拡張固有表現定義の更新と日本語 Wikipedia 分類データ 2019

関根聡[1)]　安藤まや[2)]　小林暁雄[1)]　隅田飛鳥[1)]

1）理研 AIP　2）ランゲージ・クラフト

{satoshi.sekine, akio.kobayashi, asuka.sumida}@riken.jp, ando@languagecraft.com

1. はじめに

質問応答システムや機械翻訳等、高度な自然言語処理のためには大規模な知識が必要である。特に固有表現の整備は重要で、これまでも CYC、DBpedia、YAGO、Freebase、Wikidata 等が計算機利用を目的として構築されてきたが、規模や知識体系の首尾一貫性といった問題を抱えている[関根ら 17]。そこで「拡張固有表現」[Sekine 08]に即して Wikipedia の項目を分類し、「拡張固有表現」で定義されている属性情報を抽出し、計算機に利用可能な大規模データを構築する Wikipedia 構造化プロジェクト「森羅」を実行している[小林ら 19]。その過程で、百科事典や新聞記事をもとに定義された「拡張固有表現」を Wikipedia により即したものに更新していった。本稿では「拡張固有表現」のバージョンアップと、それを用いた Wikipedia の項目分類について述べる。分類対象としたのは Wikipedia の 2019 年 1 月 20 日のデータで、マイナーな項目やメタページを除いた 920,444 件である。効率的なデータ作成を目指し、機械学習と人手によるアノテートを併用した。さらに、現在実行中である Wikipedia の分類を 30 か国語に展開するタスクについても紹介する。

2. 拡張固有表現

2.1. 拡張固有表現とは

「拡張固有表現」とは、[Sekine 08]によって定義された固有表現に関する定義であり階層構造を持つ。「人名」、「地名」、「組織名」のみならず、「イベント名」、「地位職業名」、「芸術作品名」などを含む。また、例えば「地名」には「河川名」等の「地形名」や、「星座名」等の「天体名」を含む等、幅広い種類の下位カテゴリが含まれる。バージョン 8.0 は最大 3 階層で、219 種類の「拡張固有表現」が定義されている。

