Two-Stage Fine-Tuning for Variable Definition Extraction from Chemical Process-related Papers

Shota Kato, Makishi Yamamoto, Kotaro Nagayama, Manabu Kano

Graduate School of Informatics, Kyoto University, Yoshida-Honmachi, Sakyo-ku, Kyoto, 606-8501, Japan.

*Corresponding author(s). E-mail(s): shota@human.sys.i.kyoto-u.ac.jp;

Abstract

Extracting variable definitions is challenging due to variability across fields. This study focuses on chemical process-related documents, proposing a two-stage fine-tuning method: first, learning common description patterns from multiple domains, and second, specializing in chemical processes. We tested our method on five BERT models, and it outperformed single-stage fine-tuning in all but DeBERTaV3$_{\mathrm{large}}$. DeBERTaV3$_{\mathrm{large}}$ fine-tuned with limited data, achieved 85.4% accuracy and 84.0% F1 score. Our results suggest that the proposed method is effective under limited data and computational resources and that only small datasets are sufficient to fine-tune DeBERTaV3$_{\mathrm{large}}$.

Keywords: information extraction, mathematical language processing, domain adaptation, BERT model

1 Introduction

The automated extraction of mathematical expressions, including variables and formulas, from scientific documents is crucial for diverse applications ranging from document summarization to mathematical information retrieval. As the volume of scientific literature grows, the demand for efficient and accurate processing of these documents becomes increasingly critical. This study focuses on variable definition extraction from chemical process-related documents, which aims to extract a definition from a given text associated with a variable symbol. Recent variable definition extraction methods have employed large-scale models like Bidirectional Encoder Representations from

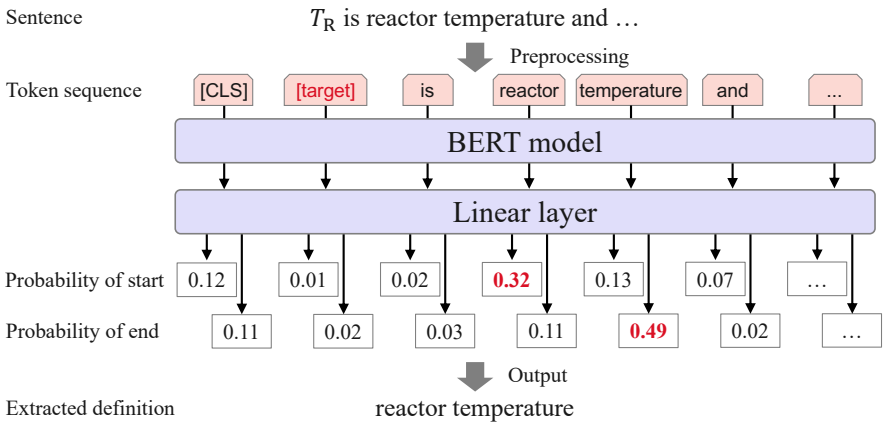

Fig. 1 A schematic diagram of the variable definition extraction method using a BERT model.

Transformers (BERT) [1] and SciBERT [2], which enhance the capability to handle complex language patterns [3]. However, existing methods still struggle with precisely extracting variable definitions, often due to the limited domain-specific data.

This study proposes a two-stage fine-tuning method to bridge the gap between general language understanding and domain-specific needs. The contributions of this paper are threefold. First, we introduce a high-performance method for extracting variable definitions from chemical process-related documents. Second, we demonstrate the effectiveness and limitations of two-stage fine-tuning. Third, we identify and discuss the problems of our methods, proposing potential enhancements for future research.

2 Method

The variable definition extraction method utilizes a BERT model to extract definitions associated with specific variables from scientific texts. The schematic diagram of the proposed method is illustrated in Figure 1. The method begins by preprocessing the input text to replace the target variable symbol with a special token [target], ensuring the model can focus on the context of these tokens during training. The text is then tokenized using BERT's tokenizer and fed into the model. The BERT model processes the input and output probabilities for each token, being the start and end of a definition. The sequence with the highest combined probability under the constraint that the start token precedes the end token is selected as the extracted definition.

