日本語法律分野文書に特化した BERT の構築

宮崎 桂輔　菅原 祐太　山田 寛章　徳永 健伸
東京工業大学 情報理工学院
{miyazaki.k.am@m, sugawara.y.ag@m, yamada.h.ax@m, take@c}.titech.ac.jp

概要

本論文では日本語の法律分野に特化した BERT モデルを提案する．民事事件判決書コーパスを用い，BERT を一から事前学習するモデルと，既存の汎用日本語 BERT に追加事前学習するモデルを作成した．実験より，民事事件判決書を用いた Masked Language Model，Next Sentence Prediction タスクについては既存の汎用日本語 BERT に追加事前学習する手法が最も良い正解率を示すことがわかった．

1　はじめに

近年，単言語，特定分野の文書に関するタスクにおいて，言語にのみ特化した汎用言語モデルに代わりドメインに特化した言語モデルが注目されている．英語の科学論文 [1]，生物医学論文 [2]，法律文書 [3] を用いて構築したドメイン特化 BERT が複数の後段タスクで汎用 BERT を超える性能を示しており，タスクによっては最先端の結果を示している．日本語においても，医学分野 [4]，金融分野 [5] などでドメインに特化した BERT が汎用日本語 BERT を超える性能を示している．

本論文では日本語の法律分野に特化した BERT モデルを提案する．なお，英語 [3] やフランス語 [6]，ルーマニア語 [7] などで法律分野に特化した BERT を構築する先行研究がある．また，最高裁判所の民事事件の判例を使用して BERT を事前学習する先行研究 [8] があるが，本論文では文書数にして 10 倍以上のより大規模なコーパスを用いて事前学習を行った．

2　モデルの構築

2.1　日本語法律分野コーパス

本研究では，日本語法律分野コーパスとして株式会社 LIC より提供された日本の民事事件の判決書 (以下，民事事件判決書コーパスと記載) を用いた．コーパス中の文書数は 170,320，文書当たりの平均文字数は約 11,300 字，コーパス全体のデータサイズは約 5.4GB である．

2.2　作成するモデル

英語法律分野文書に特化した BERT を構築する先行研究 [3] に倣い，二種類のモデルを作成した．どちらもモデルの構造として BERT-BASE [9] を用いた．

JLBERT-SC　BERT を民事事件判決書コーパスを用いて一から事前学習したモデルである．語彙は民事事件判決書コーパスから作成した．

JLBERT-FP　既存の汎用日本語 BERT に民事事件判決書コーパスによる追加事前学習を行ったモデルである．語彙は既存の汎用日本語 BERT のものをそのまま使用した．

3　実験設定

3.1　訓練データ，テストデータへの分割

NVIDIA による BERT の実装 [10] を参考に，後述の文分割を行った後の文の数の比がおよそ 9:1 となるように，コーパス内の文書を訓練データ，テストデータに分割した．

3.2　民事事件判決書コーパスのトークン化

民事事件判決書コーパスに対し，データの前処理および文分割，形態素分割，サブワード分割によるトークン化を行った．

データの前処理　前処理として，民事事件判決書コーパスに含まれる判決書の先頭に記載されている以下の情報は削除した．

- コーパス提供者が割り振った判例番号
- 民事事件記録符号規程により裁判所が付ける事件番号
- 事件名

また，インデントに使われている全角スペースはすべて削除し，それ以外の全角スペースが複数並んでいる箇所についてはスペース一つに置換した．

文分割　次に文分割を行った．句点の直後に括弧閉じが存在する場合のみを例外として扱い，それ以外のすべての改行及び句点を文末とした．文分割したコーパスの特徴は付録の表 4 の通りである．

形態素分割　次に JUMAN++ [11] を用いて文を形態素に分割した．JUMAN++の入力文長制限である 4,096 bytes を超える文については，4,096 bytes を超えないように読点の位置で文を分割した．この処理後も 4,096 bytes を超える文については，文頭から 4,096 bytes を超えないように貪欲に文字列を取って一文とみなすことを繰り返すことで文分割を実施した．

サブワード分割　最後に，BPE [12] によるサブワード分割を行った．JLBERT-FP では，柴田らが公開している既存の汎用日本語 BERT [13] の語彙を用いて BPE を適用した．また，JLBERT-SC では，訓練データから語彙を学習し，学習した語彙を用いて BPE を適用した．なお，JLBERT-SC の語彙は汎用日本語 BERT に合わせ，語彙数を 32,000 とした．サブワード分割を行ったコーパスの特徴は付録の表 5, 6 に示す通りである．

