Vol.2019-NL-241 No.5
2019/8/29

議会会議録に含まれる法律名の表記揺れ問題解決に向けたエンティティリンキングの試み

桧森 拓真[1,a)]　木村 泰知[2]　荒木 健治[3]

概要：
国会では，委員会や本会議において法律案に関する議論が行われている．数多くの議員が法律案の名称を何度も発言することから，法律案の名称は，省略されることや異なる表現で呼ばれることがあり，同一の法律案を示しているのかを判断することが困難な場合がある．例えば，「働き方改革法案」には「働き方改革関連法」「働き方改革一括法」などの表記揺れが存在する．そこで，本研究では，議会会議録に含まれる法律名の表記揺れの問題を解決するために，エンティティリンキングを行う．実験では，辞書ベース，Wikipedia2Vec をベースラインとし，提案手法との比較を行う．提案手法では，ベースラインである Wikipedia2Vec に加え，メンションの各候補エンティティと、メンションを含む一文の分散表現とのコサイン類似度，メンションとエンティティの文字列の差分の LenghScore，メンションとエンティティの間で一致している文字数に応じた Penalty に基づきスコアを算出し，最もスコアの高いエンティティを出力する．実験の結果，国会データでは，提案手法が F 値において 0.713 を示し，0.198 ポイントベースラインを上回り，地方議会会議録データでは，F 値において 0.719 を示し，0.030 ポイントベースラインを上回る結果となった．

1. はじめに

国会や地方議会では，法律や条例の議論が頻繁に行われている．例えば，衆議院における法律案審議から公布までの流れは，委員会および本会議において法律案の審議・採決が行われ，参議院でも委員会および本会議において審議・可決されることで，法律案の成立・公布となる[*1]．このように法律の審議は，衆議院と参議院の両議院において委員会および本会議の議論が行われており，それらのすべてが会議録として記録されている．一つの法律案が成立・公布されるまでの議論では，数多くの議員が法律案の名称を何度も発言することから，同一の法律案に対して，省略されることや異なる表現で呼ばれることがあり，表記揺れが生じる．

従来研究には，表記揺れ，曖昧性解消を解決するために，テキスト中の固有表現と知識ベースとを結びつけて利用する「エンティティリンキング」と呼ばれるタスクが存在する．エンティティリンキングとは，テキスト中の固有表現を，知識ベースのレコード (エンティティ) に対応付けるタスクである．知識ベースに Wikipedia を用いる場合，特に wikification と呼ばれ，Wikipedia ページがエンティティとなる [1].

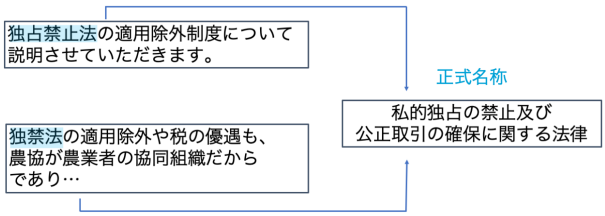

図 1　エンティティリンキングの例

エンティティリンキングは一般的に，大きく分けて二つのタスクに分けられる．一つはメンションの抽出であり，もう一つは曖昧性解消である．ここで，メンションとは，エンティティへの言及のことであり，エンティティと結びつけたい表現を指す．メンション抽出タスクでは，テキスト中からエンティティと結びつけたい表現を抽出する．一般的には，固有表現抽出の技術が用いられており，IOB2 タグ [2] などを用いてテキスト中にメンションを表す範囲にタグ付けを行う．曖昧性解消タスクでは，まずメンション

[1] 北海道大学大学院情報科学院
[2] 小樽商科大学
[3] 北海道大学大学院情報科学研究院
[a)] takuma.himori@gmail.com
[*1] 参議院で否決あるいは修正が行われた場合，衆議院にて 3 分の 2 以上の多数あるいは，両院協議会において両院が可決すると，法律案の成立・公布が行われる．

