Wikipediaのカテゴリ情報を用いた 地方議会会議録における発言に対する情報理解支援

桧森 拓真　　谷内 健太　　木村 泰知
小樽商科大学
takuma.himori@gmail.com

1 はじめに

近年，地方自治体では，行政が保有する公共データをコンピュータが扱いやすい形式で，二次利用できるように公開するオープンデータ化の取り組みが進められている．オープンデータは，ビジネスや公共サービスへの活用が望まれているものの，有効に活用されていないのが現状である．その理由として，数値データやテキストなどを結びつけて活用しづらい点が挙げられる．このようにバラバラに公開されている構造化されたデータを結びつけることをタスクとした研究が行われている．例えば，Text Analysis Conference(TAC) では，構造化された知識ベースを結びつけるエンティティリンキングと呼ばれる研究が行われている [1]．特に，文章に含まれる専門用語などを Wikipedia に結びつけるエンティティリンキングのことを Wikification と呼んでいる．Wikification に関する研究としては，松田らの研究があり，情報検索システムの性能向上などを目的として，日本語に対して自由に使える Wikification ソフトウェアの作成を進めている [2]．しかし，従来研究を自然文に対し適用するだけでは得られる情報が少ない．

そこで，本研究では，自治体から公開されているオープンデータの一つである地方議会会議録に Wikification を適用する事で，会議録内の単語 (エンティティ) を取得し，エンティティと関連している単語を一緒に提示することで，読み手の理解を向上させる支援を行うことを目的とする．Wikipedia の各記事には属するカテゴリが存在する．例えば，図 1 の「成立」という記事には,「法」というカテゴリに属している．本研究では，カテゴリ情報を用いることで，表層情報のみでは取得できない，エンティティに類似した単語を取得することである．以下，2 章では Wikification およびカテゴリ情報を用いた，類似単語の取得の説明，3 章では類似単語の取得実験，4 章では取得した類似単語の評価実験を行う．

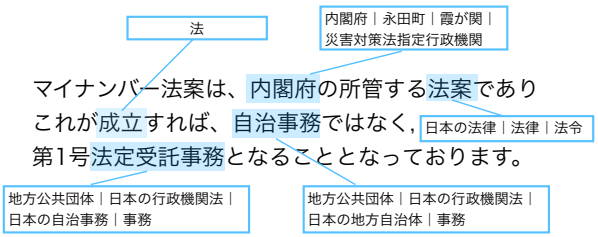

図 1: 単語と所属カテゴリ

2 Wikificationの流れ

本節では，Wikification を行い，類似単語を出力するまでの流れを示す．図 2 に Wikification の流れ図を示す．まず，地方議会会議録から抽出した議員の発言に対し，形態素解析を行う．次に，形態素解析によって得られた形態素に対しフィルターをかける．形態素の文字数が 2 文字以下である場合，その形態素は利用しない．これによって得られた形態素を入力とし，Wikification を行う．Wikification には，word2vec-wikification-pyion[1]を用いる．ここでは，入力された形態素から Wikipedia の記事候補を作成する．次に，記事候補の組み合わせからグラフを作成する．そして，グラフから最適なパスを選び出すことによって，Wikipedia の記事 (エンティティ) を取得する．最適なパスの計算の際には，鈴木らの日本語 Wikipedia エンティティベクトル[2][3] を用いている．得られたエンティティから，エンティティが属する Wikipedia のカテゴリを取得する．取得したカテゴリに含まれる記事を全て取得する．これによって得られた記事の中で，2 つ以上重複したものを類似単語として出力する．これを各エンティティごとに行う．

[1]https://github.com/Kensuke-Mitsuzawa/word2vec-wikification-py
[2]http://www.cl.ecei.tohoku.ac.jp/ m-suzuki/jawiki$_v$ector/

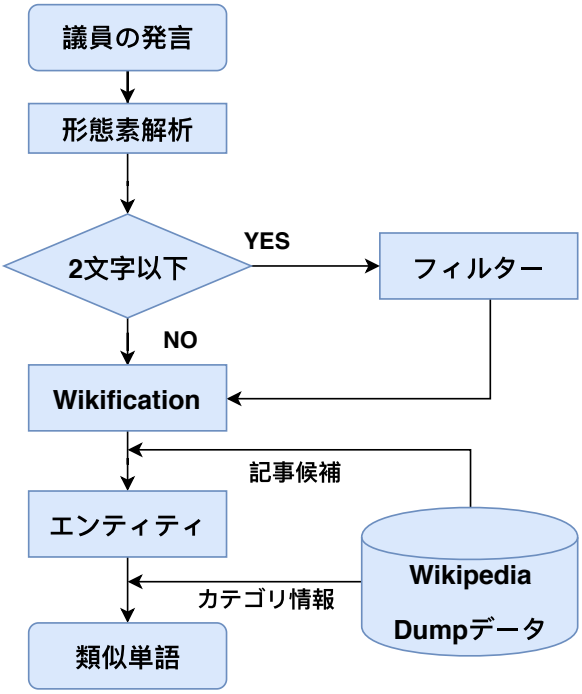

図 2: Wikification の流れ図

3 類似単語の抽出実験

3.1 実験の目的

本実験の目的は，地方議会会議録における発言に対し，Wikification を行いう．これによって得られたエンティティに対し，カテゴリ情報を用いることで，元のエンティティとの類似単語を，得られるかどうかを確認することである．

