Ferryman at SemEval-2020 Task 5: Optimized BERT for Detecting Counterfactuals

Weilong Chen, Yan Zhuang, Peng Wang, Feng Hong, Yan Wang, and Yanru Zhang*
University of Electronic Science and Technology of China
{chenweilong1995,delecisz}@std.uestc.edu.cn
{wangpeng3314,fengfengis}@std.uestc.edu.cn
{yanbo1990,yanruzhang}@uestc.edu.cn

Abstract

The main purpose of this article is to state the effect of using different methods and models for counterfactual determination and detection of causal knowledge. Nowadays, counterfactual reasoning has been widely used in various fields. In the realm of natural language process(NLP), counterfactual reasoning has huge potential to improve the correctness of a sentence. In the shared Task 5 of detecting counterfactual in SemEval 2020, we pre-process the officially given dataset according to case conversion, extract stem and abbreviation replacement. We use last-5 bidirectional encoder representation from bidirectional encoder representation from transformer (BERT)and term frequency–inverse document frequency (TF-IDF) vectorizer for counterfactual detection. Meanwhile, multi-sample dropout and cross validation are used to improve versatility and prevent problems such as poor generosity caused by overfitting. Finally, our team Ferryman ranked the 8th place in the sub-task 1 of this competition.

1 Introduction

Counterfactual refers to a conditional statement in which the first clause is a past tense subjunctive statement expressing something contrary to fact, as in "If I had studied harder, I might have passed the exam". Sometimes it can be used in practice to support the assumption that if events occurred differently in the past, what might have happened (Gentner and Yeh, 2005), as in "If Rose had not accepted doctor's advice to cut the tumor, she might have been dead". Since Rose is still alive, and it is impossible to come back to ask Rose to reject the advice. While in other cases, as in " If a drop in crude oil prices had been factored out, the stock price would not be so low", it is not the counterfactual that the drop in the price of crude oil, more than anything else, caused the stock market to plummet. Judging counterfactual is hard for the reason that it involves so many aspects of life such as economy, medical treatment. Besides, the exploration of causation is necessary.

Reasoning is a very important and challenge task in the fields of the Natural Language Process and aims to analyze the internal causality of the sentence to find out whether the sentence is correct or not. In the causal relationship, the cause is partly responsible for the result, while the result partly depends on the cause. There is an inherent causal relationship among objective things. Through grasping the causal relationship of things, people can understand the nature of things comprehensively. As an important research direction of causal reasoning, counterfactual reasoning has been employed in many cognitive processes.

According to the order, counterfactual thinking can be divided into upward counterfactual thinking and downward counterfactual thinking. The former one is the assumption that something has happened in the past, that if certain conditions are met, there is possibility that there will be a better outcome than the real outcome. For example, "If we had been on the pitch before the game, we would not have lost the game". The latter one means that the alternative result is worse than the real one, like "Fortunately, I went to the field for adaptation training before the game, otherwise I would have lost the game today". Through

*All the corresponding to Yanru Zhang.

Proceedings of the 14th International Workshop on Semantic Evaluation, pages 653–657
Barcelona, Spain (Online), December 12, 2020.

counterfactual reasoning, people can not only have the joy of success, the regret of failure and other emotional values, but more importantly, they can enhance the confidence of future personal decisions, as well as enable us to find the source of wrong decisions and correct them. In addition, these counterfactual reasoning experience, for others in a similar situation, also have important reference significance.

Many researchers put in the effort and came up with a series of methods to explore the causation. Uplift Modelling and causal Tree has been applied (Radcliffe and Surry, 2011; Rzepakowski and Jaroszewicz, 2012; Zhao et al., 2017; Athey and Imbens, 2015; Athey and Imbens, 2016; Tran and Zheleva, 2019).

In this paper, we propose a model combining Bidirectional Encoder Representation from Transformer (BERT) with multi sample dropout. After adopting feature frequency–inverse document frequency (TF-IDF) vectorizer, the model performs better. Different models are compared and we get the best F1 score by combining BERT with multi sample dropout, TF-IDF vectorizer ensemble and last-5-[CLS] token. Our model is based on real situations, so when counterfactual occur, they can be compared and identified. Multi sample dropout helps accelerate the training process and TF-IDF vectorizer with linear regression improves the robustness of the model.

Rest of the paper are organised as follows. In section 2, we show the overview of the counterfactual. The details of our model are explained in section 3 and results are shown in section 4. We summarize the whole model in section 5.