と結びつけるエンティティの候補生成を行う．その後，生成した候補に対し，ランキング付けを行い，最も高いものをメンションと結びつけるエンティティとする．候補生成の際に，候補が存在しなかった場合には，メンションに対するエンティティが無いものとして NIL とする．曖昧性解消タスクのみ行うものを Disambiguation-Only-Aproach，メンション抽出と曖昧性解消タスクの二つを行うものを End-to-End-Aproach という．従来のエンティティリンキングタスクの場合，同一の表記のメンションから文脈に相応しい候補エンティティを選択する必要がある．

しかしながら，法律名のエンティティリンキングを行う場合，候補エンティティが同一のものを示す場合がある．例えば，図 1 の「独占禁止法」というメンションの候補エンティティには，略称である「独占禁止法」というエンティティと，正式名称である「私的独占の禁止及び公正取引の確保に関する法律」が存在する．これらはいずれも意味的には同一の法律を示すものである．このように法律名のエンティティリンキングタスクの場合，意味的に同一のエンティティから正式名称を正しく推定する必要がある．

そこで，本研究では，法律名の表記揺れ問題を解決するためにエンティティリンキングを行う．最初に，法律名の正式名称と表記揺れしたもののペアを集め，法律名辞書を構築する．次に，構築した辞書をもとに，国会と地方議会会議録から法律名を含む文を抽出したものに，IOB2 タグを付与し，曖昧性解消のためのデータセットを作成する．最後に，作成したデータセットに対し，曖昧性解消の実験を行い，提案手法とベースラインの結果を比較する．

2. 関連研究

本章では，エンティティリンキングについて，Disambiguation-Only-Aproach と End-to-End-Aproach に分けて述べる．

Disambiguation-Only-Aproach

Disambiguation-Only-Aproach の研究として，山田らの研究 [3] が挙げられる．山田らは skip-gram モデル [4][5] を Link graph モデルと Anchor context モデルの 2 つのモデルへ拡張，学習を行いエンティティの曖昧性解消を行っている．Link graph モデルでは，Wikipedia におけるエンティティのリンク構造に基づき、近傍のエンティティを推定することによりエンティティの関係性の学習を行う．Anchor context モデルでは，Wikipedia 上で，エンティティへのリンクを示すアンカーテキストとその文脈に着目し、エンティティの近傍ワードを予測することで学習を行う．山田らは，日本語を含む 12 の言語で学習を行ったモデル (Wikipedia2Vec[*2]) を公開しているが，日本語のデータセットに対しては実験を行っていない．

[*2] https://wikipedia2vec.github.io/wikipedia2vec/pretrained/

End-to-End-Aproach

End-to-End-Aproach の研究として，松田らの研究 [1] が挙げられる．松田らは，日本語テキストに対して自由に使えるエンティティリンキングソフトウェア「jawikify[*3]」を公開している．「jawikify」は，メンション抽出に CRFsuite[6] を用いて IOB2 タグでタグ付けを行っている．曖昧性解消では，生成した候補エンティティに対し，文字列類似度，大域文脈，事前確率を素性としたランキング学習を行うことで尤もらしいエンティティを決定している．ランキング関数の学習には，線形カーネルのランキング SVM(SVM^{rank})[7] を用いている．

3. 法律名データセットの構築

本章では，法律名の曖昧性解消を目的としたデータセットの構築を行う．データセットの構築を行うために，まず表記揺れが存在する法律の辞書を作成する．その後，作成した法律名辞書に存在する表記をもとに，国会と地方議会会議録からデータセットの構築を行う．

3.1 法律について

国会で審議される法律は「新たな法律の制定」と「既存の法律の改正若しくは廃止」の二つに分けることができる[*4]．特に「既存の法律の改正若しくは廃止」では「既存の法律」と「既存の法律の一部を改正する法律」の違いを認識する必要がある．ここで，法律が改正されるまでの流れを「私的独占の禁止及び公正取引の確保に関する法律」を例にして説明する．法律改正までの法律は，次の 3 つに分けることができる．

1 私的独占の禁止及び公正取引の確保に関する法律
2 私的独占の禁止及び公正取引の確保に関する法律の一部を改正する法律
3 (改正後の) 私的独占の禁止及び公正取引の確保に関する法律

