# 森羅タスクと森羅公開データ

関根聡[1] 中山功太[1,2] 隅田飛鳥[1] 渋木英潔[3] 門脇一真[4] 三浦明波[5] 宇佐美佑[6] 安藤まや[7]
[1]理化学研究所 AIP [2]筑波大学 [3]株式会社 BESNA 研究所
[4]株式会社日本総合研究所 [5]株式会社アティード [6]Usami LCC [7]フリー
{satoshi.sekine,kouta.nakayama,asuka.sumida}@riken.jp

## 概要

Wikipedia に書かれている世界知識を計算機が扱えるような形に変換することを目的として、2017 年より Wikipedia を構造化する「森羅」プロジェクトを推進している。本プロジェクトは「協働による知識構築(Resource by Collaborative Contribution)」のスキームに基づき、評価型ワークショップを開催し、参加したシステムの結果を統合してより良い知識にまとめ上げ、それを公開していくことを目指している。森羅 2022 では、日本語の Wikipedia ページを分類し、ページから指定された属性値情報を抽出し、その情報を該当ページにリンクする 3 つのタスクを行なった。また、本論文ではこれまで森羅プロジェクトで構築した全データを概説するとともに、森羅 2023 の計画を紹介する。

## 1 背景と目的

自然言語理解を実現するためには、言語的及び意味的な知識が必要なことは論を待たない。しかし、大規模な知識の作成は非常に膨大なコストがかかり、メンテナンスも難しい問題である。名前を中心とした知識において、クラウドソーシングによって作成されている Wikipedia はコストの面でもメンテナンスの面でもそれ以前の百科事典の概念を一新したが、Wikipedia を構造化知識として活用しようとすると障壁は高い。Wikipedia は人が読んで理解できるように書かれており、計算機が活用できるような形ではないためである。計算機の利用を念頭においた知識ベースには、CYC[11]、DBpedia[12]、YAGO[13]、Freebase[14]、Wikidata[15]などがあるが、それぞれに解決すべき課題がある。特に CYC ではカバレージの問題、他の知識ベースでは、首尾一貫した知識体系に基づいていない構造化の問題がある。この課題を解決するため、私たちは、名前のオントロジー「拡張固有表現」[2][8][9]に Wikipedia 記事を分類し、属性情報を抽出しリンクすることで計算機が利用可能な Wikipedia の構造化を進めている[1][3][4][5][6][7]。

## 2 協働による知識構築

Wikipedia の全データの構造化を人手で行うことはほぼ不可能に近い。特に、日々更新される Wikipedia を対象にしているため、更新作業を考えると現実的ではない。しかし、記事の分類や属性値抽出のような知識構築は様々な機械学習手法によってある程度の精度で自動化できている。そこで、森羅プロジェクトでは多くの機械学習システムが協力し、より精度の高い構造化知識を半自動的に構築することを目標としている。現在の自然言語処理では「評価型ワークショップ」が数多く行われている。これらは既成のタスクに対する機械学習システムの最適化競争の側面はあるが、これを逆に利用し、構造化知識の作成を行う。つまり、運営者側で訓練データとテストデータを用意し、多くのシステムに評価型ワークショップに参加していただく。この時にテストデータを参加者には知らせないことで、参加者には訓練データ以外の全データを対象に結果を出すという仕組みを取り入れる。また、その結果は共有することを事前に約束してもらう。このようにして出力された結果をアンサンブル学習の手法を用いて、より信頼できる知識を作る。また、信頼度の低いものを人手で確認訂正する手法や、出力結果を次の学習時の訓練データにするアクティブ・ラーニング、訓練データの作成とシステムの実行を幾度も繰り返すブートストラッピング手法を取り入れることで、多くの参加者と協力しあって、精度の高いリソース作成を実現していくことを目標としている。本スキームは Resource by Collaborative Contribution (RbCC: 協働による知識構築)と名付けられ、森羅プロジェクトの最も重要な骨格となっている。

# 3 森羅タスク

森羅による Wikipedia の構造化のためには、それぞれのエンティティーの情報構造を規定する必要がある。そのために「拡張固有表現」を利用する。

## 3.1 拡張固有表現

「拡張固有表現」とは、[2][8][9]によって発表された固有表現に関する定義であり、各固有表現のカテゴリーを表す階層構造と、その固有表現が持ちうる属性情報からなる。階層構造は、「人名」、「地名」、「組織名」、「イベント名」、「地位職業名」といった幅広い種類を含む。また、例えば「地名」には「河川名」「湖沼名」等の「地形名」や、「星座名」等の「天体名」を含む等、最大 4 階層の下位カテゴリーが含まれる。一方、属性情報はカテゴリー毎に定義された固有表現が持つ属性である。例えば「人名」カテゴリーでは異表記、本名、別名・旧称、国籍などの属性が定義される一方で、「空港名」のカテゴリーでは、国、母都市、年間利用者数、別名、名前の謂れなどが定義される。最新のバージョン 9.0 は森羅タスクのために Wikipedia に即して調整したものであり、246 種類の末端カテゴリーが定義され、このうち 178 種類のカテゴリーには特有の属性情報も定義されている。階層構造の全カテゴリーと属性定義の抜粋を、付録の図 1 および表 1 に掲載する。すべての定義を含む詳細は[2]を参照されたい。

