Proceedings of the 2nd International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/103/20241128

Causal Inference in Natural Language Processing: Application Status and Future Outlook

Yifan Chu

College of Computer and Information Science College of Software, Southwest University, Chongqing, China

chuyifan@email.swu.edu.cn

Abstract. Causal reasoning has received much attention in recent years. Unlike statistical learning that focuses on the correlation between variables, causal inference can analyze the causality between variables and avoid false causal relationships caused by confounding factors. The use of causal inference to help solve Natural Language Processing (NLP) problems has made great progress. Most existing studies have focused on exploring the application of causal relationships in downstream tasks of NLP, achieving good results. This article conducted a comprehensive survey, focusing on investigating the advantages of causal inference methods in solving specific NLP problems and applications, as well as their practical advantages compared to deep learning methods. Intended to provide researchers in the NLP field with more detailed perspectives and recommendations. Specifically, this article first introduces the relevant concepts of causal inference, including causal inference and related concepts of causal relationships, as well as the differences between causal inference and statistical machine learning; Then, existing research on the use of causal relationships in downstream applications was summarized, including bias removal, stock price prediction, document dialogue based, and continuous few shot learning. Finally, a summary was made on the development of future causal relationships.

Keywords: Causal Inference, Natural Language Processing, Debias, Few shot learning.

1. Introduction

Natural Language Processing (NLP) integrates linguistics, computer science, and mathematics. It involves the interaction and interaction between computers and human natural language. Its main goal is to enable computers to understand, interpret, operate, and generate natural languages used by humans, such as English, Chinese, etc. The development of NLP enables computers to process and understand human language more intelligently to accomplish specific tasks. By pre-training the Transformer model using a large-scale dataset, pre-trained language models (PLMs) have been developed with excellent performance in NLP tasks; With the release of chatGPT, the large-scale language model (LLM) has begun to enter the human eye. LLM is an extension of PLM in terms of models and data.

In recent years, NLP has made great progress, with deep learning based NLP being particularly significant. Transformers, as a powerful neural network architecture, have demonstrated excellent performance in NLP tasks [1]. Bidirectional Encoder Representations from Transformers (BERT) is a pre-trained language model based on Transformers, which uses a bidirectional Transformer encoder to

DOI: 10.54254/2755-2721/103/20241128

learn contextual information of the language through pre-training. BERT outperforms previously proposed language models in General Language Understanding Assessment (GLUE) [2], and demonstrates excellent performance in tasks such as natural language inference, text implication, sentiment analysis, and sentence similarity modeling [3]. Meanwhile, BERT is also widely used in various applications related to language comprehension. However, unlike Bert's bidirectional training method, Generative Pre-Trained Transformer (GPT) uses autoregressive generation. GPT4 also adopts the Transformer encoder structure. It uses the dataset for pre training and fine-tuning, and has good performance in text generation [4]. It allows for dual modal input of text and image, demonstrating the potential of general artificial intelligence.

However, deep learning methods also have certain limitations.

Firstly, the model learns the form of language rather than its meaning. When training with a large amount of data, natural language models can learn the surface form of language. However, it is not possible to understand the meaning and communication intention of language like real human communication, and there are also difficulties in connecting language with objective objects. Deep learning, in a sense, is about calculating the probability of occurrence between words. When generating answers, combine words with a high probability of appearing together to create a sentence, which outputs the general distribution of each word in the word list. This will also bring problems: the model does not understand the meaning of the generated words, nor does it understand what the word represents in the real world, so the model cannot explore the intention of humans to say these words. In the octopus experiment proposed by the authors, it is assumed that an intelligent octopus O is eavesdropping on the conversation between human A and human B [5]. Intelligent Octopus O can imitate B and A to have a conversation after eavesdropping on the conversation. But if A mentions a completely new topic (such as making a coconut catapult), O will give a ridiculous answer because he does not understand the content of the topic and cannot pass the Turing test [5]. Through octopus experiments, it can be concluded that deep learning models can only generate content based on probability calculations and cannot understand the actual meaning of the content.

