Vol.2021-IS-158 No.2
2021/12/4

自然言語処理技術を用いた地域防災計画における災害廃棄物処理の特徴抽出

冨江 伸太朗[1]　廣井　慧[2]　畑山　満則[2]

概要：地方公共団体における災害対応は地域防災計画に基づいて行われる．そのため，計画を常に改定し，不備をなくしておくことは非常に重要である．しかし，計画の改定に際して，当該地公共団体の過去の災害の教訓を含めることは見られても，他の地方公共団体における知見を反映することは難しく，それゆえ，災害時に問題が発生して始めて，計画の問題点が顕在化する事例は少なくない．これは，防災計画改定のために他の地方公共団体が得た教訓を分析することの困難さが一因である．そこで本研究では，近年問題となっている廃棄物処理の分野に着目して，防災計画の不備や特長を発見することを目的とした自然言語処理による分析手法に関して考察を行う．

キーワード：災害廃棄物処理, 地域防災計画, 自然言語処理, tf-Ridf, BM25, doc2vec

Feature Extraction of Emergency Debris Operation in Local Disaster Management Plan with NLP Methods

Abstract: Disaster response in local governments is based on Local Disaster Management Plan. Therefore, it is very important to constantly revise the plan and eliminate any deficiencies. However, even if the lessons learned from past disasters of the local government are included in the revision of the plan, it is difficult to reflect the knowledge of other local governments. Therefore, sometimes, the problems of the plan will not become apparent until a problem occurs in the event of a disaster. This is partly due to the difficulty in analyzing the lessons learned by other local governments for the revision of disaster risk reduction plans. Therefore, in this study, we focus on the field of waste treatment, which has become a problem in recent years, and we consider the deficiencies and features of disaster prevention plans using an analysis method using natural language processing.

Keywords: Emergency Debris Operation, Local Disaster Management Plan, Natural language processing, tf-Ridf, BM25, doc2vec

1. 背景

1.1 地域防災計画改定における課題

地域防災計画は，防災のために処理すべき業務を具体的に定めた計画であり，全国の各市町村で作成されている．日本の防災対策の基本となる，災害対策基本法によると，地域防災計画については，毎年検討を加え，必要があると認めるときは，これを修正しなければいけないこととされているが，発災後になって始めて，地域防災計画の問題点が顕在化するという事態が災害の度に発生している．また，ある自治体においては記述に不備が存在している一方で，他の自治体においては既に不備の修正が行われている場合もある．直近では，令和元年台風第 15 号 (房総半島台風) の際，千葉県災害対策本部の設置基準に関し，千葉県地域防災計画の不備が存在したために，災害対策本部設置が被災翌日となるという問題が発生した [1]．一方で，高知，和歌山，兵庫，鹿児島県等ではこの問題を事前に把握しており，速やかに対策本部を設置できるよう防災計画に記述がなされていた [1]．もしも，千葉県が他の自治体における記述を参考にしていれば，この不備は修正できていた可能性がある．しかし，計画の改定に際して，他の地方公共団

[1] 京都大学 情報学研究科
Graduate School of Informatics, Kyoto University.
[2] 京都大学 防災研究所
Disaster Prevention Research Institute, Kyoto University.

ⓒ 2021 Information Processing Society of Japan

体の知見を適切に取り入れようとすれば，防災政策に精通した専門家により，複数の計画を比較検討することになるが，それほど数の多くない専門家が，時間をかけて検討を行い，改定支援を行うことは非常に難しい．本研究では，

- 地域防災計画改定時に，他市町村の知見を取り入れることで，不備を修正できる可能性がある
- しかし，他市町村の知見を取り入れようとするためには，必要な人員を確保することが難しい

という二つの背景を踏まえ，様々な市町村の知見を得た，実効性の高い計画への改定支援を目指す．そのためには，実効性の高い防災計画とはどのようなものであるかを定義し，防災計画に必要な内容について分析するとともに，少ない人員であっても計画改定の検討を行えるようなシステム開発が必要である．本研究では，その第一段階として，防災計画に必要な内容の不足を発見できる手法を，自然言語処理技術を用いて開発する．

1.2 計画の実効性と網羅性，模倣性

本研究では，実効性の高い計画改定を支援すべく，自然言語処理技術による防災計画の分析を行う．そのためには，計画の実効性が低いとはどのような状態なのかを定義する必要がある．防災対応における実効性は，計画と運用の実効性の高さにより達成されるものであるが，本研究においては，計画の実効性の高さのみに着目する．これは，運用において担保される実効性は情報として残りづらく分析が難しいためである．また，計画の実効性の高さを分析・検討可能となれば，計画の実効性が高まるだけでなく，防災対応への意識も向上し，運用面での実効性も引き上げられると考えられる．

