

機械学習を用いた黒塗り文書作成支援ツールの作製

金沢工業大学 工学部 情報工学科
中沢研究室 大北航暉

研究背景

[現状]

- 誰でも閲覧できる文書に含まれる個人情報の匿名化は全て手作業で行われている。
- 誤った方法で作られた匿名化処理をした文書から本来読み取ることができない情報（個人情報など）が流出してしまうことある。

[問題点]

- 手作業で行われているため時間がかかる。
- 匿名化処理が何らかの方法で解除されたときの対策がされていない

目的

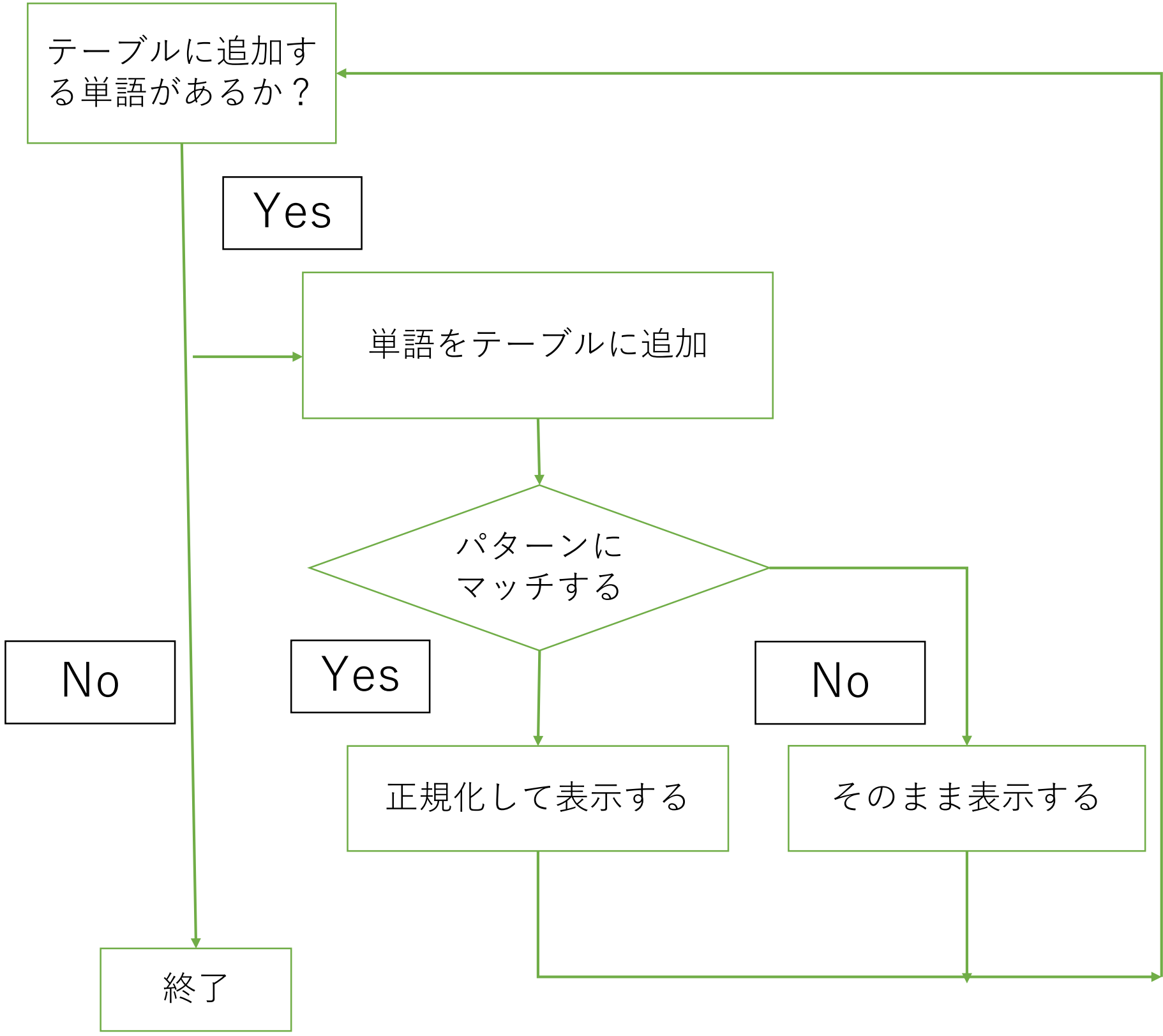
文書に含まれる個人情報の匿名化を行うシステムの作製

システム構成

1. 文書を読み込み、自然言語処理ライブラリGiNZAで単語ごとに分割し、単語ごとに品詞やベクトルなどのラベルを付ける。

```
田中太郎は、大学生です。
# text = 田中太郎は、大学生です。
1   田中   田中   PROP  名詞-固有名詞-人名-姓   -   2   compound
2   太郎   太郎   PROP  名詞-固有名詞-人名-名   -   5   nsubj   -
3   は     は     ADP   助詞-係助詞         -   2   case     -   B
4   、     、     PUNCT 補助記号-読点       -   2   punct    -   B
5   大学生 大学生  NOUN  名詞-普通名詞-一般   -   0   root     -
6   です   です   AUX   助動詞             -   5   cop      -   BunsetuBI
7   。     。     PUNCT 補助記号-句点       -   5   punct    -   B
```

2. ラベルが付いた単語を先頭から読み込み、単語のテーブルに追加していく。テーブル内の単語の並びと単語に含まれる品詞や依存構造の並びを確認し、文字や品詞などの並びが特定のパターンになった時に、正規表現を使い文字列を別の文字に変換し出力する。これを単語がなくなるまで続ける。




使用したテキスト
衆議院 厚生労働委員会 第200回国会
令和1年11月13日 第5号 (テキスト1)
令和1年11月22日 第6号 (テキスト2)

提案手法の違いについて

今回の手法

・品詞タグを使った匿名化



先行事例

・NERタグを使った匿名化

公明党の樹屋敬悟でございます。

図 NERで人名を認識できなかった場合

