日本語法律 BERT を用いた判決書からの重要箇所抽出

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

概要

本研究では判決書からの重要箇所抽出タスクにおいて，法律分野の文書のみで事前学習を行った BERT，日本語 Wikipedia で事前学習された BERT から追加の事前学習を行なった BERT を用い，その性能を汎用日本語 BERT と比較検証した．実験より，法律分野に特化した BERT モデルを用いることで，汎用日本語 BERT を超える性能があることを確認した．

1　はじめに

裁判の記録である判決書は情報技術の発達により電子化され，人々が参照できるようになっている．判決書にはその判決に至るまでの詳細な議論が記録されており，他の新聞記事や科学技術文書に比べて内包される文が長く複雑な文となっている．法律に関する仕事している人々は関連事件の調査に膨大な時間を費やしており，判決書に対する情報アクセスの容易化・効率化は重要な問題となっている．判決書から重要箇所を自動的に抽出することができれば，判決書の効率的な調査や判決書要旨の作成の大幅な効率化が見込める．また，判決書はその長い文書長から，全文を自動要約モデルに入力することは現実的ではないため，本研究で行うような重要箇所抽出は自動要約の前処理としても有用である．

近年，大規模な事前学習によって汎用的な言語モデルを構築するアプローチが普及している．なかでも BERT は様々なタスクで利用され，各タスクに応じて fine-tuning を行うことで高い性能を実現している [1]．更に，BERT は事前学習において，特定ドメインのデータを用いて学習することで，同ドメインでの性能が高まることが報告されている [2, 3]．Chalkidis らは、EU、英国、米国における英語で書かれた法律ドメインの文書のみで事前学習を行なった BERT，汎用ドメインのデータで事前学習された BERT に対してさらに法律ドメインの文書で追加事前学習を行なった BERT の 2 つの LEGAL-BERT を構築した [4]．これらの LEGAL-BERT は，法律ドメインのテキスト分類と系列ラベリングのタスクにおいて，汎用ドメインで事前学習を行った BERT よりも高い性能を達成している．日本の判決書からの重要箇所抽出においても，法律分野の文書データによって BERT の事前学習を行った上で fine-tuning を行うことで，Wikipedia 等で事前学習された BERT よりも高い性能を発揮することが期待できる．

そこで本研究では，(1) 法律分野の文書のみで事前学習を行った BERT，(2) 日本語 Wikipedia で事前学習された BERT をさらに法律分野の文書を用いて追加事前学習を行った BERT を用い，日本の判決書からの重要箇所抽出タスクにおける性能を，汎用日本語 BERT と比較検証する．

2　関連研究

日本における判決書からの重要箇所抽出の研究として，阪野らは SVM を用いて要旨の内容に該当する箇所を判決文から自動的に抽出する手法を提案している [5]．山田らは抽出型要約の一環として判決書の修辞役割（例：結論，法律の引用等）分類を行い，文脈を考慮した階層型 RNN をベースとするモデルを用いる手法を提案している [6]．

英国における判決書を対象として，Hachey らは抽出型要約において，機械学習を用いて修辞役割分類を分別したうえで要約生成を行っている [7]．Tran らは修辞役割分類に対し，CNN または BiLSTM を用いたモデルを提案している [8]．

また，ドメイン固有のテキストデータに関するこれまでの研究により，ドメイン固有の言語モデルを作成することの重要性が指摘されている．汎用ドメインのデータで事前学習された BERT のモデルに対して，さらに特定ドメインに特化したデータで追加事前学習を行うことで，様々な専門ドメインに対応したモデルが作られている．Lee らは大規模なバイオ分野の文献を用いて事前学習させること

で，バイオ分野に特化した BioBERT を作成した [2]．BioBERT は Wikipedia 等で事前学習された BERT や他のモデルに比べて，バイオ分野のテキストマイニングのタスクでより高い性能を達成している．また，Beltangy らは科学関連のコーパスに対して事前学習を行った SciBERT を導入し，テキスト分類と系列レベリングのタスクで性能改善が報告されている [3]．法律分野において，Chalkidis らは法律ドメインの文書のみで事前学習を行なった BERT，汎用ドメインのデータで事前学習された BERT に対してさらに法律分野の文書で追加事前学習を行なった BERT の 2 つの LEGAL-BERT を構築した．2 つの LEGAL-BERT を用いて，3 つの法律分野のデータセットに対しテキスト分類と系列ラベリングのタスクを解いた結果，ほとんどのタスクで LEGAL-BERT が汎用ドメインで事前学習した BERT より高い性能を達成している．これらは全て英語に限定したものであるが，近年は他言語に対して事前学習を行なったモデルが構築されている．特に法律文書に関しては，Stella らがフランス語の法律文書に特化した juriBERT，Mihai らはルーマニア語の法律文書に特化した jurBERT をそれぞれ構築している [9, 10]．