本研究においては，「災害発生時，計画に従って対応した場合に，発生する問題が少なく，かつ深刻な問題が発生しない」状態を，計画の実効性が高い状態であると定義する．計画の実効性は，計画の網羅性，計画検討の程度，自治体職員や住民の理解度，実態との乖離度，他組織との整合性等，複数の要素が絡み合って生じているが，本研究では，最もデータ化しやすく，解析を行いやすいという理由で，まずこれらの要素の中でも網羅性に着目する．また，防災計画のような行政文書は，他の行政文書を参考にして作成されることが多い．他の行政文書を参考にすること自体は知見共有の上でも非常に重要であるが，内容を検討せず他文書をただ模倣している場合，その記述が各自治体独自の特性に合致していないなどの問題が発生する可能性がある．このことから，検討の程度が少ない可能性がある文書発見のため，計画の模倣性にも着目する．本研究では，網羅性を分析するため，特徴語分析による書くべき内容の発見を，また模倣性を分析するために類似性分析による類似文書組の発見を行う．

1.3 災害廃棄物処理

廃棄物処理の課題は，災害発生の度に大きな問題となっている．直近では，令和元年東日本台風の際の水戸市において，ごみを自宅前の路上に出すよう周知したことで収集が追いつかなくなるという問題が発生した [2]．しかしこの課題は，東日本大震災の際に判明済みであり，栃木県那須烏山市など，仮置き場を開設するという形で既に対策済みの自治体も存在していた [2]．このように，廃棄物処理計画においても，近隣の地方公共団体における知見を取り入れれば修正できる可能性のある不備が存在する．

本研究では，地域防災計画改定における課題を対象とする．しかし，地域防災計画の扱う範囲は多岐にわたるため，まず災害廃棄物処理分野のみに着目して解析を行う．各地方公共団体の防災計画には，災害時に発生する廃棄物（災害廃棄物）の収集と処分を適切に行うため，災害廃棄物処理についての記述が存在する．廃棄物処理は，被災地域によって発生量やその内訳が異なる上に，市町村ごとに処理方針が異なるため [3]，市町村ごとの特性が出やすい．また，廃棄物処理特有の専門用語があり，自然言語処理での課題抽出に適していると考えられる．こうした特徴から，本研究では，防災計画改定における課題解決の最初の段階として，災害廃棄物処理の特徴抽出に取り組む．

1.4 本研究の目的

本研究では，実効性の高い防災計画改定を支援する手法作成の第一段階として，まず廃棄物処理文書の網羅性の低さ，模倣性の高さを分析する．これらの目的を達成するため，全国規模での廃棄物処理文書の収集を行い，3000 以上の文書により構成されるデータセットを作成する．そして，これらのデータセットに対し，自然言語処理技術を用い，文書の特徴語抽出や類似度分析を行うことにより，廃棄物処理文書における網羅性・模倣性の分析を行う．

1.5 先行研究

本研究に関連する先行研究としては，劉ら [4] による廃棄物処理関連章・節からのキーワード抽出を用いた分析研究が存在する．この研究においては，特徴語分析により，三自治体においてキーワード抽出が行われたが，二自治体については，広域連合離脱に伴う広域連合名の消失，また東日本大震災の発生に伴う原子力事故関連単語の増加といった結果にとどまり，地域防災計画の実行時に発生しうる問題点発見手法としては不安定さが残る．またデータセット作成も近似的な自動抽出を用いており，1,182 文書と数は多いものの精度面で問題のある結果に終わっている．さらに陳ら [5] の研究においては，災害廃棄物処理計画より頻出キーワードを抽出し，いくつかの自治体の記述の多寡について分析を行っているものの，103 市の文書を分析対象としており，データ数が少ない．本研究においては，全国

ⓒ 2021 Information Processing Society of Japan

の市町村より，人手によって抽出を行うことで，広範囲に収集した高い精度のデータセットを作成する．このデータセットに適用した特徴語分析より，廃棄物処理において重要な観点であると想定される内容を数多く抽出し，網羅性分析に用いることで，劉ら [4] の研究に比べ精度が高く，また陳ら [5] の研究に比べ広範囲を対象とした問題点発見を行った．さらに，問題点の指摘のみでは，防災対応において重要な独自性が失われてしまうと考え，模倣性にも着目した分析を行った．

