ドメインに特化した比較的少量のデータによる事前学習済み BERT の利用可能性：鉄鋼業における事例

岩月憲一
日本製鉄株式会社
iwatsuki.pz4.kenichi@jp.nipponsteel.com

概要

BERT を特定の業種に特化した場面で用いる際には、汎用のモデル、汎用のモデルに追加学習を施したモデル、フルスクラッチで事前学習したモデルのいずれかの選択肢がある。事前学習を行えば分野に特化したモデルができるが、分野が狭くなるほど用意できるデータは少なくなる。本研究では、鉄鋼業に特化したデータを用いて追加学習・事前学習を行い、ファインチューニングを要する特許分類タスクと、要しない語義曖昧性解消タスクで評価した。実験の結果、学習データが比較的少ないとしても、事前学習したモデルが優位になる場合があることが分かった。

1 はじめに

デジタルトランスフォーメーションが叫ばれる中、製造業においても自然言語処理の需要は高まっている。BERT[1]をはじめとする Transformer ベースの言語モデルは、タスクに応じてアーキテクチャを大きく変更することなく高い性能が発揮できる点で魅力的である。

しかし、Wikipedia やウェブコーパス等で事前学習された汎用の言語モデルを特定の業種におけるタスクに適用した場合に十分な性能が発揮されるとは限らない。そのため、学術論文[2]や法律文書[3]を用いて事前学習されたモデルが提案されている。この観点から、業種に特有の文書によって事前学習を行うことが望まれる。

その一方で、データサイズの問題が生じ得る。BERT 以降、事前学習に用いられるデータサイズは増加している。しかし業種を狭めれば狭めるほど利用可能なデータは少なくなる。したがって、より多くのデータで学習された汎用のモデルを用いるか、より少ないデータで学習された業種特化型のモデルを用いるということになる。

さらに、汎用の事前学習済みモデルに対してドメインに特化したデータで追加学習を行うことにより性能の向上が見られたという報告がある[4]。よって 3 つ目の選択肢として、追加学習によるモデルを用いることが挙げられる。

本研究では、いずれの選択肢をとるべきか検討するため、鉄鋼業に特化した文書を用いて BERT の事前学習と、日本語の汎用 BERT モデルに対する追加学習を行った。事前学習にあたっては、専門用語の分割が不必要に行われることを防ぐため、SentencePiece[5]を用いてトークン化を行った。

評価のために、鉄鋼分野の知識を必要とするタスクを 2 つ用意した。1 つ目は、特許分類タスクである。これは、ラベル付きの特許抄録を用いてファインチューニングをし、分類の精度を比較するものである。

2 つ目は、教師なし語義曖昧性解消タスク[6]である。BERT はもともとファインチューニングをすることを前提として提案されたモデルであるが、事前学習済みモデルをそのまま用いて文ベクトルを得ることも行われている[7]。企業活動の中で産出される文書に何らかの正解ラベルが付与されていることは少ない上に、専門性が高いためアノテーションコストが大きい。こうした背景から、ファインチューニングを要しないタスクでの評価も行うことにした。このタスクは、BERT が Masked Language Model（MLM）で学習されていることを踏まえ、曖昧性を持つ語を[MASK]に置き換えて同分野の語が推定されるかを確認するものである。鉄鋼分野において使われる略語のうち、語義が複数あるものを対象にした。

実験の結果、特許分類については事前学習モデル、追加学習モデルともに汎用モデルと同等以上の性能を発揮した。また、語義曖昧性解消については、事前学習モデルが最も良い性能を、追加学習モデルがそれに次ぐ性能を示した。

以上の結果から、比較的少ないデータであったとしても、業種に特化したデータを用いて事前学習モデルを作成することの有用性が示唆された。

2 手法

2.1 モデル

本研究では、汎用モデル、追加学習モデル、事前学習モデルの 3 つの BERT モデルを使用した。汎用モデルには、cl-tohoku/bert-base-japanese-whole-word-masking を使用した。このモデルに鉄鋼分野のデータ（後述）で追加学習を行ったモデルを追加学習モデルとした。

事前学習モデルには、SentencePiece によるトークン化を採用した。これは既存の辞書ベースのトークナイザでは専門用語への対応が難しいと判断したためである。語彙の大きさは 30,000 である（汎用モデルの語彙は 32,000 語でありほぼ同等である）。

