法令文の可読性向上のための定義規定・略称規定における文型定義及びパターンベースの正式名称・略称抽出手法

北野尚樹 [1] 西山大輝 [2,3]
[1] 筑波大学 情報学群 情報科学類 産学間連携推進室
[2] 東京科学大学 情報理工学院　[3] 理化学研究所 AIP
s2212117@u.tsukuba.ac.jp　nishiyama.d.2d7f@m.isct.ac.jp

概要

法令文では定義規定・略称規定が頻繁に使用されている（例：新型インフルエンザ等対策の推進を図るため、内閣に、新型インフルエンザ等対策推進会議（以下「会議」という。）を置く。）．これらの規定文は複雑になりやすく，読者の混乱や負担を生じさせうる．そこで本研究では，既存法では不十分であった定義規定・略称規定に関する文型を導入し，これに基づき正式名称と略称のペアを抽出するパターンベースの手法を提案する．結果，提案法は複雑な構造や文型に対応し，既存の手法や大規模言語モデルとの比較で高い精度を達成した．また，今後の法令解析研究の促進のために，実験で使用した正式名称と略称のペアのデータセットを公開する．

1　はじめに

1.1　背景

法令や契約文書などの法律文書には，特定の用語を定義したり長い用語に対して略称を定めたりする文が多数含まれる．これらの定義規定や略称規定は，文書全体で高頻度に使用されるため，読者が初めて法律文書を読む際に混乱や負担を生じる原因の一つとして指摘されている [1, 2]．一方で，こうした定義規定や略称規定を自動的に特定し，それらに注釈情報を付与できれば，法律文書の可読性向上や読者の理解促進に大きく寄与すると期待される．

しかし，法律文書内における定義規定や略称規定は書式が多様であり，正確に自動検出・抽出を行うことは容易ではない．そこで本研究では，法令文に含まれる定義規定・略称規定を対象として，正式名称と略称のペアをより網羅的かつ高精度に抽出する手法の開発に取り組む．

1.2　関連研究

中村ら [3, 4] は，法令文中の定義規定や略称規定に焦点を当て，正規表現やパターン処理により正式名称と略称を抽出する手法を提案した．具体的には，定義・略称規定分の代表的な書式を定義し，それに対応する正規表現を開発することで自動抽出を実現した．しかし整理したパターンが限られているため，抽出できる正式名称と略称の種類やパターンに制限があり，抽出漏れの発生が課題である．

金子 [5] は，法令文中の読み替え規則を対象として，パターンに基づく手法を示している．そのパターンの構築手法は，定義規定・略称規定の解析にも応用可能な面がある．しかし，読み替え規則と定義規定・略称規定の文型は文法的にも目的上も大きく異なるため，金子の手法を直接適用することは難しく，新たなアプローチが求められる．

近年では，大規模言語モデル（LLM）が自然言語処理のさまざまな分野で高い性能を示しており [6, 7]，法令文書解析においても大きな可能性があると考えられる [8, 9, 10, 11]．しかし，LLM を用いて定義規定・略称規定のペアを安定的に抽出するには，学習データの妥当性や日本語法令文特有の文型への対応が不十分となる場合が多く，さらには推論の再現性や信頼性の確保も課題となる．

1.3　貢献

このように既存手法では，法令文の構文や複雑な括弧書きの扱いなどによって，抽出漏れや誤検出が起こりやすいという課題が残っている．そこで本研究では，法令文に含まれる定義規定・略称規定に対し，網羅的なパターンベース手法を提案し，その有効性を検証した．特に，大規模言語モデル（LLM）を含む他の手法と比較を行った結果，我々の提案手