Secondly, deep learning models are difficult to explain. Deep neural networks (DNNs) are neural networks with a large number of layers, and their data processing methods are complex and difficult to understand, making it difficult to associate with a set of observable variables [6]. Commonly referred to as' black box technology '. Although the performance has been improved, the problem is that when the model makes errors in the prediction task, researchers find it difficult to identify where the errors are.

Causal inference can alleviate the above-mentioned problems [7]. Find out what are the necessary assumptions for drawing conclusions by discovering causal relationships between variables and identifying their true correlations. This makes the model's predictions more accurate and robust. Many important studies in the field of NLP are related to the inference of causal relationships. Many existing studies are focusing on how to apply causal relationships to downstream tasks in NLP.

This article compares the differences between causal inference and deep learning in NLP problems, then summarizes the good performance of existing causal inference methods in downstream NLP applications. This article aims to investigate the advantages of causal inference compared to deep learning methods, as well as how causal inference can solve various problems and advantages in the NLP field. Then, the existing research on causal inference in downstream applications of NLP was summarized, providing more detailed explanations and suggestions for researchers in the NLP field. Organizational structure of this article: Section 2 introduces the relevant concepts of causal inference. Section 3 introduces existing research on using causal inference to enhance performance in downstream applications. Section 4 provides a prospect for future development. Section 5 summarizes this paper.

2. Overview of the Concept of Causal Inference

2.1. Causal inference and causal relationship

Causal relationship refers to the existence of a causal relationship between A and B if a certain result B is caused by a certain reason A. A is the cause, B is the result. Causal inference is the use of causal

DOI: 10.54254/2755-2721/103/20241128

relationships between variables for deduction. Causal inference and machine learning models are different. Statistical machine learning models calculate, statistically analyze, and train large amounts of data to derive a possible probability distribution. Causal inference is based on discovering the causal relationship between two variables, rather than correlation and probability of occurrence. For example, variable A and variable B are not causally related, and variable A is not the cause of variable B. However, due to some confounding factors, it is easy to associate A with B in machine statistical learning. To ensure that the model does not mistakenly use false correlations in its judgments, using causal relationships to infer relationships between variables is a feasible and effective method. Many existing studies are also focusing on how to use causal inference to achieve good performance of NLP in downstream application tasks [8-11].

2.2. The ladder of causality

The ladder of causality is divided into three levels. The first layer is correlation, aimed at finding relationships and patterns between variables through observation. Based on observed phenomena, make predictions and explore the correlation and probability between different behaviors. The second layer is intervention, which is higher than association. While observing, it is also necessary to take proactive actions to change the current situation or predict possible future situations. The third layer is counterfactual. People need to detach themselves from facts and engage in imagination and reflection. Understand the reasons and possibilities behind the event, and consider what consequences different choices or actions may lead to. The ladder of causality is shown in Figure 1.

