北海道大学
HOKKAIDO UNIVERSITY

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

桧森　拓真[1]　　木村　泰知[2]　　荒木　健治[3]

[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

研究背景

国会では、法律に関する議論が頻繁に行われ

法律名(案)は省略など正式名称とは異なる表現で呼ばれることがある

ホワイトカラー・エグゼンプションは
過労死促進法案であり……

独禁法の適用除外や税の優遇も、
農協が農業者の協同組織だからであり……

表記揺れした表現がどの法律名(案)を示しているのか

判断が困難な場合がある

研究背景

ホワイトカラー・エグゼンプションは

過労死促進法案であり……

私的独占の禁止及び公正

取引の確保に関する法律

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

独禁法の適用除外や税の優遇も、

農協が農業者の協同組織だからであり……

表記揺れした法律名を正式名称と結びつけ曖昧性を解消したい

研究目的

ホワイトカラー・エグゼンプションは
過労死促進法案であり……

私的独占の禁止及び公正
取引の確保に関する法律

エンティティ
リンキング

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

独禁法の適用除外や税の優遇も、
農協が農業者の協同組織だからであり……

タグ付けのされていない文書から法律名を抽出し、法律名の曖昧性解消をする

End-to-Endなエンティティリンキングの性能評価を行う

エンティティリンキングとは

メンション

一昨　年　の　IR　推進　法　の　審議　の　際　に　は　、　…

特定複合観光施設区域の整備の推進に関する法律

建設工事従事者の安全及び健康の確保の推進に関する法律

候補エンティティ

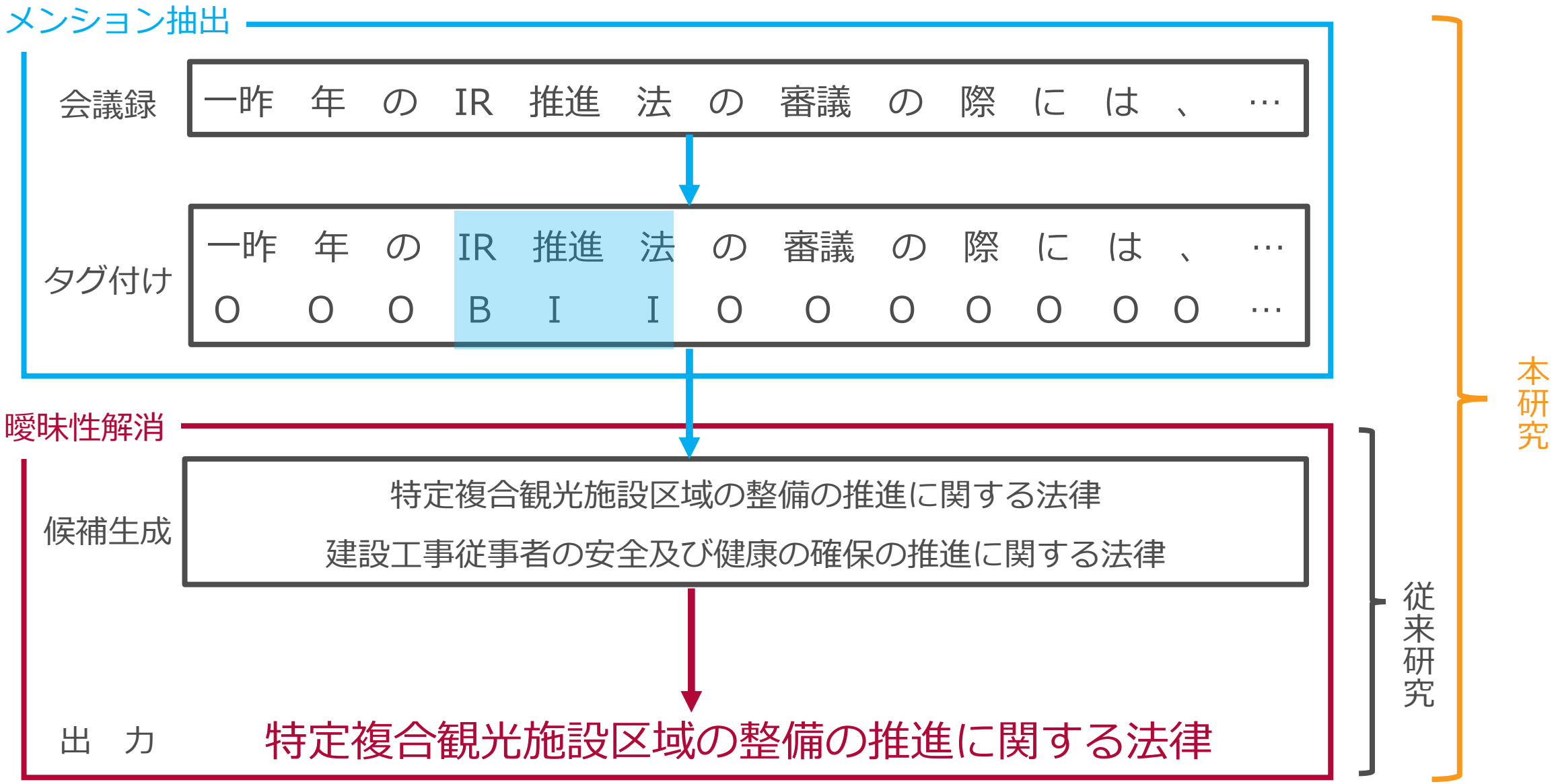
テキスト中のメンション(≒固有表現)を

知識ベースのエンティティ(エントリ)に結びつけるタスク

候補エンティティを生成し、その中から尤もらしいエンティティを決定する

本研究では、メンションを法律を示す表現と定義

法律名のエンティティリンキングの流れ



メンション抽出

テキスト中のメンションを抽出するタスク

メンションの表現にはIOB2タグがよく使われる

入力

一昨	年	の	IR	推進	法	の	審議	の	際	に	は	、	…

出力

一昨	年	の	IR	推進	法	の	審議	の	際	に	は	、	…
O	O	O	B	I	I	O	O	O	O	O	O	O	…

B(Begin)：メンションの始まりを示す

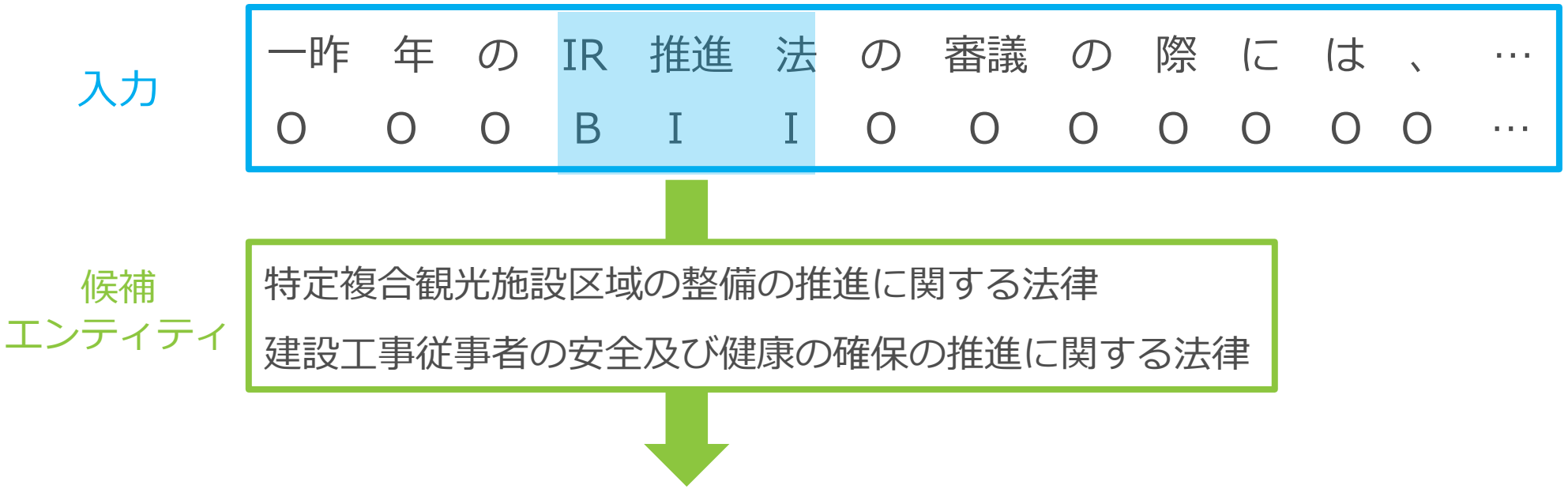
I(Inside)：メンションが続いていることを示す

O(Outside)：メンション以外を示す

メンションの曖昧性解消

抽出したメンションに対し、知識ベースから候補エンティティを生成

候補エンティティの中から尤もらしいエンティティを結びつける



出力　特定複合観光施設区域の整備の推進に関する法律

関連研究

Mozharova[1]らの研究

オープンソースライブラリDeepPavlov[1]の固有表現抽出モデル

機械学習を用いて固有表現を抽出、トークンのラベル情報を特徴量として取得