法の方が高い精度を達成している．本研究の主な貢献は以下のとおりである．

- 我々は従来の研究で扱いきれていなかった，定義規定・略称規定特有の新たな文型を導入した．そしてそれらの文型に包括的に対応できるパターンベースの手法を開発した．
- 既存のパターンベース手法や，LLM を利用した手法と比較し，精度面で提案手法の優位性を明らかにした．
- 実際の法令文から抽出した定義規定・略称規定に手動で正答データをつけたデータセットを公開し[1)]，法令文解析の研究促進のための基盤[2)]を提供した．

2 対象と課題の概要

2.1 定義規定・略称規定

定義規定・略称規定とは，法令文中で使用する単語に対して定義を与えたり，長い単語に略称を与えたりする文のことである．例えば，図 1 の法令文では，再生医療等の安全性の確保等に関する法律全体で使う「細胞」という単語が「細胞加工物の原材料となる人又は動物の細胞」のことであることを定義している．これらの規定により，法令文中の冗長な表現を短くすることができ，読みやすい法令文の作成が可能になっている．いくつかの法制執務関連の書籍 [12, 13] によると，定義規定・略称規定の文の形の一部として次のものがあるとしている．

- この法律において「○○」とは、△△～をいう。
- (・・・をいう。以下同じ。)
- (以下「・・・」をいう。)

2.2 問題設定

本研究では，法令文全体のうち定義規定・略称規定が含まれる法令文を対象とする．定義規定・略称規定が含まれる法令文を $T_i (i = 1, \ldots, N)$ とし，ここで $N \geq 1$ はそのような法令文の総数を表す．各 T_i には 1 つ以上の正式名称と略称のペア $y_i = \{(t_{\text{formal}_j}, t_{\text{abbr}_j})\}_{j=1}^{M_i}$ が含まれており，$M_i \geq 1$ は T_i に含まれるペアの数である．

定義規定・略称規定を解析する手法によって予測されたペアの集合を $\hat{y}_i = \{(\hat{t}_{\text{formal}_j}, \hat{t}_{\text{abbr}_j})\}_{j=1}^{\hat{M}_i}$ とする．ここで $\hat{M}_i \geq 1$ は抽出結果に含まれるペアの数である．

最終的に，真のペア集合 y_i に対して，抽出されたペア集合 $\hat{y}_i$ がどの程度一致するかを本研究が取り組む問題として定義する．

2.3 既存方法

中村らの手法 [3, 4] 中村らは，定義規定・略称規定を**法令の総則に置かれるもの**と，法令文中に出現する**トイウ形**と**ヲイウ形**の 3 種類に整理し，それぞれについて正式名称と略称を抽出する手法を提案した．まず法令の総則に置かれる定義規定「この法律において「○○○」とは、A をいう。」について，「(.+)」とは、(.+)(を、|をいい、|をいう。|といい、|という。|とする。) という正規表現を使い，正式名称と略称を抽出する [3]．また，法令文中に出現するトイウ形とヲイウ形は次のように定義された．

- トイウ形：A、B 及び C (以下「○○○」という。)
- ヲイウ形：○○○（A、B 及び C をいう。以下同じ。)

トイウ形・ヲイウ形に対しては，次の方法により抽出される．

1. 各法令文から，トイウ形，もしくはヲイウ形に該当する条文を抜き出す．
2. 定義規定ではない括弧書きを除去する．これには定義規定が括弧文内にあるときも含む．
3. 文頭から数えて 100 文字以内に定義規定の括弧があり，その間に読点（「、」）がなければ，以下の処理を行う．
   - トイウ形ならば，括弧内の「カギ括弧」が略称，文頭から括弧までが正式名称．
   - ヲイウ形ならば，文頭から括弧までが略称，括弧内の「をいう。」までが正式名称．

LLM を用いた手法 正式名称と対応する略称の抽出という自然言語処理タスクに対しては，LLM によるアプローチも考えられる．この時，このタスクの意図や出力形式を示すための Few-Shot プロンプト [14] により，高い精度での抽出が期待される．具体的には，入力する法令文に対して「定義規定・略称規定とみられる部分をすべて検出し，{"formal": "○○", "abbr": "△△"} の形式で正式名称と略称のペアを列挙する」ように求めるテキストを事前に指示できる．

