E-033

日本語 WordNet と Wikidata 語彙の相互補完検証
Complementary Verification of Vocabulary between Japanese WordNet and Wikdiata

米持 幸寿† 大場 みち子‡
Yukihisa Yonemochi Michiko Oba

1. はじめに

近年，スマートスピーカーやチャットボットの普及により人間が自由形式で入力する文の解析の需要が高まっている．その際，アプリケーション機能に関連する語の抽出は重要なタスクである．語の抽出には機械学習が用いられることが増えてきているが，未知語の処理は重要な課題の一つである．自由形式入力においてユーザーが非常に幅広い語彙を持っていることにより，アプリケーションシステムが想定している範囲外の語が入力されてくることがある．その際，語の抽出に失敗すると対話破綻というシステムエラーを引き起こす．この問題に対応するためにシステムに，一般常識として言語資源を取り込む試みがされている[1].

本研究では，完全な言語資源は現存せず，一つの言語資源を取り込んだシステムでは語彙が不足していると考える．複数の言語資源を統合することが語彙増強に有用と仮説を立て，シソーラスの日本語 WordNet および Wikidata の語彙集合を比較することで補完関係にあることを検証する．

2. 関連研究

IT システムが広範囲な語彙を獲得する資源としてシソーラスや Linked Open Data(LoD)の活用が考えられる．テキストからの語の抽出を，自然言語処理では固有表現抽出といい，シソーラス WordNet と Wikidata を統合する研究がある[2]．しかし，どのように補完関係にあるか示されていない．

本研究では WordNet と Wikidata には登録語彙は補完関係にあると仮説する．補完関係に関する先行研究は調査した限りでは見当たらない．そこで独自に特定のテーマについて語彙を網羅的に調査し，傾向などをまとめる．

3. 対象言語リソース

本研究では入手しやすい言語リソースとして日本語 WordNet 1.1[3]と Wikidata[4]を利用する。本章でそれらを簡単に説明する．

3.1 日本語 WordNet 1.1

日本語 WordNet はシソーラスデータベースの一つである．シソーラスとは類語辞書，あるいは対語辞書と呼ばれる辞典の一種で，もともとは人が読むための書物である．プリンストン大学の心理学研究において最初の Princeton WordNet[3] が作られ．のちに NICT によって，WordNet に日本語を追加する形で日本語 WordNet が作られた．現存する最新の日本語 WordNet はバージョン 1.1 で，ベースになっている Princeton WordNet のバージョンは 3.0 である．

登録語彙数は以下の通りである．

- 57,238 概念 (synset 数)
- 93,834 words 語
- 158,058 語義 (synset と単語のペア)
- 135,692 定義文
- 48,276 例文

†公立はこだて未来大学大学院 Future University Hakodate Graduate School
‡公立はこだて未来大学　Future University Hakodate

3.2 Wikidata

Wikidata は WikiPedia などのデータをもとにしたナリッジデータであり，RDF 形式の SPARQL エンドポイント[6]としてアクセスすることができる．

データは日々追加更新されているが，2021 年 4 月 8 日時点でのダンプデータから数えた日本語ラベルの個数は 2,525,494 個である．

3.3 データ構造

図 1 は双方のデータ構造のうち今回関係するものを簡略化した構造図である．書かれているデータはダミーであり，実際のものではない．

WordNet は概念構造を Synset というレコードで管理しており，そこには Synset-id という識別子がある．Synset は互いにいくつかの関係をもつが，上位概念を hypernym，下位概念を hyponym と呼ぶ．語表現は word というレコードで別に管理されており，概念と語を Link というレコードで結びつける．Synset と word はゼロないし多対多の関係を持っている．

Wikidata は概念を Item として管理しており Qnnnn 形式の．識別子がある．Item 同士はいくつかのプロパティ関係で接続される．本研究では instance of(P31)，subclass of(P171)，parent taxon(P279)を使う．語は label あるいは altLabel というプロパティの値として保持されている．

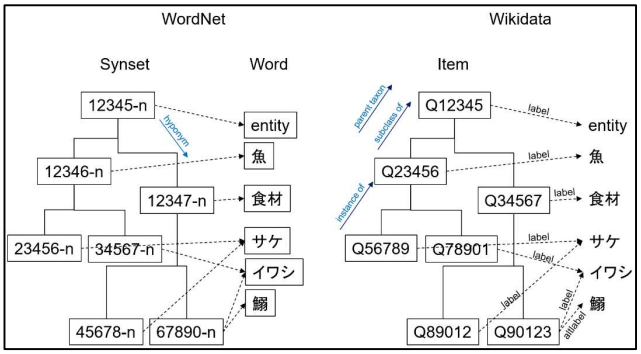

図 1　WordNet と Wikidata の概念構造

4. 検証手順

日本語 WordNet 1.1 は sqlite のデータベースファイルとして配布されており，いくつかの API ライブラリが存在する．今回は JAWJAW[7]を使い，Java によってアクセスする．Wikidata は SPARQL エンドポイントのため HTTP で Web 経由でアクセスできる．Apache Jena[8]を使い Java から呼び出すことで利用する．

全レコードを比較対象とすることもできるが，数量が膨大なうえ，一つの語表現が他分野にわたって異義語として登録されているものを処理することは処理が複雑になりすぎる．本研究では分野を絞って比較することとする．対象は「料理」「食材」「魚介類」の名詞とする．語の検索には上位・下位概念関係を使う．魚介類は生物分類から追跡するが，ひとつの概念から追跡することが困難なため，いくつかのグループにわける．表 1 に WordNet，Wikidata それぞれで利用する Synset および Item の識別子を示す．

