会議録に含まれる法律名を対象とした End-to-End のエンティティリンキングの性能評価

[1] 桧森 拓真　[2] 木村 泰知　[3] 荒木 健治

[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

himori@eis.hokudai.ac.jp

1 はじめに

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 を用いることが多く，その場合，特に wikification と呼ばれる．

法律名のエンティティリンキングの例を図 1 に示す．最初に出現する「カジノ法」という表現は「特定複合観光施設区域整備法」という法律の正式名称を示している．一方，最後に出現する「カジノ法」は「特定複合観光施設区域の整備の推進に関する法律」という法律の正式名称を示している．このように，同一の表記で，異なる正式名称を示す場合がある．また,「特定複合観光施設区域の整備の推進に関する法律」には,「カジノ解禁法」「IR 推進法」などの異なる略称が複数存在している．さらに，議会では法律に対し，否定的な立場から呼称する際に，皮肉表現が使われる場合がある．例えば,「働き方改革関連法」のことを「過労死促進法」と呼称していることがある．このように，法律名の表記揺れ問題には，文脈から判断しなければ略称が何を指しているのかわからない問題がある．また，法律の知識がなければ人手でも判断の難しい場合も存在する．

そこで，本研究ではこれらの問題を解決するため，メンションを「法律名を表す語句」と定義し，新たに法律名曖昧性解消のデータセットを作成する．桧森 [1] らの研究では，法律名の曖昧性解消のみを行っていたが，本研究ではメンション抽出及び曖昧性解消実験を End-to-End で行い評価を行う．

2 関連研究

エンティティリンキングは一般的に，メンション抽出と曖昧性解消タスクに分けられる．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．エンティティリンキングにおいて，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．

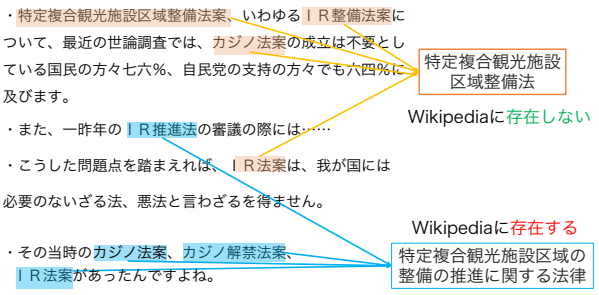

図 1: 法律名のエンティティリンキング

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．近年では，BERT[2] を用いた固有表現抽出モデルが提案されている．オープンソースライブラリである deeppavlov[3] は，マルチリンガルの固有表現抽出モデルを公開しており，そのモデルには BERT が採用されている．

曖昧性解消タスクでは，メンションと結びつけるエンティティの候補生成を行う．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．

Disambiguation-Only-Aproach の研究として，山田らの研究 [4] が挙げられる．山田らは skip-gram モデルを 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．

End-to-End-Aproach の研究として，松田らの研究 [5] 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[1]」を公開している.「jawikify」は，メンション抽出に CRFsuite 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，尤もらしいエンティティを決定している．

3　データセット構築

本章では，End-to-End-Aproach による法律名エンティティリンキングタスクに向けたデータセット構築について述べる．データセット構築には，アノテーション作業を二段階に分けて行った．

初めに，国会･地方議会会議録上に存在する法律名に IOB2 タグでメンションの注釈付けを行う．アノテーションする会議録には，人手でも文脈を見なければ判断の難しい検索語を含むものを 10 日分選んだ．検索語として,「過労死促進法」「クリーンウッド法」のように皮肉表現や正式名称と略称の表層が全く異なるもの,「戦争法」「カジノ法」「円滑化法」のように略称が複数の正式名称を指すものを用いて，それぞれの検索語ごとに 2 日分の会議録の抽出を行った．

次に，注釈付けを行った法律名に関連する Wikipedia 記事が存在する場合，記事のタイトルとその URL を注釈としてつける．データセットのフォーマットは，AIDA CoNLL-YAGO Dataset format[6] に基づいて設定した．

3.1　メンションのアノテーション

メンションのアノテーションは，20 代の文系大学生，女性 1 名と男性 1 名，20 代の理系大学院生，男性 1 名が行った．初めに，2 名が会議録 10 日分に対し，アノテーションを行う．その後，残りの 1 名は，2 名の注釈結果に揺れがあった場合，文脈を読んでメンションかどうかを判断し，最終的な注釈付けを行う．