まず，改正される前の法律 1 に対して，1 の法律の一部を改正する法律 2 が公布される．次に，2 が施工されると，改正後の法律 3 となる．

法律の分野において，1 および 3 は同一の「私的独占の禁止及び公正取引の確保に関する法律」として扱い，2 の「私的独占の禁止及び公正取引の確保に関する法律の一部を改正する法律」とは区別される．したがって，本研究では，1 と 3 を同一の正式名称，2 を別の正式名称として扱う．

3.2 法律名辞書の構築方法

本節では，法律名の表記揺れを調査するとともに，法律名の正式名称と表記揺れのペアを人手で獲得し，辞書とする．以下に辞書の作成手順と辞書に追加する条件を示す．

[*3] https://github.com/conditional/jawikify
[*4] https://www.clb.go.jp/law/process.html

手順 1　「正式名称」は内閣法制局のホームページ[*5]をもとに，平成 25 年 3 月 6 日から平成 30 年 12 月 14 日までに公布された法律一覧から選定

手順 2　全ての法律名を Wikipedia で検索し，ページ内に略称・別称が掲載されている場合はそれらを正式名称とともに辞書へ登録する．この時，公布年月日の前後を問わず法律名が変更されていた場合は，変更前後の法律名も別称として登録する

手順 3　全ての法律名を Google のニュース検索で検索をし，1 ページ目に表示された記事タイトル・記事本文で略称・別称・俗称が使用されている場合は、辞書へ登録する

手順 4　辞書に登録された正式名称・略称・別称を Twitter で検索をし，”すべてのツイート”から他の略称・別称・俗称を登録する

手順 5　それぞれの表記を略称・別称・俗称に分類し記載する

上記の手順に加え，以下の条件に基づき辞書に登録する略称・別称・俗称を決定する．

条件 1　ある法律名において，略称・別称・俗称の合計が二つ未満の場合は辞書から削除する

条件 2　「A 及び B に関する法律」といったような一つの法律名の中に A と B という法律名が含まれている場合，Wikipedia・Google ニュース検索で検索をする際「A 及び B に関する法律」,「A」,「B」の法律名で検索をし，それぞれ別の項目として登録する

条件 3　「C の一部を改正する法律」のような法律名に関して，検索結果に「改正 C」という略称が出現した場合はこれを登録しないが,”C”の部分でさらに省略され，直前に”改正”が付属している場合は辞書へ登録する

条件 4　略称・別称・俗称の分類は次のように定める

略称　正式名称で用いられている単語のみで構成されたもの

別称　正式名称で用いられていない単語を使用しているもの，言い換え表現であると判断できるもの，あるいは変更前後の法律名も含むもの

俗称　皮肉や揶揄が含まれる表現や間接的な表現であるもの

3.2.1　法律名辞書構築の結果

3.2 の手順により対象となった法律の数を表 1 に示す．対象となった 633 の法律名から，表記揺れの存在する法律名という条件に基づき，人手で辞書構築を行った結果，142 の法律名が対象となり，481 の表記揺れが登録された．表 2 に「略称」「別称」「俗称」のそれぞれの数を示す．

次に，法律名辞書の具体例を表 3 に示す．

[*5] https://www.clb.go.jp/contents/index.html

表 1　対象の法律の数（H25.3 から H30.12 まで）

H25	H26	H27	H28	H29	H30	合計
112	137	78	115	86	105	633

表 2　正式名称と表記揺れ（略称・別称 ・俗称）の数

正式名称	略称	別称	俗称	合計
142	291	152	38	481

表 3 に示すように，法律名辞書は，「正式名称」「略称・別称・俗称」「正式名称の Wikipedia ページ」の 3 項目で構成される．「正式名称の Wikipedia ページ」が空欄となっているものは，Wikipedia 上に正式名称の記事が無いことを意味する．Wikipedia 上に記事が存在しないものは，大きく分けて三つ存在する．一つは「働き方改革を推進するための関係法律の整備に関する法律」のように略称の記事が存在するものである．二つ目は，「健康増進法の一部を改正する法律」のように，すでにある法律を改正する法律は Wikipedia に存在しない．三つ目は，「スポーツにおけるドーピングの防止活動の推進に関する法律」のように新しく公布されたばかりの法律は Wikipedia に記事が無いものが多い．また，「カジノ法」のように，同一の略称が「特定複合観光施設区域の整備の推進に関する法律」と「特定複合観光施設区域整備法」のそれぞれ別の法律の正式名称を指す場合がある．このように，正式名称の表記揺れだけではなく，略称・別称・俗称の中にも表記揺れが存在する．