2. データセット作成と分析手法

2.1 災害廃棄物処理文書データセットの作成

廃棄物処理計画の課題を言語処理により分析するため，全国の市町村の防災計画より，廃棄物処理文書を抽出し，データセットを作成する．廃棄物処理文書の抽出は，以下のルールに基づき，人手で行った．

- 「災害廃棄物キーワード」を題名に含む「ひとかたまり」を，廃棄物処理文書とする
- 「災害廃棄物キーワード」とは，次の単語群を指す：災害廃棄物，廃棄物，ごみ，ゴミ，がれき，瓦礫，ガレキ，し尿
- 「ひとかたまり」とは，計画の目次において，分割されている最小単位のことを指す．市町村によって，これは節であったり，章であったりする．目次のない計画においては，人の目でこれを判別する

ただし，本研究においては，pdf 形式であってテキストデータが抽出可能な文書のみを対象とし，さらにスキャン画像で構成されていたり，抽出を行うと文字化けする pdf，資料編とタイトルに記述された pdf は除外した．現在，47 都道府県中，43 都道府県の抽出が完了しており，全 1781 市町村中，889 市町村の 3252 文書を含むデータセットが作成済みである．

2.2 前処理

2.2.1 クリーニング

防災文書を含む，行政文書においては，改行による体裁の調整がしばしば行われる．この改行により，tokenize がうまくいかないことがあるので，改行を削除するクリーニング処理を行う．また，全角・半角の統一による正規化後に，半角スペースを削除した．

2.2.2 正規化

全角・半角の統一等を行うため，neologdn[6] による正規化を行った．数字については，残しておくと大量の未知語を生成してしまうため，どのような桁の数値も，0 に置換した．

2.2.3 tokenize

tokenizer には MeCab[7] を使用し，辞書には mecab-ipadic-NEologd[8, 9] を使用した．また，以下の単語を独自にユーザ辞書に追加した．

- 仮置き場, 仮置場, 災害廃棄物処理計画, 防災計画, クリーンセンター, 簡易トイレ, 死亡獣畜, 衛生班, 対策班, 清掃班, 運搬車両, 発災

2.2.4 StopWord 除去

本研究では，「名詞・形容詞」かつ「サ変接続・一般・自立語・固有名詞」である単語のみを分析対象とする．また，出現頻度が一回の単語のうち目視で不要と判断された単語や，解析中に不要と判断された単語を StopWord 辞書に登録した．

2.3 予備解析

nlplot[10] を用いて，データセットより作成したコーパスを分析，可視化することで，データセット全体の特徴を分析した．これらは研究目的に直接役立つわけではないものの，コーパスの特徴を理解するという意味で重要である．

2.3.1 頻度分析

まず，文書にどのような単語が出現しているかを調べるため，頻出上位単語の bar chart と WordCloud[11] を作成した．これにより，コーパスを代表する単語とその分布を分析する．また，単語の出現頻度分布のヒストグラムを作成し，コーパス内の単語分布の特徴を分析する．

2.3.2 N-gram 分析

文書を構成する単語 (uni-gram) だけでなく，bi-gram, tri-gram といった N-gram[12] の分析を行う．これにより，一層文脈を理解したコーパス内容の分析を行う．

2.4 Advantage CWs

網羅性の不足，つまり計画に書くべき内容の不足を発見するためには，まず計画に書くべき内容そのものを発見する必要がある．本研究では，書くべき内容を見つけるために，以下の考え方を採用する．

- 文書の特長を発見すれば，その中に書くべき内容が含まれる
- 文書の特長は，文書の特長語 (Advantage Words) を分析することで発見できる
- 文書の特長語は，文書の特徴語であって，コーパス内の他文書の特徴語であることが少ない単語（Advantage Candidate Words : Advantage CWs）を分析することにより発見できる

この考え方は，書くべき内容の共有が以下の過程で進むという仮定を根拠としている．

- まず，国や専門家等の要請に従って，すばやく一部の自治体 (災害を受けた経験のある自治体，防災意識の高い自治体等) が計画に盛り込む
- 徐々に，多くの自治体に広がっていく

この仮定は直観的に理解しやすいものであり，また現実の廃棄物処理計画の策定率の現状とも合致する．廃棄物処理

ⓒ 2021 Information Processing Society of Japan

計画の策定については，2014 年に国が作成を求め，2015 年に 20%程度の市町村が作成したが，それ以後，2018 年時点でも，28%の市町村しか作成できていない [2]．このように，まず国や専門家の要請を受けて，一部の市町村が対応し，その後徐々に知見が広まっていく，という仮定は，現実とも合致するものである．
本研究では，書くべき内容を見つけるため，特徴語を抽出する手法として，tf-Ridf と BM25 を用いる．

