ProcessBERT: A Pre-trained Language Model for Judging Equivalence of Variable Definitions in Process Models ⋆

Shota Kato * **Kazuki Kanegami** * **Manabu Kano** *

* *Department of Systems Science, Kyoto University, Yoshida-honmachi, Sakyo-ku, Kyoto 606-8501, Japan (e-mail: manabu@human.sys.i.kyoto-u.ac.jp).*

Abstract: Digital twins are expected to play a pivotal role in digital transformation. Although process informatics has attracted much attention, physical models are essential to realizing the digital twins. However, building a physical model of an industrial process takes much toil. We aim to facilitate the physical model building by developing an automated physical model building AI, named AutoPMoB, which performs five tasks: 1) retrieving documents about a target process from literature databases, 2) converting the format of each document to HTML format, 3) extracting information required for building a physical model from the documents, such as variables, equations, and experimental data, 4) judging the equivalence of the information extracted from different documents, and 5) reorganizing the information to output a desired physical model. This study focuses on task 4, especially judging the equivalence of variable definitions, i.e., whether two noun phrases represent the same variable. We created a large-scale corpus consisting of papers on chemical engineering, and built ProcessBERT, which is a domain-specific language model pre-trained on the corpus. We proposed a method for judging the equivalence of variable definitions based on ProcessBERT. When judging the equivalence, our proposed method first uses ProcessBERT to obtain the embeddings of the variable definitions. Then, the method calculates the cosine similarity between the embeddings. The method judges that the two definitions are equivalent when the similarity is larger than a threshold. Our proposed method judged the equivalence with higher accuracy than the method based on original BERT and SciBERT.

Keywords: Artificial intelligence, Physical models, Process models, Modelling, Systems engineering, Natural languages

1. INTRODUCTION

In the process industry, physical models are indispensable for process design and operation. Conventional physical model building heavily relies on engineers with a deep understanding of a target process. They need to survey a vast amount of documents and continue to improve the model by trial-and-error until a desired model is obtained. Hence, it takes a lot of time and effort to build a physical model that meets users' requirements.

To free the engineers from physical model building, we aim to develop an automated physical model building system, named AutoPMoB. AutoPMoB automatically collects relevant documents from literature databases, extracts necessary information from them, and builds a desired physical model by combining the information. Several fundamental technologies need to be developed to realize AutoPMoB. In the present study, we proposed a method for judging the equivalence of variable definitions: whether two noun phrases represent the same variable or not.

A language model trained with a large corpus, such as BERT (Devlin et al., 2019), can achieve high performance on natural language processing (NLP) tasks. Several language models using in-domain corpora, such as BioBERT (Lee et al., 2020) and SciBERT (Beltagy et al., 2019), perform better than original BERT for NLP tasks in a specific domain. These results indicate that the model's performance varies depending on the corpus used for training. Inspired by this observation, we assume that the pre-trained model with a corpus related to chemical engineering will benefit the equivalence judgment of variable definitions.

Many of the BERT-based models in the previous studies shown in Table 1 have used Wikipedia articles or papers in arXiv.org. The Wikipedia articles cover various fields, and arXiv.org contains a few papers related to chemical engineering. Such documents seem to have little useful information for equivalence judgment of the variable definitions appearing in physical models of processes. However, to our best knowledge, there has been no corpus specific to chemical engineering; thus, we first built such a corpus. We then pre-trained ProcessBERT using the corpus. The model was evaluated by judging the equivalence of two variable definitions in papers on a continuous stirred tank

⋆ This work was supported by JST SPRING Grant Number JPMJSP2110 and JSPS KAKENHI Grant Number JP21K18849.

Table 1. Summary of various BERT models using the following text corpora: English Wikipedia (Wiki), BookCorpus (Books), Web-crawled data (Web), PubMed abstracts (PubMed), Semantic Scholar (SS), MIMIC-III database (MIMIC), and a chemical engineering corpus (ChemECorpus). Pre-training type means whether the model was initialized from original BERT (Continual) or not (From scratch). This table is based on Gu et al. (2021).

