Prototype of Automated Physical Model Builder: Challenges and Opportunities

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

\*shota@human.sys.i.kyoto-u.ac.jp

Abstract

In the process industry, physical models are indispensable, yet current models sometimes compromise accuracy or incur substantial computational costs. Such cases require new physical models, but the traditional approach to building physical models is reliant on expert knowledge and is time-consuming. This necessitates the development of a new efficient physical model building methodology. Our research aims to establish automated physical model builder, AutoPMoB, which builds physical models from manufacturing process literature. The realization of AutoPMoB requires developing several methods, including those for collecting documents related to the target process and accurately extracting information for physical model building from the documents. In this study, we develop an AutoPMoB prototype, employing a large language model alongside a model building approach previously proposed in our research. The prototype's application to a continuous stirred tank reactor showed its capability to extract necessary data accurately, although the initial attempts did not yield the anticipated models. Subsequent modifications in unifying expressions led to successful model building, underscoring the effectiveness of our system in leveraging literature for physical model building. Advancing AutoPMoB towards practical deployment necessitates specific enhancements, particularly in methods for equivalence judgment of definitions, retrieval of relevant documents, integration of non-documentary information, and domain-specific adaptation.

**Keywords**: Artificial intelligence, Physical model, Digital twin, Natural language processing, Process modelling.

* 1. Introduction

Physical models are crucial in the process industry and serve multiple purposes, including process design and process operation. Although these models can be incorporated into existing tools, there are times when their accuracy is lacking or they are computationally expensive. In such instances, a new physical model must be built. To overcome the time-consuming and costly nature of traditional iterative model building, we aim to develop automated physical model builder, AutoPMoB [Kato and Kano (2022)].

Figure 1 presents an overview of AutoPMoB. The system automatically builds a physical model of the target process in three steps: 1) it retrieves relevant documents, 2) extracts and standardizes the necessary information for model building from these documents, and 3) builds, validates, and ranks model candidates based on this standardized information. To actualize AutoPMoB, we have proposed methods to optimize performance at each step, including a method to judge the equivalence of two differential algebraic equations [Kato et al. (2023)] and a method for automatically generating physical model candidates from multiple equations [Kato and Kano (2023)].

Information extraction from documents has been facilitated by advancements in natural language processing (NLP), particularly methods that leverage the Transformer architecture [Vaswani et al. (2017)]. Recently, large language models (LLMs), which have extensive parameters trained on vast datasets, have emerged as a particularly potent technology. LLMs have exhibited superior performance in various tasks, including information extraction [Han et al. (2023)].

Figure 1 A schematic diagram of AutoPMoB.

This study integrates an LLM with our model candidate building method [Kato and Kano (2023)] to develop a prototype. Besides, we apply this prototype in a case study to assess its usefulness, identify current challenges, and propose solutions for further enhancement.

* 1. Methods
     1. Prototype of AutoPMoB

The prototype receives information necessary for building a physical model from the user: the target process and variables to be included in the model. Users must prepare TeX-formatted documents about the target process. The prototype then extracts mathematical equations and variables, which include variable symbols and definitions, from these documents, judges the equivalence of the extracted information to standardize expressions, and generates physical model candidates. We use an LLM for extracting information and judging the equivalence of variable definitions. After this extraction and standardization, we build model candidates using the methods we previously proposed [Kato and Kano (2023)].

* + - 1. Information Extraction and Equivalence Judgment of Variable Definitions

This study targets TeX documents, wherein mathematical equations are incorporated within equation environments, and variables outside of equations are displayed in inline formats (e.g., enclosed in $, like $t$). We extract mathematical equations and variable symbols from these documents using a pattern-matching method capable of accurate extraction. Since variable definitions are typically phrases within a sentence, they are challenging to extract using only a rule-based method [Schubotz et al. (2017)]; therefore, we employ an LLM-based approach that understands word meanings for accurate extraction of variable definitions from the documents.

