複数文献間の変数の同義性判定に向けた ProcessBERT の構築

金上 和毅　加藤 祥太　加納 学
京都大学大学院情報学研究科
{kanegami.kazuki.55z,katou.shouta.23v}@st.kyoto-u.ac.jp
manabu@human.sys.i.kyoto-u.ac.jp

概要

プロセス産業において重要な役割を果たす物理モデルの構築には，膨大な量の文献調査や精度向上のための試行錯誤などの多大な労力が要求される．そこで著者らは，物理モデルを自動で構築するシステムの開発を目指し，複数の要素技術の開発に取り組んでいる．本研究では，その内の一つである複数文献間における変数の同義性を判定する手法を開発するとともに，その精度向上に向け，化学工学ドメインに特化した言語モデル（ProcessBERT）を構築した．化学プロセスに関する論文計 28 報を用いて，ProcessBERT と他の言語モデルの同義性判定の性能を比較した結果，ProcessBERT が他モデルを上回る性能を発揮した．

1　はじめに

化学や鉄鋼などのプロセス産業では，プロセスの設計や運転に物理モデルに基づくプロセスシミュレータが用いられている．しかし，物理モデルの構築にはプロセスに関する深い理解と専門知識だけでなく，精度向上のための試行錯誤的な取り組みが必要とされる．本研究の最終目的は，文献データベースから必要な文献を収集し，それらからモデル構築に必要な情報（数式，変数，実験データ）を抽出し，所望のモデルを構築するという工程を自動でおこなうシステム（Automated Physical Model Builder; AutoPMoB）を開発することである．著者らは，AutoPMoB の実現に向けて，複数の要素技術の開発に取り組んでいる [1]．本研究では，その内の一つである複数文献間における変数の同義性を判定する手法を開発する．

2018 年に Devlin らによって提案された BERT[2] は，様々な自然言語処理タスクにおいて当時最高の精度を達成した．以降，精度を維持しつつ大幅なパラメータ削減を実現した ALBERT[3] や，より大規模なコーパスを用いた学習をおこなうことで大幅な精度向上を達成した RoBERTa[4] など，様々な改良型モデルが提案されている．また，ある特定のドメインにおける自然言語処理タスクを解く場合，Wikipedia などの一般的なドメインのテキストを用いて構築された言語モデルよりも，その特定のドメインのテキストを用いて構築された言語モデルの方が優れた精度を達成できることが報告されている [5, 6, 7, 8]．

本稿では，化学工学ドメインに特化した言語モデルである ProcessBERT を構築する．そして，言語モデルの埋め込みベクトルを用いる手法とファインチューニング済みモデルを用いる手法で同義性判定をおこない，ProcessBERT の性能を検証する．また，オリジナルの BERT と科学技術関連のテキストで学習した SciBERT[7] と性能を比較する．

2　ProcessBERT

2.1　コーパス

Elsevier 社が提供する Elsevier Research Product APIs[9] を用いて，化学工学に関連する 17 のジャーナルから約 13 万報の論文を収集した．使用したジャーナルとジャーナルごとの論文数を表 1 に示す．収集した論文から論文の抄録と本文（図表を除く）にあたる部分のテキストを抽出して化学工学ドメインに特化したコーパス（ChemECorpus）を構築した．ChemECorpus にはタイトル，著者情報，キーワード，参考文献，付録は含まれない．ChemECorpus の総単語数は約 7 億語（4.0GB）である．

2.2　事前学習

事前学習済みの $\text{BERT}_{\text{BASE}}$[2] に対し，ChemECorpus を用いて以下の手順で追加の事前学習をおこなった．

表 1　ChemECorpus 構築に用いたジャーナルと論文数

ジャーナル	論文数
Applied Catalysis B Environmental	10,727
Carbohydrate Polymers	16,361
Chemical Engineering and Processing - Process Intensification	3,935
Chemical Engineering Journal	27,222
Chemical Engineering Research and Design	5,375
Chemical Engineering Science	13,527
Chinese Journal & Catalysis	2,709
Computers & Chemical Engineering	6,584
Current Opinion in Chemical Biology	2,201
Journal of Catalysis	10,248
Journal of Cleaner Production	26,994
Journal of Energy Chemistry	2,236
Journal of Process Control	2,744
Progress in Crystal Growth and Characterization of Materials	256
Progress in Polymer Science	1,017
South African Journal of Chemical Engineering	233
Thermal Science and Engineering Progress	950
合計	133,319