2.4.1 tf-Ridf

tf-Ridf[13] は，文書から特徴語を抽出する手法である tf-idf[14] をベースとした手法である.tf(term frequency) は，単語 t の文書 d における出現頻度を表す．また，idf(inverse document frequency) は，単語 t の希少さを表す．カテゴリに依らず多くの文書に現れる一般語は，非常に大きな $\mathrm{idf}(t)$ スコアを持つ．一般語の $\mathrm{idf}(t)$ スコアを下げるために，ポアソン分布より推定した $\widehat{\mathrm{idf}}(t)$ スコアを $\mathrm{idf}(t)$ スコアより除した値が，$\mathrm{Ridf}(t)$ である．この二つの値を掛け合わせたものが，$\mathrm{tfRidf}(t,d) = \mathrm{tf}(t,d) \cdot \mathrm{Ridf}(t)$ である．

2.4.2 Okapi BM25

本研究では，tf-Ridf に加え，Okapi BM25[15] を使用する．$\mathrm{tf}(t,d)$ には，文書に含まれる全単語数が多いと，値が小さくなるという欠点がある．この欠点を改善するため，文書の単語数について標準化する項を追加したモデルが BM25 である．

2.4.3 Advantage CWs

これらの特徴語抽出手法を用いて，特長語の候補を発見する．この特長語の候補を Advantage Candidate Words(Advantage CWs) と呼ぶ．Advantage CWs の抽出は，以下の流れで行う．

- $\mathrm{tfRidf}(t,d)$,$\mathrm{BM25}(t,d)$ を計算し，各文書について，tfRidf,BM25 の各スコアの上位 20 単語の集合を取り出す
- 各文書について，tfRidf,BM25 の各スコアの上位 20 単語の集合を取り出し，共通部分（AND）を文書の特徴単語とする
- tfRidf,BM25 それぞれについて，コーパス全体で，上位 20 単語に含まれる回数が (文書数/8) 回以上である単語を取り出し，頻出単語とする
- 頻出単語を，各文書の特徴単語から除く

よって，Advantage CWs の定義は，以下の式で表される．

> Advantage CWs =「tfRidf 上位 20 単語」AND「BM25 上位 20 単語」- 頻出単語

この Advantage CWs を抽出し，人の目で分析していくことで，特長語 (Advantage Words) を発見し，特長語リストを作成する．この特長語リストに入っている単語が含まれるかどうかを検証することで，網羅性の不足をチェックする．

2.5 類似性分析

模倣性の高い文書を発見するには，文書同士の類似性を定量評価する必要がある．本研究では，文書のベクトル表現取得手法である doc2vec を用いて文書同士の類似性分析を行う．

2.5.1 doc2vec

文書同士の類似度を計算する一つの手法として，分散表現化した文書ベクトル同士の cos 類似度が存在する．本研究では，文書の分散表現を取得する方法として，doc2vec の一種である Distributed Bag of Words(DBoW)[16] を用いる．

2.5.2 類似文書群の発見

各文書に対し，doc2vec を用いて分散表現化した文書ベクトル同士の cos 類似度を計算し，0.9 以上である文書群を抽出する．この類似文書群を人の目で分析し，酷似している文書群を見つける．この文書群の中には，文書に関する検討の不足した文書が存在する可能性がある．

3. 分析結果

現在作成済みの，43 都道府県における市町村の廃棄物処理文書を含むデータセットを用いて分析を行った．その結果を次に示す．

3.1 予備解析結果

3.1.1 頻度分析

文書にどんな単語が出現しているかを調べるため, 頻出上位 50 単語の bar chart，WordCloud，そして単語の出現頻度のヒストグラムを作成した．これにより，コーパスを代表する単語とその出現頻度分布について分析する．頻出上位 50 単語の bar chart を図 1 に示す．bar chart を見ると，‘仮置場’ や ‘集積’，‘分別’，‘協力’，‘要請’ といった単語が上位単語として出現していることがわかる．また，‘分別’ の出現回数が 3692 回であるのに対し，‘収集’ の出現回数は 17509 回であることから，一部の単語の出現回数が突出して多いことがわかる．

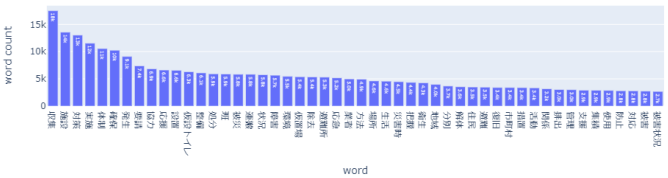

図 1 頻出上位単語の bar chart

図 2 上位 200 単語の WordCloud

上位 200 単語の WordCloud を図 2 に示す．WordCloud は，文書に出現する単語を可視化した図であり，文字の大きさは出現回数の多さを示している．また，図 3 は，出現回数 300 回までの単語の出現頻度のヒストグラムを示したものである．この図より，単語の出現回数は指数関数的に減少していることがわかる．