3.3　法律名データセットの構築

本節では，**3.2** で構築した法律名辞書をもとに，法律名の曖昧性解消に向けたデータセットの構築を行う．データセットの構築には，平成 27 年 1 月 1 日から平成 30 年 12 月 31 日までの国会の会議録四年分と平成 23 年 4 月から平成 27 年 3 月までの 47 都道府県の会議録四年分を対象とする．これらのデータに対し，**3.2** で作成した辞書をもとに，法律名の正式名称・表記揺れを含む文を抽出する．抽出した文に対し，形態素解析器 MeCab[8] で形態素解析を行い，IOB2 タグを付与する．MeCab の辞書には UniDic[9] の短単位を用いて，形態素解析を行った．実験データのフォーマットは，CoNLL-2003 データセット [10] および AIDA-CoNLL データセット [11] をもとに設定した．

次に，実験データの例を表 4,5 に示す．表 4 のようにメンションが指す正式名称の Wikipedia ページがあれば，そのリンクを記載する．表 5 のように，正式名称の Wikipedia ページがなければ記載しない．また，「子育て妨害法」のように作成した法律名辞書にない表記の場合，法律名であってもメンションとはみなさない．

4 章の実験では，本節で作成したデータセットを用いて曖昧性解消を行う．入力には，メンションあるいは，メンションとそれを含む一文全てを入力し，尤もらしいエンティティを出力する．出力したエンティティがメンション

表 3 法律名辞書の例

正式名称	略称・別称・俗称	正式名称の Wikipedia ページ
働き方改革を推進するための関係法律の整備に関する法律	働き方改革関連法	
働き方改革を推進するための関係法律の整備に関する法律	働き方改革一括法	
働き方改革を推進するための関係法律の整備に関する法律	働き方改革推進法	
働き方改革を推進するための関係法律の整備に関する法律	奴隷法	
働き方改革を推進するための関係法律の整備に関する法律	働き方改革法	
働き方改革を推進するための関係法律の整備に関する法律	ブラック企業支援法	
働き方改革を推進するための関係法律の整備に関する法律	過労死促進法	
私的独占の禁止及び公正取引の確保に関する法律	独占禁止法	https://ja.Wikipedia.org/wiki/私的独占の禁止及び公正取引の確保に関する法律
私的独占の禁止及び公正取引の確保に関する法律	独禁法	https://ja.Wikipedia.org/wiki/私的独占の禁止及び公正取引の確保に関する法律
私的独占の禁止及び公正取引の確保に関する法律	ドキンちゃん	https://ja.Wikipedia.org/wiki/私的独占の禁止及び公正取引の確保に関する法律
特定複合観光施設区域の整備の推進に関する法律	カジノ法	https://ja.Wikipedia.org/wiki/特定複合観光施設区域の整備の推進に関する法律
特定複合観光施設区域の整備の推進に関する法律	IR 法	https://ja.Wikipedia.org/wiki/特定複合観光施設区域の整備の推進に関する法律
特定複合観光施設区域の整備の推進に関する法律	IR 推進法	https://ja.Wikipedia.org/wiki/特定複合観光施設区域の整備の推進に関する法律
特定複合観光施設区域の整備の推進に関する法律	IR 整備推進法	https://ja.Wikipedia.org/wiki/特定複合観光施設区域の整備の推進に関する法律
特定複合観光施設区域整備法	カジノ法	
特定複合観光施設区域整備法	カジノ整備法	
特定複合観光施設区域整備法	IR 法	
特定複合観光施設区域整備法	IR 整備法	
健康増進法の一部を改正する法律	受動喫煙防止法	
健康増進法の一部を改正する法律	受動喫煙法	
健康増進法の一部を改正する法律	屋内禁煙法	
スポーツにおけるドーピングの防止活動の推進に関する法律	アンチドーピング法	
スポーツにおけるドーピングの防止活動の推進に関する法律	ドーピング防止活動推進法	
スポーツにおけるドーピングの防止活動の推進に関する法律	反ドーピング法	