この特徴量を学習データとし、新しくモデルの学習を行う2段階のアプローチ

山田[2]らの研究（Wikipedia2Vec)

skip-gramモデルを

Wikipediaのリンク構造に着目したLink graphモデルと

アンカーテキストの文脈に着目したAnchor contextモデルに拡張

Wikipedia上の単語とエンティティの類似度計算が行える

[1] https://github.com/deepmipt/DeepPavlov

[1] 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.

[2] 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.

データセット作成

形態素	IOB2タグ	メンション	Wikipediaタイトル	WikipediaURL
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				
など				
の				
…				

対象データ

設定した検索語を含む会議録10日分

検索語

人手でも難しいメンションを検索語とする

- 皮肉表現・表層が正式名称と全く異なる

  「過労死促進法」「クリーンウッド法」

- 略称が複数の正式名称を指すもの

  「カジノ法」「戦争法」「円滑化法」

1. メンションのアノテーション
2. Wikipedia情報のアノテーション

メンションのアノテーション

作業者

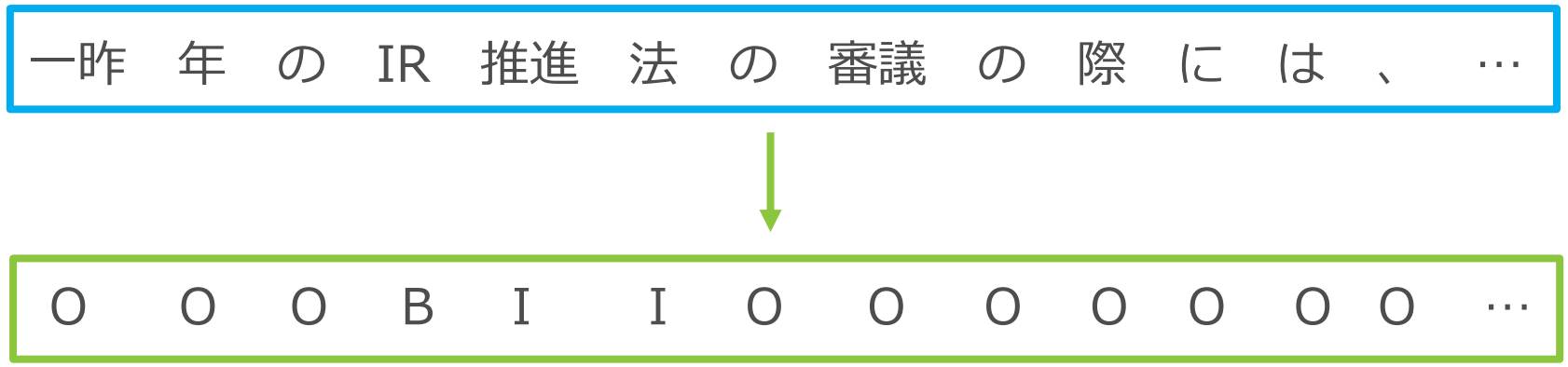
20代の文系大学生　男性1名（A)、女性1名（B)