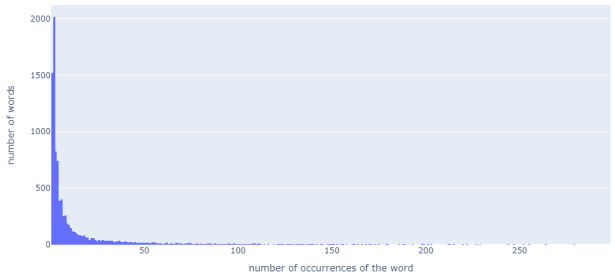

図 3 出現頻度分布のヒストグラム

3.1.2 N-gram 分析

図 4 に bi-gram, 図 5 に tri-gram の上位 50 フレーズを示す．

図 4 bi-gram の bar chart

図 4 より，‘仮置場 搬入’,‘広域 支援’ ,‘協定 締結’,‘建築物 解体’ といった bi-gram が見られることから，多くの市町村で，仮置場や広域支援，家屋の解体についての記述が存在していることがわかる．

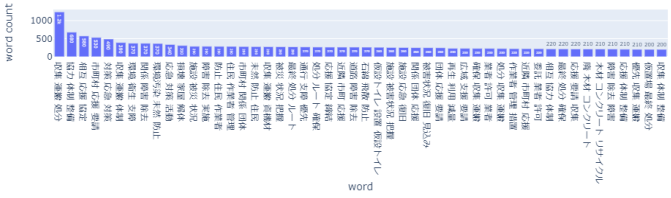

図 5 tri-gram の bar chart

また図 5 より，‘石綿 飛散 防止’,‘木材 コンクリート リサイクル’,‘環境汚染 未然 防止’,‘広域 支援 要請’ といった tri-gram が見られることから，多くの市町村で，廃棄物のリサイクルや，環境，健康への被害についての記述が存在していることがわかる．

3.2 Advantage CWs 抽出結果

表 1 に．Advantage CWs の抽出結果の一部を示す．抜粋した文書と比較すると，‘リサイクル法’ や ‘オープンスペース’,‘処理手数料’‘減免’,‘死亡獣畜’,‘フロン’,‘石綿’,‘耐震’ といった，廃棄物処理にとって重要な単語を抽出できていることがわかる．一方で，京都府宇治市の Advantage CWs に見られるように，‘城南’ のような，廃棄物処理とは関係のない単語も抽出されてしまっていることがわかる．つまり，Advantage CWs であるからといって，重要単語であるとは限らない．

また，京都府京都市の文書には，‘仮置場’ という単語が含まれているにも関わらず，Advantage CWs として抽出されていない．このように，重要な単語であっても，Advantage CWs として検出されない可能性がある．よって，Advantage CWs として抽出されていないからといって，重要な内容を記述していないことにはならない．特に，埼玉県吉川市の Advantage CWs を見てもわかる通り，‘仮置場’ には’ 仮置き場’ という表記揺れが存在するため，仮置場についての記述がある市町村であっても，‘仮置場’ という表記であれば Advantage CWs に含まれず，‘仮置き場’ という表記であれば Advantage CWs に含まれるということになってしまう．

3.3 特長語リストチェック結果

この Advantage CWs を人の目で分析することにより，特長語リストを作成した．

> ‘オープンスペース’,‘フロン’,‘分別’,‘周知’,‘木くず’,‘水産’,‘津波堆積物’

この特長語リストの中の単語が各市町村の文書内に存在しているかどうかをチェックし，網羅性が不足しているかを確かめる．また，表記揺れが存在した場合は，同じ内容でも，特長単語が異なる場合も考えられる．以下のように，or が含まれた単語対の場合は，単語対のうち一つでも文書

