マスク言語モデルにおける重点的なマスク選択での追加的学習を用いた法律文書による人物役割語の推測

翁長駿光[1] 藤田真伎[2] 狩野芳伸[1,2]
[1]静岡大学 情報学部 [2]静岡大学 総合科学技術研究科情報学専攻
{tonaga, mfujita, kano}@kanolab.net

概要

人物の文中での役割の把握は，自然言語処理において広く重要な課題であり，特に法律分野の文書の処理では，その分析と利用のために必須と言える．本研究では，日本語の事前学習済み BERT モデルに対し，あらかじめ設定した人物語を重点的にマスクしたMasked Language Modelによる追加的な学習を，法律分野の文書を用いて行った．司法試験の問題文における人物役割語の推論に適用した結果，提案手法による性能向上を確認できた．

1. はじめに

人物の文中での役割の把握は，自然言語処理において広く重要な課題であるが，役割間の関係性や階層など背後にある複雑な知識を反映させる必要があり，難易度の高いタスクといえる．

法律分野は自然言語処理技術の利用が大きく期待されている応用分野の一つであり，そこでの人物役割把握は必須要素である.法律分野の文書処理では，こうした人物関係および役割の把握や専門性の高い用語といった難しい処理が重要になるが，そのための自然言語処理技術の性能は未だ不十分である．

人物役割の先行研究にはたとえば物語テキストを対象とした登場人物の関係抽出 [1]などがある．日本語法律文書を対象にした研究には，日本語法律BERT [2]を用いた重要箇所抽出 [3]や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 [4]があるが，人物役割を推測するものは見当たらない．

本研究では，法律分野の文章における人物関係および人物の役割の推測を目標とする． そのために，日本語事前学習済み BERT に対して，あらかじめ設定した人物役割語を他の語よりもマスクされやすくした処理を追加したMasked Language Model (MLM)による追加学習を，日本語の法律文書を用いて行った．我が国の司法試験自動解答を題材とするCOLIEE Task4 で提供された問題を用いて，問題文中の人物役割語を推測する人物推論タスクを実行し，提案手法による性能向上を確認した．

<問題番号 : H18-32-5, 正解ラベル : No >
<関連条文>
第百九十二条 取引行為によって，平穏に，かつ，公然と動産の占有を始めた者は，善意であり，かつ，過失がないときは，即時にその動産について行使する権利を取得する．
<問題文>
強迫を受けてした動産売買契約を取り消した売主は，取消し前に買主から当該動産を善意かつ無過失で買い受けた者に対して，所有権に基づいて，当該動産の返還を求めることができる．

図 1 COLIEE Task4 の提供データ例

2. 関連研究

法律分野の自然言語処理技術の性能向上とコミュニティ構築を図るため，国際コンテスト型ワークショップ COLIEE (Competition on Legal Information Extraction and Entailment) [5] [6] [7] [8]が毎年開催されている．COLIEE にはいくつかのサブタスクがあるが，そのうち Task 3,4 は日本の司法試験民法短答式問題を題材にした自動解答タスクである．Task4 は図 1 のように，司法試験に過去に出題された問題とその問題を解く上で手掛かりとなる関連民法条文の二つが与えられ，その関連する民法条文に対して問題文が含意関係にあるかを二値で判定するタスクである．

COLIEE 参加者の自動解答器には，BERT [9]をはじめ深層学習ベースの解答器が多数みられる.このような深層学習を用いた司法試験の自動解答について，清田ら [10]は人手の問題分類を通して深層学習モデルが司法試験を解けているかを分析している．分類の一つは，図 2 のような問題文中における A や B

といったアルファベットで表された人物語の関係性や，それぞれの人物が関連する民法条文中におけるどの役割にあてはまるのかが回答に必要なタイプの問題である．このような，人物関係や人物がどのような役割を担うか明らかにする必要がある問題の推測は，既存の回答では不十分であることが示唆されている．