3.2 実験方法

本実験では，東京都都議会会議録の定例会のデータを用いる．実験の流れは図 2 の Wikification の流れに示した示した通りである．形態素解析には，Mecab[4] を用いる．ここで名詞すべてを形態素とした場合と，固有名詞のみを形態素とした場合を比較する．Mecab の単語辞書には ipadic と mecab-ipadic-neologd[3] を用いる．実験の具体例を図 3，図 4 に示す．

[3]https://github.com/neologd/mecab-ipadic-neologd

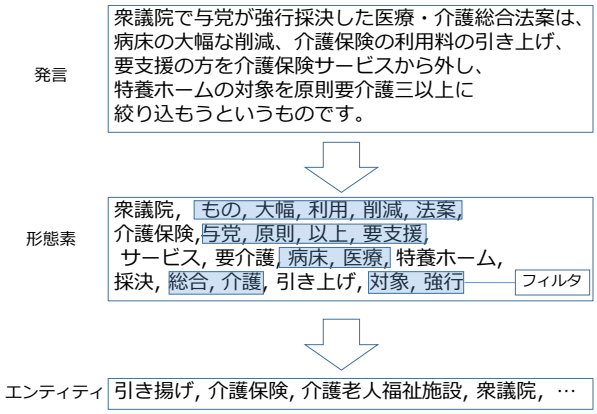

図 3: エンティティの取得例

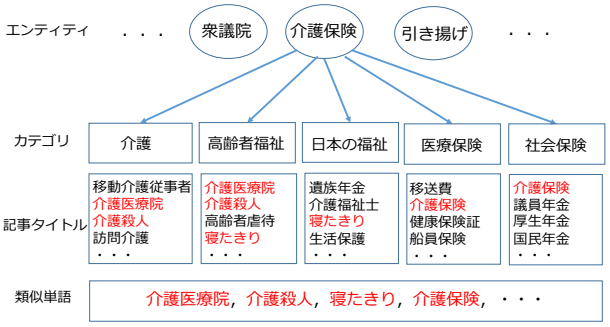

図 4: 類似単語の取得例

3.3 実験データ

実験データを表 1 に示す．実験データには，平成 26 年度東京都都議会第二回定例会のデータを用いる．データの中には，発言と非発言が存在する．ここで，発言とは議員が実際に発言したものである．非発言とは，状況説明や資料の添付など議員の発言以外の文のことである．

3.4 実験結果

実験結果を表 2 に示す．ここで，エンティティ数および類似単語数とはエンティティおよび類似単語を重複せずに数えたものである．平均エンティティ数および平均類似単語数は，一文あたりに含まれるエンティティ数および，類似単語数である．品詞情報として，名詞のみとした場合に ipadic と neologd の結果を比較すると，エンティティ数は，neologd が 1,666 と ipadic の 763 よりも多くなった．同様に類似単語数も，neologd が 415,842 と ipadic の 37,672 よりも多い結果となっ

表 1: 実験データ

対象自治体	対象年度	対象会議録	発言	非発言	総文数
東京都	平成 26 年度	第二回定例会	4,308	1,970	6,278

た．品詞情報を固有名詞とした場合にも，同様の傾向があることを確認した．また，ipadic で固有名詞のみを利用した場合，名詞全てを利用した場合と比較し，エンティティ数 763 から 81 に減少し，類似単語数も 37,672 から 4,397 に減少した．同様に neologd で固有名詞のみを用いた場合，エンティティ数が 1,666 から 1,048 に減少し，類似単語数も，415,842 から 389,154 へと減少した．平均エンティティ数および平均類似単語数は，neologd で名詞全てを利用した場合が最大となり，1.3 と 152.2 という数値になった．

3.5 考察

本節では，実験によって得られたエンティティおよび，類似単語について考察を行う．実験結果に示した通り，辞書として ipadic を用いた場合 neologd を用いた場合に比べ，エンティティ数，類似単語数は減少した．例えば,「介護保険」を形態素解析した場合，ipadic では,「介護」,「保険」のように分割されてしまう．そのため，2 文字以下になる形態素が多かったため，neologd を用いた場合と比べ，エンティティが減少した．エンティティが減少するに伴い，エンティティを利用する類似単語数も減少した．

次に，各エンティティの類似単語数について考察を行う．neologd を用いた場合の各エンティティの類似単語数上位 10 件はいずれも「野村萬斎」などの人名であった.「野村萬斎」というエンティティが持つカテゴリ数は 11 であり，類似単語数は 66,306 であった．エンティティが人名である場合，カテゴリを多く持つことが多く，そのため，2 つ以上重複する記事が多くなるため，類似単語も多くなったと考えられる．ipadic を用いた場合,「野村萬斎」という文字列は「野村」「萬」「斎」に分割され，フィルターによって除去される．そのため，neologd の類似単語数は多くなり，ipadic の類似単語数は少なくなったと考えられる．