To address variations across different domains, we employ a two-stage fine-tuning approach. Initially, the model is trained on a diverse dataset across several disciplines. This dataset is chosen for its broad coverage of scientific terminology and concepts, providing the model with a general understanding of variable definitions across various fields. This stage lays the foundation for the specialized training that follows. In the second stage, the fine-tuning focuses specifically on chemical process-related documents. This dataset comprises specialized terms and variable definitions prevalent in chemical engineering literature. Training on this dataset allows the model to adapt

to the specific linguistic and structural nuances of chemical process documentation, enhancing its ability to extract relevant information accurately.

Our experiments utilize two datasets, $\mathcal{D}_{\text{Symlink}}$ and $\mathcal{D}_{\text{Process}}$. $\mathcal{D}_{\text{Symlink}}$ is compiled from ~~101~~ papers across five fields: information science, biology, physics, mathematics, and economics [4]. It encompasses a total of 16,642 variables, of which 11,462 have definitions. $\mathcal{D}_{\text{Process}}$ includes 47 papers from five chemical processes: biodiesel production (BD), crystallization (CRYST), continuous stirred tank reactor (CSTR), Czochralski (CZ), and shell and tube heat exchanger (STHE). It consists of 2,028 variables, 1,276 of which have definitions. We split $\mathcal{D}_{\text{Symlink}}$ and $\mathcal{D}_{\text{Process}}$ into training, validation, and test datasets and evaluate the performance of models on the test datasets of $\mathcal{D}_{\text{Process}}$.

Our experiments used five pre-trained models: $\text{BERT}_{\text{BASE}}$ [1], $\text{BERT}_{\text{LARGE}}$ [1], SciBERT [2], $\text{DeBERTaV3}_{\text{base}}$ [5], and $\text{DeBERTaV3}_{\text{large}}$ [5]. Each model is fine-tuned for three epochs using the Adam optimizer [6] with a learning rate of 5×10^{-5}, a batch size of 16. At the end of each stage, the best-performing models based on validation loss are selected for further training and final evaluations. Results are averaged over five random splits of the datasets to ensure the robustness of the performance metrics.

3 Results and Discussion

Table 1 presents the results, showing performance improvement with the proposed method over single-stage fine-tuning using $\mathcal{D}_{\text{Symlink}}$ and $\mathcal{D}_{\text{Process}}$. When utilizing $\text{DeBERTaV3}_{\text{base}}$ as the base model, the two-stage fine-tuned model and the model fine-tuned on $\mathcal{D}_{\text{Process}}$ yield comparable results. Notably, $\text{DeBERTaV3}_{\text{large}}$ fine-tuned on $\mathcal{D}_{\text{Process}}$ achieves the highest performance, with an accuracy of 85.4% and an F1 score of 84.0%. In one of five experiments, $\text{DeBERTaV3}_{\text{large}}$ fine-tuned using the proposed method surpassed that fine-tuned on $\mathcal{D}_{\text{Process}}$. $\text{DeBERTaV3}_{\text{large}}$ fine-tuned on $\mathcal{D}_{\text{Process}}$ generally outperforms two-stage fine-tuned models, indicating its effectiveness with a small dataset and potential improvements with different fine-tuning strategies. It can be concluded that $\text{DeBERTaV3}_{\text{large}}$ is ideal without size constraints, while $\text{DeBERTaV3}_{\text{base}}$ fine-tuned with the proposed method is best for lightweight needs for variable definition extraction.

4 Conclusion

This study introduces a high-performance method for extracting variable definitions from chemical process-related documents using a two-stage fine-tuning approach. The proposed method outperformed single-stage fine-tuning except for $\text{DeBERTaV3}_{\text{large}}$. $\text{DeBERTaV3}_{\text{large}}$ fine-tuned on $\mathcal{D}_{\text{Process}}$ achieved the highest performance, indicating its effectiveness with small datasets. Our findings suggest that appropriate data utilization strategies may differ with model size in the standard setting of downstream tasks, where both general large and domain-specific small datasets are available. Future research should explore strategies to optimize performance across various models and domains.