3. 提案手法

提案するマスク手法と従来手法との比較を容易にするため，深層学習モデルとしてBERTを用いる．事前学習済みの日本語 BERT モデルを前提に，Masked Language Model (MLM)による追加的な学習を行う．その追加学習時に，より重点的に学習させたい語句をマスクして学習させる．本研究で利用するトークナイザはこの事前学習済みモデルのもので統一する．

重点的に学習させたい語句は，本研究では，法律分野における人物役割語とし，民法条文に記載されている「人，者，主」で終わる漢字で構成される単語と「相手方」を合わせて 74 語を人手で列挙した(74 語の詳細については付録を参照)．

3.1 事前学習モデルへの追加的学習

追加的な学習では，以下の 3 つの前処理に応じたモデルを作成する．それぞれの前処理で得られるトークン列について，[MASK]トークンを推測させる学習によりモデルを訓練する．

ランダム 15%マスクモデル（15%） 従来の BERT のMLMと同様に,トークン全体の15%をランダムで選択する．ランダムに選択されたトークンのうち，80%を[MASK]トークンに,10%はランダムに選ばれた別のトークンに置きかえ，残りの 10%は元のトークンのままモデルに与える．

ランダムマスク + 人物語マスクモデル（15%＋人物語） 入力にあらかじめ定めた人物役割語が含まれている場合，それらのトークンを[MASK]トークンに置き換える.次に前述のランダム 15%マスクモデルと同様の処理を行うが，選ばれたトークンが人物役割語に対応するものであった場合は，[MASK]のままにし，何もしない．

人物語マスクモデル（人物語） 人物役割語のトークンのみを[MASK]トークンに置き換える．人物役割語が含まれていないサンプルについては学習に使用しない．

上の 3 つのモデルと，追加学習をしていない日本語事前学習済み BERT モデルの 4 つを，次節の人物役割語推測タスクによって比較する．

3.2 人物役割語の推測

人物役割語の推測は[MASK]を言語モデルにより推測して実行するため，前節で説明した以上の学習は行わない．

人物役割語は語によって 2 トークン以上のサブワードに分割されることがある．このため，単にトークン単位で[MASK]に置き換えると，[MASK]トークンの連続する数が人物語を予測するヒントになる可能性がある．人物役割語の推論にあたっては，推測対象の役割語についてマスク前のトークン数を既知として固定する場合と，トークン数を未知として任意のトークン数で推測するより実用的な設定の 2 種類を試みる．

人物役割語あたりのトークン数固定 人物役割語に該当するトークンをひとつずつ[MASK]トークンに置き換え，その[MASK]トークンにあてはまるトークンを予測させる．

人物役割語あたりのトークン数可変 各人物語ごとに，1 から 4 までの[MASK]トークンの連続数それぞれで連続する[MASK]トークンを推測する．たとえば，人物語が 1 つの場合は[MASK]トークンの数が異なる 4 パターン，人物語が 2 つの場合は 8 パターンの推測を行う．それぞれの推測結果について，[MASK]を推測結果で置換したうえで，MLMScoring [11]により各文章のスコアを計算し，その値が最も大きいものを最終的な予測結果とする．

いずれの場合でも，予測したトークン列が元の文章で該当する人物役割語と同じであるかどうかを，元の人物役割語の単位でカウントして評価する．

<問題番号 ： R02-9-U, 正解ラベル ： Yes>
<関連条文>
第二百条 占有者がその占有を奪われたときは，占有回収の訴えにより，その物の返還及び損害の賠償を請求することができる．
<問題文>
A は自己の所有する工作機械を B に賃貸していたが，B は，工作機械の賃貸借契約継続中に工作機械を C に窃取された．この場合，B は，A から独立して，C に対して占有回収の訴えを提起することができる．