1. バッチサイズを 64 とし，最大長が 128 のシーケンスを用い，2 つの事前学習タスク（Masked Language Model，Next Sentence Prediction）を 900,000 ステップにわたり実行した．
2. バッチサイズを 8 とし，最大長が 512 の長いシーケンスを用い，1. と同じタスクを 100,000 ステップにわたり追加で実行した．

また，学習回数の違いによるモデルの性能差を検証するため，ステップ数のみを 2 倍にして事前学習をおこなったモデル（$\text{ProcessBERT}_{\text{double}}$）も構築した．計算には Google Cloud Platform[10] で利用できる 8 コアの TPU v3 を使用した．ProcessBERT の事前学習には約 13 時間を要した．

事前学習には，Devlin らが Github 上で公開しているプログラム [11]（run_pretraining.py）を使用した．また，事前学習時に指定する語彙やハイパーパラメータは，$\text{BERT}_{\text{BASE}}$ と同じものを使用した．

3　実験

3.1　データセット

晶析プロセス（Crystallization; CRYST），連続槽型反応器（Continuous Stirred Tank Reactor; CSTR），多管式熱交換器（Shell and Tube Heat Exchanger; STHE）に関する論文をそれぞれ 11，10，7 報収集し，本文中から変数定義に相当する名詞句を抽出した．次に，同一プロセスの異なる 2 つの論文間ですべての変数定義のペアを作成し，ペアごとに同義（1）もしくは非同義（0）のラベルを手動で付与した．プロセスごとの同義と非同義のペアの数を表 2 に示す．いずれのプロセスも，同義のペアの割合が小さい不均衡データであるため，以下のようにして各プロセスのトレーニング用およびテスト用のデータセットを作成した．

トレーニング用　3.2.2 の検証における学習回数を一定にするため，データの総数が 2,500 になるように非同義のペアをランダムにサンプリングする．

テスト用　同義のペアの数が全体の 10%になるように非同義のペアをランダムにサンプリングする．

表 2　各プロセスの論文数と同義および非同義のペア数

	論文数	同義	非同義
CRYST	11	54	12,200
CSTR	10	122	4,693
STHE	7	61	13,720

3.2 検証方法

ProcessBERT および ProcessBERT$_{\text{double}}$ の性能を検証するため，BERT$_{\text{BASE}}$ および SciBERT[7] を比較対象とし，以下の 2 通りの検証をおこなう．

3.2.1 変数定義の類似度による同義性判定

図 1 に示すように，モデルから得られる単語の埋め込みベクトルを使用して 2 つの変数定義の類似度を計算し，その同義性を判定する．

まず，2 つの変数定義を表すベクトルを以下の手順で得る．

1. 変数定義にあたる名詞句をモデルに入力し，モデル内に 12 層ある Transformer Encoder から，ストップワード（冠詞，前置詞，等位接続詞など）に該当しない単語の埋め込みベクトルを抽出する．抽出した単語の数を n 個，j 層目の Transformer Encoder から出力された i 番目の単語の埋め込みベクトルを $\boldsymbol{v}_{i,j}$ とする．$(1 \leq i \leq n, 1 \leq j \leq 12)$
2. 変数定義を表すベクトル $\boldsymbol{d}$ を以下の式にしたがって計算する．

$$\boldsymbol{d} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{12}\sum_{j=1}^{12}\boldsymbol{v}_{i,j}\right)$$

次に，得られた 2 つの変数定義を表すベクトルのコサイン類似度（変数定義の類似度）を算出する．この値が閾値を超えた場合，2 つの変数が同義であると判定する．

3.2.2 ファインチューニング済みモデルによる同義性判定

3.1 のトレーニング用データセットのうち 2 プロセス分を用いてモデルのファインチューニングをおこなう．次に，得られたモデルを用いて残り 1 プロセスのテスト用データセットに対する性能を検証する．これを計 3 回おこない，すべてのプロセスで検証する．

なお，ファインチューニングには，Devlin らが提供するプログラム [11]（run_classifier.py）を使用する．2 つの名詞句が同義であるか否かを分類するというタスクは，Microsoft Research Paraphrase Corpus（MRPC）[12] を用いたタスクとほぼ同様の内容であるため，run_classifier.py 実行時の引数（TASK_NAME）には"MRPC"を指定する．