BERT の実装には huggingface/transformers[8]を使用した。また、事前学習と追加学習ともに MLM のみであり、Next Sentence Prediction は行っていない。

2.2 事前学習用データ

事前学習用のデータとして、鉄鋼・非鉄各社の技報および鉄鋼各社の特許公報を使用した。テキストの総量は 767MB であった。汎用モデルは日本語版 Wikipedia によって事前学習されているが、そのデータサイズは 2.6GB である[i]。従って、事前学習モデルのデータは汎用モデルのデータの 3 割程度の大きさしかない。

日本語版 Wikipedia にも鉄鋼分野の記事は少なからず存在しているが、その総量は全体の 3 割に満たないと考えられる。

3 評価タスク

3.1 特許分類

本タスクは、特許抄録を入力とし、その特許が属する技術分野を出力する分類タスクである。1 つの特許抄録は複数の分野に属しうるため、マルチラベル分類問題である。

今回使用するデータとして、鉄鋼分野に属する 36,415 件の特許抄録を抽出した。このうち約 1 割を評価用データとし、残りを訓練用データとした。それぞれの抄録にラベルが専門家の手によって付与されており、「製鋼-鋼精錬」「鋼管-油井管」などがある。

なお、事前学習モデルおよび追加学習モデルはその事前／追加学習データに特許公報が含まれているが、本タスクのデータと全く同じではなく、一部重複している。

また、入力は先頭の 256 トークンのみを使用しており、それ以降のテキストがある場合は無視されている。

訓練用データを用いて、5 分割交差検証によりパラメータチューニングを行った。その後、評価用データを用いて評価を行った。

評価は、本タスクがマルチラベル分類問題であるため、次の指標により行った。1 つ目は、正解率である。1 つの文書に対し過不足なくすべてのラベルが正しく推定された場合に正解とみなした。2 つ目は、各分類ラベルの F 値である。これについてはミクロ平均とマクロ平均の両方を算出した。

ファインチューニングは、3 種類の BERT それぞれに線形層を加えて行った。BERT の出力のうち[CLS]トークンを線形層に入力した。

3.2 語義曖昧性解消

本タスクは、同じ表記であるが語義の異なる言葉の語義を推定するために、ファインチューニングを行わずに BERT の MLM を用いるというものである[6]。先行研究[6]では一般の日本語についての語義曖昧性解消を目的としていたが、本研究ではこれを鉄鋼分野に適用するため、改変を加えた。

タスクの詳細を実例で示す。例えば、ASR という単語には、「自動速度制御」（Automatic Speed Regulation）という意味と、「自動車破砕後の残留物」（Automobile Shredder Residue）という意味がある。いずれの語義であるかは文脈によって判断される。この 2 語は属する技術分野が異なるため、語義というよりもどの分野の語であるかが分かれば十分である。これを踏まえ、この語を含む文のうち当該語を[MASK]トークンに置き換え、BERT に[MASK]に入る語を予測させる。予測結果として得られた語の示す技術分野が、元の語の属する技術分野であると推定する。上記の例では、ASR に対して、制御に関す

[i] https://huggingface.co/cl-tohoku/bert-base-japanese-whole-word-masking

る語「AGC」（Automatic Gauge Control）が予測されれば制御分野として推定される。

今回は、[MASK]を埋める語として推定された語のうち最も確率の高い 1 語の示す分野が同じである場合に正解とみなした。ただし、元の語と全く同じ語が推定された場合は、どの技術分野かわからないため、次点の語を用いた。

本タスクの対象となる曖昧な語には、略語を用いた。「鉄鋼実務用語辞典」[9]、「鉄の事典」[10]等に収録されている略語から、表記が同じであるが、属する技術分野の異なる略語を 10 件選択した。1 件につき、2 つの技術分野が該当するようにした。

略語を含む文は、各語各技術分野につき 5 文を論文や社内技術文書から抽出した。合計 100 文を用いて本タスクを行った。

本タスクはファインチューニングを必要としないため、事前学習／追加学習済みのモデルに対してそのまま文を入力した。

4 結果と考察

4.1 特許分類

特許分類の結果を表 1 に示す。いずれの指標でみても、汎用モデルよりも追加学習・事前学習モデルの方がやや上回る結果となった。

表 1 特許分類の結果

モデル	正解率	Mic. F 値	Mac. F 値
汎用	0.563	0.716	0.478
追加学習	0.585	0.726	0.505
事前学習	0.592	0.738	0.506