3.3　モデルの事前学習の設定

NVIDIA による BERT の実装 [10] を参考に，事前学習の設定を行った．

3.3.1　事前学習で扱うタスク

BERT の事前学習では，以下の二つのタスクで学習を行う．

Masked Language Model (MLM)　ランダムにマスキングされた文中のトークンについて，マスキングされる前の正しいトークンを予測するタスクである．なお，マスキングの対象とされたトークンのうち，80%は [MASK] トークンに，10%はランダムなトークンに置き換えて，残りの 10%は元のトークンのままモデルに入力する．

Next Sentence Prediction (NSP)　与えられた二文が文書内で連続していたかどうかを予測するタスクである．文書内に実在する連続する 2 文を正例，文書内からランダムに抽出した連続しない 2 文を負例とし，正例と負例を 50%ずつの割合にして学習を行う．

3.3.2　事前学習の流れ

事前学習は，phase 1 と phase 2 の二段階に分けて行った．phase 1 では，入力文の長さの上限を 128 トークンとし，MLM タスクでは一文あたり min(20, 文中のトークン数の 15%) トークンをマスクした．phase 2 では，入力文の長さの上限を 512 トークンとし，MLM タスクでは一文あたり min(80, 文中のトークン数の 15%) トークンをマスクした．

3.3.3　モデルの詳細，ハイパーパラメータ

JLBERT-SC，JLBERT-FP 共にモデルの構造は BERT-BASE [9] とした．JLBERT-FP では，初期パラメータとして柴田らが公開している汎用日本語 BERT [13] の BASE WWM 版のパラメータを使用した．optimizer には LAMB [14] を使用した．その他のハイパーパラメータは付録の表 8, 9 に記す．なお，本設定では phase 1 における 1 epoch はミニバッチ入力回数約 68 万回，約 5,300steps(1GPU の場合)，phase 2 における 1 epoch はミニバッチ入力回数約 192 万回，約 3,700steps(1GPU の場合) に対応する．

3.3.4　実行環境

BERT の事前学習には，NVIDIA RTX A6000 GPU を 4 つ使用した．モデルの事前学習にかかった時間は，phase 1 が約 28 時間，phase 2 が約 18 時間，合わせて約 46 時間であった．

3.4　JLBERT-FP の学習データ数削減による性能の変化

追加実験として，学習データ量の違いによる性能の変化を検証するため，学習データの数を段階的に削減して JLBERT-FP の事前学習を行い，MLM，NSP の各タスクにおける性能を比較した．

3.4.1　学習データ

学習データとして，JLBERT-FP の学習データを文書単位でおよそ 50%，25%に削減したデータセットを用意した．これらのデータセットに関する情報は付録の表 7 の通りである．これらのデータセットを用いて事前学習を行ったモデルを以下 JLBERT-FP-50%，JLBERT-FP-25%と略記する．

3.4.2　実験設定

実験設定は以下に記載する事項を除き 3.3 と同じとした．JLBERT-FP-50%，JLBERT-FP-25%の事前学習の設定には，3.3 と比較して以下の変更がある．

- GPU を 1 つ用いて学習を行った．実効バッチサイズは 1/4 となっている．
- 学習 step 数は変わらず実効バッチサイズが 1/4 となっていることから，ミニバッチ入力回数は 1/4 となっている．
- 学習データ数がそれぞれおよそ 1/2，1/4，ミニバッチ入力回数が 1/4 となっていることから，epoch 数はそれぞれおよそ 1/2 倍，等倍となっている．

3.5 非法律ドメインの文書分類

法律ドメインに特化して事前学習を行った BERT の，法律以外のドメインにおけるタスクの性能を検証する．ここでは，livedoor ニュースコーパス [15] を用いて，JLBERT-SC，JLBERT-FP と汎用日本語 BERT の間で性能比較を行った．

3.5.1 データセット，タスク

livedoor ニュースコーパスは，「livedoor ニュース」内の 9 つのカテゴリのニュース記事からなるコーパスである．本研究では，記事の内容からその記事のカテゴリを予測する文書分類タスクを行った．記事数は 7,367 件であり，データセットに占める各カテゴリの記事の割合が等しく記事数がおよそ 8:1:1 となるように学習，テスト，検証データセットに分割した．学習，テスト，検証データセットそれぞれについて，各カテゴリに属する記事の件数は付録の表 10 の通りである．