MRPC タスクにおける BERT モデルのファインチューニング時の処理を図 2 に示す．まず，2 つの変数定義の名詞句を [SEP] トークンで繋ぎ，先頭に [CLS] トークンを付加した単語列をモデルに入力する．その後，[CLS] トークンに対応する最終層の埋め込みベクトル $C(\in \mathbb{R}^{768})$ と重み行列 $W(\in \mathbb{R}^{2\times768})$ から，「同義」および「非同義」の 2 クラスの予測値 $\text{softmax}(CW^T)$ を計算する．「同義」のクラスの予測値が「非同義」のクラスの予測値より大きい場合，2 つの変数が同義であると判定する．

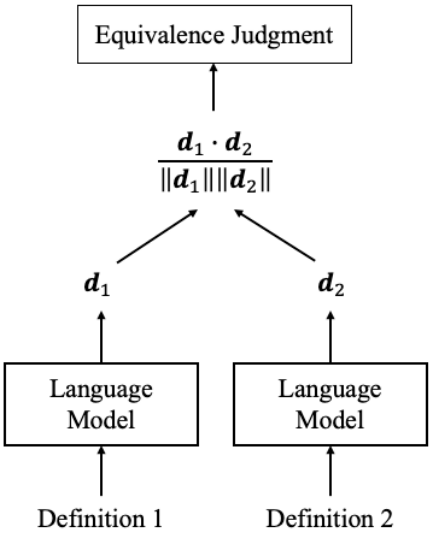

図 1 変数定義の類似度による同義性判定

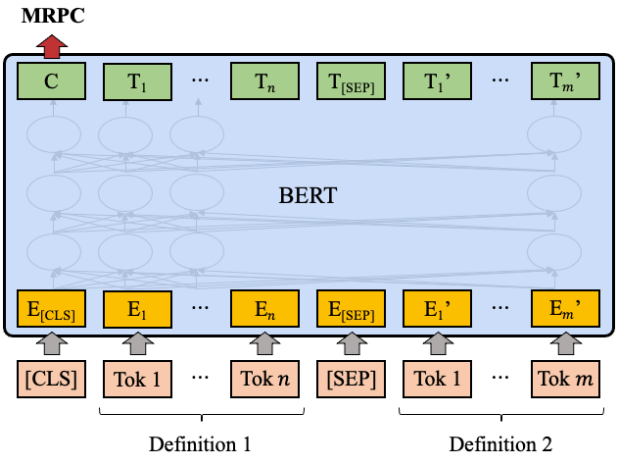

図 2 MRPC タスクにおける BERT モデルのファインチューニング時の処理 [2]

3.3 結果と考察

3.3.1 結果

変数定義の類似度による同義性判定の検証結果として，Youden's Index[13] を閾値とした時の F1 値を表 3 に示す．CRYST のデータセットに対しては，

SciBERT が最高値となったが，他の 2 プロセスおよび全プロセスをまとめたデータセット（All）に対しては，ProcessBERT が最高値を達成した．

ファインチューニング済みモデルによる同義性判定の検証結果として，F1 値を表 4 に示す．いずれのデータセットに対しても，ProcessBERT，$\text{ProcessBERT}_{\text{double}}$ ともに，最高値が他モデルを下回った．

表 3 変数定義の類似度による同義性判定の検証における各モデルの F1 値

Model	CRYST	CSTR	STHE	All
ProcessBERT	0.752	**0.642**	**0.726**	**0.653**
$\text{ProcessBERT}_{\text{double}}$	0.658	0.569	0.667	0.590
$\text{BERT}_{\text{BASE}}$	0.730	0.537	0.671	0.567
SciBERT	**0.766**	0.631	0.699	0.622

表 4 ファインチューニング済みモデルによる同義性判定の検証における各モデルの F1 値

Model	CRYST	CSTR	STHE	All
ProcessBERT	0.660	0.270	0.790	0.552
$\text{ProcessBERT}_{\text{double}}$	0.725	0.137	0.842	0.557
$\text{BERT}_{\text{BASE}}$	0.652	**0.336**	0.825	**0.583**
SciBERT	**0.827**	0.094	**0.855**	0.579

