民事判決のオープンデータ化へ向けた機械処理による判例仮名化の検証

久本空海　城戸祐亮　津金澤佳亨　八木田樹
株式会社 Legalscape
info@legalscape.co.jp

概要

民事判決を公開するためには、個人情報や秘匿情報への配慮が課題となっており、それらの匿名加工（仮名化）が必要とされる。しかし、大量の判決全てを人手のみで匿名加工することは現実的に難しい。

我々は判例仮名化を固有表現抽出問題の拡張と捉え、その自動処理の実現性を検証した。実際の判例データを用いた実験により、事前学習済み言語モデルを使った系列ラベリングで一定以上の性能（適合率 93.4%、再現率 94.5%）が実現可能なことを確認した。加えて、人手修正の方向性についても検討した。

なお、当発表に関する情報はウェブ上の記事としても公開している[i]。

1 はじめに

法治国家において判例（裁判所による過去の法律的判断）は、法的安全性のために重要な情報である。またそのようなデータは、様々な研究や新たなサービスの開発にも有益なものである。

日本国における判例は現状、裁判所によって社会的関心が高いと判断されたものがウェブサイト掲載されているほか、法律系出版社などが一部を独自に収集し公開している。町村の調査[1]によれば、2017年に全国の地方裁判所の民事・行政事件の処理済み件数は約 16 万件だったが、裁判所がウェブ掲載した判決数はそのうち 44 件（0.03%）のみであり、民間企業である Westlaw Japan の判例データベースでも 5,033 件（3.02%）にすぎない。

日本国における民事司法制度改革の流れから 2020 年 3 月、「民事判決のオープンデータ化検討プロジェクトチーム（PT）[ii]」が日弁連法務研究財団により設置された。これは、政府による民事裁判手続き IT 化の取り組みに併せ、民事判決を電子化し公開することを検討するプロジェクトである。この取り組みには、日本弁護士連合会の関係者や法律系出版社、法学研究者などが構成員として関わり、加えてオブザーバーとして内閣官房や法務省、最高裁も参加している。

判例のオープンデータ化へ向けて大きな課題となるのが、個人情報や秘匿情報の取り扱いである。センシティブな情報が含まれうる判決文は、そのままの形で公開することは難しく、これらの匿名加工（仮名化）が必要となる。

日本において年間数十万件とも言われる民事判決の全てを、人手のみで匿名加工することは現実的に難しい。そのときに、その作業を自然言語処理技術で（半）自動化するということが考えられる。

我々は判例の仮名化を固有表現抽出問題の拡張として機械処理による検証を行い、一定以上の性能が実現可能であることを確認した。当発表では各国の状況や、実際の日本語判例文を用いた実験、そしてそれを踏まえた民事判決オープンデータ公開を実現するに当たっての課題について述べる。

2 判例の仮名処理

判例の匿名加工処理においては、単に個人情報や秘匿情報など文中のセンシティブな部分を“黒塗り”すれば十分というわけではない。例えば「**原告山田**」「**被告佐藤**」「**佐藤が～**」といった文中箇所を、一律に「**原告■**」「**被告■**」「**■が～**」と黒塗りしてしまうと、各箇所がどの実体を指すかがわからず、閲覧者にとっての利用価値が損なわれてしまう。そのために、それらが同一実体であることを示すために“仮名記号”を割り振り「**原告 X**」「**被告 Y**」「**Y が～**」などと置き換える処理が施されることが多い。これらの処理を総称して仮名化と呼ぶ。

裁判所や法律系出版社などから現在公開されている判例は、人手作業による仮名化が行われている。

[i] https://note.com/legalscape/n/nf6341940deaa
[ii] https://www.jlf.or.jp/work/hanketsuopendata-pt/

しかし、もしこれが年間数十万件というスケールになると、それら全てを人手のみで実施するのは現実的ではない。そこで検討したいのが、自然言語処理技術の活用による（半）自動化である。

2.1 他国での取り組み

米国においては既に Public Access to Court Electronic Records （PACER）[iii]という公的データベースが整備されており、典型的には裁判後 24 時間以内に情報が閲覧可能となる。これは、多くにおいて仮名化がなされていないために可能だと考えられる（ただし刑事事件では一部の個人情報が除去・編集される[iv]）。

多くの国では日本と同様に仮名化の問題が存在する。近年、その機械処理に取り組んでいる事例もいくつか公開され始めているが、未だどの国においてもその仕組みが大規模に実運用されるまでは至ってないようである。