本研究では日本語 Wikipedia による BERT [11]，宮崎らが構築した法律文書に特化した BERT[12] をそれぞれ用いて，性能を比較する．

3 実験

3.1 用いたコーパス

本研究では，株式会社 LIC より提供された日本の民事事件判決書を日本語法律分野コーパスとして用いる．各判決書には法律知識を有する作業者が判例として重要な記述と思われる箇所を特定した注釈（重要箇所）が付与されている．コーパスは全 84,900 件の判決書から成る．実験のためにコーパスを訓練データ，開発データ，テストデータに 8:1:1 の割合で分割した．重要箇所が元から付与されていない，または適切に付与されていなかった文書を除外し，最終的に訓練データ 67,870 件，開発データ 8,490 件，テストデータ 8,490 件を得た．

3.2 前処理

民事事件判決書のコーパスに対し，データの前処理および文分割，形態素分割を行った．まず，行頭に出てくる見出し番号（例：（一）や（ア）など）を削除し，インデントに使われている全角スペースもすべて削除した．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．次に JUMAN++ [13] を用いて文を形態素に分割した．

表 1 コーパス中の重要箇所を含む文の割合 [%]

文中の重要箇所の割合 (文字)	訓練	開発	テスト
100% (一文全体)	89.6	89.7	90.0
≥ 90%	96.5	96.3	96.9
≥ 70%	97.9	97.6	98.3
≥ 50%	98.7	99.0	99.1

表 2 データセットの詳細 (重要文抽出)

対象	訓練	開発	テスト
文書数	67,870	8,490	8,490
文書毎の平均文数	114.6	116.8	115.6
平均文書長 (形態素)	5,815	5,881	5,766
文書毎の平均重要文数	7.26	7.27	7.23

3.3 重要文抽出タスクの設定

一文中の重要箇所の割合と，それらの文がコーパス全体に占める割合の関係を表 1 に示す．この表から，重要箇所が付与された文のうち 9 割は文全体が重要箇所となっており，文の一部だけが重要箇所となっていても，重要箇所が文に占める割合が十分に大きいことが分かる．そのため，本実験では重要箇所を含む文全体を重要文とみなして，重要文抽出のデータセットを構築する．

作成したデータセットの特徴を表 2 に示す．１文書あたりの平均文数が大きく，平均文書長が長いことが分かる．

3.4 実験設定

本研究では，ある文が重要箇所か否かを当てる二値分類タスクを解く．評価指標として精度，再現率，F 値を用いる．

実験には 3 つの異なる BERT の事前学習済みモデルを用い，それぞれを本重要箇所抽出タスクに対して fine-tuning した上で実施する．用いる事前学習済みモデルは，京都大学が公開した日本語 Wikipedia によって学習した日本語 BERT[11]，宮崎ら [12] によって BERT を日本語法律分野コーパスを用いて一から事前学習した JLBERT-SC，既存の日本語 BERT に日本語法律分野コーパスによる追加事前学習をおこなった JLBERT-FP である．なお，JLBERT-SC 及

表 3 サブワード化時のデータセット（日本語 BERT）

対象	訓練	開発	テスト
文書数	67,870	8,490	8,490
平均文書長（サブワード）	6,414	6,564	6,439
文書毎の長文数の平均	0.15	0.16	0.144

* 長文：512 トークンを越える文

表 4 サブワード化時のデータセット（JLBERT-SC）

対象	訓練	開発	テスト
文書数	67,870	8,490	8,490
平均文書長（サブワード）	6,136	6,277	6,165
文書毎の長文数の平均	0.12	0.13	0.12

* 長文：512 トークンを越える文

び JLBERT-FP の事前学習に用いられた文書と本実験で用いるテストデータ中の文書との間に重複はない．

訓練，開発，テストデータにおいて，重要箇所となる文の総数が文全体の総数の約 6%ほどとなっておりデータセットに含まれるインスタンスのクラスが占める割合に偏りがある．これに対処するために損失関数の重みの値を工夫する．損失関数として使う BCEWithLogitsLoss は

$$l_c = -w_c[y_c \cdot log\sigma(x_c) + (1-y_c) \cdot log(1-\sigma(x_c))] \quad (1)$$

と表される．w_c はクラス c における重み（ハイパーパラメータ），y_c はラベル（今回の場合 0 か 1），x_c は推定値を表す．この時，クラスにおける重み w_c を

$$w_c = \frac{\text{クラスの学習データ数}}{\text{全クラスの学習データ数の和}} \quad (2)$$

とすることで損失への寄与率に差が出ないようにすることができると考えられる．

またサブワードによるトークン化を行う際，日本語 BERT と JLBERT-FP では元の日本語 Wikipedia を用いた事前学習時に構築された既存の語彙を用いた．JLBERT-SC では日本語法律分野コーパス [12] を用いて事前学習した際に構築した語彙を用いた．この時のデータセットの詳細を表 3, 4 に示す．法律文書に特化した語彙リストを参照した方が日本語 BERT の語彙リストを参照した場合よりも，平均文書長が小さくなっていることが分かる．学習と推論の際，512 トークンを超える文に関しては文の初めから 512 トークンまでを入力として用い，以降は削除した．