1) https://github.com/japanese-law-analysis/data_set
2) https://github.com/japanese-law-analysis

3　新たな文型の導入

正確に正式名称と略称のペアを抽出するために，既存法がサポートできない定義規定・略称規定における文型を，新たに導入する．法令文を観察した結果，定義規定・略称規定には次のような特徴があることがわかった．

- ヲイウ形には「○○○（A に規定する○○○をいう。以下同じ。）」という，用語を定義する規定文の情報を補足する形が含まれる（本稿においてこれを**ヲイウ-規定補足形**とする）．
- ヲイウ形には「○○等（○○又は△△をいう。）」という，複数の概念をまとめて「○○等」にする形が含まれる（本稿においてこれを**ヲイウ-トウ圧縮形**とする）．
- 総則に置かれる定義規定は「「○○○」とは、Aをいう。」という形になる（本稿においてこれを**トハ形**とする）．
- トイウ形とヲイウ形ともに，括弧書きの前にある正式名称や略称は本文中の読点もしくは文頭まで続くことが多い．ただし，読点の一つ前が漢字の場合は単なる並列表現であるので正式名称や略称に含まれる．
- トハ形・ヲイウ形・トイウ形などは一つの規定文中に混ざって複数出現することがある．

中村らの手法ではヲイウ-規定補足形・ヲイウ-トウ圧縮形・ヲイウ形・トイウ形において，略称や正式名称を正しいものよりも広く取ってしまうため，定義規定・略称規定を解析することが困難である。

4　提案手法

我々は，3 節で導入した文型に対応できる，定義規定・略称規定から正式名称と略称を抽出する手法を提案する．提案手法の手順は以下である[3]．

1. ヲイウ形とその派生形，トハ形，トイウ形，に該当する法令文を抽出
2. 括弧書き中のテキストと位置を記録し除去
3. トハ形の場合，法令文中の「とは、」と「をいう。」・「をいい、」の間を正式名称と，「とは、」の直前の鍵括弧内を略称と定義（図 1）
4. ヲイウ-トウ圧縮形の場合，括弧の直前の「等」の直前であって括弧書きに含まれる最長のものを略称と，弧書きの頭から「をいう。」・「をいい、」までを正式名称と定義（図 2）
5. ヲイウ-規定補足形の場合，括弧書き内の「に規定する」と「をいう。」・「をいい、」の間を略称と，括弧書きの頭から「をいう。」・「をいい、」までを正式名称と定義（図 3）
6. ヲイウ形の場合，括弧書きから「ひらがなの直後の読点」もしくは文頭までを略称と，括弧書きの頭から「をいう。」や「をいい、」までを正式名称と定義（図 4）
7. トイウ形の場合，括弧書きから「ひらがなの直後の読点」もしくは文頭までを正式名称と，括弧文中の鉤括弧内を略称と定義（図 4）
8. 記録していた括弧書きを元の箇所に復元
9. 以上の 2-7 までの操作を括弧書き中に再帰的に適用

これにより，第 3 節で整理した定義規定・略称規定のより詳細な特徴に対応した抽出が可能になるとともに，括弧書きの中身も抽出することができるようになった。

この法律において「**細胞**」とは、**細胞加工物の原材料となる人又は動物の細胞**をいう。

図 1　提案手法の手順 3 を適用したトハ形の解析手法．[4] 墨かけが略称，囲い線が正式名称．

独立行政法人農業者年金基金法による給付の支給を受ける権利に係る**届出等**（**届出又は申出**をいう。以下この号において同じ。）の受理

図 2　提案手法の手順 4 を適用したヲイウ-トウ圧縮形の解析手法．[5] 墨かけが略称，囲い線が正式名称．

青年等就農資金（**農業経営基盤強化促進法（昭和五十五年法律第六十五号）第十四条の六第一項第一号に規定する青年等就農資金（同法の定めるところにより貸し付けられるものに限る。）**をいう。以下同じ。）