表 4 実験データの例 (エンティティの Wikipedia ページが存在する場合)

形態素	IOB2 タグ	メンション	エンティティ (正式名称)	エンティティの Wikipedia ページ
独占	B	独占禁止法	私的独占の禁止及び公正取引の確保に関する法律	https://ja.Wikipedia.org/wiki/私的独占の禁止及び公正取引の確保に関する法律
禁止	I	独占禁止法	私的独占の禁止及び公正取引の確保に関する法律	https://ja.Wikipedia.org/wiki/私的独占の禁止及び公正取引の確保に関する法律
法	I	独占禁止法	私的独占の禁止及び公正取引の確保に関する法律	https://ja.Wikipedia.org/wiki/私的独占の禁止及び公正取引の確保に関する法律
の				
適用				
除外				
制度				
に				
つい				
て				
説明				
さ				
せ				
て				
いただき				
ます				
。				

表 5 実験データの例 (エンティティの Wikipedia ページが存在しない場合)

形態素	IOB2 タグ	メンション	エンティティ (正式名称)	エンティティの Wikipedia ページ
ホワイトカラーエグゼンプション				
は				
過労	B	過労死促進法	働き方改革を推進するための関係法律の整備に関する法律	
死	I	過労死促進法	働き方改革を推進するための関係法律の整備に関する法律	
促進	I	過労死促進法	働き方改革を推進するための関係法律の整備に関する法律	
法	I	過労死促進法	働き方改革を推進するための関係法律の整備に関する法律	
案				
で				
あり				
、				
子育て				
妨害				
法				
案				
で				
あり				
、				
家庭				
不仲				
法				
案				
です				
よ				
。				

の正式名称であれば正解とする．

4. 法律名の曖昧性解消実験

4.1 実験の目的

本実験では，End-to-End-Aproach に向けた，曖昧性解消タスクのみを対象とする．法律名のメンションの抽出が正確に行えた場合，どの程度の精度で法律名の曖昧性解消を行えるかを明らかにすることを目的とする．

4.2 実験方法

本実験では，メンションあるいはメンションを含む一文を入力として与え，尤もらしいエンティティを出力する．

入力　メンションあるいはメンションを含む一文
出力　エンティティ
評価　適合率，再現率，F 値
比較手法　辞書ベース，Wikipedia2Vec，
提案手法 (Wikipedia2Vec, コサイン類似度，LengthScore, Penalty)

評価には，適合率，再現率，F 値を用いる．それぞれの計算方法を式 (1)，(2)，(3) に示す．

$$\text{適合率} = \frac{\text{正しくリンクされたエンティティ数}}{\text{リンクされたメンション数}} \quad (1)$$

$$\text{再現率} = \frac{\text{正しくリンクされたエンティティ数}}{\text{メンションの総数}} \quad (2)$$

$$\text{F 値} = \frac{2 \cdot \text{適合率} \cdot \text{再現率}}{\text{適合率} + \text{再現率}} \quad (3)$$

辞書ベース以外の手法では，各候補エンティティに対し，スコアをもとにランキング付けを行い，最も高いスコアのエンティティを出力し，正しい正式名称を出力した場合正解とする．

4.2.1 ベースライン

本実験のベースラインとして，辞書ベースと山田らの Wikipedia2Vec[3] を用いる．Wikipedia2Vec のモデルには，2019 年 7 月 3 日の日本語 Wikipedia の dump データを用いて，300 次元の分散表現の学習を行った．

辞書ベース

1 Wikipedia のカテゴリ「日本の法律[*6]」に属する法律名全てを登録した辞書を自動で作成する
2 メンションの最初の形態素で始まる法律名を辞書から抽出する
3 抽出した法律名にエンティティと一致するものが存在すればそれを出力する
4 存在しなければ抽出した中で最長の法律名を出力する
5 法律名を抽出できなかった場合 NIL とする

