(18) 土木分野における事前学習モデル BERT による精度検証

箱石　健太[1]・一言　正之[2]・菅田　大輔[3]

[1]非会員　日本工営株式会社　中央研究所（〒300-1259 茨城県つくば市稲荷原 2304）
E-mail: hakoishi-kn@n-koei.jp

[2]正会員　日本工営株式会社　中央研究所（〒300-1259 茨城県つくば市稲荷原 2304）
E-mail: hitokoto-ms@n-koei.jp

[3]非会員　日本工営株式会社　中央研究所（〒300-1259 茨城県つくば市稲荷原 2304）
E-mail: sugeta-ds@n-koei.jp

国土交通データプラットフォームにより，様々なデータが集約され，それらを活用し様々なイノベーションの促進等や効率化が実現可能となってきている．しかし我々が日常的に使用する土木分野における工事情報や巡視点検に関わる文章に対して，既往の自然言語処理技術による精度は十分ではない可能性がある．本研究では，土木に関連する文章を BERT に学習させ土木 BERT を構築した．既往 BERT と土木 BERT の精度検証を実施し，土木 BERT の優位性を示し，土木の文章の学習が有効であることを確認した．

Key Words: *bert,civil engineering,natural language processing,deep larning,domain adaptation*

1. はじめに

国土交通省は，令和元年 5 月に「国土交通データプラットフォーム（仮称）整備計画」を公表し，各種観測データや維持管理データといった様々なデータの集約を計画している [1]．これらのデータを利用し，効率化やイノベーションの促進等を実現するために，機械学習技術の活用が期待される．

しかし観測データのような数値情報ではなく，我々が日常的に使用する言語，つまり公共工事や巡視点検に関わる維持管理などで使用する土木分野の文章には，「インサート」，「犬走り」，「いわし」といった一般的に使用される用語とは意味が異なる用語が出現する．これらは文脈を考慮し，その用語の意味を捉える必要があり，これらの情報を有効活用するためには自然言語処理の技術が必要となってくる．

近年，自然言語処理技術の発展により，文脈に応じた単語の分散表現を獲得できる BERT[2]が開発され，多くの自然言語処理のタスクで優れた性能が示された．日本語に対応した BERT は京都大学[3]，東北大学[4]により公開されているが，これらは日本語 Wikipedia を事前に学習したモデルであるため，土木分野への適用性が高くない可能性がある．しかし BERT は専門分野の文書を事前に学習することで，その分野における精度向上が報告されている [5][6][7]．

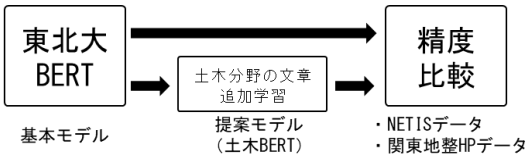

図-1　本研究の流れ

本研究では，土木に関する文章を用いて，東北大 BERT に対し再事前学習を実施したモデルを構築した（土木 BERT）．次に NETIS[8]に掲載されている新技術と関東地方整備局に掲載されている技術情報 [9]から教師データを作成し，東北大 BERT と土木 BERT の精度を比較し，土木 BERT の有効性を検証した．本研究の流れを**図-1** に示す．

表-1　事前学習のデータ件数

データ名	文章数
土木学会発行の論文	14,806
道路・河川に関わる点検要領 情報化施工	3,975
各都道府県の 土木工事共通仕様書	202,965
建設白書	688
国土交通白書	15,691
合計	238,125

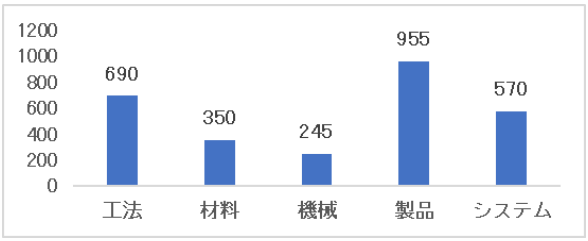

図-2　NETIS データセット件数

2.　土木 BERT

土木 BERT は日本語を事前に学習した東北大 BERT に対し，追加で事前学習を行う．使用した東北大 BERT と土木 BERT について説明する．

(1)　東北大 BERT

BERT は形態素解析後の各単語に対し，12 層の BertLayer により，文章内の離れた単語であっても，より深く文脈を考慮し，単語間の共通点や類似性などの分散表現を獲得できるモデルである．BERT は入力の一部をマスクして予測するタスクと，与えられた 2 文が連続した文であるかを判別するタスクによって事前学習を行う．

