In subtask C, the MNLI corpus from subtask B was used to examine the explanation generated by RoBERTa and GPT-2 (Radford et al., 2019) by extracting the contradiction between a sentence and their explanation. The study resulted in 88.2% and 80.5% accuracy for subtasks A and B, and a BLEU of 5.5 (out of 10) for the final subtask. 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 commonsense. The achieved accuracy score of the tasks were 95.6% and 94.9% for subtask A and B, respectively. The work in (Collins et al., 2020) uses the 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 Data

The SemEval 2020, task 4, provided the dataset containing three different sections as train, validation, and test. Each category includes three pairs of comma-separated value files corresponding to each of the mentioned tasks in the introduction section. The first task's input is two statements, and the output is to identify which one is against the commonsense (i.e., binary classification):
Statement 1: *Mohammad put a fish into the fridge.*
Statement 2: *Mohammad put a dinosaur into the fridge.*

The second task's input is a single statement and the output is to select the most satisfactory reasons why this statement is against commonsense (i.e., multiclass classification):
Statement: *Ali put a dinosaur into the fridge.*
Reasons: *1) Dinosaurs went extinct about 65 million years ago. 2) A dinosaur is much bigger than a fridge.*

The final task's input is also a single statement and the output is to generate a referential explanation based on the dataset's gold standard answers.
Statement: *Mohammad is drinking an apple.*
Referential Reasons: *1) Apples can not be drunk. 2) An apple is a whole food and unable to be drunk without being juiced. 3) He eats an apple.*

3 Methodology

In this project, we will build systems for distinguishing the commonsense statements from those that do not make sense. This task has been divided into three subtasks (A, B, and C). For subtasks A and B, we will use logistic regression and support vector machine as a baseline and will utilize multiple state-of-the-art pre-trained models consisting of BERT, ALBERT, RoBERTa, and XLNET to achieve a superior performance in both identification and explanation task (Task C). Specifically speaking, we are interested to fit ensemble machine learning and deep learning models to the problem. For instance, adding a feed forward layer and a softmax layer at the end of a language model while preserving its original structure, can be one of our approaches. Fig. 1, shows the architecture of the ensemble model.

For the pre-processing phase, we will use the following pipeline (subject to change):

This pipeline is in charge of cleaning the corpus from undesirable and worthless items. The pipeline is divided into two phases:

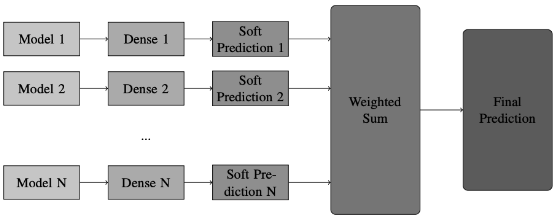

Figure 1: The architecture of the ensemble model (Pai, 2020)

1. **Removing noise:** Removing stop-words (like, and, the, etc.) is a wise decision since language models work on the concept of probability and the occurrences of stop-words are usually high, leading to a noticeable decrease of the system's accuracy. In addition, since the language is case sensitive, all the words should be converted to the lowercase format.

2. **Text normalization:** In this phase, tokenization, stemming or Lemmatisation will be applied to the text for converting sequence of words to tokens and removing their prefix and suffixes.

4 Evaluation

Based on our elementary analysis of the dataset, all three subtasks of the project have a balanced ratio among their class variables. Consequently, our primary evaluation metric for the first and second tasks, binary and multiclass classification, would be accuracy. We will also report the precision, recall, f1-score, and learning curve of our machine learning and deep learning methods. On the other hand, the quality of the generated sentences (reasons) can not be evaluated with the afore-mentioned criteria, implying the need to use the Bilingual Evaluation Understudy (BLEU) score, which is usually used in machine translation to express the quality of the machine-translated text with respect to top quality translation.

5 Resources

We have found the following Github repositories related to the 4th task of the SemEval-2020.

1. ECNU-SenseMaker (SemEval-2020 Task 4)

2. Commonsense Validation and Explanation Task

This repository belongs to the project.

References

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.

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 pretraining.

Alec Radford, Jeffrey 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 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.