20代の理系大学院生　男性1名（C）

作業内容

AとBが10回分の会議録中のメンションにアノテーションを行う

CがAとBのつけた注釈に揺れがあった場合、文脈を読んで最終的なアノテーションを行う



Wikipedia情報のアノテーション

作業者

20代の理系大学院生　男性1名（第一著者）

作業内容

アノテーションしたメンションに対して、Wikipedia情報を注釈付けする

メンション（法律名）の情報が書かれている記事のタイトルとURLを付与

そのような記事が存在しなければNILを注釈付け

対象データ

Wikipedia（2019-12-01 dump データ）

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

出典: フリー百科事典『ウィキペディア（Wikipedia）』

メンションと表記が異なる場合でも

同一のものを指していたら注釈付けする

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

形態素解析にはUnidicの短単位を用いた

メンション数：重複を含めたメンションの数

異なりメンション数：ユニークなメンションの数

実　験

メンション抽出実験

アノテーションされていない文書に対してメンション抽出を行う

End-to-Endの曖昧性解消実験

アノテーションされていない文書に対してメンション抽出を行い

抽出したメンションをそれぞれエンティティと結びつける

メンション抽出実験

実験方法

アノテーションされていない文書を入力し、IOB2タグでタグ付けしたものを出力する

入力　一昨　年　の　IR　推進　法　の　審議　の　際　に　は　、　…

出力　O　O　O　B　I　I　O　O　O　O　O　O　O　…

B(Begin)：メンションの始まりを示す
I(Inside)：メンションが続いていることを示す
O(Outside)：メンション以外を示す

評価方法

CoNLL2003 に基づき、適合率・再現率・F値を評価

		正解 (注釈結果)	
		B, I	O
予測結果	B, I	TP	FP
	O	FN	TN

メンション抽出実験

手　法

辞書ベース：e-Gov[2]に登録されている略称・正式名称を辞書に登録

文中に完全一致するものがあれば、それをメンションとする

正式法令名	略称法令名1	略称法令名2	略称法令名3
外務省設置法	中央省庁等改革関連法		
オウム真理教犯罪被害者等を救済するための給付金の支給に関する法律	オウム真理教犯罪被害者救済法		
ハンセン病問題の解決の促進に関する法律	ハンセン病問題基本法	ハンセン病基本法	ハンセン病問題解決促進法
いじめ防止対策推進法	いじめ対策法		