[*6] https://ja.Wikipedia.org/wiki/Category:日本の法律

Wikipedia2Vec

1 メンションを Wikipedia2Vec の入力とする
2 メンションと最も類似度の高いエンティティを出力とする
3 入力したメンションに対し，エンティティが存在しなければ NIL とする

4.2.2 提案手法

Wikipedia2Vec を用いて，各メンションに対し，候補エンティティを最大 5 つまで列挙する．候補エンティティを 5 つよりも多くするとメンションと関係のない法律も含まれるため今回は候補エンティティを 5 つまでとした．各候補エンティティに対し，Wikipedia2Vec の類似度，コサイン類似度，LengthScore，Penalty を用いてスコアを計算し，最もスコアの高いエンティティを出力とする．それぞれのスコアの計算方法を以下に示す．メンションを m，エンティティを e，メンションの文字の集合を $c \in m$ と表す．

Wikipedia2Vec

1 メンションを Wikipedia2Vec の入力とする
2 各候補エンティティのメンションとの類似度をスコア S_{W2V} とする

コサイン類似度

1 doc2vec[12] を用いて，メンションを含む一文を 300 次元のベクトルに変換する
2 doc2vec のモデルは，2019 年 7 月 3 日の日本語 Wikipedia の dump データを用いて作成した
3 学習の際のパラメータは Lau ら [13] のものを用いる
4 変換したベクトルと各候補エンティティの Wikipedia2Vec の 300 次元のベクトルとのコサイン類似度をスコア S_{COS} とする

LengthScore

1 メンションとエンティティの文字列の長さに基づきスコアの計算を行う
2 LengthScore を S_L とし，スコアの計算方法を式 (4) に示す

$$S_L = \frac{\text{length(e)} - \text{length(m)}}{\max\{\text{length(e)},\text{length(m)}\}} \quad (4)$$

Penalty

1 LengthScore のみを用いた場合，一番長い法律名のスコアが高くなる
2 そこで，メンションと候補エンティティの一致している文字数に着目し，ペナルティを付与する
3 ペナルティの計算方法を式 (5)，(6)，(7) に示す

$$\text{Penalty}(P) = \frac{\text{length(m)} - \text{Count(m,e)}}{\text{length(m)}} \quad (5)$$

$$\text{Count(m,e)} = \sum_{c \in m} f(c, e) \quad (6)$$

$$f(c, e) = \begin{cases} 1 & (c \in e) \\ 0 & (otherwise) \end{cases} \quad (7)$$

スコアの計算方法

1 候補エンティティのスコアの計算方法を式 (8)，(9) に示す
2 上記の 4 つのスコアに対し，それぞれ重みをかけ合わせ，総和を取ったものをエンティティのスコアとする
3 重みは 0 から 1.0 まで 0.1 刻みの値を取り，各スコアごとに重みを付与する
4 重みが 0 のときは，そのスコアを使わないこととする
5 スコアの最大が 0 以下になった場合，NIL とする

$$\text{Score} = \alpha \cdot S_{W2V} + \beta \cdot S_{COS} + \gamma \cdot S_L - \delta \cdot P \quad (8)$$

$$\alpha, \beta, \gamma, \delta = \{0.0, 0.1, 0.2, ..., 1.0\} \quad (9)$$

4.3 実験データ

本実験で用いる実験データを表 6 に示す．実験データには，**3.3** で作成した法律名データセットを用いる．

表 6 実験データ

データ	文数	メンション	正式名称	表記揺れ
国会	3,631	4,023	1,019	3,004
地方議会会議録	10,188	11,318	2,112	9,026

4.4 国会の実験結果

国会データに対する実験結果を表 7 に示す．実験結果から，適合率のみで比較すると辞書ベースが 0.879 と最も高いことがわかる．これは，メンションの約 8 割を辞書に存在しない NIL としているためである．そのため，辞書ベースでは再現率が 0.181，F 値が 0.300 となりどちらも最も低い値となっていることがわかる．Wikipedia2Vec では，NIL が約 3 割ほどに減少しており，F 値も辞書ベースと比較し，0.215 ポイント向上した．これは，Wikipedia のリンク構造情報を利用することで，メンションに関連したエンティティが上手く取得できているためである．また，Wikipedia2Vec にコサイン類似度，LengthScore，Penalty をそれぞれ候補エンティティのスコアに加えることで精度の向上が確認された．F 値において最も高いもので 0.713 を達成し，Wikipedia2Vec のみと比較し 0.198 ポイント向上した．