3.3.2 考察

事前学習用に構築した ChemECorpus のサイズ（7 億語）は，過去に提案された多くのドメイン特化 BERT モデルの事前学習に使われたコーパスと比較して小さい（SciBERT[7]: 32 億語，BioBERT[5]: 45 億語，PubMedBERT[8]: 31 億語）．サイズの小さい ChemECorpus では，化学工学ドメインにおける専門的な知識を十分に学習できていない可能性が考えられる．また，先行研究 [8] では，事前学習において一般的なドメインのコーパスで学習された $\text{BERT}_{\text{BASE}}$ を用いずに，一から事前学習をおこなうことで，より性能の高いモデルが得られるとされている．今後，十分なサイズのコーパスを構築し，一から事前学習をおこなうことでより高性能な言語モデルが得られる可能性がある．

ファインチューニング済みモデルによる同義性判定の検証において，CSTR のデータセットに対する F1 値が，他の 2 プロセスの結果と比較して明確に小さくなっている．結果を精査したところ，この検証では 2 つの名詞句を構成する単語がほぼ同一であるペアのみを同義と判定する傾向があった．同義である 2 つの名詞句を構成する単語が大きく異なるペアが他プロセスと比較して多かったことが，CSTR のデータセットにおける性能が低下した原因であると考えられる．これについては，トレーニング用データとテスト用データの分割を各プロセスのデータセットごとにおこない，各プロセスにおける変数定義の表記揺れをモデルが学習することで対応できる可能性がある．その検証をおこなうには，現時点では正例のデータ数が不足しているため，今後は正例のデータ数が十分に確保されたデータセットの構築を進める必要がある．

また，ProcessBERT と $\text{ProcessBERT}_{\text{double}}$ の結果から，学習回数を増やしても ProcessBERT の性能は向上しないことが分かった．この結果は，先行研究 [6] の内容と符合するものである．

4 おわりに

本研究では，複数文献間における変数の同義性判定に向けた一つのアプローチとして，約 13 万報の化学工学関連の論文から構築された ChemECorpus を用いて ProcessBERT を構築した．$\text{BERT}_{\text{BASE}}$ と SciBERT を比較対象として，変数定義の類似度を用いる手法とファインチューニング済みモデルを用いる手法による 2 通りの検証をおこなったところ，前者の手法において，ProcessBERT を用いた場合の同義性判定が最高の精度を達成することが示された．

今後は，正例のデータ数が十分に確保されたデータセットの構築を進めていく．また，化学工学関連のテキストを追加で収集して ChemECorpus を拡張した上で，$\text{BERT}_{\text{BASE}}$ から追加で学習するのではなく一から事前学習をおこなうことで，ProcessBERT の更なる改良を図る．

謝辞

本研究は JSPS 科研費 JP21K18849 および JST 次世代研究者挑戦的研究プログラム JPMJSP2110 の助成を受けたものです．

参考文献

[1] S. Kato and M. Kano. Identifier information based variable extraction method from scientific papers for automatic physical model building. *PSE Asia Paper No. 210043*, 2020.

[2] J. Devlin, M. Chang, K. Lee and K. Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, **1**, pp.4171-4186, 2019.

[3] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma and R. Soricut. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. *arXiv preprint arXiv:1909.11942*, 2019.

[4] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer and V. Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*, 2019.

[5] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C.H. So and J. Kang. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, **36**(4), pp.1234-1240, 2020.

[6] E. Alsentzer, J. Murphy, W. Boag, W.H. Weng, D. Jindi, T. Naumann and M. McDermott. Publicly Available Clinical BERT Embeddings. *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pp.72-78, 2019.

[7] I. Beltagy, K. Lo and A. Cohan. 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)*, pp.3615-3620, 2019.

[8] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao and H. Poon. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *ACM Transactions on Computing for Healthcare*, pp.1-23, 2021.

[9] Elsevier Developer Portal. https://dev.elsevier.com/. (Accessed on 2022/1/12).

[10] Cloud Computing Services | Google Cloud. https://cloud.google.com/. (Accessed on 2022/1/12).

[11] GitHub-google-research/bert: TensorFlow code and pre-trained models for BERT. https://github.com/google-research/bert. (Accessed on 2022/1/12).

[12] W.B. Dolan and C. Brockett. Automatically constructing a corpus of sentential paraphrases. *In Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*, 2005.

[13] W.J. Youden. Index for rating diagnostic tests. *Cancer*, **3**(1), pp. 32-35, 1950.