2 Related Work

There are a lot of measures to do the reasoning, which mainly divided into two groups, statics learning and deep learning. The most basic assumption in statistical learning theory is that training data and test data come from the same distribution. However, in most practical cases, the test data is extracted from a distribution that is only relevant to, but not identical to, the distribution of the training data, which is a big challenge to the reasoning. Besides, counterfactual distributions tend to differ from fact distributions.

Johansson (2016) redefines the counterfactual problem as the problem of covariate transformation and later as domain adaptation problem, for the reason that actual distribution and the counterfactual result distribution are not the same. Besides, the following points are considered: a) the minimum error rate of factual outcomes; b) the use of relevant factual results to guide counterfactual results is done by constraining the results of similar interventions; c) the distribution of the interventions is similar and is overcome by minimizing the discrepancy distance. Here the discrepancy distance refers to the difference between the fact distribution and the counterfactual distribution in the representation space. Later Shalit (2017) improves the discrepancy distance to be the joint distribution. The new model as a whole is similar to the previous one but overcomes the following limitation: a) the need of a two-step optimization and linear hypotheses of the learned representation and the lack of supporting deep neural networks, b) the loss of the treatment indicator due to high-dimensional learned representation. However, the weight compensating for the difference in treatment group size in the sample is given.

Hassanpour (2019) improves the former weight to be learned and combines representation learning with re-weight. Firstly, representation learning was used to minimize selection bias and ensure that the factual results were as correct as possible. Re-weight could adjust the weight of samples to make the distribution of observed data and counterfactual data as consistent as possible.

In contrast to previous approaches to domain adaptation, Alaa (2017) considers counterfactual reasoning to be a multi task framework that mitigates selection bias through dropout bias scoring: each iteration has a certain dropout probability that depends on bias scoring. Their model uses a deep multi task network with a series layers to model potential (factual and counterfactual) outcomes.

3 Methdology and data

3.1 Task Description

Task 5 mainly consists of two sub-tasks, Sub-task 1 is detecting counterfactual statements. According to official definitions, counter-factual refer to things that did not actually happen or cannot happen. In Sub-task 1, we need to determine whether each statement in the official dataset is counterfactual. For example: “If you prescribe a combination of paroxetine and exposure therapy two months ago, you can

avoid her post-traumatic stress", we need to give a judgment. Sub-task 2 is to locate the cause and effect in counterfactual statements. For example:" Because she did not avoid her post-traumatic stress, (we know) no combination of paroxetine and exposure therapy". Some statements only have the first part without the latter part, which is incomplete and we need to assign "-1" to the index.(Yang et al., 2020) The rest of the papers focuses on the Sub-task 1.

3.2 Data Pre-processing

Abbreviation Replacement - It is widely known that people on social platforms usually use abbreviations to comment. For example, 'you're' is the abbreviation of 'you are'. In order to make our model more accurate, a dictionary is needed to substitute abbreviations in the English data set.
Word Stemming - Word stemming can remove the affix to obtain the root word. For instance, the Word stemming operation move 'loving', 'loves', 'loved' to a common root 'love'. We use this method to map related words to the same stem generally gives satisfactory results in the English data set.
Other Normalization Approaches - We lower the characters for BERT_uncased and delete the stop words, which is meaningless in the English data set. Besides, TD-IDF Vectorizer has been adopted to transform the original text to be the feature matrix.

3.3 Methodology

Our model is based on BERT and last-5-[CLS] token to tackle the issue of counter-factual detection. Before the training process, TF-IDF vectorizer have been adopted to empowering each work with their importance. We divided the training data into 5 folds and use the cross validation to improve the training process. To avoid overfitting, as well as accelerating training and improving generalization, we have applied the multi-sample dropout in our model. The details are as follows.

- Light Gradient Boosting Machine(LightGBM) (Ke et al., 2017) is a framework that implements the gradient boosting decision tree (GBDT) algorithm and supports efficient parallel training. It has several advantages, faster training speed, lower memory consumption, distribution sport (i.e. it can quickly process massive data.)

- TF-IDF measures the importance of a word to a document, which is often used as a weight vector. The TF-IDF value increases as the number of times a word appears in the document, and is offset by the number of documents that contain the word, which helps to adjust for the fact that some words appear more frequently. This measurement is widely used in NLP models, and it also performs well in counterfactual detection.

- BERT is one powerful model proposed by Google research team (Devlin et al., 2018), which has two steps, pre-training and fine-tuning. During pre-training stage, BERT is trained on unlabeled online news over pre-training tasks. For fine-tuning, all of the parameters that have been initialized in the pre-training process, will then be fine-tuned in the counterfactual detecting task. In this task, we conduct experiments with different models, BERT has better performance than other models on detecting counterfactual. We compare the F1 value between BERT and non-BERT model as shown in Table 1, and choose BERT as our basic model.