フィンランドでは GDPR（EU 一般データ保護規則）に従うため、統計的手法とルールベース手法を組み合わせた半自動の個人情報匿名加工システムの開発が研究機関と法務省により進められている[2]。デンマークからは、法律系出版社による判例の自動匿名加工に関する報告がある[3]。ウルグアイでも、法文書の自動匿名加工に関する検証が報告されている[4]。フランス最高裁判所からは、判例に対する固有表現抽出の精度改善について報告がある[5]。また、匿名加工に直接利用されてはいないが、法領域での固有表現抽出についてドイツからの報告がある[6,7]。

3 日本語判例仮名化の実証実験

当節では 2020 年度に前述の PT で実施した機械処理による判例仮名化の検証について述べる。

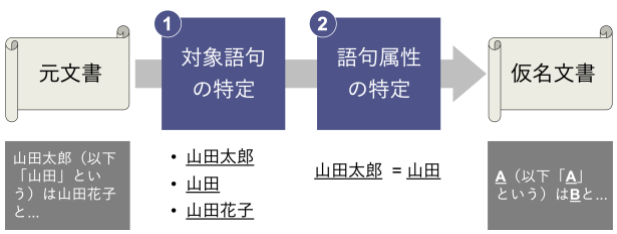

図 1 機械処理の流れ

[iii] https://pacer.uscourts.gov/

[iv] https://pacer.uscourts.gov/help/faqs/what-information-available-through-pacer

3.1 問題設定

今回の検証では機械処理による仮名化を「1.対象語句の特定」と「2. 語句属性の特定」という 2 つのステップに分けて処理を行った（図 1）。

1.は、いわゆる固有表現抽出の問題である。このとき、仮名が「漏れた」時にはプライバシーリスクに影響し、逆に仮名「し過ぎた」時には、閲覧者の利用性や権利へ影響する。またこれは単純な分類問題と違い「部分的な誤り（部分的な正解）」もありえる。このようなケースは、例えばもし機械処理後に人手修正をするのであれば、その生産性へ影響することが想定される。これらの観点を踏まえて、単に「精度 95%」などというのではなく、適合率・再現率それぞれの値や、完全一致と部分一致での性能の違いなどを確認していく必要がある。

2.は、関係抽出やエンティティ・リンキング、共参照解析といったタスクと類似したものであり、仮名対象語句ごとに、適切な仮名記号を選択する処理である。そのとき単に同じ文字列を同じ記号とするのではなく、指し示す実体や、語句の種類を考慮する必要がある。例えば、同じ「山田」という語句でも、それが同性の別人物であれば別記号を付与する必要がある。他にも、住所が本籍かそれ以外かによって仮名化の粒度を区別することもある。これは仮名基準次第であり、今回はある一定の基準をもとに検証したが、法的要件や利用ニーズによって基準は変化する。この基準については別途 PT で、法学者らなどにより議論が進められている。

3.2 データ

今回の検証では、実際の判例原文と、それを人手により仮名化した文書のペア、計 1,642 件を用意した。そのうち 100 件ずつを検証セット、評価セットとし、残り 1,442 件（約 15 万行、1,200 万文字）をモデル学習およびルール作成に利用した。

3.3 手法

予備調査を踏まえ当検証では最終的に、処理ステップ 1.にはファインチューンした BERT [8]、2.にはルール処理を用いて仮名化を実施した。

処理ステップ 1.について、法文書は一般的な自然言語文書に比べて規則的に書かれていることからまずはルール記述による処理を試みたが、今回試した

範囲では高い性能には至らなかった。次に spaCy[v][9]による固有表現抽出モデルの学習と検証を行ったが、結果として再現率 85%、適合率 92%程度に留まった。加えて spaCy モデルでは言語資源として法律用語一覧（約 1.2 万語）の活用も試みたが、性能への影響は確認できなかった。これには、そもそも対象判例データにおいて約 0.9%の語のみが該当する法令用語であること、また法律用語自体が仮名対象になることはないこと、などが理由として考えられる。

上記の予備検証を経て、最終的に利用したのは次のようなモデルである。ベースには東北大学による日本語の事前学習済み BERT（Whole Word Masking）[vi]を用い、それを別途用意した判例文書（仮名後のみ）約 2.5 万件により MaskedLM で 10 万ステップの継続学習を行った。その後、学習用の判例データにより固有表現抽出モデルをファインチューニングした。また BERT モデルに加えて、少しの簡単な後処理（当事者セクションのルール解析による原告・被告名の把握、日付や電話番号のパターンによる抽出、など）も行った。

今回は時間やリソースの都合上、各設定で一つのモデルのみを学習した。異なる初期値から複数のモデルを学習することによる平均値の確認などをしたわけではなく、あくまでも大まかにこの問題設定と実データでどの程度の性能が達成可能か、どのような誤り傾向になるかを確認したのみであることに注意が必要である。