図 2 人物を表すアルファベットを含む問題例

4. 実験

4.1 追加学習用訓練データセット

訓練データとして，日本語の法律関連文書をいくつか用意し，十分なデータ量が得られるようにした．具体的には，民法の全条文テキスト，最高裁判所の裁判例検索システム [i]より収集した日本の民事事件判例，弁護士ドットコムのデータセット [12]の 3 種類である．それぞれ以下の前処理を適用した．

民法条文　「第一篇，第一章，第一条，第一節，第一項」といった条や項の番号，および見出しは削除した．結果，1,381 件の条文を得た．

民事判例　事件名と判決日，ページ番号，空白および空白行は削除した．また，文頭にある見出し番号のカタカナやアルファベット，数字も削除した．また，1 文をトークナイズした際に，512 トークンを超えるような文は除外した．結果，23,854 件の民事事件判例を得た．

弁護士ドットコム　弁護士ドットコムのデータは，質問者による質問文と弁護士による質問への回答文で構成されている．このうち，弁護士による回答の文章のみを抽出した後，その文章内に前述の人手で列挙した人物役割語のうちのいずれかが 1 つでも含まれている回答のみを残し，含まれていない回答文は除外した．また，民法条文の処理と同様に，1 文が 512 トークンを超えるような文は削除した．結果，38,127 件の回答を得た，

4.2 人物推論用データセット

関連研究の章で紹介した COLIEE 2022 Task4 では，民法分野短答式問題の問題文とその問題を解くために必要な関連条文群のペアが与えられた(図 1)．このデータの分量は他のデータセットよりも少ないが，文章が最も整っているうえ，関連条文とペアになっており法律上の役割と紐づけされている．さらにその紐づけにより人間にも与えられたペアのみで解くことができるため，テストデータに適している．

まずこれらの問題文のうち，図 2 のように問題文中に A や B といったアルファベットで表された人物が登場する問題と含まれない問題とに，アルファベットが出現するかどうかで自動分類した．次に，人物役割語を含むデータを増やすため，アルファベットを人物役割語に置換した．すなわち，「～である A」「～B」「～の C」のいずれかのパターンで，「～」にリストの単語がマッチする場合，その問題文中にあるアルファベットを対応する人物語に置き換えた．置き換えられないアルファベットが残った場合は除外した．全て置き換えられた場合，テストデータの一部とする．

アルファベット人物が含まれない問題において，あらかじめ定めた人物役割語が一つでも含まれる問題をテストデータとして利用する．

テストデータについて，関連する民法条文と問題文を[SEP]トークンでつなぎ，文章をトークナイズした際に（512 － 置き換えた人物語の数 ＊ 4 トークン）を超えてしまうものは最大トークン数の制約のため除外する．

最後に，問題文中に含まれる人物役割語を[MASK]トークンに置き換え，計 458 件（置き換えた人物役割語数:1,082）のテストデータを作成した．この作成過程と例を図 3 に示す．

4.3 学習

日本語事前学習済み BERT モデルとして，東北大学の bert-base-japanese-whole-word-masking[ii]モデルを用いた．学習時のパラメータ設定は付録に記載する．

5. 実験結果

表 1 に評価結果を示す．トークン数固定の場合，人物役割語を中心にマスクさせたモデルの方がより高い性能が得られた．トークン数可変（未知）の場合，トークン数固定よりも難易度の高い設定になるため全体的に性能は低いが，トークン数固定の場合と同様に，人物語を追加でマスクするようにして学習したモデルの方が高い性能が得られた．

表 1 人物役割語推論の評価結果（正答率）

トークン数	固定	可変
ベースライン	0.50	0.18
15%	0.48	0.16
15%+**人物語**	0.51	**0.20**
人物語	**0.53**	**0.20**

[i] https://www.courts.go.jp/app/hanrei_jp/search1?reload=1
[ii] https://huggingface.co/cl-tohoku/bert-base-japanese-whole-word-masking