3.5.2 livedoor コーパスのトークン化

livedoor ニュースコーパスに対し，データの前処理および文分割，形態素分割，サブワード分割によるトークン化を行った．本研究では，コーパス内の各文書の先頭三行 (記事のリンク，記事作成日時，記事の題名) を削除することとした．また，文書内の半角英数字，記号については，全て全角のものに変換した．文分割は改行のみを文末とした．形態素分割，サブワード分割については，3.2 と同様に JUMAN++および BPE を用いた．モデルに入力するトークン列の長さの上限は 512 とした．

3.5.3 モデルの詳細，ハイパーパラメータ

比較するモデルとして，JLBERT-SC，JLBERT-FP および柴田らが公開している汎用日本語 BERT [13] の BASE WWM 版を用いた．optimizer は Adam [16] とし，学習率を $\{10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$ の中で探索した．バッチサイズは 32 とした．

表 1 事前学習の各タスクの正解率

	MLM		NSP	
タスク	train	test	train	test
JLBERT-SC	0.8140	0.7891	0.9977	0.9914
JLBERT-FP	0.8319	**0.8054**	0.9979	**0.9921**
JLBERT-FP-50%	0.8119	0.7834	0.9958	0.9893
JLBERT-FP-25%	0.8158	0.7806	0.9976	0.9888

表 2 livedoor ニュース文書分類タスクの正解率

	日本語 BERT	JLBERT-SC	JLBERT-FP
正解率	**0.9512**	**0.9512**	0.9322

4 実験結果

4.1 JLBERT-SC，JLBERT-FP の事前学習

作成した JLBERT-SC，JLBERT-FP，JLBERT-FP-50%，JLBERT-FP-25%それぞれについて，学習データ全体に対する MLM，NSP の正解率を表 1 に示す．この結果から，民事事件判決書の MLM，NSP では，一から民事事件判決書コーパスで事前学習を行う JLBERT-SC より汎用日本語 BERT に民事事件判決書コーパスによる追加事前学習を行う JLBERT-FP の方が優れた手法であると言える．また，追加事前学習に用いるデータの量は多ければ多いほど良いことがわかる．したがって，コーパスサイズを増やすことで更なる性能向上が期待できる．

また，各ステップにおけるミニバッチに対する loss および MLM，NSP の正解率の推移をそれぞれ図 1, 2, 3 に示す．なお，図 3 中の loss とは，MLM，NSP それぞれの loss の和である．この図から，MLM については学習データからさらに学習する余地があることがわかる．

さらに，phase 2 におけるテストデータに対する MLM，NSP の正解率の推移を表 3 に示す．この結果より，JLBERT-FP，JLBERT-SC 共に phase 2 の step 数を増やすほどモデルの性能が良いことがわかる．したがって，step 数を増やすことで更なるモデルの性能向上が期待できる．

表 3　phase 2 における各タスクの正解率推移

手法	steps	test MLM	test NSP
JLBERT-FP	0	0.2929	0.9126
	1190	0.7997	0.9914
	1390	0.8024	0.9918
	1563	0.8054	0.9921
JLBERT-SC	0	0.2824	0.9213
	1190	0.7835	0.9906
	1390	0.7863	0.9902
	1563	0.7891	0.9914