Model	Vocabulary	Pre-training type	Corpus	Corpus size
BERT (Devlin et al., 2019)	Wiki + Books	From scratch	Wiki + Books	3.3B words / 16GB
RoBERTa (Liu et al., 2019)	Wiki + Books + Web	From scratch	Wiki + Books + Web	160GB
BioBERT (Lee et al., 2020)	Wiki + Books	Continual	PubMed	4.5B words
SciBERT (Beltagy et al., 2019)	SS	From scratch	SS	3.2B words
ClinicalBERT (Alsentzer et al., 2019)	Wiki + Books	Continual	MIMIC	0.5B words / 3.7GB
BlueBERT (Peng et al., 2019)	Wiki + Books	Continual	PubMed + MIMIC	4.5B words
PubMedBERT (Gu et al., 2021)	PubMed	From scratch	PubMed	3.1B words / 21GB
ProcessBERT	Wiki + Books	Continual	ChemECorpus	0.68B words / 4.0GB

reactor (CSTR). We finally compared the model's performance with original BERT and SciBERT.

2. RELATED WORK

BERT (Devlin et al., 2019) is a major language model which utilizes a transformer network and has achieved the highest accuracy on various NLP tasks. Table 1 summarizes various BERT models. Most of them were pre-trained using "general-domain" text corpora created from newswire and Web domains. For example, the original BERT model was pre-trained on Wikipedia and BookCorpus (Zhu et al., 2015). RoBERTa (Liu et al., 2019) performed even larger-scale pre-training using a large amount of additional text data: CC-NEWS collected from the English portion of the CommonCrawl News dataset (Nagel, 2016), OPENWEBTEXT (Gokaslan and Cohen, 2019), and STORIES (Trinh and Le, 2018).

Previous studies have shown that language models using in-domain corpora perform better than those using general-domain corpora when solving NLP tasks in a specific domain. BioBERT (Lee et al., 2020) was pre-trained on PubMed abstracts and outperformed previous models on biomedical text mining tasks such as named entity recognition, relation extraction, and question answering. Alsentzer et al. (2019) pre-trained ClinicalBERT on clinical text from the approximately 2 million notes in the MIMIC-III v1.4 database (Johnson et al., 2016) and found that its embeddings are superior to those of other models for clinical NLP tasks. SciBERT (Beltagy et al., 2019) was trained on the full text of biomedical and computer science papers from Semantic Scholar corpus (Ammar et al., 2018). Moreover, SciBERT used an in-domain vocabulary while other models used the original BERT vocabulary. The model achieved new state-of-the-art results on several tasks from scientific domains.

3. METHODS

We build a language model specific to chemical engineering and judge the equivalence of variable definitions using the model.

3.1 Corpus

We collected papers related to chemical engineering using Elsevier Research Product APIs available at `https://dev.elsevier.com/` from 18 journals as shown in Table 2. We first obtained a list of DOI and then downloaded documents. We then removed some of the documents that were not journal articles and finally obtained about 130,000 papers. The numbers of DOIs and obtained papers are shown in Table 2.

The obtained papers are in XML format and contain tags that have nothing to do with chemical engineering. Pre-training with the papers without preprocessing will result in a language model with poor performance. Since the goal of this study is to judge the equivalence of two variable definitions correctly, we decided to use the abstract and the full text, i.e., "ce:simple-para" element in "ce:abstract" element and "ce:param" elements in the papers in XML format. We did not use the text in the keywords, references, and nomenclature (notation). We removed all of the tags from the abstract and full text. It is difficult to hold the information of figures and tables without tags; hence, they are removed and not used for pre-training. Then, we split the sentences in the papers using ScispaCy (Neumann et al., 2019), a Python library for practical biomedical/scientific text processing. We finally obtained a chemical engineering corpus (ChemECorpus) with a total word count of approximately 0.68 billion (4.0GB).

3.2 Pre-training

The various model parameters, including the model dimensionality, were initialized to $\text{BERT}_{\text{BASE}}$ parameters: the number of layers (i.e., Transformer blocks) is 12, the number of units in each layer is 768, and the number of self-attention heads is 12 (Devlin et al., 2019). The pre-training tasks also followed those used for $\text{BERT}_{\text{BASE}}$: masked language model and next sentence prediction. We used their recommended hyperparameters and Adam (Kingma and Ba, 2015) for the optimizer with a learning rate of 2×10^{-5}.

We used the original BERT code (available at `https://github.com/google-research/bert`). We trained a language model using a maximum sequence length of 128 and a batch size of 64 for 900,000 steps and then using longer sequences of a maximum length of 512 with a batch size of 8 for additional 100,000 steps. Here, we call the pre-trained language model ProcessBERT. Pre-training of ProcessBERT was performed on a single TPU v3 with 8

Table 2. Journal and the number of DOIs and the obtained papers. Some documents with DOIs were not papers, and thus were not included in ChemECorpus.