3.2　Wikipedia 情報のアノテーション

Wikipedia 情報のアノテーションは，理系大学院生 1 名が行った．**3.1** で注釈付けを行った法律名（メンション）に対し，関連する Wikipedia の記事が存在する場合，記事タイトルとその URL を注釈としてつける．例えば,「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」という正式名称が会議録中に出てきた場合，Wikipedia には「働き方改革関連法」という記事が存在するため，これを注釈として付ける．

このように Wikipedia の記事タイトルがメンションである法律名の正式名称とは異なる場合でも注釈付けを行っている．ここで，Wikipedia の記事にメンションに関連したものがなかった場合，NIL とする．

3.3　データセット

3.1，**3.2** でアノテーションを行い作成したデータセットの詳細を表 1 に示す．アノテーションを行った会議録 10 日分は，訓練データに 4 日分，開発データに 2 日分，テストデータに 4 日分を用いる．

ここで，異なりメンション数とは，重複を省いたメンションの総数である．重複を含めた場合のメンションの総数をメンション数として示す．

本データセットの実例を図 2 に示す．

形態素	IOB2 タグ	メンション	Wikipediaタイトル	Wikipedia URL
現在				
の				
犯	B	犯収法	犯罪による収益の移転防止に関する法律	https://ja.wiki...
収	I	犯収法	犯罪による収益の移転防止に関する法律	https://ja.wiki...
法	I	犯収法	犯罪による収益の移転防止に関する法律	https://ja.wiki...
、				
犯罪	B	犯罪収益移転防止法	犯罪による収益の移転防止に関する法律	https://ja.wiki...
収益	I	犯罪収益移転防止法	犯罪による収益の移転防止に関する法律	https://ja.wiki...
移転	I	犯罪収益移転防止法	犯罪による収益の移転防止に関する法律	https://ja.wiki...
防止	I	犯罪収益移転防止法	犯罪による収益の移転防止に関する法律	https://ja.wiki...
法	I	犯罪収益移転防止法	犯罪による収益の移転防止に関する法律	https://ja.wiki...
の				
改正				
等				
など				
の				
…				

図 2: データセットの例

4　実験

本章では **3** 章で作成したデータセットを用いて，2 つの実験を行う．実験では，メンション抽出のみを行った場合の結果と，メンション抽出及び曖昧性解消を End-to-End で行った場合の結果を示す．

4.1　メンション抽出実験

本節では，メンション抽出の実験を行う．メンション抽出には，オープンソースライブラリである

[1]https://github.com/conditional/jawikify

表 1: データセット

データ	形態素数	メンション数	異なりメンション数	検索語
訓練データ	89,095	83	26	カジノ法
	33,259	47	22	戦争法
	57,510	38	22	円滑化法
	19,468	55	23	過労死促進法
開発データ	66,785	26	18	クリーンウッド法
	20,128	13	5	クリーンウッド法
テストデータ	26,002	96	26	カジノ法
	92,674	70	25	戦争法
	40,916	29	7	円滑化法
	24,412	17	10	過労死促進法

deeppavlov[3] を用いる．deeppavlov は BERT を用いた，マルチリンガルの固有表現抽出モデルを公開しており，学習データを用いて新たにモデルを作成することが可能である．本実験は，作成したデータセットを用いて固有表現抽出モデルを作成し，メンションの抽出を行う．

また，ベースラインとして，辞書ベースとの比較も行う．辞書には，e-Gov[2]に登録されている法律を用いる．e-Gov には法律の正式名称に加え，略称が登録されており，これらに登録されている法律名および略称を辞書として用いる．

4.1.1 評価方法

CoNLL2003[7] の評価方法に基づいて評価を行う．メンション（B，I）を正しく認識したものを TP，メンションでないもの（O）を正しく認識したものを FT，メンションでないものをメンションと誤識別したものを FP，メンションをメンションでないと誤識別したものを TP とし，適合率，再現率，F 値で評価する．