保険局

長浜谷浩樹 **PERSON** 君、

図 NERで人名を正しく認識した場合

先行事例

Towards De-identification of Legal Texts

- 組織間で公開または共有できる個人情報が規制されているため、機密データを削除または難読化するために、文書の匿名化プロセスを行う必要がある。
- 法的なテキストの匿名化に焦点を当てており、その目的は、物語の意味を失うことなく、訴訟に関与する人名を隠すこと。
- 結果：ドキュメントの84％には、NERツールでカバーされていない名前が少なくとも1つある。

NER(Named Entity Recognition) (和)固有表現抽出：

文中から固有表現を抽出し、定義された固有表現分類へと分類する

固有表現:人名や地名などの固有名詞、日付、時間、金額、％など

実行結果

- 一部匿名化できていないが、発言者[○ 名前 役職名]や文章中に出てくる人名や役職名の匿名化ができています。
- 匿名化している部分が人名と役職名だけなので、内容が失われていない。
- 一部の文字が文中から消えている。

○**盛山委員長** これより会議を開きます。

厚生労働関係の基本施策に関する件について調査を進めます。

この際、去る十一月十一日、ハンセン病問題対策に関する調査のため、国立療養所多磨全生園及び国立ハンセン病資料館の視察を行いましたので、参加委員を代表して、私から調査の概要を御報告申し上げます。

まず、東京都東村山市の国立療養所多磨全生園において、石井園長から概況説明を聴取するとともに、全国ハンセン病療養所入所者協議会の藤崎事務局長及び多磨全生園入所者自治会の平沢会長を始めとする皆様と懇談し、療養所に勤務する職員の定数削減問題やハンセン病に対する差別根絶に向けたさらなる普及啓発の促進等について要望を受けました。

その後、園内の各施設について説明を聴取した後、入所者の御遺骨が安置されている納骨堂において、亡くなられた方々の御冥福をお祈りするとともに、献花を行いました。

次に、国立ハンセン病資料館に向かい、ハンセン病にまつわる歴史や、過酷な状況の中で生活をしてこられた入所者の方々の体験を示す展示資料などについて、職員から説明を聴取しました。

以上が視察の概要であります。

最後に、今回の視察に御協力をいただきました皆様に心から御礼を申し上げ、視察の報告とさせていただきます。

○**盛山委員長** 次に、第百九十八回国会、内閣提出、医薬品、医療機器等の品質、有効性及び安全性の確保等に関する法律等の一部を改正する法律案を議題といたします。