Journal	Number of DOIs	Number of papers in ChemECorpus
Applied Catalysis B Environmental	11,369	10,727
Carbohydrate Polymers	17,280	16,361
Chemical Engineering and Processing - Process Intensification	4,200	3,935
Chemical Engineering Journal	27,818	27,222
Chemical Engineering Research and Design	6,000	5,375
Chemical Engineering Science	14,572	13,527
Chinese Journal of Catalysis	2,731	2,709
Computer Aided Chemical Engineering	8,213	1,344
Computers & Chemical Engineering	5,610	5,240
Current Opinion in Chemical Biology	2,605	2,201
Journal of Catalysis	10,849	10,248
Journal of Cleaner Production	27,814	26,994
Journal of Energy Chemistry	2,251	2,236
Journal of Process Control	3,048	2,744
Progress in Crystal Growth and Characterization of Materials	332	256
Progress in Polymer Science	1,252	1,017
South African Journal of Chemical Engineering	242	233
Thermal Science and Engineering Progress	986	950
Total	147,172	133,319

cores, which took about 13 hours to complete. We created another model ($\text{ProcessBERT}_{\text{double}}$) trained with double the number of steps (1,800,000 and additional 200,000 steps).

3.3 Equivalence Judgment

We use the word embeddings from language models to calculate the similarity of two definitions and judge their equivalence, as shown in Fig. 1. First, two variable definitions to be judged are inputted to the language model, and then embedding vectors (768 dimensions) for each word output from Transformer Encoders in the model are obtained. Next, a vector obtained by averaging the embedding vectors of each word is used as the one representing the definition. Finally, the cosine similarity between the vectors is calculated as the similarity between the definitions. The two definitions are judged to be equivalent when the similarity exceeds a threshold.

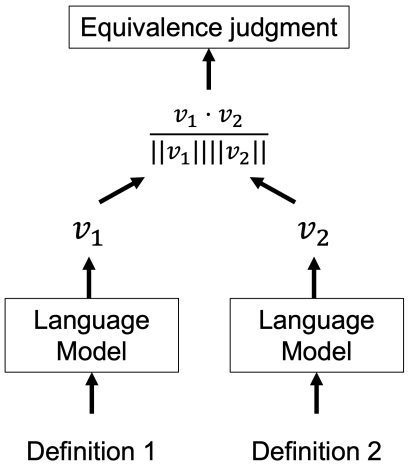

Fig. 1. An architecture of judgment of equivalence of two variable definitions. The two definitions are vectorized by the same language model. v_1 and v_2 denote the embedding vectors of definitions 1 and 2, respectively.

4. EXPERIMENT

We compared the performance of ProcessBERT with that of original BERT and SciBERT.

4.1 Evaluation Dataset

We prepared three papers (Botre et al., 2016; Sundari and Nachiappan, 2017; Wang et al., 2016) and extracted noun phrases representing variable definitions from the section describing a CSTR model. We then created a combination of variable definitions in different papers and manually assigned a label: equivalent (1) or not (0). We finally obtained a dataset including 21 equivalent pairs and 539 non-equivalent pairs. Part of the samples in the dataset are shown in Table 3.

4.2 Evaluation Metrics

A receiver operating characteristic (ROC) curve and the area under the ROC curve (ROC-AUC) are commonly used to assess the performance of a binary classifier (Davis and Goadrich, 2006). The ROC curve plots the true positive rate (TPR) versus the false positive rate (FPR)

Table 3. Examples in evaluation dataset. Each sample has two definitions and a label that indicates whether the definitions are equivalent (1) or not (0).

Definition 1	Definition 2	Label
constant reaction volume	reactor volume	1
flow rate in and out	feed flow rate	1
constant reaction volume	feed flow rate	0
feed temperature	feed flow rate	0

under various thresholds. The TPR, which is also known as recall, and FPR are calculated as follows:

$$TPR = \frac{TP}{TP + FN}, \tag{1}$$

$$FPR = \frac{FP}{FP + TN}, \tag{2}$$

where true positives (TP) and true negatives (TN) are samples correctly predicted as positive and negative, respectively. False positives (FP) and false negatives (FN) are samples incorrectly predicted as positive and negative, respectively.

A precision-recall (PR) curve and the area under the PR curve (PR-AUC) are often used to evaluate the performance of a binary classifier for imbalanced data. The PR curve is a plot of precision versus recall under various thresholds. Precision is calculated as follows:

$$Precision = \frac{TP}{TP + FP}. \tag{3}$$

In imbalanced data, the number of negative samples far exceeds the number of positive samples. As a result, the FPR used in ROC analysis changes little even when the number of FP is large. On the other hand, precision compares FP to TP rather than TN, and thereby PR-AUC is better suited for evaluating the classifier's performance for a minority positive class.