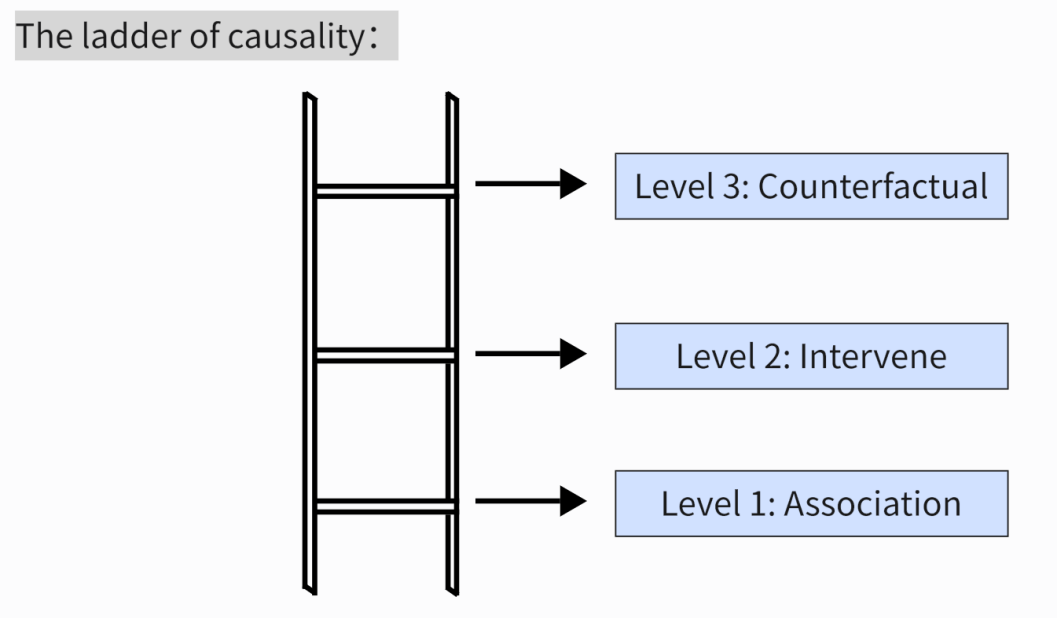

Figure 1. The ladder of causality.

2.3. Identification Hypothesis of Causal Reasoning

In order to estimate the intervention effect, the following assumptions are commonly used in causal inference research: 1) SUTVA: When conducting intervention, the dosage of intervention will not affect the results, and each unit is independent of each other. 2) Ignorability: The background variables and intervention strategies are independent of the outcomes. If the background variables are the same, the results should be the same regardless of the intervention strategy; similarly, if the background variables are the same, regardless of the outcome, the intervention strategy should be consistent. 3) Positivity: Any intervention is possible and uncertain, and the probability of each intervention strategy occurring is greater than zero.

DOI: 10.54254/2755-2721/103/20241128

3. Causal inference application analysis

The large language model is a hot research topic today. However, big language models also have limitations, such as social bias and the need for large datasets. Existing research has shown that causal reasoning can effectively solve existing problems. This section summarizes the existing research on using causal inference methods to solve downstream NLP application tasks.

3.1. Debias

An urgent issue that LLMs currently need to address is the problem of social bias. When applying pre trained model PLMs to downstream tasks, it is necessary to use domain specific data for training and fine tune the pre trained language model PLMs. When using corpus for fine-tuning, board impressions and biases may reappear, and may even be magnified. This can have serious consequences when the model is applied in practical scenarios.

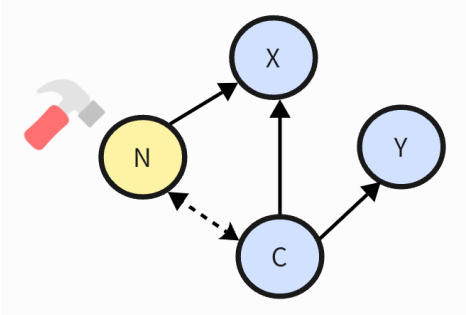

Figure 2. Reasons for Debias

Debias enters fine-tuned PLM due to the association between non causal factors and downstream labels. As shown in Figure 2, X represents the original sentence, N represents non causal factors, C represents causal factors, and Y is the truth label. Causal factor C and non causal factor N jointly produce X, but the truth label Y is only related to causal factor C. Hammer represents intervention in non causal factors. However, due to the additional probability dependency between C and N, a false association has also been established between N and Y through C. Therefore, non causal factor N may re-enter PLM through the path of N ->C ->Y. [8]

The existing methods are divided into two categories. (1) Use depolarization method during pre training [12]. However, when the model that removes bias is fine tuned in downstream datasets, bias will reappear. This is because prejudice has not been truly eliminated, it has only been covered up. (2) Delete potentially biased sentences from downstream datasets [13]. But it is also possible to have bias. In addition, studies have found that existing depolarization methods may lead to a decrease in performance of downstream tasks [14].