4.5 地方議会会議録の実験結果

地方議会会議録データに対する実験結果を表 8 に示す．実験結果から，国会データと同様に辞書ベースでは，適合率が 0.851 と高く，再現率が 0.156，F 値が 0.264 といずれも最も低い値となっていることがわかる．また，NIL が約 8 割と国会データと同程度の割合を示した．Wikipedia2Vec では，NIL が約 3 割ほどに減少しており，こちらも国会データと同程度の割合を示している．辞書ベースと比較した場合，国会データと同様の理由から，F 値は 0.425 ポイント向上した．しかしながら，Wikipedia2Vec にコサイン類似度，LengthScore をスコアに加えた場合，Wikipedia2Vec と同様の結果を示している．これは，コサイン類似度および LengthScore の重みが 0 の時，最も精度が良くなったため，Wikipedia2Vec と同様の結果となっている．最後に，Penalty をスコアに加えることで，F 値において 0.719 と最も高い数値を示したが，コサイン類似度を考慮した場合と考慮しないいずれの場合においても同様の結果を示した．

4.6 考察

本節では，国会データおよび地方議会会議録データの実験結果の例を挙げ，考察を行う．表 9 にメンションと各手法の出力を示す．

まず初めに「独占禁止法」を例に挙げて考察を行う．Wikipedia 上には，略称である「独占禁止法」と正式名称である「私的独占の禁止及び公正取引の確保に関する法律」がエンティティとして存在する．Wikipedia2Vec のみを用いた場合，メンション「独占禁止法」はエンティティ「独占禁止法」を出力する．これは，メンションとエンティティが完全一致するため，正式名称のエンティティよりも，略称である「独占禁止法」というエンティティを出力してしまうためである．LengthScore および Penalty をスコアに加えた場合，正式名称であるエンティティ「私的独占の禁止及び公正取引の確保に関する法律の一部を改正する法律」を正しく出力できていることがわかった．これは，文字列の長さ，および，メンションとエンティティの間の一致している文字数を考慮することで，このような例に対し，有効に働いていることがわかる．

次に「円滑化法」を例に挙げる．「円滑化法」の正式名称は「マンションの建替えの円滑化等に関する法律」であり，Wikipedia のエンティティとして存在する．しかしながら，いずれの手法においても，NIL を出力していることがわかる．これは Wikipedia2Vec を用いて候補エンティティの生成を行った際に，エンティティ「マンションの建替えの円滑化等に関する法律」が候補となっていないことが原因である．そのため，Wikipedia のエンティティには存在しているにも関わらず，出力はすべて NIL となってしまっているため，法律名の曖昧性解消を行う際には，候補エンティティを生成する際に課題があることがわかる．

表 7 国会の実験結果

手法	NIL	適合率	再現率	F 値
辞書ベース	3,195	0.879	0.181	0.300
Wikipedia2Vec	1,096	0.601	0.448	0.515
Wikipedia2Vec + コサイン類似度	1,096	0.769	0.559	0.647
Wikipedia2Vec + コサイン類似度 + LengthScore	1,096	0.775	0.564	0.652
Wikipedia2Vec + LengthScore	1,096	0.811	0.590	0.683
Wikipedia2Vec + LengthScore - Penalty	1,172	0.848	0.601	0.704
Wikipedia2Vec + コサイン類似度 + LengthScore - Penalty	1,194	0.875	0.603	0.713

表 8 地方議会会議録の実験結果

手法	NIL	適合率	再現率	F 値
辞書ベース	9,238	0.851	0.156	0.264
Wikipedia2Vec	3,395	0.836	0.585	0.689
Wikipedia2Vec + コサイン類似度	3,395	0.836	0.585	0.689
Wikipedia2Vec + コサイン類似度 + LengthScore	3,395	0.836	0.585	0.689
Wikipedia2Vec + LengthScore	3,395	0.836	0.585	0.689
Wikipedia2Vec + LengthScore - Penalty	3,428	0.875	0.610	0.719
Wikipedia2Vec + コサイン類似度 + LengthScore - Penalty	3,428	0.875	0.610	0.719