表 1 Advantage CWs 抽出結果

自治体名	Advantage CWs	文書（抜粋）
佐賀県唐津市	[‘地震災害’, ‘避難場所’, ‘リサイクル法’, ‘津波災害’, ‘実施方針’, ‘建築物’]	(4) 廃棄物の処理には、各種リサイクル法（家電リサイクル法、パソコンリサイクル法、自動車リサイクル法、容器包装リサイクル法、建設リサイクル法）に配慮し方針を立てる。
京都府京都市	[‘推計’, ‘環境政策’, ‘オープンスペース’, ‘チーム’, ‘調整’, ‘受入’, ‘実行’]	オープンスペース調整チーム事務局は, 災害廃棄物仮置場として暫定的にオープンスペースを割り当てる。
京都府宇治市	[‘管理組合’, ‘城南’, ‘伝達’,‘情報’,‘対象’, ‘くみ取り’, ‘減免’, ‘浸水’, ‘処理手数料’]	ごみ処理手数料の減免についての手続きは原則としてごみ減量推進課（生活環境班）が担当するものとするが、状況に応じて関係処理施設と協議のうえ、必要な措置を講じるものとする。
大阪府島本町	[‘死亡獣畜’, ‘見込み’, ‘危機管理’, ‘総務’, ‘臨時’, ‘業務内容’, ‘創造’, ‘建築物’]	（ア）都市創造対策部環境課は、死亡獣畜発見の連絡を受けた場合は、直ちに収集し、消毒その他の衛生処理を行う。
愛知県幸田町	[‘産業廃棄物’, ‘リサイクル’, ‘フロン’, ‘支援要請’]	また、フロン使用機器の廃棄処理にあたっては、適切なフロン回収を行う。
埼玉県吉川市	[‘回収’, ‘仮置き場’, ‘石綿’, ‘半壊’, ‘個人’]	○倒壊建築物の解体にあたっては石綿含有建材の使用の有無を確認し、石綿を使用している場合には散水等の飛散防止対策を徹底する。
神奈川県川崎市	[‘立上’, ‘耐震’, ‘焼却炉’,‘基準’, ‘耐震設計’, ‘建築基準法’,‘プラント’, ‘耐震性’]	廃棄物処理関連施設の地域における災害廃棄物処理体制上の役割を明確にしつつ、今後、耐震、耐水及び耐浪性の確保をはじめ、特に、ごみ焼却処理施設の建設時には、商用電源を確保できない場合でも焼却炉立上げを可能とするための始動用電源や燃料保管設備等の配備や薬剤等の備蓄を行うなど、災害対策を講じるよう努める。

内に存在していれば，その内容が文書内に含まれるものとしてチェックを行う．

- ‘アスベスト’ or ‘石綿’

また，二つの単語が両方含まれなければ，書くべき内容が含まれると見なせない場合もある．この場合は，以下のように and を用いて表す．

- ‘耐震’ and ‘焼却炉’

ただし，以下のように片方の単語に表記揺れが存在する場合は，or と and を両方用いてチェックする．

- (‘処理手数料’ or ‘手数料’) and ‘減免’

これらの特長単語をチェックした結果，特長単語が存在している市町村数は，表 2 のようになった．このリストを用いて，各市町村に特長単語が存在しているかどうかを確かめた．たとえば，福知山市では，‘アスベスト’ or ‘石綿’ が含まれていなかった．実際，福知山市地域防災計画の全 pdf 内を検索しても，アスベストや石綿についての記述は存在していない．一方，福知山市の文書には，‘死亡獣畜’ or ‘獣畜’ は含まれていなかったが，‘動物の死体’ という記述が防災文書には存在していた．このように，特長単語リ

表 2 特長語リストチェック結果

特長語リスト	特長語の存在している市町村数
‘水産’	36
‘オープンスペース’	68
‘津波堆積物’	76
‘フロン’	150
‘木くず’	162
‘周知’	774
‘分別’	1402
‘家電リサイクル法’ or ‘リサイクル法’	106
‘死亡獣畜’ or ‘獣畜’	436
‘アスベスト’ or ‘石綿’	854
‘推計’ or ‘推定’	930
‘仮置場’ or ‘仮置き場’	1510
‘広域’ or ‘連携’ or ‘協定’	1532
‘耐震’ and ‘焼却炉’	10
(‘処理手数料’ or ‘手数料’) and ‘減免’	10

ⓒ 2021 Information Processing Society of Japan

ストを用いて，書くべき内容の不足が確認できる場合もあれば，できない場合も存在する．

3.4 類似性分析結果

doc2vec による類似性分析を行った結果，異なる都道府県内にあるにも関わらず，非常に類似した記述をしている文書対が見つかった．図 6 に示すように，神奈川県 箱根町の廃棄物処理文書 [17] と山形県 遊佐町の廃棄物処理文書 [18] は，両自治体が全く異なる都道府県に所属しているにも関わらず，言い回しや文書の構成、表の作り方，章立てが酷似している．また，箱根町と神奈川県の廃棄物処理文書は似ておらず，また遊佐町と山形県の廃棄物処理文書も似てはいなかった．つまり，二つの自治体の廃棄物処理文書は，都道府県の廃棄物処理文書とは関係なく，非常に類似しているということがわかる．