図 1 に例としてある特許抄録に各モデルのアテンションの重み（最終層のものの合計）を可視化したものを示す。上から順に汎用モデル、追加学習モデル、事前学習モデルである。背景色が濃い方が重みが大きいことを示している。また、同時にトークン化の違いも見える。事前学習モデルのみ SentencePiece を用いており、「電着塗装後の」や「耐食性に優れた」が 1 つのトークンになっている。これに対し、汎用モデルのトークナイザによるトークンは「電/着/塗装/後/の」や「耐/食/性/に/優れ/た」とかなり細かくなっている。結果的にこうした重要語の重みが小さくなっている。

事前学習に用いたデータが比較的小さいながらも、特許文献に頻出する語句をうまく扱えたことで汎用 BERT と同等以上の性能が出ていると考えられる。

4.2 語義曖昧性解消

表 2 に語義曖昧性解消の結果を示す。汎用モデルより追加学習モデル、さらに事前学習モデルが良い性能を示した。

表 3 は､事前学習モデルのみが正解した例である。汎用モデルは、「事故」という語を予測している。「事故が発生する」という日本語は自然な表現であるが、鉄鋼に関する文脈であることが十分に認識されていないためにこうした推定がなされたものと考えられる。

表 4 は、すべてのモデルが正解した例である。Barrel per day という単位に対して、3 つとも単位を推定できている。このように、鉄鋼分野に限らず広く多分野で使われる語句については、汎用モデルでも十分に正解できることが分かる。

表 5 も、すべてのモデルが正解した例であるが、表 4 の例に比べて鉄鋼分野に寄っている文である。このように専門的な文脈であっても汎用モデルが正解できる例もある。逆に、表 6 は全てのモデルが不正解となった例であるが、事前学習モデルと追加学習モデルは元素記号を推定しているのに対し、汎用モデルは企業名の文脈と認識している。「インテル」という語は事前学習モデルの中にはないため、間違えるにしても分野を大きく外すことはない。このように、入力文から分野を推定しにくい場合に汎用モデルは不利である。こうした点から、特定の分野のデータで学習させることの意義が確認できる。

では、追加学習モデルと事前学習モデルではどうか。表 3 や表 6 の誤りの例を見ると、汎用モデルと比較して追加学習モデルは全く関係ない別の分野の語を出してしまうことはある程度避けられていることが分かる。ただし、それでも事前学習モデルには及んでおらず、中間的な性能である。

以上の議論を踏まえると、とりわけファインチューニングを行わずに BERT の文・単語ベクトルを使用する場面にあっては、必ずしも汎用モデルや追加学習モデルが優れるとは限らないと言える。特定の分野に特化した文脈では、全く異なる分野の文脈として捉えられてしまう可能性がある。特定業種に特化したデータで事前学習を行うことで回避できると考えられる。

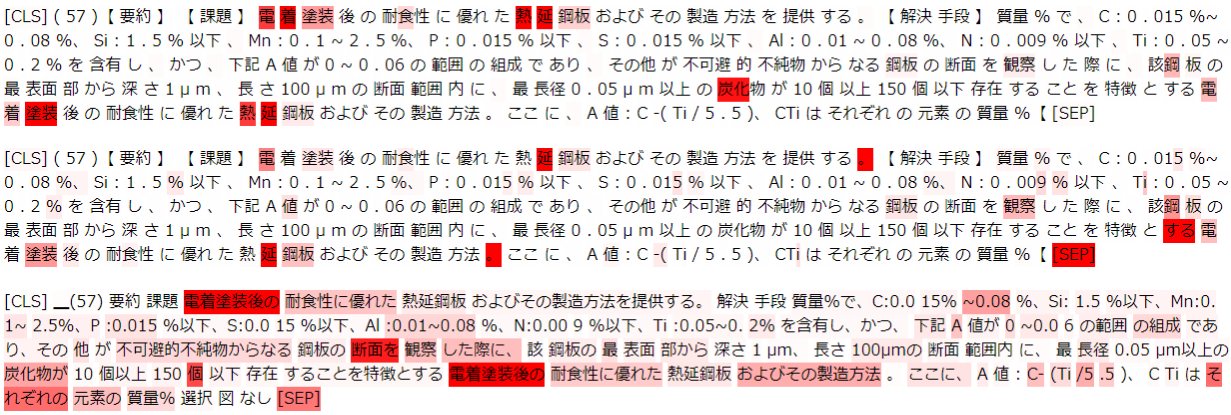

図 1　特許抄録[11]に対する各モデルのアテンションの重みとトークン化の結果

表 2　語義曖昧性解消の結果

モデル	正解率
汎用	0.15 (15/100)
追加学習	0.24 (24/100)
事前学習	0.37 (37/100)