BERT　：DeepPavlovのBERTモデルを訓練データでファインチューニングを行い

法律名用のメンション抽出モデルを作成

「法」を含まない入力は全てOと出力

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

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%

辞書ベース：辞書に登録したもののみを出力するため、適合率は100.00%

既存の辞書に無いものには対応できず、再現率・F値はBERTに劣る

BERT　　：適合率は辞書ベースに劣る

再現率・F値は辞書ベースを上回りF値は75.67%で30.94ポイント向上した

適合率が再現率に比べ低いのは、

メンションでないもの（O)をメンション（B，I)と出力する傾向があったため

エラー分析

メンション	改定	PKO	法
BERT 出力	O	B	I
正　解	B	I	I

改定PKO法はPKO法の出現頻度よりも低いため

PKO法のみがメンションとして学習された

メンション	料理	法
BERT 出力	B	I
正　解	O	O

法律ではない○○法をメンションと誤認識

正解のメンションは○○法という

表記なため○○法と表記されているものを

メンションと出力

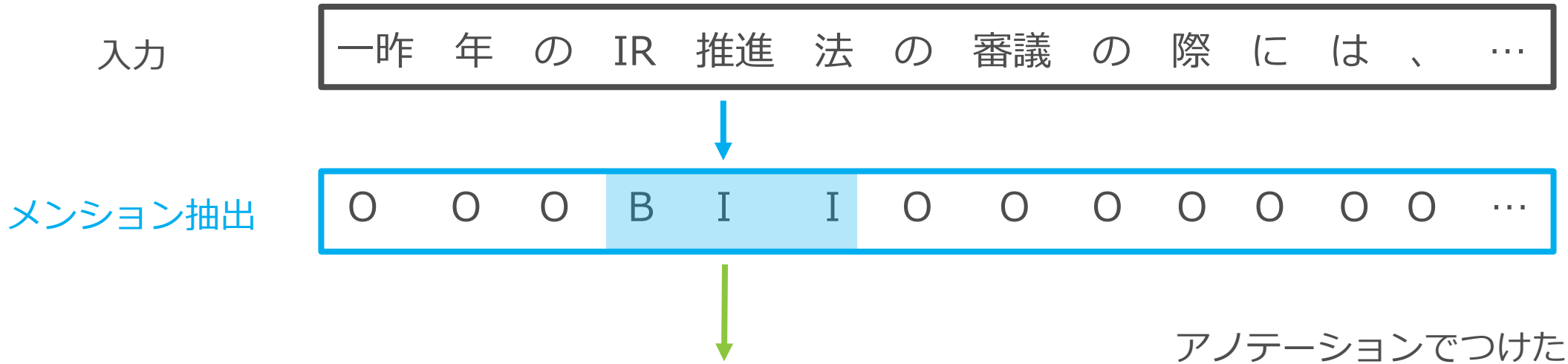
End-to-Endの曖昧性解消実験

実験方法

アノテーションされていない文書を入力し、メンション抽出を行う

抽出されたメンションに対し、それぞれエンティティを出力する

メンション抽出にはメンション抽出実験で精度の良かったBERTを用いる



アノテーションでつけた

Wikipediaタイトルを正解とする

NILの場合、NILを出力すると正解

End-to-Endの曖昧性解消実験

手　法

従来研究である、桧森ら[3]の法律名の曖昧性解消手法を用いる
Wikipedia2vecでメンションの候補エンティティを生成
候補エンティティに対し、Wikipedia2Vecの類似度、コサイン類似度、文字列の長さ、
メンションと候補エンティティの文字の一致度に応じてスコア付けをする
最もスコアの高い候補エンティティをエンティティとして出力する

評価方法

$$\text{適合率} = \frac{\text{正しくリンクされたエンティティ数}}{\text{リンクされたメンション数} + \text{メンション抽出の誤り}}$$

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

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

メンション抽出の誤りは
曖昧性解消を誤ったものとして扱う

[3] 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, 木村泰知, 荒木健治, 情報処理学会 自然言語処理研究会, 2019年08月