2.2. ENE Version 8.0 とは

これまでの「拡張固有表現定義書」は百科事典、新聞記事を対象とした質問応答システムや WordNet 等のオントロジーを参考に構築されてきたが、今回の更新

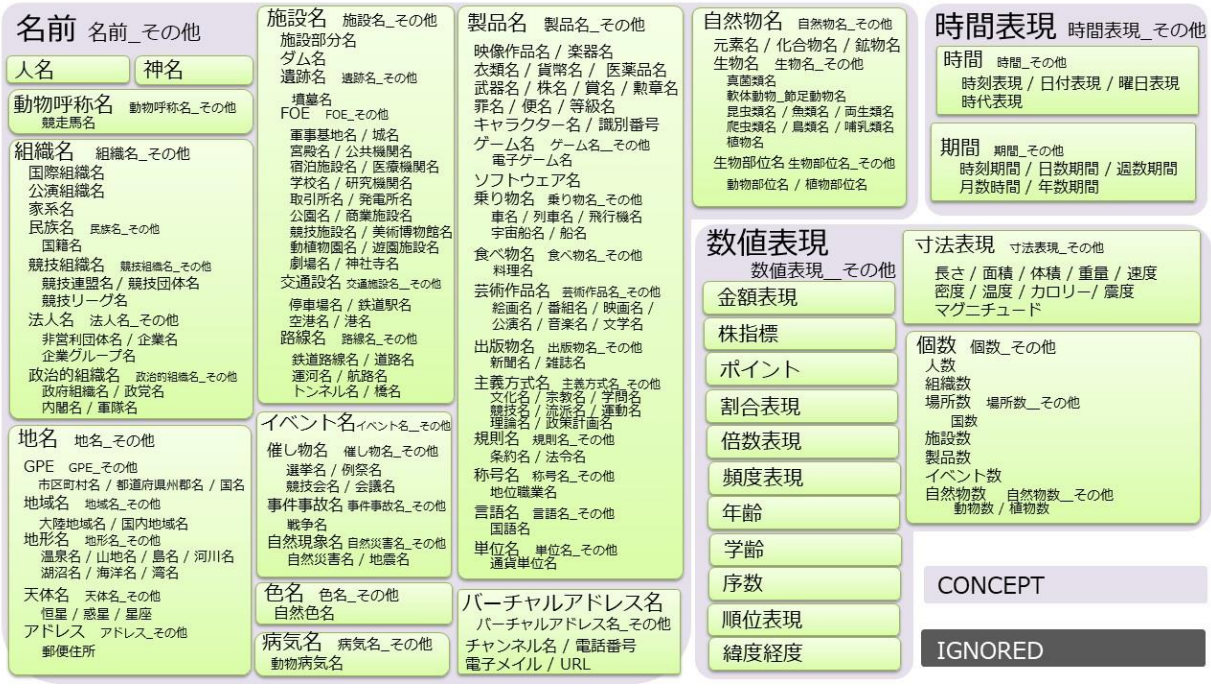

図 1 拡張固有表現の階層 バージョン 8.0

ではより Wikipedia に即した定義となっている。Wikipedia は情報の更新頻度が高く、一般的に幅広く利用されているためである。拡張固有表現の旧定義書をもとに Wikipedia の項目を分類したところ、3 つの問題点が発見された。第 1 に、本来定義しきれない語が分類されるはずの「*__その他」というカテゴリに多くの項目が分類されたこと、第 2 に項目の分類の過程で、判断に迷うカテゴリが存在したこと、第 3 に Wikipedia 特有のメタページなど拡張固有表現に分類できないページが存在したことである。Wikipedia には、同じような表記の項目が複数ある場合の「曖昧さ回避」、項目に異表記がある場合等の「転送元」、「1930 年の日本公開映画」のように項目になるような言葉が並ぶ「一覧」といったページが存在する。これらは分類対象ではない項目となる。以降、順に述べる。

1)「*__その他」への対応

「拡張固有表現」のカテゴリのうち末端でないカテゴリには「*__その他」というカテゴリが存在する。すべてのカテゴリを定義することは困難であるため、定義しきれないものは「*__その他」となる。例えば、旧バージョンの定義では、「施設名」には「学校名」、「劇場名」、「電車駅名」など、24 のカテゴリがあった。しかしながら Wikipedia の項目として比較的多く見られる「医療機関」や「病院」を分類するカテゴリは存在しなかったため「施設名__その他」に分類されていた。そこで、「医療機関名」というカテゴリを新設した。このようにして、「施設名」以下には 9 カテゴリを新設し、全体では 21 カテゴリを新設した（表 1）。

具体的な手順を述べる。大部分の Wikipedia の第 1 文には項目を定義する名詞類（定義語と呼ぶ）が存在する。例えば、「東京大学医学部附属病院は、東京都文京区本郷七丁目にある東京大学医学部附属の大学病院である」の場合は、「大学病院」が定義語となる。これらの定義語をパターンにより抽出した結果 82%の項目から定義語を得ることができた。「医療機関」や「病院」等の類似性も考慮の上、特定の意味の定義語で説明される項目が 200 語以上存在した場合には、カテゴリを新設することとした。

表 1 新設カテゴリ

施設名	製品名
ダム名	映像作品名
軍事基地名	楽器名
城名	ゲーム名__その他
宮殿名	電子ゲーム名
宿泊施設名	ソフトウェア名
医療機関名	バーチャルアドレス名__その他
発電所名	チャンネル名
商業施設名	**イベント名**
交通施設名__その他	選挙名
動物呼称名	**組織名**
動物呼称名__その他	競技連盟名
競走馬名	非営利団体名

2) 第 1 文に則った定義の変更

Wikipedia 項目の分類作業において、判断に迷うことが生じた。これまでも同じ属性をもつ項目が同じカテゴリに属すという方針のもと、設計してきたが、より Wikipedia の項目に適切なカテゴリ定義を作成した。例えば「競技組織名」である（表 2）。

Ver.7.1.2 の「競技組織名」には「競技リーグ名」、「プロ競技組織名」、「競技組織名__その他」が含まれていた。しかし、Ver.8.0 では、「競技団体名」、「競技リーグ名」、「競技連盟名」、「競技組織名__その他」となっている。これは 2 つの問題を解決するために改変した。第 1 の問題は、Ver.7.1.2 で Wikipedia の項目を分類する過程で、「プロ競技組織名」に該当するか否かの判断が困難だったことである。Wikipedia の項目が「プロ」の組織か否かが明確に記載されていない項目が多くあった。一方、アマチュアの競技組織は「競技組織名__その他」に分類されていた。プロ、アマチュアに限らず競技団体は同様の属性をもつため、「プロ競技組織名」を廃止し、プロ、アマチュアを問わず分類できる「競技団体名」を作成した。第 2 の問題は「競技組織名__その他」に競技をする団体と、競技組織や選手を統括する団体という、全く異なる属性をもつ項目が分類されていたことである。競技組織や選手を統括する団体も数多く見られたため、「競技連盟名」を新設した。表 3 にはその他のカテゴリ変更を示す。この作業により更新され