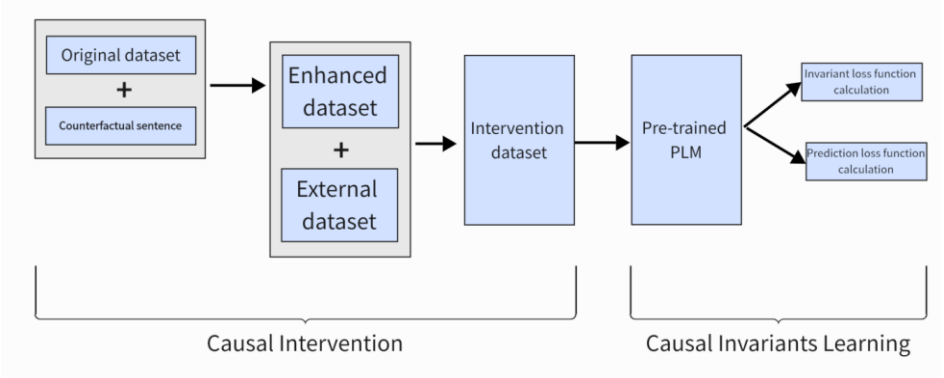

Figure 3. Overview of Causal- Debias. Divided into two parts: causal intervention and causal invariant learning

DOI: 10.54254/2755-2721/103/20241128

The existing method utilizes causal relationships to remove bias by utilizing causal relationships in downstream datasets [8]. As shown in Figure 3. Specifically, first distinguish between causal factors and non causal factors. In social debias, non causal factors refer to different ethnic groups and genders. The author first obtains raw datasets from different groups, and then generates corresponding counterfactual sentences: the attribute words in the sentences of the raw dataset are replaced with their corresponding pairs to create counterfactual sentences. For example, he ->she. Thus, the original dataset and counterfactual sentences were combined to obtain an enhanced dataset. Then perform semantic matching between the enhanced dataset obtained and the external dataset to obtain intervention data. The intervention dataset was obtained by combining the intervention data and the enhanced dataset. Example of intervention dataset shown in Table 1.

Table 1. Example of intervention dataset.

Sentence Type	Sentence Example
Original	demonstrates that **he** remains at the top of **his** game.
Counterfactual	demonstrates that **she** remains at the top of **her** game.
Expansion	Tom is a **guy** who dares to challenge difficulties and persevere to the end.

Then perform causal invariant learning. Hand over the intervention dataset to PLM for causal invariant learning. Combining invariant loss with specific downstream tasks ensures that the performance of the model does not decrease while removing bias, in order to achieve the goal of fine-tuning the language model.

3.2. Few-shot learning

Continuous Few-shot Relation Learning (CFRL) refers to a limited number of training samples and a continuous training mode [15]. If traditional machine statistical learning is used, it is prone to catastrophic forgetting due to the limited number of training samples. And because machine statistical learning calculates the probability distribution between variables, it is difficult to obtain an accurate probability distribution with a small number of data samples. In few-shot learning, memory based methods have been proven to be an effective solution for catastrophic forgetting [16]. However, memory based methods also have limitations as they rely on a large amount of training data. For many tasks, it is difficult to obtain a large amount of labeled datasets. Ye et al. used a causal perspective to explain the problem of catastrophic forgetting: Forgetting is due to the lack of a causal path from old data to predictions [9].Subsequently, Ye et al. proposed a causal effect continuous few shot framework.This framework utilizes the principle of causal inference to effectively address catastrophic forgetting [9].

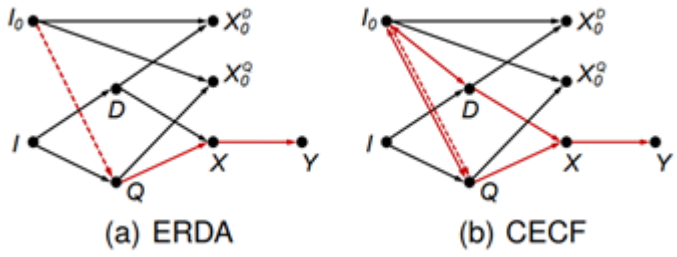

Figure 4. The causal graph for CFRL [9]