We evaluated the performance of four models with ROC-AUC and PR-AUC. PR-AUC was used because the evaluation dataset in this experiment has a small number of equivalent pairs ($21/560 = 4\%$).

Since BERT architecture has 12 layers (i.e., Transformer blocks), we can obtain 12 embedding vectors for each word. In this study, we compared the performance of all 12 layers of embedding vectors separately to validate which layer should be used for our purpose.

4.3 Results and Discussions

Tables 4 and 5 show ROC-AUC and PR-AUC for each layer of each model. The ROC-AUC of ProcessBERT and $\text{ProcessBERT}_{\text{double}}$ is equal to or larger than that of BERT and SciBERT. In layers 1–3, and 12, $\text{ProcessBERT}_{\text{double}}$ achieved the highest PR-AUC. In layers 7–11, ProcessBERT achieved the highest PR-AUC, and BERT achieved the highest PR-AUC in the other layer. The highest PR-AUC of ProcessBERT, $\text{ProcessBERT}_{\text{double}}$, BERT, and SciBERT are 0.468, 0.482, 0.442, and 0.403. These results indicate that ProcessBERT or $\text{ProcessBERT}_{\text{double}}$ performed better than BERT or SciBERT on average. The results also indicate that the performance of the layer closer to the input is higher, and $\text{ProcessBERT}_{\text{double}}$ using Layer 1 performed the best performance.

However, the value of PR-AUC itself is small, and ProcessBERT and $\text{ProcessBERT}_{\text{double}}$ are still not practical. In the present study, the dataset is created from only three papers— this is not enough to conclude that the proposed method is useful. We need to evaluate the performance using a dataset that contains more pairs of variable definitions.

5. CONCLUSION

This study developed ProcessBERT, a chemical engineering domain-specific language model, and a method for variable definitions equivalence judgment based on ProcessBERT. The proposed method calculates the cosine similarity between two variable definitions and judges equivalent when the similarity exceeds a threshold. The results have shown that the method based on ProcessBERT achieved higher ROC-AUC and PR-AUC than the methods based on original BERT and SciBERT.

However, the current performance of the proposed method is insufficient for practical use. In future work, we will improve the performance by applying fine-tuning because fine-tuning can significantly improve the performance of NLP tasks. Devlin et al. (2019) fine-tuned the BERT model on a task judging whether each sentence pair constitutes a paraphrase using Microsoft Research Paraphrase Corpus (MRPC, (Dolan and Brockett, 2005)). The same architecture can be applied to the equivalence judgment of variable definitions.

We will also increase the number of test data to validate the proposed method in various situations.

Furthermore, although the 4th task was addressed in this study, the fundamental technologies to accomplish other tasks have been under development. We will combine these modules to release the prototype of AutoPMoB.

ACKNOWLEDGEMENTS

This work was supported by JSPS KAKENHI Grant Number JP21K18849 and JST SPRING Grant Number JPMJSP2110.

REFERENCES

Alsentzer, E., Murphy, J., Boag, W., Weng, W.H., Jindi, D., Naumann, T., and McDermott, M. (2019). Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, 72–78. doi:10.18653/v1/W19-1909.

Ammar, W., Groeneveld, D., Bhagavatula, C., Beltagy, I., Crawford, M., Downey, D., Dunkelberger, J., Elgohary, A., Feldman, S., Ha, V., Kinney, R., Kohlmeier, S., Lo, K., Murray, T., Ooi, H.H., Peters, M., Power, J., Skjonsberg, S., Wang, L., Wilhelm, C., Yuan, Z., van Zuylen, M., and Etzioni, O. (2018). Construction of the literature graph in semantic scholar. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*, 84–91. doi:10.18653/v1/N18-3011.

Beltagy, I., Lo, K., and Cohan, A. (2019). 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 (EMNLP-IJCNLP)*, 3615–3620. doi:10.18653/v1/D19-1371.

Botre, C., Mansouri, M., Nounou, M., Nounou, H., and Karim, M.N. (2016). Kernel PLS-based GLRT method for fault detection of chemical processes. *Journal of Loss Prevention in the Process Industries*, 43, 212–224. doi:10.1016/j.jlp.2016.05.023.

Table 4. Area under the receiver operating characteristic curve (ROC-AUC). The column means which layer (Transformer block) was used for obtaining embedding vectors. The bold numbers represent the maximum value in each layer.