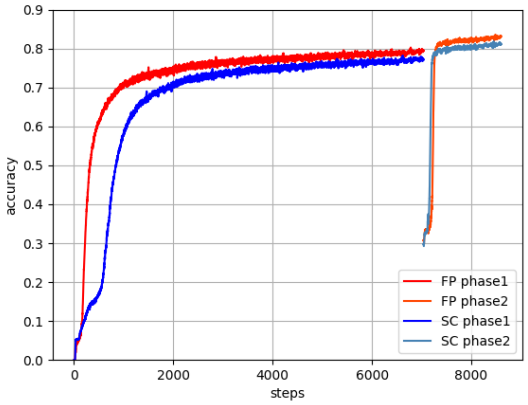

図 1　実効バッチに対する MLM 正解率の推移

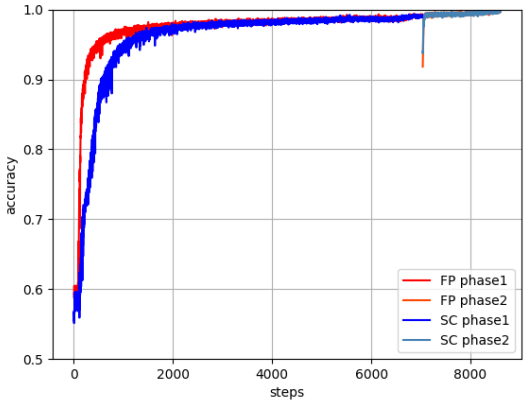

図 2　実効バッチに対する NSP 正解率の推移

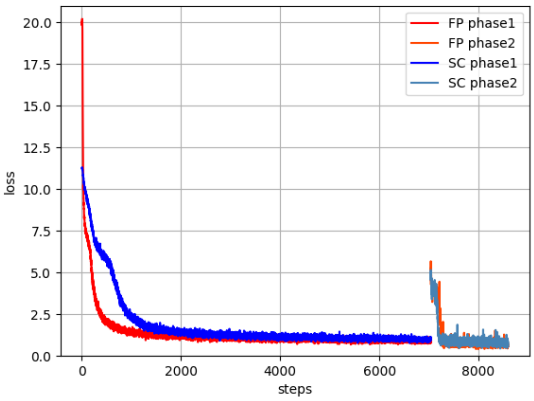

図 3　実効バッチに対する loss の推移

4.2　livedoor ニュース文書分類結果

各モデルの文書分類タスクにおける正解率を表 2 に示す．民事事件判決書コーパスのみから語彙の学習，事前学習を行ったモデルである JLBERT-SC が汎用日本語 BERT に並ぶ正解率を示している．今回使用した汎用日本語 BERT が事前学習時に最大入力トークン長を 128 とした学習のみ行なっていることを踏まえると，提案手法において事前学習時に最大入力トークン長を 128 とした学習 (phase 1) に加えて最大入力トークン長を 512 とした学習 (phase 2) を実施していることの有効性が推測できる．また，汎用日本語 BERT に対する民事事件判決書コーパスによる追加の事前学習は，法律分野以外のドメインのタスクにおいて BERT の fine-tuning に悪影響を及ぼすことがわかる．

5　おわりに

本論文では日本語の法律分野に特化した BERT モデルを作成した．実験の結果，既存の汎用日本語 BERT に民事事件判決書コーパスによる追加事前学習を行う手法により，テストデータに対する Masked Language Model，Next Sentence Prediction タスクの正解率が最も高くなることがわかった．また，学習データ数や学習 step 数を増やすことによりモデルの更なる性能向上が期待できることが示された．本モデルの後段タスクにおける有用性を実証するために，JLBERT-SC，JLBERT-FP を用いて判決書の重要箇所抽出タスクを行う [17].

謝辞

本研究で使用した判決書データは株式会社 LIC から提供を受けたものである．本研究は，JST, ACT-X，JPMJAX20AM の支援を受けたものである．

参考文献

[1] Iz Beltagy, Kyle Lo, and Arman Cohan. SciBERT: Pretrained language model for scientific text. In **EMNLP**, 2019.

[2] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. 09 2019.

[3] Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. LEGAL-BERT: The muppets straight out of law school. In **Findings of the Association for Computational Linguistics: EMNLP 2020**, pp. 2898–2904, Online, November 2020. Association for Computational Linguistics.

[4] Yoshimasa Kawazoe, Daisaku Shibata, Emiko Shinohara, Eiji Aramaki, and Kazuhiko Ohe. A clinical specific BERT developed using a huge japanese clinical text corpus. **PLOS ONE**, Vol. 16, No. 11, pp. 1–11, 11 2021.

[5] 鈴木雅弘, 坂地泰紀, 平野正徳, 和泉潔. 金融文書を用いた事前学習言語モデルの構築と検証. 人工知能学会第 27 回金融情報学研究会 (SIG-FIN), 2021.