表 2 「競技組織名」の改変

Version 7.1.2		Version 8.0
競技組織名__その他	アマチュアの競技団体	競技組織名__その他
	統括団体	→ 競技連盟名
プロ競技組織名	プロの競技団体のみ	→ 競技団体名
競技リーグ名		→ 競技リーグ名

Version 7.1.2	Version 8.0	説明
郡名 都道府県州名	都道府県州郡名	統合
ＧＯＥ__その他	ＦＯＥ__その他	名称変更
古墳名	墳墓名	新しい墓までを含む
電車駅名	鉄道駅名	新交通システム等を含む
電車路線名	鉄道路線名	新交通システム等を含む
材料名	化合物名 製品名__その他	左記カテゴリとの重複分類も多く、線引きも困難なため削除。 左記カテゴリに分類。

た拡張固有表現定義書は、より Wikipedia に即したものとなったと考える。

3) 「IGNORED」と「CONCEPT」

Wikipedia には、「曖昧さ回避」、「転送元」、「一覧」といったページが存在する。「曖昧さ回避」や「転送元」は通常 Wikipedia の中でメタ情報が付与されるものであるが少なからずの数のライターがメタ情報の付与をしておらず、分類対象データに紛れ込んでいる。それらの項目が分類対象とならないことを示すために用意したのが「IGNORED」である。「CONCEPT」は「ペン」や「空港」といったより一般的な意味をもつことば、つまり「拡張固有表現」には該当しない項目を分類するカテゴリである。従来の固有表現の定義ではこのような一般的な概念を示す項目は対象外となっているが、本研究では「CONCPET」というカテゴリを作成しそこに分類する。

3. Wikipedia の分類作業

Wikipedia の項目「拡張固有表現階層 Ver.8.0」に沿って分類した。分類対象は、2019 年 1 月 20 日の日本語 Wikipedia、1,831,059 ページである。曖昧さ回避や転送元ページ、Wikipedia のメタページは除外し、マイナーなページを除くために Wikipedia 内での非リンク数が 5 より大きいものを分類対象とした。対象項目は 920,444 個となった（表 4）。

項目をカテゴリに分類する際の判断には、原則として Wikipedia の本文の第 1 文を使用する。また、ひとつの項目に対して、複数のカテゴリに分類することもできる。例えば小説や映画作品となっている「風と共に去りぬ」は「文学名」と「映画名」に分類される。1 つ以上のカテゴリに分類するか否かの判断は、第 1 文に記載があるか否かとしている。例えば、映画化された記述がページの最後に記載されている場合は「映画名」には分類しない。

表 4 Wikipedia 項目の統計データ

タイプ	被リンク数	データ数
通常ページ	6以上	**920,444**
	5以下	137,447
	合計	*1,057,891*
転送		694,545
一覧		9,105
曖昧さ回避		69,518
メタページの合計		*773,168*
すべてのページの合計		**1,831,059**

大量の項目を分類するのは、労力がかかる。そのため、機械学習による分類およびアノテーターによる分類という 2 つの手法を併用することで効率化を図った。機械学習による分類では、「拡張固有表現階層 Ver.7.1.2」による Wikipedia 分類データと新たに用意した新設・変更カテゴリのデータを用いて残りのページを分類した[Suzuki et al. 18]。この分類では信頼性を表すスコアが付与されており、そのスコアが低いものを中心に人手でのチェックを行うことができる。

分類対象となった 920,444 件のうち、何らかの「拡張固有表現」カテゴリに分類されたデータ数は 856,064 件、IGNORED は 13,360 件、CONCEPT は 51,020 件だった（表 5）。

表 5 分類データの詳細

ラベル	データ数
ENEに分類されたデータ	856,064
IGNORED	*13,360*
CONCEPT	51,020
合計	**920,444**

分類された「拡張固有表現」カテゴリごとの頻度を見ると、「人名」が圧倒的に多く、次に「市区町村名」、「音楽名」、「番組名」、「企業名」と続いている（表 6）。一方、頻度の少ないカテゴリには数値表現が多くみられた（表 7）。「名前」、「数値」、「時間」の 3 つのトップカテゴリのうち、「名前」の直下にある 13 のカテゴリごとに割合を見ると「人名」が依然として最も多く、「製品名」、「施設名」、「地名」、「組織名」と続いている（表 8）。Message Understanding Conference[Grishman and Sundheim, 1996]において発表され、自然言語処理システムで採用されてきた固有表現は「person」、「location」、「organization」、「time」、「date」、「money」、「percentage」の 7 つで構成されているが、ここでは採用されていない「製品名」と「施設名」が Wikipedia においては多く含まれていることがわかる。