Model	F1
LightGBM	0.795
TF-IDF	0.812
BERT	**0.835**

Table 1: Performance among BERT and other models.

- Dropout is a commonly used regularization method in deep neural network (DNN). During the training process, dropout randomly ignores some neural to avoid overfitting. The proposed method in this work adopts multi-sample dropout (Inoue, 2019), which both accelerate the training process

of neural network and improve generalization over the traditional dropout. Multi-sample dropout can be easily implemented into our language model. After adding multi-dropout to BERT, the stability of the model has improved a lot.

- We call the beginning of sentence a token. [CLS] is a token, which can represent the whole sentences. Last-5-[CLS] has a good performance in the Internet news sentiment analysis, which is based on A Robustly Optimized BERT Pretraining Approach (ROBERTa) (Liu et al., 2019). We use this token to better represent each sentence so that can improve the accuracy on identifying counter-factual.

4 Result

Our best results for Task5 are summarized in Table 2. Precision rate measures the ability of identifying positive samples, and recall rate measures the ability of identifying negative samples. F1 score considers recall rate, precision rate together. High F1 score shows good performance on identifying counterfactual. Our method reaches the highest F1 score at 0.856, which shows powerful identification on counterfactual detection. In our experiments, we found multi sample dropout can improve the robustness compared to the dropout. TF-IDF vectorizer ensemble and last-5-[CLS] token can improve correctness rate by extracting more detailed information.

Based-Model	Method	F1
BERT	Dropout	0.835
BERT	Multi sample dropout	0.839
BERT	TF-IDF vectorizer ensemble	0.845
BERT	Last-5-[CLS]	0.843
BERT	Multi sample dropout, TF-IDF vectorizer ensemble and last-5-[CLS]	**0.856**

Table 2: Performance between BERT and other models

5 Conclusion

To detect counterfactual online effectively, we have applied the last-5-[CLS] token and BERT to deal with the multiple types of news in Task 5. To reduce the computation cost, we have also applied the TF-IDF to measure the importance of each word. In the training process, we have used the multi-sample dropout to avoid overfitting and accelerate training speed. Over all, our work has show competitive results when comparing with the others. We use the multi sample dropout to improve the robustness compared with dropout. Also, the combination of TF-IDF vectorizer and lat-5-[CLS] token is also a creative way to improve the accuracy rate by extracting more detailed information.

References

Ahmed M Alaa, Michael Weisz, and Mihaela Van Der Schaar. 2017. Deep counterfactual networks with propensity-dropout. *arXiv preprint arXiv:1706.05966*.

Susan Athey and Guido W Imbens. 2015. Machine learning methods for estimating heterogeneous causal effects. *stat*, 1050(5):1–26.

Susan Athey and Guido Imbens. 2016. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360.

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*.

Dedre Gentner and David Yeh. 2005. Reasoning counterfactually in chinese: Picking up the pieces. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 27.

Negar Hassanpour and Russell Greiner. 2019. Counterfactual regression with importance sampling weights. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 5880–5887.

Hiroshi Inoue. 2019. Multi-sample dropout for accelerated training and better generalization. *arXiv preprint arXiv:1905.09788*.

Fredrik Johansson, Uri Shalit, and David Sontag. 2016. Learning representations for counterfactual inference. In *International conference on machine learning*, pages 3020–3029.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in neural information processing systems*, pages 3146–3154.

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. *CoRR*, abs/1907.11692.

Nicholas J Radcliffe and Patrick D Surry. 2011. Real-world uplift modelling with significance-based uplift trees. *White Paper TR-2011-1, Stochastic Solutions*, pages 1–33.

Piotr Rzepakowski and Szymon Jaroszewicz. 2012. Decision trees for uplift modeling with single and multiple treatments. *Knowledge and Information Systems*, 32(2):303–327.

Uri Shalit, Fredrik D Johansson, and David Sontag. 2017. Estimating individual treatment effect: generalization bounds and algorithms. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3076–3085. JMLR. org.

Christopher Tran and Elena Zheleva. 2019. Learning triggers for heterogeneous treatment effects. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 5183–5190.

Xiaoyu Yang, Stephen Obadinma, Huasha Zhao, Qiong Zhang, Stan Matwin, and Xiaodan Zhu. 2020. SemEval-2020 task 5: Counterfactual recognition. In *Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval-2020)*, Barcelona, Spain.

Yan Zhao, Xiao Fang, and David Simchi-Levi. 2017. Uplift modeling with multiple treatments and general response types. In *Proceedings of the 2017 SIAM International Conference on Data Mining*, pages 588–596. SIAM.