図 3　提案手法の手順 5 を適用したヲイウ-規定補足形の解析手法．[6] 墨かけが略称，囲い線が正式名称．

3）実　　装：https://github.com/japanese-law-analysis/pattern_based_formal_and_abbr_extraction

4）再生医療等の安全性の確保等に関する法律（平成二十五年法律第八十五号）第二条

5）住民基本台帳法別表第一から別表第六までの総務省令で定める事務を定める省令（平成十四年総務省令第十三号）第一条第百九十九項第五号

6）農業信用保証保険法（昭和三十六年法律第二百四号）第二条第三項第三号

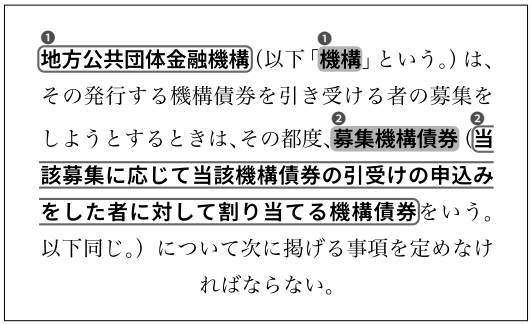

図 4　提案手法の手順 6 を適用したイトウ形（図中 1 に対応）及び手順 7 を適用したヲイウ形（図中 2 に対応）の解析手法．[7] 墨かけが略称，囲い線が正式名称．

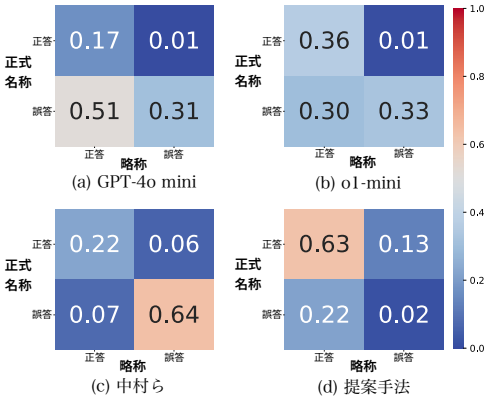

図 5　各モデルにおける，正式名称と略称それぞれの正誤の各組み合わせにおける頻度割合．横軸は略称の正答・誤答，縦軸は正式名称をそれぞれ表す．

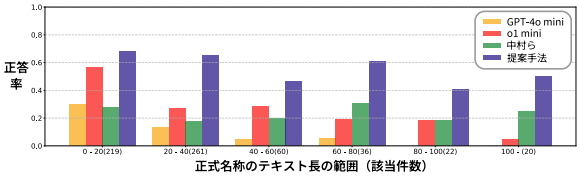

図 6　正式名称のテキスト長別の各手法の正答率．

5　実験

5.1　実験方法・内容

日本政府が法令データを公開している e-GOV 法令検索[8] では，現在施行されている法令の XML データを取得することができる．実験では，2024 年 12 月 17 日時点でのデータを対象とする．この法令データから定義規定・略称規定の文を 401 個ランダムに抽出した．この集合を L とする．L に対して，i 番目の法令文における M_i 組の正式名称と略称のペア $y_{i,j} = (t_{\text{formal}\,j}, t_{\text{abbr}\,j})(j = 1, \ldots, M_i)$ を手動で抽出することで、正答データを作成した．

L 中の各法令文 T_i から正式名称と略称のペア $\hat{y}_{i,j}$ を抽出した．抽出のための手法として，LLM の GPT-4o mini[15]，o1-mini[16]，中村らの手法 [3, 4]，提案手法を用いて，正式名称および略称ごとの正答率により性能を比較した．また正式名称のテキスト長の違いが精度に及ぼす影響を観察した．なお，LLM で用いたプロンプトは付録に示す．