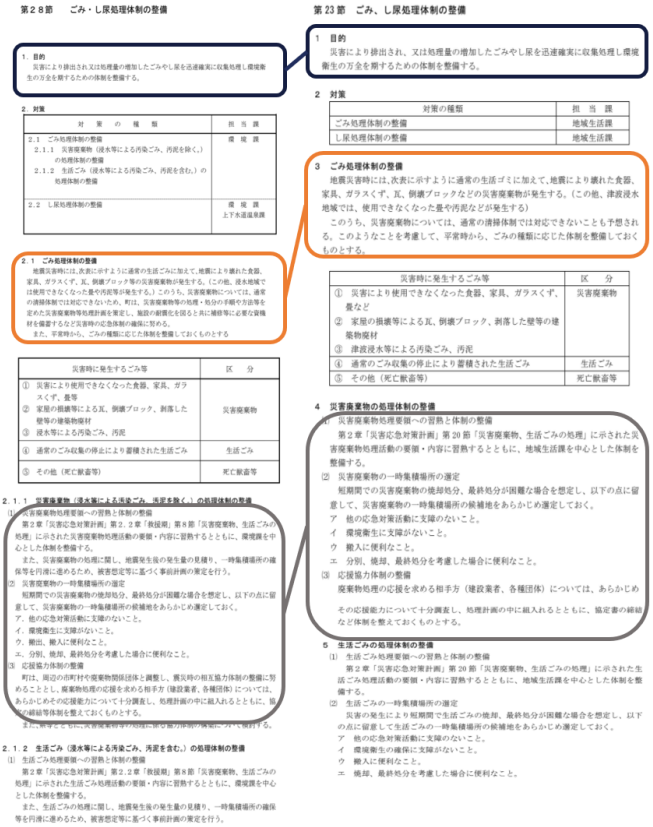

図 6 類似文書対の比較
左：神奈川県箱根町, 右：山形県遊佐町

4. 考察と今後の課題

4.1 Advantage CWs 抽出

3.2 に示したように，京都府京都市の文書には，‘仮置場’という単語が含まれているにも関わらず，Advantage CWs として抽出されていない．これは，図 2 に示されるように，‘仮置場’という単語がどの文書でも多く使われている単語であり，idf 値が小さくなってしまっているためと考えられる．このように，頻度を用いて特長を探す手法では，重要な単語であっても見逃してしまう可能性がある．
また表 2 に示すように，Advantage CWs より作成した特長語リストにより，網羅性の不足のチェックを行った．例として挙げた福知山市の場合では，特長語リストを用いて網羅性が不足していると判断したにも関わらず，実際は防災文書内に記載があった．これは表記揺れにより発生する．また，表記揺れ以外にも，抽出が適切に行われなかった際に，同様の事態が発生すると推察される．

4.2 類似性分析

類似性分析により，異なる都道府県内にあるにも関わらず，非常に類似した構造を持つ文書対が見つかった．この文書対が非常に似ている理由として，以下の 3 点が考えられる．

(1) 同一コンサルタント会社を利用していたため
(2) 同一文書を参考にしていたため
(3) 市町村間に何らかの交流が存在していたため

神奈川県 箱根町と山形県 遊佐町の間には，現在何ら防災上の関係性が見られず，また連携するとしてもより近隣の自治体を優先するであろうことから，(3) については考慮しない．(1),(2) のどちらかを特定するためには，まずは各自治体の文書作成に携わったコンサルタント会社を調査する必要がある．しかし，遊佐町については入札会社が公表されていたものの，箱根町については公表がなされていなかったため，これ以上の調査については，箱根町へのヒアリングが必要である．

5. 結論

本研究では，地域防災計画の改定時に必要な人員を確保できておらず，他市町村の知見を取り入れることが難しいという課題の解決を目指し，その第一歩として，まず計画の網羅性と検討の程度に着目した．そして，網羅性の不足した文書・模倣性の高い文書の発見を目指すべく，自然言語処理技術を用いた解析手法を提案した．

5.1 網羅性の不足発見について

網羅性の不足を発見するために，まず，Advantage CWs を用いて，どんな市町村にも必要な内容と，その内容に関係する単語を探し，特長語リストを作成した．その結果，特長語リストを用いたチェックにより，書くべき内容の不足した文書を発見することができた．しかし，表記揺れや抽出過程での漏れにより，特長語リストだけでは網羅性が不足しているかどうか特定することはできず，最終的には文書を精読する必要がある．この問題を解決するため，今後表記揺れの抑制に取り組む．