公開されている日本語学習済み BERT のうち，本研究では今後の汎用性や拡張性を考慮し，自然言語処理に特化した深層学習フレームワーク HuggingFace[10]に対応した東北大学乾研究室が公開しているモデル[4]（cl-tohoku/bert-base-japanese-whole-word-masking）を使用した．

(2)　土木 BERT

東北大 BERT に対して，土木分野に関わる文章の再事前学習を行った．一般的な文章を学習した BERT に対して，特定分野の文章を再事前学習することでモデルの精度の改善が報告されており[5][6]，土木 BERT が再事前学習により土木の知識を獲得することが期待できる．

再事前学習に使用したデータと内訳を**表-1** に示す．なお，図表に含まれる文字や極端に文章が短く意味を成さない文章があったため，それらを除外するため，有価証券報告書のテキストデータ抽出を参考に[11]，試行錯誤的に以下の条件を設定し，文章を抽出した．

・「名詞」で始まり，「句点」で終わる．
・形態素が 20 以上で構成されている．
・文字化けしていない．

a)　土木学会発行の論文

J-STAGE[12]で公開された土木学会が発行している 71 誌の日本語で執筆された論文の抄録を対象とした．論文中の計算式や PDF からテキストファイルに変換する過程で文章が正しく抽出できなかった論文があったため，対象は抄録に留まっている．

b)　道路・河川に関わる点検要領，情報化施工

国土交通省の HP[13]に掲載されている維持管理に関わる施策，提言，点検要領，監督・検査要領，ガイドライン等を対象とした．

c)　各都道府県の土木工事共通仕様書

各都道府県の HP に掲載されている土木工事に関わる共通仕様書，施工管理基準等を対象とした．

d)　建設白書，国土交通白書

国土交通省の HP[13]に掲載されている平成 8 年から令和 3 年度までの建設白書，国土交通白書を対象とした．建設白書は既にテキスト化されたデータが公開されているため，そのデータを使用した．国土交通白書は PDF 版と HTML 版があり，テキストデータの抽出効率の観点から HTML 版に掲載されている文章を対象とした．

なお再事前学習時のハイパーパラメータは，東北大 BERT のパラメータを設定し，表-1 の各データ毎に 9 : 1 の割合で学習，テストデータに分割し学習を行った．GPU（NVIDIA GeForce RTX 3090）を使用し，学習回数は EarlyStopping 機構により，テストデータの損失関数の値が上昇し始めた 170 回時点で停止させた．なお計算時間は約 27 時間を要した．

3.　精度検証用のデータセット

東北大 BERT と土木 BERT の精度を比較するため，以下のデータセットを作成した．

(1)　NETIS

NETIS には様々な新技術の名称，概要文，区分，受賞歴などが掲載されている．概要文は特徴が明確にわかるよう 127 文字以内にまとめられており，区分は「工法，材料，機械，製品，システム」の 5 つのいずれかに分類され，概要文を読むことでどの区分に分類されるかが大凡想定できる．この概要文と区分の組合せに着目し，教師データとして 2,810 件作成した（**図-2**）．

具体例として，「システム」に区分される概要文を下記に示す．

表-2　NETISによる精度検証結果

	土木 BERT			東北大 BERT		
	precision	recall	f1-score	precision	recall	f1-score
工法	0.779	0.786	0.782	0.776	0.764	0.769
材料	0.719	0.663	0.689	0.703	0.634	0.665
機械	0.714	0.624	0.656	0.693	0.629	0.648
製品	0.764	0.747	0.754	0.751	0.758	0.754
システム	0.816	0.909	0.860	0.813	0.879	0.844
平均	0.759	0.746	0.748	0.747	0.733	0.736
正解率			0.768			0.757