5.2　実験結果・考察

図 5 に，抽出性能の実験結果を示す．各図の左上が，正式名称と略称の両方を正確に抽出できた割合，正答率を意味する．提案法の正答率 63%は他の手法に比べて大きく上まわっていることがわかる．一方で，提案手法で抽出に失敗した法令文を観察すると，次のようなことが原因となり，抽出には成功しているものの抽出するべき対象と地の文の区切り位置の決定に失敗したと考えられる．

- ヲイウ-規定補足形の中の「△△（○○に規定するものをいう。）」という例外の存在
- 「～に係る○○（△△をいう。）」など正式名称と地の文の区切りに読点が出現しない文の存在
- 「又は」などの並列表現のある文の場合，抽出するべき対象が並列の一部なのか全体なのかを判定することの困難性

続いて図 6 に，正式名称のテキスト長の正答率への影響を示す．LLM はテキスト長が長くなると正答率は低下する一方で，中村らの手法や提案法は正答率変化が少ない．これはパターンベースの手法がテキスト長の変化により頑健であることを示す．

6　結論

我々が定義規定・略称規定の複雑な文型を整理して作成したパターンベースの解析手法によって，既存手法や LLM を利用した手法よりも精度の良い定義規定・略称規定の解析を可能とした．一方で，パターン化できない文型の存在や，意味情報が必要な並列表現の解析が必要なケースでは不十分であることがわかった．今後は，単純なパターンベースでの処理のみならず，字句解析や掛かり受け解析などの意味情報が付与されたデータを使った処理も検討する．さらに，略称に対する正式名称の注釈を実際にユーザーに提示することで，法令文の可読性が向上するかを検証していく．

7）　地方公共団体金融機構法施行令（平成十九年政令第三百八十四号）第四条

8）　https://laws.e-gov.go.jp/bulkdownload/

謝辞

本研究は JSPS 科研費 JP24KJ1049 の助成，及び筑波大学情報科学類産学間連携推進室による研究環境の提供と助言を受けた．

参考文献

[1] Eric Martínez, Francis Mollica, and Edward Gibson. Poor writing, not specialized concepts, drives processing difficulty in legal language. **Cognition**, Vol. 224, p. 105070, 2022.

[2] 竹井直樹, 柴田文明. 保険約款と保険商品のわかりやすさの向上について. 損害保険研究, Vol. 72, No. 2, pp. 109–127, 2010.

[3] Makoto Nakamura, Ryusei Kobayashi, Yasuhiro Ogawa, and Katsuhiko Toyama. A pattern-based approach to hyponymy relation acquisition for the agricultural thesaurus. In **Proceedings of International Symposium on Agricultural Ontology Service 2012 (AOS2012)**, 2012.

[4] 中村誠, 小川泰弘, 外山勝彦. 法令文中において括弧書きで定義されている法令用語とその語釈文の抽出. 言語処理学会 第 19 回年次大会 発表論文集, 2013.

[5] 金子尚樹. 単語の出現規則に着目した法令中の読み変え規定文の解析手法の提案. 情報処理学会第 85 回全国大会, 2023.

[6] Muhammad Usman Hadi, Qasem Al Tashi, Abbas Shah, Rizwan Qureshi, Amgad Muneer, Muhammad Irfan, Anas Zafar, Muhammad Bilal Shaikh, Naveed Akhtar, Jia Wu, et al. Large language models: a comprehensive survey of its applications, challenges, limitations, and future prospects. **Authorea Preprints**, 2024.

[7] Mohaimenul Azam Khan Raiaan, Md Saddam Hossain Mukta, Kaniz Fatema, Nur Mohammad Fahad, Sadman Sakib, Most Marufatul Jannat Mim, Jubaer Ahmad, Mohammed Eunus Ali, and Sami Azam. A review on large language models: Architectures, applications, taxonomies, open issues and challenges. **IEEE Access**, 2024.