この際、お諮りいたします。

本案審査のため、本日、政府参考人として内閣官房全世代型社会保障検討室次長河西康之君、内閣府大臣官房長大塚幸寛君、公正取引委員会事務総局経済取引局長菅久修一君、財務省大臣官房審議官小野平八郎君、文部科学省大臣官房審議官森見憲君、厚生労働省大臣官房長土生栄二君、大臣官房年金管理審議官日原知己君、医政局長吉田学君、健康局長宮寄雅則君、医薬・生活衛生局長榎見英樹君、子ども家庭局長渡辺由美子君、老健局長大島一博君、保険局長浜谷浩樹君、経済産業省大臣官房審議官上田洋二君の出席を求め、説明を聴取し、また、会計検査院事務総局第一局長三田啓君の出席を求め、説明を聴取したいと存じますが、御異議ありませんか。

〔「異議なし」と呼ぶ者あり〕

○**盛山委員長** 御異議なしと認めます。よって、そのように決しました。

○**盛山委員長** 質疑の申出がありますので、順次これを許します。三ッ林裕巳君。

○三ッ林委員 おはようございます。自由民主党の三ッ林裕巳でございます。

変換前のテキスト(一部)

○**■■■■■** より会議を開きます。

厚生労働関係の基本施策に関する件について調査を進めます。

この際、去る十一月十一日、ハンセン病問題対策に関する調査のため、国立療養所多磨全生園及び国立ハンセン病資料館の視察を行いましたので、参加委員を代表して、私から調査の概要を御報告申し上げます。

まず、東京都東村山市の国立療養所多磨全生園において、■■■■■概況説明を聴取するとともに、全国ハンセン病療養所入所者協議会の■■■■■多磨全生園入所者自治会の■■■■■始めとする皆様と懇談し、療養所に勤務する職員の定数削減問題やハンセン病に対する差別根絶に向けたさらなる普及啓発の促進等について要望を受けました。

その後、園内の各施設について説明を聴取した後、入所者の御遺骨が安置されている納骨堂において、亡くなられた方々の御冥福をお祈りするとともに、献花を行いました。

次に、国立ハンセン病資料館に向かい、ハンセン病にまつわる歴史や、過酷な状況の中で生活をしてこられた入所者の方々の体験を示す展示資料などについて、職員から説明を聴取しました。

以上が視察の概要であります。

最後に、今回の視察に御協力をいただきました皆様に心から御礼を申し上げ、視察の報告とさせていただきます。

○**■■■■■** 第百九十八回国会、内閣提出、医薬品、医療機器等の品質、有効性及び安全性の確保等に関する法律等の一部を改正する法律案を議題といたします。

この際、お諮りいたします。

本案審査のため、本日、政府参考人として■■■■■君、■■■■■君、■■■■■君、■■■■■君、■■■■■君、■■■■■君、大臣官房年金管理審議官日原知己君、■■■■■君、健康局長宮寄雅則君、医薬・生活衛生局長榎見英樹君、子ども家庭局長渡辺由美子君、老健局長大島一博君、保険局長浜谷浩樹君、経済産業省大臣官房審議官上田洋二君の出席を求め、説明を聴取し、また、会計検査院事務総局第一局長三田啓君の出席を求め、説明を聴取したいと存じますが、御異議ありませんか。

〔「異議なし」と呼ぶ者あり〕

○**盛山委員長**御異議なしと認めます。よって、そのように決しました。

○**■■■■■** ■■申出がありますので、順次これを許します。三ッ林裕巳君。

○三ッ林委員 おはようございます。自由民主党の三ッ林裕巳でございます。

変換後のテキスト(一部)

評価方法

- システムで作成した役職と人名を別の文字に変換した文書と手作業で役職と人名を別の記号に変換した文書の2つを、n-gramを用いて類似度を求める。

評価方針

- システムで作成した文書に含まれる人物名や役職名がどれだけ別の文字に変換できたか。

今後の課題

- 一部の匿名化ができなかった人物名や役職名の匿名化
- 匿名化したテキストの改行の改善

以上のことから、変換処理に関するアルゴリズムの改善を行う。