Acknowledgements. This research was supported by JSPS KAKENHI Grant Number JP23K13595.

Table 1 Variable definition extraction results. #Fine-tuning refers to the number of fine-tuning stages. The parentheses indicate the datasets used for fine-tuning. Each value is shown in %.

Base model	#Fine-tuning	Acc.	Rec.	Pre.	F1
$\text{BERT}_{\text{BASE}}$	one ($\mathcal{D}_{\text{Process}}$)	73.3	68.9	67.7	68.3
	one ($\mathcal{D}_{\text{Symlink}} + \mathcal{D}_{\text{Process}}$)	66.9	66.6	60.8	63.5
	two ($\mathcal{D}_{\text{Symlink}} + \mathcal{D}_{\text{Process}}$)	74.4	75.7	67.8	71.4
$\text{BERT}_{\text{LARGE}}$	one ($\mathcal{D}_{\text{Process}}$)	74.6	68.3	69.6	68.9
	one ($\mathcal{D}_{\text{Symlink}} + \mathcal{D}_{\text{Process}}$)	70.5	70.3	63.6	66.8
	two ($\mathcal{D}_{\text{Symlink}} + \mathcal{D}_{\text{Process}}$)	77.8	76.5	73.0	74.6
SciBERT	one ($\mathcal{D}_{\text{Process}}$)	78.6	76.3	73.2	74.7
	one ($\mathcal{D}_{\text{Symlink}} + \mathcal{D}_{\text{Process}}$)	71.5	72.8	66.0	69.2
	two ($\mathcal{D}_{\text{Symlink}} + \mathcal{D}_{\text{Process}}$)	79.7	78.5	74.9	76.6
$\text{DeBERTaV3}_{\text{base}}$	one ($\mathcal{D}_{\text{Process}}$)	82.1	82.3	78.0	80.1
	one ($\mathcal{D}_{\text{Symlink}} + \mathcal{D}_{\text{Process}}$)	74.3	80.8	68.6	74.2
	two ($\mathcal{D}_{\text{Symlink}} + \mathcal{D}_{\text{Process}}$)	81.9	83.3	77.9	80.5
$\text{DeBERTaV3}_{\text{large}}$	one ($\mathcal{D}_{\text{Process}}$)	85.4	84.6	83.5	84.0
	one ($\mathcal{D}_{\text{Symlink}} + \mathcal{D}_{\text{Process}}$)	76.4	80.6	71.3	75.7
	two ($\mathcal{D}_{\text{Symlink}} + \mathcal{D}_{\text{Process}}$)	82.0	84.1	78.2	80.9

References

[1] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: 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), pp. 4171–4186 (2019)

[2] Beltagy, I., Lo, K., Cohan, A.: SciBERT: A pretrained language model for scientific text. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pp. 3615–3620 (2019)

[3] Kang, D., Head, A., Sidhu, R., Lo, K., Weld, D., Hearst, M.A.: Document-Level definition detection in scholarly documents: Existing models, error analyses, and future directions. In: Proceedings of the First Workshop on Scholarly Document Processing, pp. 196–206 (2020)

[4] Lai, V., Pouran Ben Veyseh, A., Dernoncourt, F., Nguyen, T.: SemEval 2022 task 12: Symlink - linking mathematical symbols to their descriptions. In: Proceedings of the 16th International Workshop on Semantic Evaluation, pp. 1671–1678 (2022)

[5] He, P., Gao, J., Chen, W.: DeBERTaV3: Improving deBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In: The Eleventh International Conference on Learning Representations (2023)

[6] Kingma, D.P., Ba, J.: Adam: A Method for Stochastic Optimization. Preprint arXiv:1412.6980 (2014)