ⓒ 2021 Information Processing Society of Japan

5.2 検討の不足した文書発見について

検討の不足した文書を発見するため，本研究では，

- 非常に類似した計画が発見された場合，検討が不足している可能性がある

との考えに基づき，模倣性の高い文書の発見に取り組んだ．その結果，異なる都道府県にも関わらず，非常に類似した文書対の発見に至った．しかし，その類似の理由を明確にしなければ，検討の不足や，計画の実効性を分析することは難しい．この問題を解決するため，今後，該当自治体へのヒアリングを通じ，類似理由の分析に取り組む．

6. 今後の計画

Advantage CWs を用いた，共通内容の分析においては，表記揺れにより，Advantage CWs が検出されない，特長単語が不足していると誤って判断してしまうといった問題が発生した．この問題を解決するために，類似単語の発見による表記揺れの抑制を行う．また，類似性分析において，異なる都道府県にも関わらず，非常に酷似した文書対が発見された．しかし，文書同士を比較するだけでは，この文書対がなぜ類似しているのかが明確にはならない．類似理由の明確化のため，ヒアリングを行う必要がある．さらに，データセット作成にも取り組む．現在，データセットは 43 都道府県の文書のみを含んでいる．今後，4 都道府県のデータ抽出を進めることで，全国データセットの完成を目指す．

References

[1] 千葉県. **令和元年房総半島台風等への対応に関する検証報告書**. 2020.

[2] 毎日新聞. **災害編／上　自治体の処理計画整備、急務　台風１９号被害からみえた課題**. 2019.

[3] 西川 貴則, 日比野 直彦, and 森地 茂. “災害廃棄物等の処理に関する課題とその対応”. In: **土木学会論文集** *D3* **（土木計画学）** Vol.72.No.5 (Dec. 2016), pp. I_103–I_110.

[4] 英楠 劉 and 満則 畑山. **キーワード抽出を用いた地域防災計画における災害廃棄物管理に関する比較分析**. Tech. rep. 4. 京都大学, 京都大学防災研究所, Feb. 2019.

[5] 陳 唐伊伊, 大窪 和明, and 劉 庭秀. “災害廃棄物処理計画の特徴と課題分析：キーワードから見えたもの”. In: **廃棄物資源循環学会研究発表会講演集** 30 (2019), p. 129.

[6] 池上 有希乃. *neologdn*. URL: `https://github.com/ikegami-yukino/neologdn`. (accessed: 08.09.2021).

[7] 工藤拓. *MeCab: Yet Another Part-of-Speech and Morphological Analyzer*. URL: `https://taku910.github.io/mecab/`. (accessed: 08.09.2021).

[8] Sato Toshinori. *Neologism dictionary based on the language resources on the Web for Mecab*. 2015. URL: `https://github.com/neologd/mecab-ipadic-neologd`.

[9] Taiichi Hashimoto Toshinori Sato and Manabu Okumura. “Operation of a word segmentation dictionary generation system called NEologd (in Japanese)”. In: *Information Processing Society of Japan, Special Interest Group on Natural Language Processing (IPSJ-SIGNL)*. Information Processing Society of Japan, 2016, NL-229–15.

[10] takapy. *nlplot*. URL: `https://github.com/takapy0210/nlplot`. (accessed: 08.09.2021).

[11] Layla Oesper et al. “WordCloud: a Cytoscape plugin to create a visual semantic summary of networks”. In: *Source code for biology and medicine* 6.1 (2011), p. 7.

[12] 森信介, 山地治, and 長尾眞. **予測単位の変更による** *n - gram* **モデルの改善**. Tech. rep. 120(1997-SLP-019). 京都大学工学研究科, 京都大学工学研究科, 京都大学工学研究科, Dec. 1997.

[13] Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. The MIT Press., 1999.

[14] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill., 1983.

[15] Stephen Robertson and Hugo Zaragoza. “The Probabilistic Relevance Framework: BM25 and Beyond”. In: *Foundations and Trends® in Information Retrieval* 3.4 (2009), pp. 333–389.

[16] Quoc Le and Tomas Mikolov. “Distributed Representations of Sentences and Documents”. In: *Proceedings of the 31st International Conference on Machine Learning*. Vol. 32. Proceedings of Machine Learning Research 2. 22–24 Jun 2014, pp. 1188–1196.

[17] 神奈川県箱根町. **箱根町地域防災計画**. 2020.

[18] 山形県遊佐町. **遊佐町地域防災計画**. 2017.

ⓒ 2021 Information Processing Society of Japan