As shown in Figure 4, there is only one path between l_0 and Y in (a). Ye et al. established a new causal path between l_0 and Y by colliding memory data with old data. This new causal path enhances the impact of old data on predictions. In Figure 3 (b), there are two paths from l_0 to Y. Then use adaptive weights to balance these two paths, that is, to balance the model's ability to learn new and old

DOI: 10.54254/2755-2721/103/20241128

relationships. Previous methods only had one causal path between old data and new predictions, which came from data replay; But this framework additionally constructs a causal path between old data and new predictions, using collision effects. Two causal paths are more effective in preventing forgetting compared to one causal path. Ye et al.'s experiment also demonstrated that compared with other models, the causal effect continuous few shot framework model performed better.

3.3. Downstream tasks

The two most popular downstream tasks for LLMs currently are conversational systems and financial forecasting systems. Causal reasoning can solve the problem of inability to capture causal relationships in dialogue systems, as well as the problem of excessive attention to false correlations in stock price prediction.

3.3.1. Dialogue system. Dialogue systems are currently one of the most important application directions of NLP. There are currently two challenges that must be overcome [10]: (1) Lack of a large-scale causal complete document dialogue dataset; (2) Unable to capture causal relationships.(1) The existing solution omits the pre training process for DocGD and instead chooses to use language models to initialize parameters, but does not address the issue of lacking a complete dataset.For problem (2), traditional likelihood objective tables are insufficient to capture causal relationships between variables [17], so an algorithm is needed to capture causal relationships.

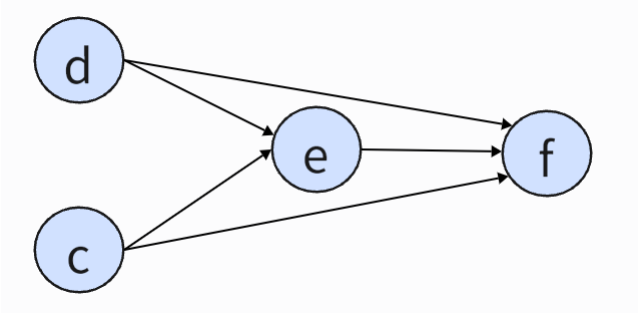

Figure 5. The relationship between various parts of DocGD.

For the first question, Zhao et al. proposed a new strategy for constructing a DocGD pre training corpus [10]. As shown in Figure 5, d represents a document containing content. c represents the user's question. e represents the evidence in the document. r represents the answer generated by the model to the user's question. Document based dialogue task refers to finding the corresponding text content in the document as evidence based on the content in the document and the user's questions, and then the model answers based on the evidence.

Use Wikipedia documents as the data source d in Figure 5. Each sentence in the document is considered as evidence e.Use inpainter to generate conversations, so that the generated conversations can ensure that the content is authentic and accurate. On the Reddit social platform, users' answers to a question usually come with a URL link pointing to a web page. The author captures conversations containing URL links in the platform, takes the content pointed to by the URL as document d, takes the user's question as c, and takes the user's answer to this question as r. The dataset constructed in this way is authentic and accurate, and can be used for the pre training process of the model.

For the second question, Zhao et al. proposed a causal perturbation pre training strategy [10]. Modeling causal relationships. On the basis of the original document, use perturbation document d - containing evidence e to enhance the model's ability to recover from irrelevant perturbations. Meanwhile, using perturbation documents d+that do not contain evidence e to promote reliance on useful evidence. As shown in Figure 6.