全ての BERT において optimizer は AdamW，学習率 $1e-5$，バッチサイズ $\in \{16, 32\}$，エポック数 $\in \{3, 4\}$，ドロップアウト 0.5 に設定し学習した．

表 5 各モデルの性能（マクロ平均値）

モデル	精度	再現率	F 値
BiLSTM	0.309	0.865	0.455
日本語 BERT	0.510	0.438	0.471
JLBERT-SC	0.522	0.520	0.521
JLBERT-FP	0.526	0.512	0.519

表 6 各モデルの性能（ミクロ平均値）

モデル	精度	再現率	F 値
BiLSTM	0.157	0.874	0.266
日本語 BERT	0.463	0.300	0.361
JLBERT-SC	0.459	0.377	0.412
JLBERT-FP	0.472	0.365	0.412

ベースラインモデルとして，2 層からなる BiLSTM を用いた重要箇所抽出器を構築した．optimizer は SGD，学習率 0.01，バッチサイズ $\in \{32, 64\}$，エポック数 15 で学習した．トークン化は日本語 BERT と JLBERT-FP に習い既存の日本語 BERT の語彙を用いて BPE を適用した．単語埋め込みの次元数 200, 中間層の次元数を 100 とした．また，損失関数である CrossEntropyLoss に対し BERT と同様の重みづけを行った．

4 実験結果

重要箇所抽出の結果を表 5,6 に記す．ベースラインである BiLSTM に比べて，全ての事前学習済みモデルが高い性能を出していることが分かる．また，マクロ平均に関して，F 値が JLBERT-SC と JLBERT-FP は日本語 BERT と比べて高い性能を出している（JLBERT-SC では 0.15，JLBERT-FP では 0.148 の上昇）．同様にミクロ平均に関しても高い性能を出している．これは英語法律文書に特化した BERT においても同様の性能向上が見られている [4]．今回はクラス間の比率の差が大きい不均衡データであったが，損失関数の重み付けの工夫を行いドメインに特化した BERT を構築することでドメイン固有のタスクにおいて性能が向上することがわかる．

また，入力制限である 512 トークンを超える文に対して BERT が上手く推論できているかどうかを評価した．文書全体から 512 トークン以下の文と 512 トークンを越える文をそれぞれ 500 個ずつランダムにサンプリングして，各 BERT におけるミクロ平均の F 値を算出する方法を計 10 回繰り返しその平均

表 7 文長におけるモデルの性能差（F 値）

モデル	512 トークン以下	512 トークン超
日本語 BERT	0.33	0.45
JLBERT-SC	0.41	0.46
JLBERT-FP	0.41	0.48

を取った．その結果を表 7 に示す．表 5,6 に示したものと同様に，JLBERT-SC と JLBERT-FP が汎用日本語 BERT よりも高い性能が出ていることが分かる．また，512 トークンを超える文における F 値が 512 トークン以下の文と表 6 の F 値よりも大きな値となっており，512 トークンを超える文に対して十分な認識性能を持っていることが分かる．

5 おわりに

本研究では判決書からの重要箇所抽出タスクにおいて，汎用日本語 BERT と法律分野に特化した BERT モデルを適用しその性能を検証した．その結果，法律分野の文書のみで事前学習を行った BERT，日本語 Wikipedia で事前学習された BERT をさらに追加事前学習を行なった BERT の両方で汎用日本語 BERT を越える性能があることを確認した．今後の課題としては，今回のモデルは文脈を考慮していないため，文脈を考慮した階層型モデルの提案とその実装，最終的に抽出した文を抽象型要約タスクに応用することが考えられる．

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] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. **arXiv preprint arXiv:1810.04805**, 2018.

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

[4] 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.

[5] 阪野慎司, 松原茂樹, 吉川正俊. 機械学習に基づく判決文の重要箇所特定. 言語処理学会年次大会発表論文集, pp. 1075–1078, 名古屋, 2006.

[6] 山田寛章, Simone Teufel, 徳永健伸. 見出し情報を考慮した階層型 rnn による日本語判決書のための修辞役割分類. 言語処理学会年次大会発表論文集, pp. 37–40, 2018.

[7] Ben Hachey and Claire Grover. Extractive summarisation of legal texts. **Artificial Intelligence and Law**, Vol. 14, No. 4, pp. 305–345, 2006.

[8] Vu D. Tran, Minh L. Nguyen, Kiyoaki Shirai, and Ken Satoh. An approach of rhetorical status recognition for judgments in court documents using deep learning models. In **2019 11th International Conference on Knowledge and Systems Engineering (KSE)**, pp. 1–6, 2019.

[9] 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.

[10] 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.

[11] 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.

[12] 宮崎桂輔, 菅原祐太, 山田寛章, 徳永健伸. 日本語法律分野文書に特化した bert の構築. 言語処理学会 第 28 回年次大会, 2022.

[13] 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.