End-to-Endの曖昧性解消実験結果

手法	適合率	再現率	F値
辞書ベース（e-Gov）	12.67%	23.78%	16.52%
桧森らの手法	27.44%	51.49%	35.80%

辞書ベースと比較し、桧森らの手法が適合率・再現率・F値において上回る結果

適合率

再現率と比べ24.05ポイント低い結果となった

メンション抽出の際の誤りが影響

再現率

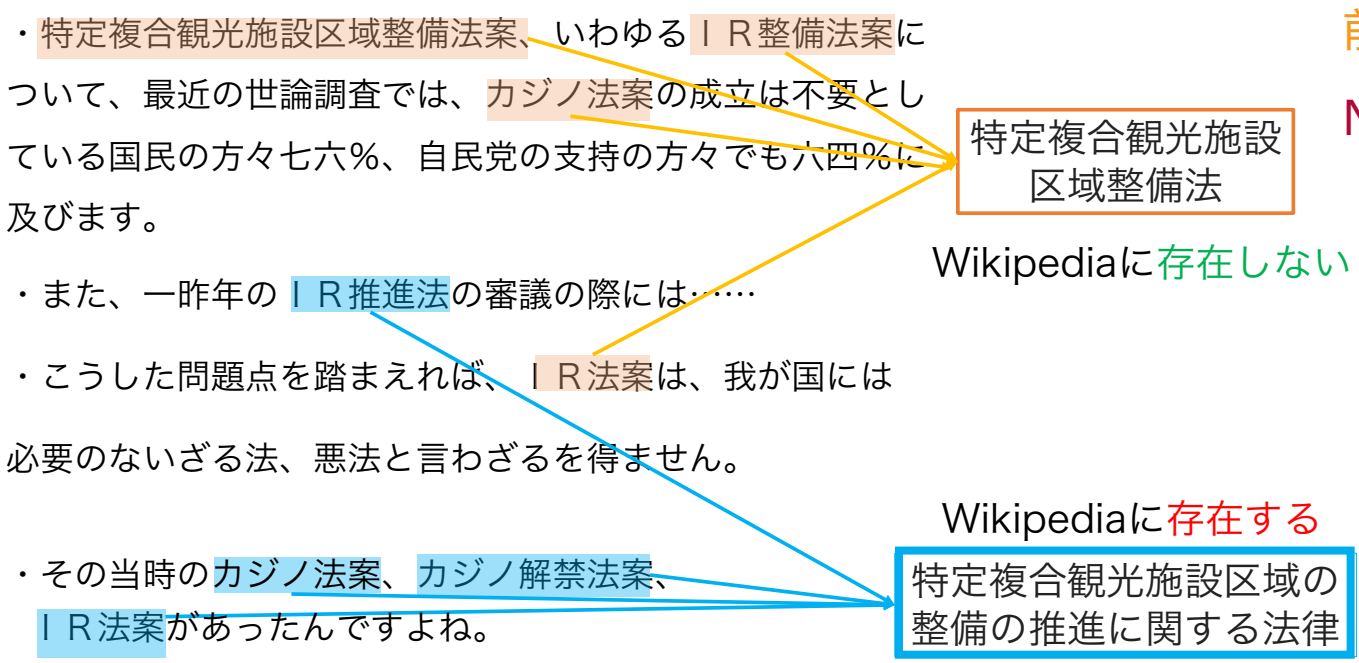
51.49%とテストデータ中のメンションの半分強が正解

文脈を考慮しないと難しいものが不正解となる場合が多かった

エラー分析

人手でも文脈を考慮しないとわからない例

「カジノ法案」は「特定複合観光施設区域整備法」と
「特定複合観光施設区域の整備の推進に関する法律」の異なる法律を指す場合がある



前者を指す場合、Wikipediaに存在しないためNILを出力する必要があるが，システムは「特定複合観光施設区域の整備の推進に関する法律」を出力している

候補エンティティを生成した場合、候補エンティティのどれかを出力してしまう

まとめと今後の課題

ま　と　め

・End-to-Endのエンティティリンキングを行うためのデータセットを作成

・メンション抽出ではBERTを用いた手法がF値75.67%を示した

・曖昧性解消では、桧森らの手法が再現率では51.49%を示したが、

メンション抽出の誤りの影響が大きくF値は35.80%を示した

今後の課題

・会議録全体の文脈を考慮したり、NILが正解のものに対して誤ったものを

出力してしまう問題を解決する必要がある

Appendix

注釈付けマニュアル

目　的

文章中に存在する法律・法案を示す表現を見つけること

作　業

ブラウザ上で以下のように注釈付けをする

しかし、総合的なＴＰＰ等関連政策大綱及びＴＰＰ整備法に基づく国内対策は、ＴＰＰ及び日ＥＵ・ＥＰＡの発効に対応したものであって、米国からの対日輸出の更なる拡大があれば全く不十分なものとなるのではないでしょうか。これでは我が国の農業にとって大打撃となります。

国民民主党は、今国会に他の野党会派と共同で、衆議院に畜産経営の安定に関する法律及び独立行政法人農畜産業振興機構法の一部を改正する法律案を提出しております。

注釈付けの基準（1/2）

・法律・法案の単位

「Aの法律」案：案までを全体に含めます

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

「Aの法律」の一部を改正する法律（案）：2つに分けず1つの法律（案）とする

労働安全衛生法の一部を改正する法律案

「Aの法律」及び「Bの法律」の一部を改正する法律（案）：同上

畜産経営の安定に関する法律及び独立行政法人農畜産業振興機構法の一部を改正する法律案

・一文に複数の法律名がある場合

それぞれの法律ごとに注釈付けを行う

その当時のカジノ法案、カジノ解禁法案、ＩＲ法案があったんですよね。

注釈付けの基準（2/2）

・実際には存在しない法律名（皮肉表現や比喩表現など）

〇〇法（案）という表現であれば実際にはない法律名にも注釈付けをする

第二に、本法案が過労死促進法であることが審議を通じて明らかになったからです。

・○○法案・〇〇法制のような表現

「〇〇法案」「○○法制」という表現も注釈付けをする

請願26年８号「「介護・医療総合確保法案」の撤回等を求める意見書の提出について」

安保法制

・憲法・刑法

「憲法」や「刑法」には注釈付けをしない

候補エンティティのスコア付け

生成した各候補エンティティごとに以下の4つのスコアに重み付けしたものを
各候補エンティティのスコアとする
候補エンティティの中で最もスコアの高いものをエンティティとして出力する