表 1　語の検索に利用する識別子

	WordNet	Wikidata
魚類	01473806-n	Q127282
頭足類（イカ,タコ等）	01971094-n	Q128257
二枚貝（ホタテ等）	01955933-n	Q25368
腹足綱（サザエ,タニシ等）	01942177-n	Q4867740
棘皮動物（ナマコ等）	02316707-n	Q44631
甲殻類（エビ,カニ等）	01974773-n	Q25364
料理（カレーライス等）	07557434-n	Q746549
食べ物、食材	07555863-n	Q25403900

日本語 WordNet および Wikidata では英語のラベルも入手できるが，今回は日本語ラベルのみを対象として集計する．

5. 結果と考察

集計結果を表 2 に示す．上から発見語数の多い順に並べてある．左右に WordNet と Wikidata の語数を示す．*Synset* は WordNet で見つかる概念の数で，N はそれに紐づく日本語のラベルの数，oNはそのうち WordNet にのみ見つかる語の数を示す．同様に，*Item* は Wikidate で見つかる概念の数，Dはそれに紐づくラベルの数，oD はそのうち Wikidata にのみ見つかる語の数を示す．C は双方に見つかる語の個数を示す．式に表すと以下のようになる。

N = WordNet に登録されている語彙集合
D = Wikidata に登録されている語彙集合

$$C = N \cap D$$
$$oN = N \cap \overline{D}$$
$$oD = \overline{N} \cap D$$

表 2　WordNet と Wikidata の語彙数

	WordNet				Wikidata		
	Synset	*N*	*oN*	*C*	*Item*	*D*	*oD*
料理	316	185	87	98	2422	4120	4022
食べ物	1121	926	680	246	1527	2867	2621
魚類	633	220	116	104	1612	2256	2152
甲殻類	66	36	15	21	421	642	621
腹足綱	34	20	10	10	333	458	448
二枚貝	35	16	7	9	201	341	332
頭足類	7	3	0	3	144	198	195
棘皮動物	14	11	3	8	77	107	99

獲得できる語彙 v は次の式で表すことができる。

$$v = N \cap D = oN + C + oD$$

これらの傾向を確認するため oN，C，oD を可視化したグラフを図 2 に示す．

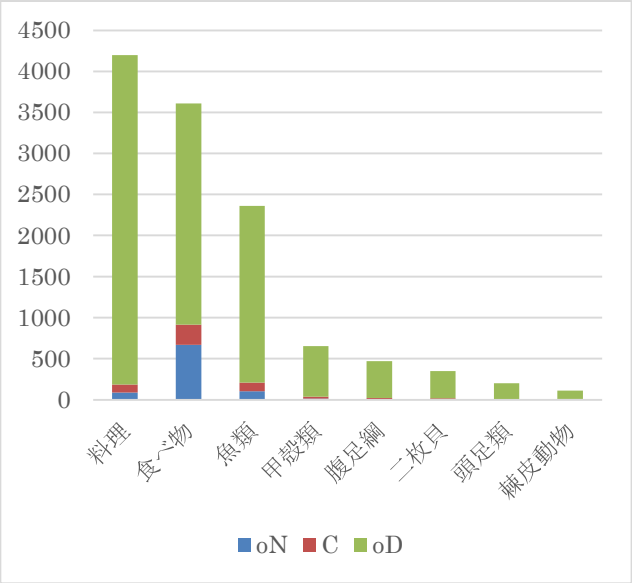

図 2　語彙獲得数

すべての項目で WordNet が少ないながらも語彙を補完できることがわかる．料理の名称は圧倒的に Wikidata が多いが，食材の名称では WordNet からの語彙補完量が多い．これは，Wikidata で「これは食材である」という関係プロパティの登録量が少ないことを表している．魚類全般に対して甲殻類，腹足類，二枚貝，頭足類，棘皮動物は登録量が少ないが，そもそも生物の数として少ないと推察できる．

6. おわりに

本研究にて，WordNet と Wikidata の語彙には相互補完性があり，統合して使うことに意義があることを示した．

今回は料理，食材、魚介類の名詞のみを扱ったが，より広範囲に登録されている語の全体像を把握する取り組みを今後行いたい．英語のラベルでも同様の検証は可能なため，英語による集計も試みたい．

参考文献

[1] Wasi, Sheeban, Madhurendra Sachan, and Manuj Darbari, “Document Classification Using Wikidata Properties”, Information and Communication Technology for Sustainable Development. Springer, Singapore, 729-737 (2020)
[2] McCrae, John Philip, and David Cillessen, “Towards a Linking between WordNet and Wikidata”, Proceedings of the 11th Global Wordnet Conference (2021)
[3] Miller, George A, “WordNet: a lexical database for English”, Communications of the ACM 38.11, pp39-41 (1995)
[4] Francis Bond, Hitoshi Isahara, Sanae Fujita, Kiyotaka Uchimoto, Takayuki Kuribayashi and Kyoko Kanzaki, “Enhancing the Japanese WordNet”, 7th Workshop on Asian Language Resources, pp. 1-8 (2009)
[5] Erxleben, Fredo, et al, “Introducing Wikidata to the linked data web”, International semantic web conference, Springer, Cham (2014)
[6] w3c, “SPARQL Query Language for RDF”, https://www.w3.org/TR/rdf-sparql-query/, (2008)
[7] Hideki Shima, “JAWJAW: Java Wrapper for Japanese WordNet”, http://www.cs.cmu.edu/~hideki/software/jawjaw/index.html(2013)
[8] Apache, “Apache Jena”, https://jena.apache.org/, (2009)