## 3.2 三種類のタスク

構造化は３つのプロセスに分解でき、評価型ワークショップではそれぞれをタスクとして実施する。

１） 分類
Wikipedia の各ページを拡張固有表現の末端カテゴリーに分類する。例えば「島崎藤村」は人名、彼の作品の「嵐」は「書籍名」などである。

２） 属性値抽出
分類されたページから拡張固有表現定義で規定された属性の値を抽出する。例えば「島崎藤村」の「本名」は「島崎春樹」で、「作品」には上記の「嵐」が含まれる、などである。

３） エンティティーリンキング（リンク）
抽出された属性値に対して、その値を表す Wikipedia のページをリンクする。例えば、上記の「島崎藤村」の「作品」として抽出された「嵐」を該当する書籍の「嵐」のページにリンクする。

## 3.3 2018 から 2022 に行った 9 タスク

森羅の評価型ワークショップは 2018 年に始まり、2022 年まで 9 タスクを実施してきた。表 2 にタスクの一覧を示す。2018 年から 3 年間は、カテゴリー数を拡張しながら属性値抽出タスクを行なった。また、2020 年から 2 年間は構造化知識の多言語化を目指して、アクティブユーザー数の多い 30 言語の Wikipedia のすべてのページを拡張固有表現に分類するタスクを実施した。2021 年にはリンクタスクを実施した。2022 年は日本語 Wikipedia を対象に 3 つのタスクを同時に実施することにより、Wikipedia の構造化を End-to-End で行うことを目指した評価型ワークショップを実施した。

表 2. 5 年間に行なった 9 タスク

| タスク名 | タスク | 言語 | 詳細 |
|---|---|---|---|
| 2018[5] | 属性値抽出 | 日本語 | 5 カテゴリー |
| 2019[6] | 属性値抽出 | 日本語 | 35 カテゴリー |
| 2020-JP | 属性値抽出 | 日本語 | 78 カテゴリー |
| 2020-ML | 分類 | 30 言語 | |
| 2021-LinkJP | リンク | 日本語 | 7 カテゴリー |
| 2021-ML | 分類 | 30 言語 | |
| 2022 | 分類<br>属性値抽出<br>リンク | 日本語 | 全(178)カテゴリー |

# 4 森羅 2022

森羅 2022 では、日本語の Wikipedia を対象に分類、属性値抽出、エンティティーリンキング（リンク）タスクの３つのタスクを実施した。基本的に End-to-End のタスクとし、エラーが含まれた前段階のタスクの出力を基にしたタスク設定とした。したがって、例えば、あるページの分類が「河川名」ではなく「湖沼名」と間違っていたとしても、その属性値抽出において共通の属性、例えば「所在地」の属性を正しく抽出できれば、その結果は正解とすることにした。この目的のため、拡張固有表現のカテゴリー横断の属性定義の共通化を行なった。分類タスクは後述のワークショップを通して多くの参加者を得たが、属性値抽出とリンクタスクは参加者が非常に少ないことが事前に分かっていたため、Wikipedia 全体を対象とした評価は実施せずに終了した。

### 4.1 スケジュールとリーダーボード

森羅 2022 は以下のスケジュールで実施した。

2022 年 5 月 12 日：キックオフミーティング＆
　　データ公開、リーダーボードオープン
2022 年 8 月 4,10 日：ワークショップ（後述）
2022 年 9 月 30 日,10 月 27 日：ワークショップ 2
2022 年 11 月 14 日：実行結果の提出締切
2023 年 1 月 18 日：ワークショップ（報告会）

なお、RbCC によるリソース作成には参加者にシステムの全出力を提出してもらうことが鍵となるが、一方で参加者にとっては全ページを予測するのは計算機資源や処理時間の面から負担となる。そこで、一部のみの提出が可能なリーダーボードを作成し、システムの評価をスムーズに行えるようにした。