DOI: 10.54254/2755-2721/103/20241128

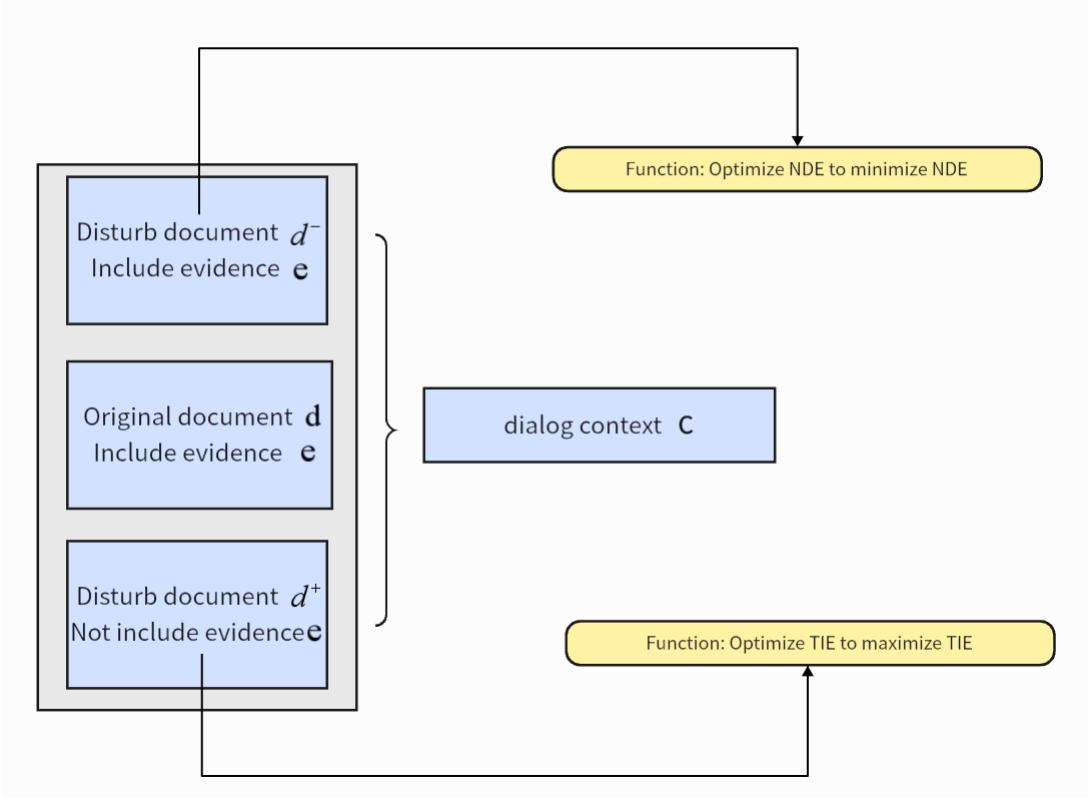

Figure 6. Causal perturbation pre training strategy for CausalDD

3.3.2. Financial forecasting system. In the financial field, predicting stock price changes is also a key topic of discussion. As a specific vertical field, the financial sector has obvious characteristics and a high demand for causal analysis. The demand for accurate prediction of stock price changes is also important. Once accurate predictions of future stock price changes can be made through various information, it will greatly drive economic development.

In recent years, research has improved the accuracy of stock price trend prediction through the correlation between stocks [18,19]. Correlation is composed of both confounding factors and causality. If correlation is used for prediction, it can lead to excessive attention to false correlations, resulting in incorrect predictions.

Currently, an effective method to enhance stock price prediction is to introduce causal inference. Luo et al. proposed a Multi-Memory Interaction Network (CMIN) [11] that can not only capture news texts in the financial field, but also enhance the correlation between stocks by capturing causal relationships. Specifically, the model consists of three main modules: feature embedding module, multi-memory networks, and multi-directional interaction module. The feature embedding module consists of a text encoder and a price encoder. The text encoder encodes and embeds financial text into words. A price encoder is used to capture the interrelationship between prices and events. In addition, there is a causal matrix that is calculated using historical closing prices. The multi memory network consists of a text memory network and a stock memory network for selection and re memory. The multi-directional interaction module allows text information and price information to interact, and this interaction mechanism not only learns the correlation between text and stocks, but also learns the correlation between stocks. Finally, use softmax to generate the final prediction. As shown in Figure 7. Unlike previous studies, this model identifies the true causal relationships between variables and captures the true interdependencies between bone fragments, rather than correlations containing confounding factors. Effectively avoiding false correlations. The test results also proved that the model effectively improved the accuracy of prediction.