表-3　関東地方整備局の技術情報による精度検証結果

No.	穴埋め問題文	正答	推論結果（各 BERT）	
			土木	東北大
1	・・（略）・・加速度応答をリアルタイムに測定して得られるデジタル情報から，土工における●●●を簡易評価する方法の開発を行う	締固め状況	施工	機能
2	・・（略）・・大型構造物である●●●の 3 次元計測を実現するための未解決問題点を解決し，簡便かつ高速に密な3次元計測を実現する，・・（略）・・	トンネル	トンネル	部品
3	橋梁の健全度を正しく予測できれば，構造物を制御する，つまり，管理シナリオを描き，健全度に応じた具体的な●●●を計画に落とし込むことができる．・・（略）・・	維持管理行為	対策	データ
4	・・（略）・・導水路トンネルの損傷や劣化が近年顕在化している．そこで，導水路トンネル●●●の効率化のため，以下の3点のICT／AI技術に関する研究開発を，・・（略）・・	維持管理	管理	工事

■例：区分「システム」の概要文

「本技術は自然言語が理解できる AI である．従来は人手で対応していたが，本技術の活用により自動化できるため，経済性・利便性が向上する．」

なお本研究では，以下の理由によりNETISのデータを利用した．

a)　データ数

平成 10 年度から現在に到るまで運用されており，土木分野に関連する文章が豊富である．

b)　信頼性

掲載されている新技術は規定の要件を満たしており，データの信頼性が高い．

c)　再利用性

CSV出力が可能であるため，再利用性が高い．

(2)　関東地方整備局の技術情報

関東地方整備局では「現場ニーズと技術シーズのマッチング」の取組みの 1 つとして技術研究開発を実施している [9]．その研究成果の一部の文章を引用し，任意の箇所を穴埋めし，その穴埋めの箇所を推測させた．この穴埋めは，土木に関連する単語である箇所を無作為に設定し，計 10 問の穴埋め問題を作成した．穴埋め問題の例を**表-3** に示す．

4.　計算条件と評価指標

精度検証を実施する上での計算条件と評価指標を示す．

(1)　NETIS

a)　計算条件

各 BERT に対しファインチューニングを実施した．最終層に分類用の全結合層（分類器）を加え，入力を「概要文」，出力を「区分」となるモデルを構築した．再事前学習で得られたパラメータは固定し，学習は 12 段ある BertLayer の最終 12 段目と追加した分類器を対象とした．最適化手法は Adam，損失関数には交差エントロピー誤差を用いた．

b)　交差検証および評価指標

5 分割交差検証を実施した．NETIS は区分ごとにデータ件数の偏りがあったため，層化抽出法により分割データ内の区分割合を揃えた．評価指標は Accuracy，Precision・Recall・F1-score とし，各区分毎の平均値を算出した．

(2)　関東地方整備局の技術情報

a)　計算条件

各 BERT に対し，穴埋め問題を推論させる．BERT は穴埋め問題のタスクによる再事前学習を実施しているため，本検証のためのファインチューニングは不要である．

b)　検証および評価指標

穴埋め問題のタスクの出力には，各語彙の分類スコアが出力される．その分類スコアの上位 1 語を用いて穴埋め問題の推論結果とした．

事前検証により，穴埋め問題の回答は，単語は異なるが同義・広義のような回答が含まれているケー

スがあり，正答ではないが正答に近いといった曖昧な条件で判断する必要があったため，定性的な評価と考察を行った．

5. 精度検証結果と考察

東北大 BERT と土木 BERT の精度検証を実施し，その結果と考察を下記に示す．

(1) NETIS

精度検証結果を**表-2** に示す．評価指標は総じて土木 BERT のほうが高い結果となった．このことから土木に関する文章に関する自然言語処理のタスクを実施する場合は，土木文書の再事前学習が有効であることが示唆された．なお仁木ら [6]は金融文書 19,232,512 文を学習し F1-score を 1.9%向上させ，山腰ら [14]は法令文書 467,382 文を学習し正解率を 0.8%向上させている．本研究では 238,125 文を再事前学習に使用しているが，今後はこの文章をさらに増やすことで精度向上が図れると考える．