[8] 新保彰人, 菅原裕太, 山田寛章, 徳永健伸. 大規模言語モデルを用いた日本語判決書の自動要約. 言語処理学会 第 30 回年次大会 発表論文集, 2024.

[9] Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, et al. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. **Advances in Neural Information Processing Systems**, Vol. 36, , 2024.

[10] John J Nay, David Karamardian, Sarah B Lawsky, Wenting Tao, Meghana Bhat, Raghav Jain, Aaron Travis Lee, Jonathan H Choi, and Jungo Kasai. Large language models as tax attorneys: a case study in legal capabilities emergence. **Philosophical Transactions of the Royal Society A**, Vol. 382, No. 2270, p. 20230159, 2024.

[11] Jens Frankenreiter and Julian Nyarko. Natural language processing in legal tech. **Legal Tech and the Future of Civil Justice (David Engstrom ed.) Forthcoming**, 2022.

[12] 礒崎陽輔. 分かりやすい 法律・条例の書き方［改訂版］. 株式会社ぎょうせい, 2017.

[13] 石毛正純. 法制執務詳解 新版 III. 株式会社ぎょうせい, 2022.

[14] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. **Advances in neural information processing systems**, Vol. 33, pp. 1877–1901, 2020.

[15] OpenAI. Gpt-4o mini: advancing cost-efficient intelligence. Accessed on January 2025.

[16] OpenAI. Openai o1-mini: Advancing cost-efficient reasoning. Accessed on January 2025.

付録

A　LLM に用いたプロンプト

LLM による定義規定・略称規定における正式名称と略称のペアの抽出を行うにあたり，そのタスクの意図や出力形式を示すための Few-Shot プロンプトを作成した．下記に，本研究の実験で使用したプロンプトを示す．サンプル 1 から 4 までが少数例（Few-Shot）として与えられており，それぞれ簡易的な法令文を例に「入力」と「期待する出力」を対応づけて示している．これらのサンプルは，提案法がサポートする定義規定・略称規定における文型を反映したものである．なお，{law_text} には，定義規定・略称規定が含まれる法令文 T_i が代入される．

以下のタスクを行ってください．

【タスク】
- 入力として与えられるテキストには，法令文独特の形で「正式名称」と「略称」が定義されている可能性があります．

- これをすべて検出し，それぞれのペアを JSON 配列で出力してください．
- 略称が複数ある場合はすべて挙げること．
- 見つからなかった場合は空の配列 `[]` を返してください．
- 出力は JSON のみで，他の説明文は一切不要です．

【JSON の形式】
下記のような配列で返してください．配列要素が複数になる場合はカンマで区切ります．

```
[
  {{
    "formal": "正式名称",
    "abbr": "略称"
  }},
  {{
    "formal": "正式名称",
    "abbr": "略称"
  }}
]
```

【サンプル 1】
入力:

「〇〇」とは、△△をいう。

期待する出力:
```
[
  {{
    "formal": "△△",
    "abbr": "〇〇"
  }}
]
```

【サンプル 2】
入力:

△△（以下「〇〇」という。）

期待する出力:
```
[
  {{
    "formal": "△△",
    "abbr": "〇〇"
  }}
]
```

【サンプル 3】
入力:

〇〇（△△をいう。以下同じ。）

期待する出力:
```
[
  {{
    "formal": "△△",
    "abbr": "〇〇"
  }}
]
```

【サンプル 4】
入力:

〇〇（△△に規定する〇〇をいう。以下同じ。）

期待する出力:
```
[
  {{
    "formal": "△△に規定する〇〇",
    "abbr": "〇〇"
  }}
]
```

【本番入力】

{law_text}

注意:
- 略称と正式名称が見つかっても，その他の要素（定義されていない言葉や補足情報）は出力しないでください．
- 出力は valid な JSON 配列のみで，前後に余計な文章を書かないでください．