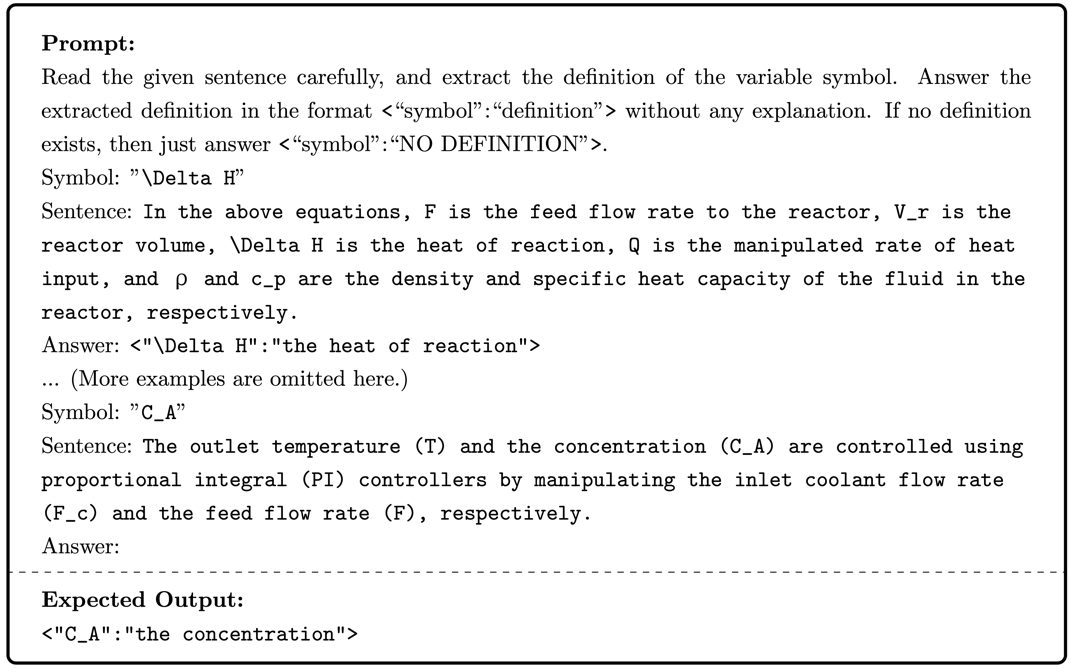
The efficiency of LLMs improves with well-designed input, known as prompts. We implement few-shot learning, which incorporates multiple examples into the prompt [Brown et al. (2020)]. As Han et al. (2023) noted that ChatGPT's information extraction performance diminishes with the increase in the variety of output information, we crafted distinct prompts for the tasks of information extraction and equivalence judgment. These prompts include a task description, output format, examples, and a test case based on the ones by Han et al. (2023). Figure 2 illustrates a prompt example for extracting variable definitions, featuring a symbol and a corresponding sentence for each instance. For judging the equivalence of variable definitions, we input the symbol, its definition, and an equation that includes them.

Figure 2 An example of prompts for variable definition extraction

* + - 1. Model Candidate Building

We apply the physical model building algorithm [Kato and Kano (2023)] to create models that fulfill the user-defined requirements. This method inputs a set of equations alongside the user's specifications and yields sets of equations that align with these requirements. The requirements consist of variables that must be included in the model and whose values can be freely determined. Desired model candidates are generated by amalgamating equations to match the degree of freedom with the difference between the number of equations and the number of variables.

* + 1. Experimental Settings

We evaluated the prototype's ability to automatically build the preferred physical model of a continuous stirred tank reactor (CSTR). We prepared nine documents using the following procedure: 1) we obtained nine papers relevant to the CSTR in PDF format, 2) converted them into TeX format with InftyReader [Suzuki et al. (2003)], and 3) manually revised them to accurately represent the original papers. Only the modeling sections from each paper were used in document preparation. Table 1 enumerates the used papers, and Figure 3 presents the equations from Papers 1 and 2.