表 6 頻度の高いカテゴリ

カテゴリ	頻度	カテゴリ	頻度
人名	269,688	学校名	25,579
市区町村名	49,028	文学名	21,093
音楽名	46,889	映画名	19,381
番組名	33,747	鉄道駅名	18,296
企業名	30,120	競技会名	17,471

表 7 頻度の低いカテゴリ

頻度	ENE数	例
0	26	重量、カロリー、電子メイル
1	10	速度、週数期間、郵便住所
2-5	12	年齢、期間__その他、電話番号
6-10	2	割合表現、曜日表現
11-20	3	時刻表現、称号名__その他、自然色名

表 8 上位カテゴリごとの割合

上位カテゴリ	%	上位カテゴリ	%
人名	30.76	バーチャルアドレス名	0.42
製品名	28.61	動物呼称名	0.40
施設名	14.49	病気名	0.25
地名	10.16	神名	0.15
組織名	8.70	色名	0.02
イベント名	3.67	名前__その他	0.00
自然物名	2.54		

分類データは下記の URL からダウンロード可能である。ライセンスは Wikipedia と同じ CC-BY-SA で公開している。
http://shinra-project.info/shinra_data

4. 30 言語での Wikipedia 項目の分類（SHINRA2020-ML）

現在、日本語の Wikipedia の項目に対しての分類が完了しているが、それを 30 言語に拡張するプロジェクト（SHINRA2020-ML）を遂行中である。英語のみではなく各言語処理においてもこのようなデータの重要性を認識しているためである。Wikipedia の項目を分類し、構造化するプロジェクト「森羅」は、「Resource by Collaborative Contribution」というスローガンのもと、評価型ワークショップとして 2017 年にスタートした。これはシステム評価をしながら、参加システムの結果を集約させ、より良いデータを構築することを目的としている。SHINRA2020-ML は、NTCIR15 のひとつとして実施しているが、参加者には結果を共有するよう求めている（システムの公開は求めていない）。このタスクの教師用データは分類された日本語 Wikipedia と Wikipedia の言語間リンクを用いて自動的に作成される。例えば、200 万項目を有するドイツ語の Wikipedia のうち、316K の項目が日本語 Wikipedia とリンクされている。したがって、ドイツ語の場合は残りの 1,700K を分類するタスクとなる。我々が提供するデータは、①分類された日本語 Wikipedia データ、②2019 年 1 月 20 日の Wikipedia の言語間リンク情報、③2019 年 1 月 20 日付けの 30 か国語の Wikipedia データ、④拡張固有表現の定義である。評価は、日本語とのリンクのない 1,000 件のテストデータで行うが、テストデータを公開することはない。したがって、参加者は 1 か国語、もしくは複数の言語のすべての出力結果を提出する必要があり、さらにその結果の共有をお願いしている。本タスクに関する情報は下記の URL を参照されたい。参加表明は 2020 年 6 月 30 日まで、結果の提出期限は、2020 年 7 月 31 日である。
http://shinra-project.info/shinra2020ml/

5. まとめ

本稿では、Wikipedia に即するように更新された拡張固有表現 Version 8.0 の紹介と、それに則った Wikipedia 項目（920K）の分類について述べた。また、30 か国語の Wikipedia の分類を行う評価型ワークショップについても紹介した。将来的にはワークショップの成果物は公開し、自然言語処理技術の発展のため、多くの人に使ってもらいたいと願っている。

参考文献

[関根ら 17] 関根聡,安藤まや,小林暁雄,乾健太郎「拡張固有表現に基づく Wikipedia 項目の分類と構造化」. 人工知能学会第 43 回セマンティックウェッブとオントロジー研究会(2017)

[Sekine 08] Satoshi Sekine. Extended Named Entity Ontology with Attribute Information. LREC08

[小林ら 19] 小林暁雄, 中山功太, 関根聡「森羅:Wikipedia 構造化プロジェクト 2018 結果の分析と考察」. 言語処理学会第 25 回年次大会(2019)

[Suzuki et al. 18] Suzuki, M, Matsuda, K, Sekine, S, Okazaki, N, and Inui, K. A joint neural model for fine-grained named entity classification of wikipedia articles. IEICE Transactions on Information and Systems, E101.D(1):73-81, 2018.
doi: 10.1587/transinf. 2017SWP0005.

[Grishman and Sundheim, 1996] Grishman, R. and Sundheim, B. (1996). Message Understanding Conference . 6: A Brief History. In Proceedings of the 16th conference on Computational linguistics - Volume 1 (COLING '96), Vol. 1. Association for Computational Linguistics, Stroudsburg, PA, USA, 466-471.