4 情報理解支援の評価実験

4.1 目的

本実験の目的は，3 章で取得した類似単語が適切であるかどうかを評価することである．

4.2 評価方法

3 章で取得した類似単語に対し,「関連性」,「有用性」,「理解支援に役立つか」の 3 項目に対して評価を行う．評価は，○ (ある，役立つ)，△ (どちらともいえない)，× (ない，役立たない) の三段階評価で筆者が行う．評価する類似単語には，ランダムに 20 件ずつ抽出した，計 80 件を用いる．

4.3 評価結果

評価結果を表 3 に示す．評価結果から，関連性については，ipadic+名詞，ipadic+固有名詞が 11 件が関連性のあるものとなり，neologd を用いた場合よりも関連性のある単語が多い結果となった．しかしながら，有用性，理解支援は，ipadic+固有名詞，neologd+固有名詞が最も高くなったものの，関連性のあるまたは，理解支援に役立つものはいずれも 4 件であった．

4.4 考察

本節では，評価結果について考察を行う．評価結果でも述べたように，ipadic を用いた場合，neologd よりも関連性のある類似単語が多くなった．実験結果のところでも述べたように，neologd を用いた場合人名を上手く取得することができた．しかしながら，人名のエンティティはカテゴリ数が多く，多くの類似単語を取得してしまうため，元々のエンティティとは関連の低いものを取得してしまったと考えられる．また，有用性，理解支援は，ipadic，neologd のいずれも少ない結果となった．これは，国名，駅名や建物名，企業名などのエンティティの場合，関連のある類似単語

表 2: 実験結果

辞書	品詞情報	エンティティ数	類似単語数	平均エンティティ数	平均類似単語数
ipadic	名詞	763	37,672	0.57	18.09
ipadic	固有名詞	81	4,397	0.03	1.38
neologd	名詞	1,666	415,842	1.27	152.16
neologd	固有名詞	1,048	389,154	0.63	124.47

表 3: 類似単語の評価

組合せ	評価	関連性	有用性	理解支援
ipadic	○	11	2	2
＋	△	6	3	3
名詞	×	3	15	15
ipadic	○	11	4	4
＋	△	2	7	6
固有名詞	×	7	9	10
neologd	○	4	3	3
＋	△	2	0	0
名詞	×	14	17	17
neologd	○	6	4	4
＋	△	1	1	1
固有名詞	×	13	15	15

を取得できたが，直接議員の発言と関係のない類似単語を取得してしまったためである．また，これらについても人名ほどカテゴリ数は多くないが，同様の傾向があることを確認した．

次に，関連性や関連性があり，理解支援に役立つ類似単語について考察を行う．例を挙げると，「公明党」というエンティティに対して，「憲法 20 条を考える会」という類似単語が得られた．「憲法 20 条を考える会」とは，自由民主党が，当時の細川連立内閣と創価学会の関係を政教一致であると批判するために結成されたものである．これは，「憲法 20 条を考える会」は「自由民主党」の会であるが，属するカテゴリには，「自由民主党」と「公明党」が含まれる．そのため，カテゴリ情報を用いることで，表層上だけではわからない，「公明党」というエンティティに関係のある「憲法 20 条を考える会」という類似単語を，上手く取得することができたと考えられる．

5 おわりに

本研究では，地方議会会議録における読み手の理解支援を目的として，カテゴリ情報を用いることで，議員の発言と類似した単語の取得し，それらの評価を行なった．抽出実験および評価実験の結果，関連性のある単語は得られたものの，有用性のある，あるいは，理解支援に役立つ単語は，少ない結果となった．しかしながら，カテゴリ情報を利用することで，表層上だけではわからない，意味的に関連のある類似単語を上手く取得できたことを確認した．今後は，類似単語のノイズとなる，人名などのエンティティにフィルターをかけることで，意味のある類似単語のみを取得することを検討する予定である．

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．

参考文献

[1] 相澤彰子，古川竜也，相良毅． 言語横断エンティティリンキングのための語義曖昧性解消．情報知識学会誌,24(2):172-177, 2014.

[2] 乾健太郎，松田耕史，岡崎直観．日本語 wikification ツールキット:jawikify． 言語処理学会 23 回年次大会, pp.250-253, 2017.

[3] 鈴木正敏, 松田耕史, 関根聡, 岡崎直観, 乾健太郎. Wikipedia 記事に対する拡張固有表現ラベルの多重付与. 言語処理学会第 22 回年次大会, 2016.

[4] Taku Kudo, Kaoru Yamamoto, Yuji Matsumoto: Applying Conditional Random Fields to Japanese Morphological Analysis． Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, pp.230-237, 2004.

[5] 小澤俊介, 内元清貴, 伝康晴． BCCWJ に基づく中・長単位解析ツール Comainu． 言語処理学会第 20 回年次大会予稿集, pp. 582585, 2014.