Model	Layer 1	Layer 2	Layer 3	Layer 4	Layer 5	Layer 6
ProcessBERT	**0.914**	0.906	**0.908**	0.901	0.911	**0.909**
ProcessBERT$_{\text{double}}$	0.905	**0.908**	0.906	**0.909**	**0.914**	0.907
BERT	0.905	0.866	0.862	0.865	0.86	0.846
SciBERT	0.877	0.865	0.859	0.866	0.875	0.872

Model	Layer 7	Layer 8	Layer 9	Layer 10	Layer 11	Layer 12
ProcessBERT	0.906	0.898	0.893	0.911	0.911	**0.925**
ProcessBERT$_{\text{double}}$	**0.912**	**0.905**	**0.911**	**0.923**	**0.917**	0.917
BERT	0.827	0.798	0.782	0.779	0.819	0.868
SciBERT	0.853	0.854	0.856	0.829	0.836	0.789

Table 5. Area under the precision-recall curve (PR-AUC). The column means which layer (Transformer block) was used for obtaining embedding vectors. The bold numbers represent the maximum value in each layer.

Model	Layer 1	Layer 2	Layer 3	Layer 4	Layer 5	Layer 6
ProcessBERT	0.468	0.397	0.377	0.290	0.311	0.294
ProcessBERT$_{\text{double}}$	**0.482**	**0.44**	**0.390**	0.298	0.302	0.282
BERT	0.442	0.385	0.353	**0.349**	**0.336**	**0.313**
SciBERT	0.403	0.367	0.339	0.345	0.322	0.273

Model	Layer 7	Layer 8	Layer 9	Layer 10	Layer 11	Layer 12
ProcessBERT	**0.313**	**0.305**	**0.301**	**0.326**	**0.314**	0.353
ProcessBERT$_{\text{double}}$	0.249	0.234	0.262	0.318	0.304	**0.362**
BERT	0.290	0.264	0.248	0.232	0.235	0.301
SciBERT	0.244	0.236	0.233	0.218	0.224	0.217

Davis, J. and Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd international conference on Machine learning*, ICML '06, 233–240. Association for Computing Machinery, New York, NY, USA. doi:10.1145/1143844.1143874.

Devlin, J., Chang, M.W., Lee, K., and Toutanova, K. (2019). 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)*, 4171–4186. doi:10.18653/v1/N19-1423.

Dolan, W.B. and Brockett, C. (2005). Automatically Constructing a Corpus of Sentential Paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Gokaslan, A. and Cohen, V. (2019). Openwebtext corpus. `http://Skylion007.github.io/OpenWebTextCorpus`.

Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., and Poon, H. (2021). Domain-Specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1), 1–23. doi:10.1145/3458754.

Johnson, A.E.W., Pollard, T.J., Shen, L., Lehman, L.W.H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L.A., and Mark, R.G. (2016). MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3, 160035. doi:10.1038/sdata.2016.35.

Kingma, D.P. and Ba, J. (2015). Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations (ICLR), Conference Track Proceedings*.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., and Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4), 1234–1240. doi:10.1093/bioinformatics/btz682.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Nagel, S. (2016). News dataset available – common crawl. `https://commoncrawl.org/2016/10/news-dataset-available/`. (Accessed on 04/05/2022).

Neumann, M., King, D., Beltagy, I., and Ammar, W. (2019). ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, 319–327. doi:10.18653/v1/W19-5034.

Peng, Y., Yan, S., and Lu, Z. (2019). Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, 58–65. doi:10.18653/v1/W19-5006.

Sundari, S. and Nachiappan, A. (2017). Decoupling based control analysis of a continuous stirred tank reactor (cstr). In *2017 International Conference on Innovative Research In Electrical Sciences (IICIRES)*, 1–5. doi:10.1109/IICIRES.2017.8078297.

Trinh, T.H. and Le, Q.V. (2018). A Simple Method for Commonsense Reasoning. *arXiv preprint arXiv:1806.02847*.

Wang, T., Gao, H., and Qiu, J. (2016). A combined adaptive neural network and nonlinear model predictive control for multirate networked industrial process control. *IEEE Trans Neural Netw Learn Syst*, 27(2), 416–425. doi:10.1109/TNNLS.2015.2411671.

Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., and Fidler, S. (2015). Aligning books and movies: Towards Story-Like visual explanations by watching movies and reading books. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 19–27. doi:10.1109/ICCV.2015.11.