- Wikipedia2Vecの類似度 S_{W2V}
- コサイン類似度　S_{COS}
- Length Score　S_L
- Penalty　P

$$S = \alpha \cdot S_{W2V} + \beta \cdot S_{COS} + \gamma \cdot S_L - \delta \cdot P$$
$$\alpha = 0.7, \beta = 0.1, \gamma = 0.2, \delta = 0.9$$

Wikipedia2Vec（S_{W2V}）

手　順

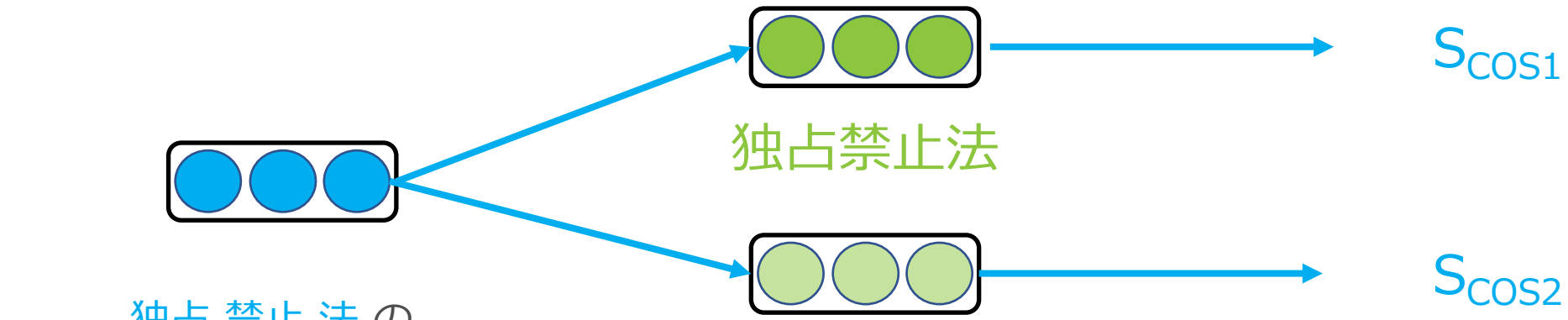
- メンションをWikipedia2Vecに入力
- 類似度の高いエンティティを最大5つまで候補エンティティとして列挙
- 各エンティティの類似度を S_{W2V} とする

コサイン類似度（S_{COS}）

手　順

・メンションを含む一文をdoc2vecでベクトルにする

・メンションのベクトルとWikipedia2Vecで出力した各候補エンティティのベクトルとのコサイン類似度を S_{COS} とする



Length Score（S_L）

手　順

- 文字列が長いものほど正式名称である場合が多い
- メンションとエンティティの文字列の長さでスコアを計算
- Wikipedia2Vecで出力した各候補エンティティごとに
  以下の計算方法による値をスコア S_L とする

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

e：候補エンティティ
m：メンション

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

$$S_L = \frac{22 - 5}{\max(5, 22)} \approx 0.772$$

Penalty（P）

手　順

・メンションとエンティティの一致している文字数に応じて
ペナルティを付与（一致しているほどペナルティを下げる）

・Wikipedia2Vecで出力した各候補エンティティごとに
以下の計算方法による値をペナルティ Pとする

$$P = \frac{length(m) - Count(m,e)}{length(m)}$$

e：候補エンティティ
m：メンション

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

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

独占禁止法　$P = \frac{5-5}{5} = 0$

私的独占の禁止及び公正
取引の確保に関する法律