同様に今回は、最高性能の追求を十分に行ったわけではない。例えば、文字単位ではなく単語単位 BERT の利用や、CRF 層の追加、BERT 以外の言語モデルの導入などにより、性能向上が達成できる可能性がある。

処理ステップ 2.（語句属性の特定）は、統計的モデルではなく、基本的にルールベースで実施した。本来、関係抽出や共参照分析といった問題は容易なものではないが、法律文書においては曖昧な記述は基本的に無く「**山田太郎（以下「山田」という。）**」と明示的に述べられているなど、ルールによる処理と相性が良いと言える。今回は具体的な処理として、前述したような明示的記述の利用や、形態素解析器による人名フルネームの姓・名への分割、ひらがなやカタカナといった異表記での対応付けなどを行った。

[v] Transformer ベースではない v2 系

[vi] https://huggingface.co/cl-tohoku/bert-base-japanese-char-whole-word-masking, 簡単の為に文字単位でのモデルを利用

3.4 結果

表 1 固有表現抽出の性能

	完全一致		部分一致	
	適合率	再現率	適合率	再現率
箇所数	93.4	94.5	96.4 (+3.0)	96.8 (+2.3)
種類数	87.6	91.8	94.5 (+6.9)	96.4 (+4.6)

処理ステップ 1.（対象語句の特定）の性能を表 1 に示す。のべ仮名箇所数ベースで、適合率 93.4%（部分正解含め 96.4%）、再現率 94.5%（部分正解含め 96.8%）となった。一方、仮名単語の種類数ベースではより低い性能となっており（適合率は 5.8 ポイント、再現率は 2.7 ポイント低下）、これは出現回数が多い単語における誤りが比較的少ないことを示している。全体としては約 5 割の判例は修正を全く必要とせず、必要とする判例でも平均 7 箇所、3 単語の修正に留まるという結果になった。

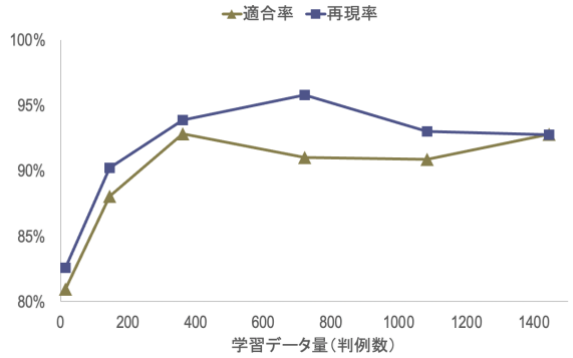

図 2 学習データ量と固有表現抽出モデル性能

また、学習データ量を減らしたモデルでは、25%（361 件）程度で最大量（1,442 件）と同程度の性能に至った（図 2）。単に学習データ量を増加させても、ここから性能が大幅に向上することは見込めない可能性が示唆される。この背景には、そもそも法律文書は規則的に記述されていることや、難しい事例は基本的に特殊な固有名詞であって判例データが増加してもそれらのカバレッジが上がるわけではないことなどが想定される。

そして 1.を完全に正解できた前提では、処理ステップ 2.（語句属性の特定）は精度[vii]98.0%となった。

[vii] 「正解の仮名語集合（仮名記号）」のうち「システムが完全に正しく出力した集合」の数の割合

3.5 エラー事例

ここでは、どのような誤りの類型があったかの偽事例を示す。これらはあくまで実際の誤り事例を参考にして人為的に作成した疑似的な例であり、実際の判例文は一切掲載していない。

表 2 固有表現抽出における誤りの疑似例

種類	誤り	人手正解と機械出力
人名	過多	正解 被告は金を要求した 出力 被告は**金**を要求した
人名	漏れ	正解 ネットネームの **tanaka@net** が 出力 ネットネームの tanaka@net が
企業名	過多	正解 被告は国立病院に 出力 被告は**国立**病院に
企業名	漏れ	正解 被告は**すみれフラワー**に 出力 被告はすみれフラワーに
固有名	漏れ	正解 ネット掲示板「**東京グルメ案内**」 出力 ネット掲示板「東京グルメ案内」
住所	部分	正解 **東京都北区**北部地域 出力 **東京都北区北部地域**
数値	漏れ	正解 ログイン ID: **12345** 出力 ログイン ID: 12345
生年月日	漏れ	正解: **1 月 1 日**に息子 A を出産 出力: 1 月 1 日に息子 A を出産

処理ステップ 1.（対象語句の特定）での誤り例を表 2 に示す。人名や企業名に関しては、それが一般名詞ともなり得る場合の誤りが多く見られた。また、インターネット上でのハンドルネームやスレッド名は多様に表記され、難易度が高い。住所はスパンが長くなり、完全な誤りは稀だが部分的な誤りが見られた。数値表現は、電話番号などのようにパターンがある程度既知でない場合の誤りが多かった。生年月日は、それが単なる日付表現かどうかは文脈から汲み取る必要があり誤るケースがあった。

