1W-04

Transformer モデルによる法律関係文書の自動匿名化

関 泰明 †　　濱上 知樹 ‡
横浜国立大学 †　　横浜国立大学 ‡

1 はじめに

裁判の判例などの法律関係文書の公開に際して，膨大なテキストに対する匿名化処理がボトルネックとなっている．最高裁判所の令和 2 年度の民事既済事件件数は 5366 件 [1] であるのに対し，裁判所 HP で公開されている判例の数は 45 件 (2021-12-11 時点) と，判例の公開が進んでいないのが現状である．したがって法律関係文書の自動匿名化が求められている．

法律関係文書では匿名化する単語の種類ごとに出現数に不均衡があり，出現数の少ない単語の学習を困難にしている．本研究では不均衡下で高精度な匿名化を行うことを目的とし，Transformer 系の自然言語処理モデル BART[2] とデータ拡張を組み合わせた手法を提案する．

2 関連研究

既存のテキスト匿名化手法として，k-匿名化がある [3]．これはテキスト内で k 回以下の出現数の表現をマスクする手法である．実装が簡便である一方で，単語の品詞を考慮できないこと，マスク量がテキストの数や長さに依存して変動することといった課題がある．

品詞や文脈を考慮できる手法として，Transformer 系の単語分類モデルである BERT を用いた固有表現抽出タスクとしての匿名化が行われている [4]．匿名化の精度は高いが，推論結果に適切な後処理を施す必要がある．また日本語のテキストに応用する場合では分かち書き手法をモデルと学習データで揃える必要がある．

3 データ拡張と BART による匿名化

本研究では，Transformer を事前学習に対応させた Seq2Seq モデルである BART を用いて匿名化を行う．Seq2Seq モデルでは学習データの分かち書き手法について考慮する必要が無く，すぐに利用できる文章がモデルの推論結果として得られるという利点がある．

また単語の出現数の不均衡への対策としてデータ拡張を組み合わせる．単語クラスごとの出現数が同程度になるように学習データの一部を複製し，複製分には過学習抑制のためのノイズとして，匿名化後文字列のアルファベットや記号と 10% の漢字 2 字以上の名詞のランダムな置換を施す．

4 使用するテキストとその分析

法律関係文書では文書の種類によって匿名化すべき単語の種類や匿名化基準が異なるため，本研究では裁判の判例テキスト [5] と弁護士への相談回答テキスト [6] という 2 種類のデータセットを用いた．これらのデータセットでは匿名化対象の単語を別のアルファベットや記号に置換する．このうち裁判の判例テキストでの匿名化の例が図 1 である．

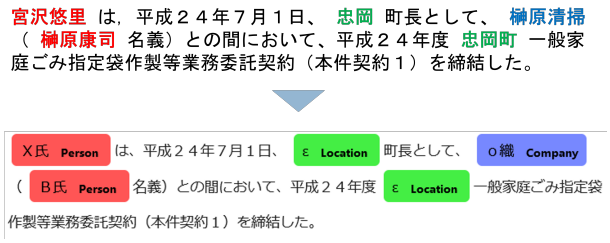

図 1　判例テキストでの匿名化の例

4.1 GiNZA による固有表現抽出

テキストの特性を可視化するために GiNZA[7] による固有表現抽出を行った．匿名化する各単語クラスに相当する固有表現を定め，その固有表現として抽出されたもののうち匿名化対象であったものの割合 (Precision) を計算した．その結果が表 1 中段である．

Precision 値の小ささは，その単語クラスに「同じ固有表現だが匿名化しない単語」が多いことを表す．したがって Precision が小さいも

Anonymization of legal documents using Transformer-based Model
† Yasuaki Seki, Yokohama National University
‡ Tomoki Hamagami, Yokohama National University

表 1　GiNZA による固有表現抽出結果 (中段) と BART による匿名化実験結果 (下段)

データセット	裁判の判例データ				弁護士への相談回答データ					
単語クラス	人名	会社名	地名	日付	人名	年齢	会社名	地名	日付	一般
出現数	14667	2301	392	246	705	130	332	174	35	283
GiNZA での固有表現	Person	Company	City	Date	Person	Age	Company	City,Country Province,	Date	School_Age, Ordinal_Number
Precision	0.9006	0.3849	0.086	0.0087	0.9781	0.5371	0.9548	0.5926	0.2074	0.0186
拡張倍率 [倍]	0.3	2.0	10.0	15.0	1.0	5.0	2.0	4.0	25.0	2.5
f1(拡張あり)	0.9448	0.4161	0.3199	0.6425	0.9509	0.8334	0.9388	0.7617	0.6289	0.6278
f1(拡張なし)	0.9119	0.2924	0.1275	0.3091	0.7645	0.0000	0.2130	0.0000	0.0000	0.0625

の程匿名化の是非を文脈やトピックから判断する必要があり，学習が困難であると考えられる．

5　匿名化の実験

2 つのテキストに対してそれぞれ「データ拡張なし」,「データ拡張あり」の 2 通りでの BART による匿名化の学習を行った．判例データでは 145 の文書を 500 字程度ごとに分割して BART 入力単位である系列を 9368 個作成し，弁護士データはテキストが長くないため分割せずに 1200 系列を用意した．精度の評価は 5 分割交差検証の平均スコア (f 値) で行った．BART での学習には日本語 Wikipedia を使用して Text Infilling のみで事前学習されたモデルを用いた．

6　実験結果

各単語クラスごとのデータ拡張の倍率と匿名化の精度を表 1 下段に示す．

拡張なしで匿名化に失敗した弁護士データの単語クラスにおいて，拡張ありでは 60% 以上の精度となった．これは単語クラス間の出現数を揃えたことの効果であると考えられる．一方で拡張ありでも精度が 70% に満たない単語クラスがあった．これは固有表現抽出の Precision が小さいもの，すなわち学習が困難なものであった．したがって，これらの学習が困難な単語クラスを重点的に学習できる手法が必要であることが明らかとなった．

7　おわりに

Transformer 系の Seq2Seq モデル BART とデータ拡張によって法律関係文書の匿名化を行う手法を提案した．2 つのテキストでの実験を行った結果，半分の単語クラスで 70% 以上の精度で匿名化ができた．一方で学習が困難な単語クラスについてはデータ拡張のみでは精度の向上が小さく，単語クラスごとの学習の困難さが精度に大きく寄与することが明らかとなった．今後はこれら学習が困難な単語クラスを重点的に学習する手法の検討を行う．

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．

参考文献

[1] 民事・行政事件数　事件の種類及び新受, 既済, 未済　最高裁判所 (オンライン), 入手先〈www.courts.go.jp/app/files/toukei/008/012008.pdf〉(参照 2021-12-11).

[2] Mike Lewis et al.: *BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension*, Association for Computational Linguistics, 7871 – 7880 (2020).

[3] 荒牧英治, 増川佐知子, 宮部真衣, 森田瑞樹: テキストの k-匿名化, 情報処理学会研究報告 Vol.2012-DBS-155 No.9 (2012).

[4] Jihang Mao, Wanli Liu: *Hadoken: a BERT-CRF Model for Medical Document Anonymization*, IberLEF, 720-726 (2019).

[5] Nishika - 判例の個人情報の自動マスキング (オンライン), 入手先〈www.nishika.com/competitions/7/summary〉(参照 2021-06-03).

[6] 弁護士ドットコム株式会社 (2020): 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

[7] 松田 寛: GiNZA - Universal Dependencies による実用的日本語解析, 自然言語処理 27(3), 695-701 (2020).