表 3　事前学習モデルのみが正解した例

破砕選別業者はプレスまたはシュレッダー処理を行い，後者では[MASK]が発生する。 [12] [MASK]＝ASR	
汎用	事故
追加学習	スクラップ
事前学習（正解）	シュレッダーダスト

表 4　すべてのモデルが正解した例

処理容量は 70,000[MASK]で出荷はシャトルタンカーにより行われる。 [13] [MASK]＝BPD（=barrel per day）	
汎用（正解）	t
追加学習（正解）	t
事前学習（正解）	m3

表 5　より専門的な文脈であってもすべてのモデルが正解している例

以上の 2 回の実験結果より、[MASK]から鋳鉄に加炭される炭素量は加炭材の約 1/4 であり、小型高周波誘導溶解炉での実験結果と差異はなかった。 [14] [MASK]＝BIC（=bio coke）	
汎用（正解）	石炭
追加学習（正解）	石炭
事前学習（正解）	コークス

表 6　すべて不正解の例

結果は表 2 の通りだが，c-AIC の選択パフォーマンスは AIC や[MASK]のそれと比べて良くなっている． [15] [MASK]＝BIC（= Bayesian information criterion）	
汎用	インテル
追加学習	Si
事前学習	Pb

5　おわりに

本稿では、特定の業種に特化した場面において BERT を用いる際に、汎用のモデル、汎用のモデルに追加学習を施したモデル、特定業種のデータで事前学習をしたモデルのいずれを使用するのが良いかを鉄鋼業を例に検討した。鉄鋼業に特化したテキストデータを用いて事前学習ならびに追加学習を行い、3 種類のモデルを用意した。

ファインチューニングを要する特許分類タスクと、要しない語義曖昧性解消タスクによって各モデルの評価を行った。実験の結果、特許分類タスクでは事前学習モデルが汎用モデルと同等以上の性能を発揮し、語義曖昧性解消タスクでは汎用モデルよりも追加学習モデルが、さらに事前学習モデルが良い性能を発揮した。

これらの結果から、学習に使えるデータが比較的少ないとしても、特定の業種に特化したモデルを事前学習させることには有用性があると言える。

今後は BERT 以外の言語モデルについても比較検討を加えたい。

参考文献

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. 2018. 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*, 4171–4186.

[2] Iz Beltagy, Kyle Lo, Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific Text. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, 3615–3620.

[3] 星野玲那, 狩野芳伸. 2020. 司法試験自動解答を題材にしたBERTによる法律分野の含意関係認識. 言語処理学会第 26 回年次大会発表論文集, 577–580.

[4] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, Noah A. Smith. 2020. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8342–8360.

[5] Taku Kudo, John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 66–71.

[6] 新納浩幸, 馬ブン. 2021. BERT の Masked Language Model を用いた教師なし語義曖昧性解消. 言語処理学会第 27 回年次大会発表論文集, 1039–1042.

[7] Han Xiao. 2018. Bert-as-service. https://github.com/hanxiao/bert-as-service

[8] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, ... Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45.

[9] 鉄鋼新聞社. 2006. 新版鉄鋼実務用語辞典.

[10] 朝倉書店. 2014. 鉄の事典.

[11] 新日本製鐵株式会社. 2006. 電着塗装後の耐食性に優れた熱延鋼板およびその製造方法. 特開2006-241539.

[12] 山末英嗣, 松八重一代, 中島謙一, 醍醐市朗, 石原慶一. 2014. 使用済み自動車から得られる鉄スクラップの関与物質総量. 鉄と鋼, *100*(6), 778–787.

[13] 橋本康正. 1984. 海洋における浮遊式･移動式生産システムの海外の現況. 石油技術協会誌, *49*(5), 332–335.

[14] 冨田義弘, 尾鼻美規, 井田民男. 2014. 高周波誘導溶解におけるバイオコークスの加炭材代替効果の検証. スマートプロセス学会誌, *3*(5), 289–294.

[15] 庄野宏. 2006. モデル選択手法の水産資源解析への応用. 計量生物学, *27*(1), 55–67.