処理ステップ 2.（語句属性の特定）では、以下に述べる 4 つの誤り類型が確認された。

- 明示的ではない略称: 「**東京タナカ病院**」→「**タナカ病院**」
- 一意に特定できない人名: 「**田中太郎**」「**田中二郎**」が登場する判例での曖昧性のある「**田中**」
- 同表記だが区別される名前: 「**田中が所有する田中ビル**」→「**A が所有する B ビル**」
- その他、明示的ではない別表記: ニックネーム（「**田中直樹**」→「**ナオ**」）や（ネットネーム（「**田中**」→「tanaka@net」）など

このステップ 2.での誤りは、1.と異なりプライバシーリスクには影響しない。一方で閲覧者の利便性には影響しうるが、現在の簡単なルール処理でも 100%に近い精度が達成できている。そのため、判例の仮名化においては、ステップ 1.がより難しく重要な問題だと言える。

4 判決オープンデータ化へ向けて

前節で述べた初期検証により、機械処理で 90%以上の自動仮名化は達成しうることが確認された。

しかし、個人情報への配慮や閲覧者の利便性を考えると、高性能だから実用可能であるとは必ずしも言えない。そして機械で処理する限り、ここから性能改善を重ねても 100%を保証することはできない。そのため、完全な仮名化には人間のチェックが不可欠となる[viii]。

完全な自動化ではなく、機械出力の活用による人間作業の高速化によって大量のデータを処理するという方向性が、オープンデータ化へ向けて PT で検討している一案である。そのためには、機械処理の適合率（仮名過多）を妥協してでも再現率（仮名漏れ）の誤りを減らすことが有用かもしれない。また、モデルへの人手修正の適切なフィードバックが効果的かもしれない。加えて、人間が作業しやすいツールの開発も非常に重要だと考えられる。

前述した検証結果を踏まえ、我々は判決のオープンデータ化へ向けて、機械出力を活用した効果的な人手修正作業の検証などに引き続き取り組んでいる。

5 おわりに

当論文では、民事判決の公開を実現するために必要な仮名化の概要や、各国の状況を解説した。また、それを固有表現抽出問題の拡張としたとき、実際の判例文を用いた検証で一定以上の性能（適合率 93.4%、再現率 94.5%）が実現可能なことを示した。

実際に日本国で民事判決をオープンデータ化するには、人手による機械出力の修正や運用スキーマの検討など様々な課題があり、その実現へ向けて更なる検証や議論を進めている。

[viii] 違う方向性として例えば、閲覧者を限定することによる完全な修正を必要としない判例公開スキームも検討の余地がある

謝辞

当検証の実施にあたって、民事判決オープンデータ化 PT の構成員にサポートと助言を頂いた。また法律分野での自然言語処理について、名古屋大学の外山勝彦教授、小川泰弘准教授に助言を頂いた。

参考文献

1.**民事判決 ネット上で提供 官民で検討、23 年度にも**. 日本経済新聞, 2020 年 6 月 7 日. https://www.nikkei.com/article/DGXMZO60029090V00C20A6000000.

2. Oksanen, A., Tamper, M., Tuominen, J., Hietanen, A., & Hyvönen, E. (2019). **ANOPPI: A Pseudonymization Service for Finnish Court Documents**. JURIX.

3. Povlsen, C., Jongejan, B., Hansen, D.H., & Simonsen, B.K. (2016). **Anonymization of Court Orders**. 2016 11th Iberian Conference on Information Systems and Technologies (CISTI), 1-4.

4. Garat, D., & Wonsever, D. (2019). **Towards De-identification of Legal Texts**. ArXiv, abs/1910.03739.

5. Barrière, V., & Fouret, A. (2019). **May I Check Again? — A simple but efficient way to generate and use contextual dictionaries for Named Entity Recognition. Application to French Legal Texts.** NODALIDA.

6. Leitner, E., Rehm, G., & Schneider, J.M. (2019). **Fine-Grained Named Entity Recognition in Legal Documents**. SEMANTiCS.

7. Leitner, E., Rehm, G., & Moreno-Schneider, J. (2020). **A Dataset of German Legal Documents for Named Entity Recognition**. LREC.

8. Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). **BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding**. ArXiv, abs/1810.04805.

9. Honnibal, M., Montani, I., Van Landeghem, S., &. Boyd, A. (2020). **spaCy: Industrial-strength Natural Language Processing in Python**. 10.5281/zenodo.121230.