4.1.2 実験結果

メンション抽出実験の結果を表 2 に示す．辞書ベースでは，適合率では 100.00%を示している．これは，辞書に登録されているもののみを出力しており，メンションでないものをメンションと誤識別することがなかったためである．しかしながら，辞書に載っていないメンションには対応できないため，F 値は 44.73%という結果になった．BERT を用いた手法では，F 値は 75.67%を示した．辞書ベースと比較し，30.94 ポイント向上する結果となった．

4.1.3 考察

メンション抽出で誤った例として，「改正 PKO 法」が挙げられる．「改正 PKO 法」がメンションとして抽出されるのが正しいが，改正「PKO 法」と「PKO 法」部分のみがメンションとして抽出されていた．これは，データセットに出現する「PKO 法」の出現頻度と比べ，「改正 PKO 法」の出現頻度が低いためである．そのため，出現頻度の低い「改定」部分がメンションとして学習されず，出現頻度の高い「PKO 法」部分のみがメンションとして抽出されたと考えられる．

表 2: メンション抽出実験結果

手法	適合率	再現率	F 値
辞書ベース	100.00%	28.81%	44.73%
BERT	63.09%	94.51%	75.67%

4.2 End-to-End-Aproach による曖昧性解消実験

本節では，End-to-End-Aproach による曖昧性解消実験を行う．まず初めに，メンション抽出を行う．その後，抽出したメンションに対し，Wikipedia 記事と結びつける曖昧性解消を行う．メンション抽出部分には，**4.1** で用いた BERT を使用する．

曖昧性解消には，文脈情報・文字列の長さ・文字の一致度に着目した，桧森らの法律名曖昧性解消の手法 [1] を用いる．

4.2.1 評価方法

曖昧性解消の評価では，桧森らが行った曖昧性解消実験 [1] を基に評価を行う．評価方法において，従来の実験では NIL をシステムの出力に含めていなかったが，本実験では NIL もシステムの出力とみなし，正解が NIL の場合に NIL を出力した時は正解，NIL 以外のものを出力した時は不正解として評価を行う．また，メンション抽出において誤ったものは，曖昧性解消において誤ったエンティティを出力したとして扱い，評価を行う．

[2]https://www.e-gov.go.jp/

4.2.2 実験結果

End-to-End-Aproach による曖昧性解消実験の結果を表 3 に示す．実験の結果，F 値 35.80%で曖昧性を解消できることを確認した．適合率 27.44%は，再現率 50.49%と比較すると，低い結果となった．これは，メンション抽出の際に誤ったメンションを含んでいるためである．

表 3: 曖昧性解消実験結果

適合率	再現率	F 値
27.44%	51.49%	35.80%

4.2.3 考察

曖昧性解消実験で誤った例として,「カジノ法案」が挙げられる.「カジノ法案」は「特定複合観光施設区域の整備の推進に関する法律」と「特定復号観光施設区間整備法」のそれぞれ異なる正式名称を指す略称である．前者は Wikipedia に存在するが，後者は存在しないため,「カジノ法案」が後者を指す場合，NIL が正解となる．このようなものに対し,「特定複合観光施設区域の整備の推進に関する法律」を誤って出力するものが多く見られた．これは，システムが会議録全体の文脈情報ではなく，一文の文脈情報のみを用いて判断を行っているため，Wikipedia に出現する「特定複合観光施設区域の整備の推進に関する法律」のスコアが高くなるためであると考えられる．

5 おわりに

本研究では，End-to-End-Aproach による法律名エンティティリンキングに向けた，データセット構築及び，実験を行った．メンション抽出実験では，BERT を用いた手法が F 値 75.67%を示した．End-to-End-Aproach による曖昧性解消実験では，F 値 35.80%を示した．今後の課題として，議題や議事日程などの会議録特有の文脈情報を用いて，曖昧性解消を行うことなどが挙げられる．

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．

参考文献

[1] 桧森拓真, 木村泰知, 荒木健治. 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. 情報処理学会第２４１回自然言語処理研究会予稿集, Vol. 2019-NL-241, No. 5, pp. 1–8, 2019.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[3] V. Mozharova and N. Loukachevitch. Two-stage approach in russian named entity recognition. In *2016 International FRUCT Conference on Intelligence, Social Media and Web (ISMW FRUCT)*, pp. 1–6, Aug 2016.

[4] 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.

[5] 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.

[6] 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.

[7] 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.