〈問題番号 ： H18-27-U，正解ラベル ： No〉
〈関連条文〉
弁済の費用について別段の意思表示がないときは，その費用は，債務者の負担とする．ただし，債権者が住所の移転その他の行為によって弁済の費用を増加させたときは，その増加額は，債権者の負担とする．
〈問題文〉
持参債務の債権者が履行期前に遠方に転居した場合，目的物の運送費は債務者の負担となる．

1．関連条文と問題文を[SEP]トークンでつなげる．
(以下，問題文のみ表記する)

2．関連条文と問題文の組をトークナイズする
[持参，債務，の，債権，者，が，履行，期，前，に，遠方，に，転居，し，た，場合，，，目的，物，の，運送，費，は，，，債務，者，の，負担，と，なる，．]

3.「債権者」,「債務者」に該当するトークンを[MASK]トークンに置き換える．
〈置き換え後の問題文〉
[持参，債務，の，[MASK]，[MASK]，が，履行，期，前，に，遠方，に，転居，し，た，場合，，，目的，物，の，運送，費，は，[MASK]，[MASK]， の，負担，と，なる，．]

図 3 テストデータの作成過程

6. 考察

人物語を追加でマスクする追加学習を行ったモデル（**15%＋人物語**）は，追加学習を行っていないモデル（**15%**やベースライン）よりも，関連条文をはじめとする周囲の文脈情報をもとにした人物役割語の予測がより正解できるようになった．例を図 4 に示す．

一方で，登場する人物役割語が複数になった場合，複数の人物間の関係性を明確にする必要があるが，このような人物役割語の予測はできていない傾向にあった．図 5 は，予測する際に債務者と債権者の関係性を捉える必要があるが，全てのモデルが「債権者」が正解となる部分に「債務者」を予測してしまった例を示している．

人物役割語を重点的にマスクすることで，周囲の文脈情報から人物役割語の推測は比較的にできるようになったと考えられるが，複数の人物間における関係性の把握は依然として課題であると考えられる．また，図 5 の予測に見られるように，本来は「第三者」が正解であるが，「者」であっても文章は自然であり間違いとはいえない．単純な文字列比較だけでなく意味的な評価指標も検討する必要がある．

<問題番号 ：H21-3-2，正解ラベル ：Yes >
<関連条文>
まだ引き渡されていない売買の目的物が果実を生じたときは，その果実は，売主に帰属する．買主は，引渡しの日から，代金の利息を支払う義務を負う．ただし，代金の支払について期限があるときは，その期限が到来するまでは，利息を支払うことを要しない．
<問題文>
【売主】は，目的物の引渡しを遅滞している場合でも，引渡しまで果実を収取することができる．
※ 【】がマスクされていた人物役割語を表す．
<予測結果>
ベース BERT ： 買主
15% のみ ： 買主
15%＋ 人物語 ： 売主
人物語のみ ： 売主

図 4 正しく予測できるようになった例

〈問題番号 ： H28-20-1，正解ラベル ： Yes〉
〈関連条文〉
債務者のために弁済をした者は，債権者に代位する．
〈問題文〉
【債務者】の意思に反することなく有効に弁済した【第三者】は，弁済によって当然に【債権者】に代位する．
※ 【】がマスクされていた人物役割語を表し，予測の 1,2,3 は問題文の登場順に対応している．
〈予測結果〉
BERT ： 1.債務者，2. 者 ，3．債務者
15%のみ ： 1.債務者，2. 者 ，3．債権者
15% +人物語 ： 1.債務者，2. 者 ，3．債務者
人物語のみ ： 1.債務者，2．第三者，3．債務者

図 5 正しく予測できていない例

7. おわりに

本研究では，MLM の追加学習を行った BERT による，法律文書における人物役割語予測を行った．追加で学習させたい語を重点的にマスクすることで，人物役割語推測の性能が向上した．今後は，人物役割語だけではなく，法律分野における専門語への適応や追加学習モデルを用いた司法試験の自動解答への活用にも取り組んでいきたい．

謝辞

本研究では，国立情報学研究所の IDR データセット提供サービスにより弁護士ドットコム株式会社から提供を受けた「弁護士ドットコムデータセット」を利用した．本研究は JSPS 科研費 JP22H00804，JP21K18115，JP20K20509，JST AIP 加速課題 JPMJCR22U4，およびセコム科学技術財団特定領域研究助成の支援をうけた．

参考文献

1. **西原 弘真，白井 清昭.** 物語テキストを対象とした登場人物の関係抽出．言語処理学会 第 21 回年次大会, 2015.