表 9 各手法の出力例

メンション	辞書	Wikipedia2vec	+コサイン類似度	+LengthScore	+Penalty
独占禁止法	独占禁止法	独占禁止法	独占禁止法	独占禁止法	私的独占の禁止及び公正取引の確保に関する法律
円滑化法	NIL	NIL	NIL	NIL	NIL

5. おわりに

本稿では，国会・地方議会会議録における法律名の表記揺れの問題を解決することを目的として，法律名データセットの構築，および，エンティティリンキングの実験結果について述べた．

法律名データセットの構築では，公布された法律名から，正式名称と表記揺れ（略称・別称・俗称）の対応関係の項目からなる辞書を作成した．また，作成した法律名辞書の項目から国会と地方議会会議録の法律名データセットを構築した．法律名の表記揺れを解決するための評価実験では，法律名データセットに対し，法律名の曖昧性解消の実験を行った．国会会議録を用いた実験では，提案手法が F 値において 0.713 示し，0.198 ポイントベースラインを上回る結果となった．また，地方議会会議録を用いた実験では，F 値において 0.719 を示し，0.030 ポイントベースラインを上回る結果となった．適合率では，国会データ，地方議会会議録データともに，0.875 を示した．

これらの結果から，国会・地方議会会議録における法律名の表記揺れ問題を解決するエンティティリンキングタスクにおいて，提案手法がベースラインを上回ることを確認した．しかしながら，国会データ，地方議会会議録データともに NIL が約 3 割ほど存在し，Wikipedia 上に正式名称のエンティティが存在しない場合があることが確認された．

今後は，Wikipedia に正式名称が存在しない場合に，正しく NIL を出力できるかの評価実験を行う予定である．

謝辞 本研究は JSPS 科研費 JP16H02912 およびセコム科学技術振興財団の助成を受けています．本研究で利用している法律名辞書を構築した佐藤栞氏に感謝いたします．

参考文献

[1] 松田耕史, 岡崎直観, 乾健太郎. 日本語 wikification ツールキット: jawikify. 言語処理学会第 23 回年次大会, 2017.

[2] Erik F. Tjong, Kim Sang, and Jorn Veenstra. Representing text chunks. In *Ninth Conference of the European Chapter of the Association for Computational Linguistics*, Bergen, Norway, June 1999. Association for Computational Linguistics.

[3] Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yoshiyasu Takefuji. Wikipedia2vec: An optimized tool for learning embeddings of words and entities from wikipedia. *arXiv preprint 1812.06280*, 2018.

[4] Tomas Mikolov, Kai Chen, Greg S. Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013.

[5] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pp. 3111–3119, USA, 2013. Curran Associates Inc.

[6] Naoaki Okazaki. Crfsuite: a fast implementation of conditional random fields (crfs), 2007.

[7] Thorsten Joachims. Training linear svms in linear time. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, pp. 217–226, New York, NY, USA, 2006. ACM.

[8] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. Applying conditional random fields to Japanese morphological analysis. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pp. 230–237, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[9] 康晴伝, 智信小木曽, 秀樹小椋, 篤山田, 信明峯松, 清貴内元, 花絵小磯. コーパス日本語学のための言語資源：形態素解析用電子化辞書の開発とその応用. 日本語科学, Vol. 22, pp. 101–123, oct 2007.

[10] Erik F. Tjong Kim Sang and Fien De Meulder. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4*, CONLL '03, pp. 142–147, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

[11] Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. Robust disambiguation of named entities in text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pp. 782–792, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[12] Quoc V. Le and Tomas Mikolov. Distributed representations of sentences and documents. *CoRR*, Vol. abs/1405.4053, , 2014.

[13] Jey Han Lau and Timothy Baldwin. An empirical evaluation of doc2vec with practical insights into document embedding generation. *CoRR*, Vol. abs/1607.05368, , 2016.