DOI: 10.54254/2755-2721/103/20241128

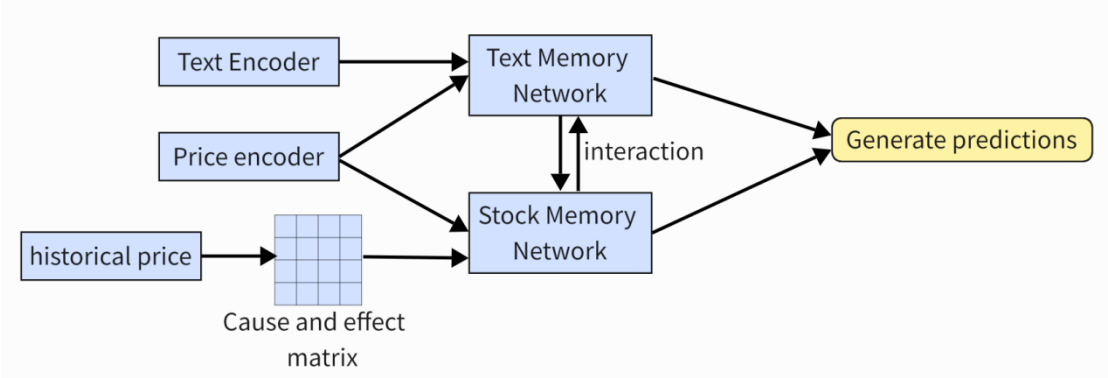

Figure 7. The structure of Causality-guided Multi-Memory Interaction Network (CMIN)

4. Prospect and Challenge

A larger language model does not necessarily mean that the model will better understand user intent. For example, large language models may generate useless answers to users, such as fabricating facts and generating biased or harmful content. This indicates that the model is not aligned with the user's intention. Simply using negative LLMs or positive causal inference methods is incorrect. researchers need to combine the two to complement each other's strengths and weaknesses, and achieve the best performance.

A feasible solution is to use a causal based human-machine collaboration approach. The causal based human-machine collaboration mechanism refers to the joint cooperation between humans and machines, using user feedback to fine tune the model. By understanding and analyzing causal relationships, more efficient decision-making and problem-solving can be achieved. Specifically, in order to make the text generated by the model more in line with human expectations, researchers fine tune the model using a training dataset with human responses as input. Divided into several steps: the human-machine first determines the same problem that needs to be solved, (1) the model generates text. (2) Human beings conduct causal analysis. After understanding the essence and potential causal relationships of the problem, write it into text and input it into the model. (3) The model learns based on the text written by humans to achieve the goal of fine-tuning the model. After fine-tuning, the model generates text again. (4) Humans rewrite again, sort and return to the model for further fine-tuning. Afterwards, the text generated by the model becomes high-quality text that meets human requirements.

Zhang et al. proposed a causal collab algorithm [20] to evaluate the style changes of various interaction strategies in human-computer collaboration. Experiments have shown that the causal collab algorithm can alleviate confounding factors and enhance counterfactual estimation. However, there is currently no optimal strategy for generating text in human-computer interaction. In future research, more attention can be paid to causal relationships in human-machine collaboration. Using causal analysis for human-machine collaboration is both present and future. There is no one size fits all solution to automation, and human-machine collaboration may be the most likely future.

5. Conclusion

This article first explains causal relationships and causal reasoning. Then, a comparison and explanation were made between causal relationships and deep learning methods. Then, the existing research and application advantages of using causal relationships in downstream applications of NLP were summarized, including removing social bias, continuous few shot learning, document based dialogue, and financial prediction systems. After using causal inference, good performance has been demonstrated in these downstream scenarios. Finally, a prospect for the future development of causal relationships was made: simply expanding the model will not bring performance improvements. Hoping that in future research, more attention can be paid to the field of human-machine collaboration based on causal

DOI: 10.54254/2755-2721/103/20241128

inference. Combining human-computer collaboration methods with causal inference and using human responses to fine tune the model may achieve optimal results in text generation.