(2) 関東地方整備局の技術情報

精度検証結果を**表-3** に示す．No.2 のケースでは土木 BERT が正解となったが，その他は東北大 BERT を含め正答はできなかった．正答と推論結果を比較すると，土木 BERT のほうが「締固め」に対し「施工」，「維持管理行為」に対し「対策」，「管理」といった正答に近い単語を回答している．このことから穴埋め箇所の前後，つまり文脈を読み解くタスクにおいて，土木文書の再事前学習が有効であることが示唆されている．なお正答と完全一致しなかった理由として，東北大・土木 BERT で共に使用している形態素解析の語彙数は 32,000 語あるが，この中に「締固め」などの土木に関連する語彙が不足していたためと考える．

6. おわりに

本研究では，東北大 BERT に対し土木に関連する文章を再事前学習させ，土木 BERT を構築した．東北大 BERT，土木 BERT に対し，NETIS のデータ，関東地方整備局の技術情報データを使用して精度検証を実施し，土木 BERT のほうが優れた予測精度を示した．以上の結果より，土木に関連する文書の再事前学習が有効であることが示唆された．今後の課題として以下を挙げる．

(1) 今後の課題

a) データセットの整備

表-1 に示す土木に関連する文章を収集，整理したが，更なるデータを用意する必要がある．

b) 精度検証用のデータセットの拡充

本研究では 2 種の精度検証用のデータセットを作成したが，土木に関連する文章の理解・精度を示すためには更なる多種多様なベンチマーク用のデータセットを拡充する必要がある．

c) 形態素解析の辞書の整備

精度検証の結果，形態素解析の辞書に専門用語不足が判明した．荒木ら [15]は発注情報から土木用語の辞書を作成している．本研究と並行して形態素解析の辞書整備を進める必要がある．

参考文献

1) 片柳貴文，藤原鉄朗，遠藤和志，高石光博，九鬼和広，竹内恭一：国土交通データプラットフォームの整備に向けて，こうえいフォーラム，pp.19-24，第 29 号 2021 .4.
2) Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in Proc. NAACL-HLT 2019, pp. 4171–4186 (2019).
3) 柴田知秀，河原大輔，黒橋禎夫：BERT による日本語構文解析の精度向上，言語処理学会 第 25 回年次大会，pp.205-208，名古屋，(2019.3).
4) 東北大学乾研究室：Pretrained Japanese BERT models released/ 日本語 BERT モデル公開，<https://www.nlp.ecei.tohoku.ac.jp/news-release/3284/>，（アクセス 2022.5.24）.
5) 柴田大作，河添悦昌，嶋本公徳，篠原恵美子，荒牧英治：診療記録で事前学習した BERT による疼痛表現の抽出，医療情報学，2020，40 巻，2 号，p.73-82.
6) 仁木裕太，坂地泰紀，和泉潔，松島裕康：再事前学習した BERT を用いた金融文書中の因果関係知識有無の判別，人工知能学会全国大会論文集，2020，JSAI2020 巻，第 34 回 (2020).
7) Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, Jaewoo Kang：BioBERT: a pre-trained biomedical language representation model for biomedical text mining，Bioinformatics 2020，36，pp.1234-1240.
8) 国土交通省：NETIS，<https://www.netis.mlit.go.jp/>，（アクセス 2022.5.24）.
9) 関東地方整備局：技術情報，<https://www.ktr.mlit.go.jp/gijyutu/gijyutu00000108.html>，（アクセス 2022.5.24）.
10) Huggin Face：Huggin Face，<https://huggingface.co/>，（アクセス 2022.5.24）.
11) 和泉潔，坂地泰紀，松島裕康：金融・経済分析のためのテキストマイニング，pp.9-19，岩波書店，2021.
12) 国立研究開発法人科学技術振興機構：J-STAGE，<https://www.jstage.jst.go.jp/>，（アクセス 2022.5.24）.
13) 国土交通省：国土交通省，<https://www.mlit.go.jp/>，（アクセス 2022.5.24）.
14) 山腰貴大，駒水孝裕，小川泰弘，外山勝彦：事前学習モデル BERT による法令用語の校正，人工知能学会全国大会論文集，2020，JSAI2020 巻，第 34 回 (2020).
15) 荒木光一,吉田龍史,田中裕二,広瀬允佳,瀧本圭,高橋浩貴：大規模テキストを用いた土木用語を含む形態素解析用辞書の作成，土木情報学シンポジウム講演集 vol.44，pp217-pp220，2019