[6] Stella Douka, Hadi Abdine, Michalis Vazirgiannis, Rajaa El Hamdani, and David Restrepo Amariles. JuriBERT: A masked-language model adaptation for French legal text. In **Proceedings of the Natural Legal Language Processing Workshop 2021**, pp. 95–101, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[7] Mihai Masala, Radu Cristian Alexandru Iacob, Ana Sabina Uban, Marina Cidota, Horia Velicu, Traian Rebedea, and Marius Popescu. jurBERT: A Romanian BERT model for legal judgement prediction. In **Proceedings of the Natural Legal Language Processing Workshop 2021**, pp. 86–94, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[8] 星野玲那, 狩野芳伸. 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, 2020.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)**, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[10] NVIDIA. BERT For PyTorch,（2022-01 閲覧）. https://github.com/NVIDIA/DeepLearningExamples/tree/master/PyTorch/LanguageModeling/BERT.

[11] Arseny Tolmachev and Sadao Kurohashi. Juman++ v2: A practical and modern morphological analyzer. 言語処理学会 第 24 回年次大会, 岡山, 2018.

[12] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In **Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 1715–1725, Berlin, Germany, August 2016. Association for Computational Linguistics.

[13] 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, 名古屋, 2019.

[14] Yang You, Jing Li, Sashank Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. Large batch optimization for deep learning: Training BERT in 76 minutes. In **International Conference on Learning Representations**, 2020.

[15] 株式会社ロンウイット. livedoor ニュースコーパス,（2022-01 閲覧）. http://www.rondhuit.com/download.html#ldcc.

[16] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, **3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings**, 2015.

[17] 菅原祐太, 宮崎桂輔, 山田寛章, 徳永健伸. 日本語法律 BERT を用いた判決書からの重要箇所抽出. 言語処理学会 第 28 回年次大会, 2022.

付録

表 4 コーパス文分割時のデータ

対象	値
文の数	23,789,799
4,096 bytes を超えた文	2,055
貪欲な文分割を行った文	21

表 5 トークン化されたコーパスのデータ (JLBERT-SC)

対象	train	test	全体
文書数	167,109	3,301	170,320
文数	21,411,914	2,378,943	23,790,857
総 token 数/文数	48.15	47.83	48.12
[UNK] 数	402418	49107	451525
[UNK] の割合	0.0390%	0.0432%	0.0394%
128 token 超の文数	1449694	143383	1593077
512 token 超の文数	15830	1365	17195

表 6 トークン化されたコーパスのデータ (JLBERT-FP)

対象	train	test	全体
文書数	167,109	3,301	170,320
文数	21,411,914	2,378,943	23,790,857
総 token 数/文数	50.15	49.28	50.06
[UNK] 数	4,543,996	533,412	5,077,408
[UNK] の割合	0.4232%	0.4550%	0.4263%
128 token 超の文数	1,604,348	155,306	1,759,654
512 token 超の文数	18,768	1,550	20,318

表 7 token 化されたコーパスのデータ (JLBERT-FP-n%)

対象	train	train-50%	train-25%
文書数	167,109	83,282	41,508
文数	21,411,914	10,706,044	5,353,046
総 token 数/文数	50.15	50.06	50.02
[UNK] 数	4,543,996	2,277,995	1,139,966
[UNK] の割合	0.4232%	0.4251%	0.4257%
128 token 超の文数	1,604,348	801,359	401,617
512 token 超の文数	18,768	9,459	4,693

表 8 モデルのハイパーパラメータ

対象	値
隠れ状態の次元	768
中間層の次元	3,072
multi-head attention の head の数	12
layer 数	12
語彙数	32,008
attention の dropout 率	0.1
隠れ状態の dropout 率	0.1

表 9 事前学習のハイパーパラメータ

対象	phase 1	phase 2
実効バッチサイズ	32,768	16,384
実効バッチサイズ/GPU	8,192	4,096
ミニバッチサイズ	64	8
勾配蓄積の回数/GPU	128	512
学習 step 数	7,038	1,563
ミニバッチ入力回数/GPU	900,864	800,256
warmup を行う step 数	2,000	200
warmup を行う step の割合	28.43%	12.80%
入力文長の上限 (トークン)	128	512
文毎の [MASK] の割合	0.15	0.15
文毎の最大 [MASK] 数	20	80
学習率	0.006	0.004

表 10 livedoor ニュースコーパスのカテゴリ

カテゴリ	総記事数	train	test	valid
dokujo-tsushin	870	696	87	87
it-life-hack	870	720	90	90
livedoor-homme	511	409	51	51
smax	870	696	87	87
topic-news	770	696	87	87
movie-enter	870	691	86	86
kaden-channel	864	616	77	77
peachy	842	696	87	87
sports-watch	900	673	85	85
合計	7,367	5,893	737	737