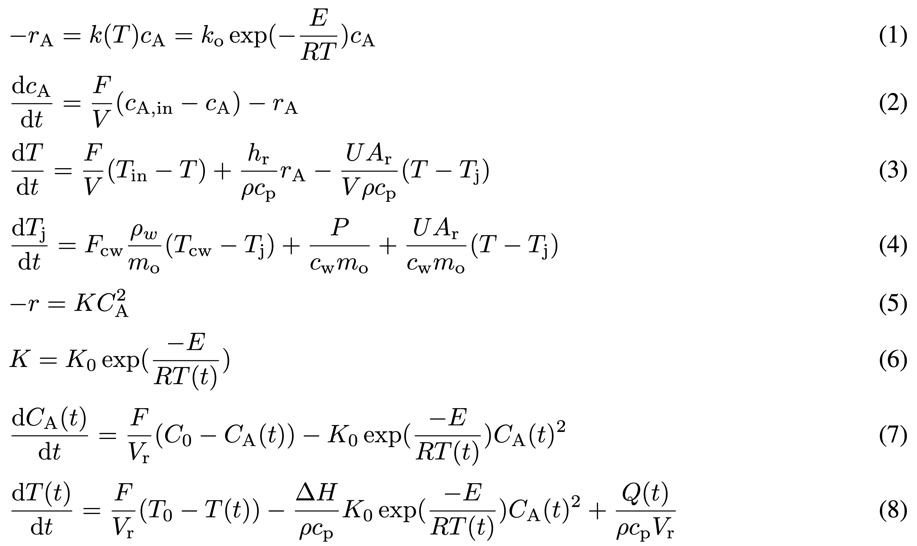
Table 1 List of papers used in case study

Figure 3 Part of equations included in documents. Equations (1)-(4) are in document 1, and equations (5)-(8) are in document 2.

|  |  |
| --- | --- |
| No. | DOI |
| 1 | 10.1016/j.cherd.2019.09.009 |
| 2 | 10.1016/j.jlp.2016.05.023 |
| 3 | 10.1016/j.jlp.2012.10.003 |
| 4 | 10.1016/S0005-1098(01)00083-8 |
| 5 | 10.1252/jcej.07WE187 |
| 6 | 10.1109/EPEPEMC.2010.5606563 |
| 7 | 10.1016/j.compchemeng.2005.11.008 |
| 8 | 10.1109/ICEFEET49149.2020.9187017 |
| 9 | 10.1109/IICIRES.2017.8078297 |

We determine that the prototype has successfully built models if it, given the nine documents and user-defined requirements, outputs models capable of computing the values of all variables in the equations. Additionally, we assess the prototype's ability to correctly extract variable symbols, definitions, and equations, and to ascertain their equivalence.

* 1. Results and Discussion
     1. Experimental Results

The accuracy, precision, and recall of variable definition extraction were 74.1%, 75.9%, and 95.3%, respectively. In the equivalence judgment of variable definitions, no equivalent pairs with different variable symbols were identified as equivalent (Recall=0%). Moreover, for pairs with the same variable symbol and definition, only about half were correctly identified as equivalent when they appeared in different equations (Recall=47.4%).

In model candidate building, the prototype successfully built a physical model from each document by using the set of variables it contained. However, it failed to generate a model candidate by integrating equations from different documents because the notations were not standardized.

* + 1. Discussion

Analyzing the LLM output for variable definition extraction revealed that most errors were due to definitions not matching the correct ones precisely. Despite the LLM occasionally altering expressions or generating definitions not present in the text, it accurately extracted variable definitions from the nine documents. Variable definition extraction performance may decline for more specialized or niche processes, as Han et al. (2023) reported that the performance of ChatGPT is lower for specialized tasks than for general ones. A method is needed to ascertain when LLMs should be employed for variable definition extraction and to devise a proficient extraction technique for expertise-demanding processes.

For variable definition equivalence judgment, we provided the symbols for the two variable definitions and the equations incorporating those symbols. Although judging equivalence solely on variable definitions proved challenging, the optimal input strategy remains undetermined. Enhancing LLM performance could involve incorporating more context of the definitions into the prompts or leveraging external knowledge, which warrants further investigation. Additionally, the equivalence judgment method becomes computationally expensive as the number of documents increases, since the method judges the equivalence of all possible combinations. Developing a more efficient method for processing a larger number of documents promptly is essential for future progress.