2. **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.

3. **宮崎 桂輔, 菅原 祐太, 山田 寛章, 徳永 健伸.** 日本語法律 BERT を用いた重要箇所抽出．言語処理学会 第 28 回年次大会, 2022.

4. **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.

5. **Mi-Young Kim, Juliano Rabelo, Randy Goebel, Masaharu Yoshioka, Yoshinobu Kano, and Ken Satoh.** COLIEE 2022 Summary: Methods for Legal Document Retrieval and Entailment. In Proceedings of the Sixteenth International Workshop on Juris-informatics (JURISIN 2022), 2022.

6. **J. Rabelo, R. Goebel, M.-Y. Kim, Y. Kano, M. Yoshioka, and K.Satoh**, Summary of the Competition on Legal Information Extraction/Entailment (COLIEE) 2021. In Proceedings of the Eighth International Competition on Legal Information Extraction/Entailment (COLIEE 2021), 2021, pp.1-7.

7. **J. Rabelo, M.-Y. Kim, R. Goebel, M. Yoshioka, Y. Kano, and K. Satoh**, COLIEE 2020: Methods for Legal Document Retrieval and Entailment. In Proceedings of the Fourteenth International Workshop on Juris-informatics (JURISIN 2020), 2020, pp. 1–15.

8. **R. Goebel, Y. Kano, M.-Y. Kim, J. Rabelo, K. Satoh, and M. Yoshioka**, COLIEE 2019 Overview. In Proceedings of the Competition on Legal Information Retrieval and Entailment Workshop (COLIEE 2019) in association with the 17th International Conference on Artificial Intelligence and Law, Jun. 2019, pp. 1–9.

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, 2019. Association for Computational Linguistics.

10. **清田 直樹, 狩野 芳伸, 藤田 真伎.** 深層学習モデルは司法試験をどこまで解いているのか：問題分類とそれに基づく分析．言語処理学会 第 28 回年次大会, 2022.

11. **Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff**. Masked Language Model Scoring. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 2699–2712, Online. 2020. Association for Computational Linguistics.

12. **弁護士ドットコム株式会社.** 弁護士ドットコムデータセット．国立情報学研究所情報学研究データリポジトリ. (データセット). https://doi.org/10.32130/idr.12.1, 2020.

A 付録

A.1 人手で選定した人物役割語のリスト

自然人，法人，本人，相手方，他人，当事者，第三者，表意者，譲受人，占有者，
譲渡人，所有者，地上権者，地役権者，永小作人，管理人，不在者，債権者，留置権者，先取特権者，
質権者，根抵当権者，抵当権者，債務者，質権設定者，根抵当権設定者，抵当権設定者，物上保証人，
連帯保証人，保証人，連帯債務者，申込人，売主，買主，贈与者，受贈者，貸主，借主，賃貸人，賃借人，
使用者，労働者，注文者，請負人，委任者，受任者，受寄者，受託者，未成年者，成年者，成年後見人，
成年被後見人，未成年被後見人，後見監督人，被後見人，後見人，保佐監督人，被保佐人，保佐人，
補助監督人，被補助人，補助人，被相続人，相続人，法定代理人，復代理人，無権代理人，代理人，
監督人，共有者，承継人，利害関係人，組合員，代位権者

A.2 BERT 学習時のパラメータ設定

max_position_embeddings : 512,
batch_size : 32,
hidden_size : 768,
hidden_dropout_prob : 0.1,
learning_rate : 1e-5
max_epochs : 10 (early_stopping による早期終了あり)