References

[1] Vaswani, A. (2017). Attention is all you need. arxiv preprint arxiv:1706.03762.
[2] Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2018). GLUE: A multi-task benchmark and analysis platform for natural language understanding. arxiv preprint arxiv:1804.07461.
[3] Wang, J., Huang, J. X., Tu, X., Wang, J., Huang, A. J., Laskar, M. T. R., & Bhuiyan, A. (2024). Utilizing BERT for Information Retrieval: Survey, Applications, Resources, and Challenges. ACM Computing Surveys, 56(7), 1-33.
[4] Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... & McGrew, B. (2023). Gpt-4 technical report. arxiv preprint arxiv:2303.08774.
[5] Bender, E. M., & Koller, A. (2020, July). Climbing towards NLU: On meaning, form, and understanding in the age of data. In Proceedings of the 58th annual meeting of the association for computational linguistics (pp. 5185-5198).
[6] Luo, S., Ivison, H., Han, S. C., & Poon, J. (2024). Local interpretations for explainable natural language processing: A survey. ACM Computing Surveys, 56(9), 1-36.
[7] Feder, A., Keith, K. A., Manzoor, E., Pryzant, R., Sridhar, D., Wood-Doughty, Z., ... & Yang, D. (2022). Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. Transactions of the Association for Computational Linguistics, 10, 1138-1158.
[8] Zhou, F., Mao, Y., Yu, L., Yang, Y., & Zhong, T. (2023, July). Causal-debias: Unifying debiasing in pretrained language models and fine-tuning via causal invariant learning. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 4227-4241).
[9] Ye, W., Zhang, P., Zhang, J., Gao, H., & Wang, M. (2024, May). Distilling Causal Effect of Data in Continual Few-shot Relation Learning. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024) (pp. 5041-5051).
[10] Zhao, Y., Yu, B., Yu, H., Li, B., Li, J., Wang, C., ... & Zhang, N. L. (2023). Causal document-grounded dialogue pre-training. arxiv preprint arxiv:2305.10927.
[11] Luo, D., Liao, W., Li, S., Cheng, X., & Yan, R. (2023, July). Causality-guided multi-memory interaction network for multivariate stock price movement prediction. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 12164-12176).
[12] Meade, N., Poole-Dayan, E., & Reddy, S. (2021). An empirical survey of the effectiveness of debiasing techniques for pre-trained language models. arxiv preprint arxiv:2110.08527.
[13] Liang, P. P., Li, I. M., Zheng, E., Lim, Y. C., Salakhutdinov, R., & Morency, L. P. (2020). Towards debiasing sentence representations. arxiv preprint arxiv:2007.08100.
[14] Meade, N., Poole-Dayan, E., & Reddy, S. (2021). An empirical survey of the effectiveness of debiasing techniques for pre-trained language models. arxiv preprint arxiv:2110.08527.
[15] Qin, C., & Joty, S. (2022). Continual few-shot relation learning via embedding space regularization and data augmentation. arxiv preprint arxiv:2203.02135.
[16] Wang, H., Yu, M., Guo, X., Chang, S., & Wang, W. Y. (2019). Sentence embedding alignment for lifelong relation extraction. arxiv preprint arxiv:1903.02588.
[17] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of machine learning research, 21(140), 1-67.
[18] Yoo, J., Soun, Y., Park, Y. C., & Kang, U. (2021, August). Accurate multivariate stock movement prediction via data-axis transformer with multi-level contexts. In Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining (pp. 2037-2045).

DOI: 10.54254/2755-2721/103/20241128

[19] Long, J., Chen, Z., He, W., Wu, T., & Ren, J. (2020). An integrated framework of deep learning and knowledge graph for prediction of stock price trend: An application in Chinese stock exchange market. Applied Soft Computing, 91, 106205.
[20] Zhang, B., Wang, Y., & Dhillon, P. S. (2024). Causal Inference for Human-Language Model Collaboration. arxiv preprint arxiv:2404.00207.