To assess model candidate building, we manually standardized the expressions in two documents (1 and 2 in Table 1). After the standardization, the prototype was able to yield the required physical models that could compute all variables within the models. Presently, the algorithm requires users to specify variables; future improvements will aim to enable the method to build models without user-specified variables.

* + 1. Challenges

Overcoming three key obstacles is essential for the realization of AutoPMoB.

* + - 1. Document Retrieval

Although this study presupposed the availability of documents relevant to physical modeling, locating such documents in practice is often arduous and financially burdensome. Furthermore, finding documents for physical model building is challenging because current search engines lack the capability to specifically search for mathematical formulas. The goal is to create a system that efficiently retrieves the necessary documents for model construction tailored to the target process and the objectives of model building.

* + - 1. Integration of Non-literature Information

Not all literature, including academic papers and textbooks, provides the complete data needed for physical model building. In instances where literature falls short, the incorporation of non-literature data, like ontologies and other databases, becomes valuable. Additionally, the nature of information, such as its novelty and reliability, varies by source, necessitating strategies to manage each type of data effectively.

* + - 1. Adaptation for unknown processes

Information extraction and equivalence judgment performance of NLP-based methods tend to deteriorate when applied to unfamiliar processes, such as new or specialized ones. Where similar data exists, transfer learning may offer a viable solution; hence, a major challenge is to identify how documents about different processes relate and to devise a method capable of delivering high performance even in the absence of target process data.

* 1. Conclusions

In pursuit of establishing automated physical model builder, AutoPMoB, we devised a prototype leveraging a large language model (LLM). This prototype is capable of extracting necessary information, such as variables and equations, standardizing their notations, and generating model candidates. We tested the prototype on a continuous stirred tank reactor and assessed its effectiveness. The prototype successfully extracted all variable symbols and equations and achieved an accuracy of 74% in variable definition extraction. Nonetheless, the process for judging the equivalence of variable definitions was largely ineffective, with over half of the equivalent definitions failing to be identified correctly. However, when the equivalence judgment was accurate and the expressions were standardized, the algorithm could yield the intended model candidates. Future efforts will focus on enhancing the efficacy of each core technology to fulfill the vision of AutoPMoB. Additionally, we plan to develop strategies for sourcing documents essential for model building, leveraging non-literature information, and maintaining the performance of fundamental technologies when dealing with unfamiliar processes.

* 1. Acknowledgments

This work was supported by JSPS KAKENHI Grant Number JP23K13595 and the Chubei Itoh Foundation.

References

S. Kato and M. Kano, 2022, Towards An Automated Physical Model Builder: CSTR Case Study, Computer Aided Chemical Engineering, Volume 49, Pages 1669–1674.

S. Kato, C. Zhang, and M. Kano, 2023, Simple Algorithm for Judging Equivalence of Differential-Algebraic Equation Systems, Scientific Reports, Volume 13, Page 11534.

S. Kato and M. Kano, 2023, Efficient Physical Model Building Algorithm Using Equations Extracted from Documents, Computer Aided Chemical Engineering, Volume 52, Pages 151–156.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, 2017, Attention Is All You Need, Advances in Neural Information Processing Systems, Volume 30.

R. Han, T. Peng, C. Yang, B. Wang, L. Liu, and X. Wan, 2023, Is Information Extraction Solved by ChatGPT? An Analysis of Performance, Evaluation Criteria, Robustness and Errors, arXiv preprint arXiv:2305.14450.

M. Schubotz, L. Krämer, N. Meuschke, F. Hamborg, and B. Gipp, 2017, Evaluating and Improving the Extraction of Mathematical Identifier Definitions, Experimental IR Meets Multilinguality, Multimodality, and Interaction, Pages 82–94.

T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, 2020, Language Models are Few-Shot Learners, Advances in Neural Information Processing Systems, Volume 33, Pages 1877–1901.

M. Suzuki, F. Tamari, R. Fukuda, S. Uchida, T. Kanahori, 2003, INFTY: An Integrated OCR System for Mathematical Documents. Proceedings of DocEng’03, Pages 95–104.