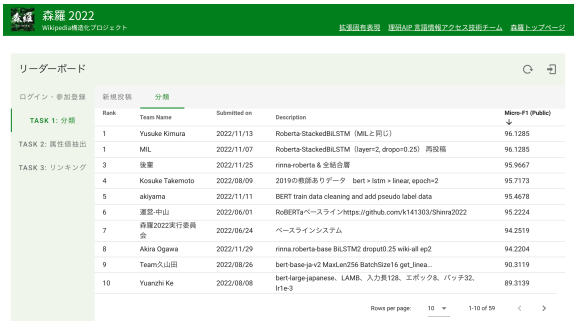


図 2. リーダーボードのスナップショット

### 4.2 ワークショップと参加者

森羅プロジェクトの一環として、深層学習を用いた自然言語処理に取り組みたい学生、研究者、社会人を対象として、サンプルプログラムを提供、解説し、1 週間から 1 ヶ月をおいて参加者が独自に改良する形のワークショップを開催した。1 回目は森羅の分類タスクを題材とし 8 月に行い、登録者は 381 人と大盛況の会となった。2 回目は「属性値抽出」を意識した「固有表現抽出」を題材とし 9〜10 月に行い、登録者は 201 人であった。ワークショップで使用した資料と当日の録画はホームページにて公開している[i]。森羅 2022 ではこのワークショップに関係した参加者が多く、主催者のシステムを含め、分類タスクでは 59 システム、属性値抽出タスク、リンクタスクではそれぞれ 3 システムの提出があった。

## 5 森羅公開データ

5 年間の森羅プロジェクトで作成した学習データ、参加者の出力データなどの大部分を森羅ホームページ[1]にて公開している。

### 5.1 日本言語分類データ

日本語 Wikipedia を拡張固有表現に分類したデータは表 3 の通り、拡張固有表現のバージョンの違い 3 種類がある。拡張固有表現は属性値抽出タスクの進展と共に変化しており、今後使われる方は「2022 分類」データを使われることをお勧めする。

表 3. 森羅日本語分類データ

| データ名 | Wikipedia バージョン | 拡張固有表現バージョン | ページ数 |
|---|---|---|---|
| 2018 分類 | 20190124 | 8.0 | 920,444 |
| 2020 分類 | 20190124 | 8.1 | 920,444 |
| 2022 分類 | 20190124 | 9.0 | 920,444 |

### 5.2 多言語分類データ

2020-ML、2021-ML の多言語分類タスクで利用、作成された 2 種類のデータを公開している。一つは 5.1 で説明した「2020 分類」データと Wikipedia の言語間リンクを辿って作った 30 言語分のトレーニングデータである。表 4 にその統計情報を記載した。公開データには「日本語からのリンク数」の合計の 502 万 9617 ページが含まれている。二つ目のデータは、30 言語の 3255 万 5929 ページに対する 12 システムの出力である。

### 5.3 属性値抽出データ

属性値抽出タスクは、森羅 2018、2019、2020-JP、2022 で実施され、その度にトレーニングデータを 5、30、43、100 カテゴリーずつ増やしてきた。それらのカテゴリーに対して拡張固有表現も改良され、データも整備された。したがって、2020 年に公開されたデータが最も一貫性があり全てを包含しているデータとなっているため、表 5 にそのデータの概要を紹介する。公開データにはそれぞれの Wikipedia 記事が付随しているが、作成時期により Wikipedia バージョンが異なっていることに注意されたい。

[i] http://shinra-project.info/bert_workshop_portal/

表 4. 30 言語の Wikipedia 統計情報　森羅公開データ（2020 多言語分類）

| 言語 | 記事数 | 日本語からのリンク数 | 言語 | 記事数 | 日本語からのリンク数 | 言語 | 記事数 | 日本語からのリンク数 |
|---|---|---|---|---|---|---|---|---|
| アラビア語 | 661,541 | 73,054 | フィンランド語 | 450,896 | 144,750 | ポーランド語 | 1,316,706 | 225,552 |
| ブルガリア語 | 249,797 | 89,017 | フランス語 | 2,075,813 | 318,828 | ポルトガル語 | 1,015,295 | 217,896 |
| カタルーニャ語 | 601,745 | 139,032 | ヘブライ語 | 237,245 | 96,434 | ルーマニア語 | 391,414 | 92,002 |
| チェコ語 | 430,268 | 125,959 | ヒンディー語 | 132,194 | 30,547 | ロシア語 | 1,524,136 | 253,012 |
| デンマーク語 | 242,592 | 86,238 | ハンガリー語 | 443,256 | 120,295 | スウェーデン語 | 3,759,113 | 180,948 |
| ドイツ語 | 2,263,769 | 274,732 | インドネシア語 | 453,624 | 115,643 | タイ語 | 129,693 | 59,791 |
| ギリシア語 | 360,833 | 60,513 | イタリア語 | 1,498,102 | 270,193 | トルコ語 | 325,672 | 111,592 |
| 英語 | 5,793,197 | 439,352 | 朝鮮語 | 440,331 | 190,807 | ウクライナ語 | 882,756 | 167,237 |
| スペイン語 | 1,514,362 | 257,835 | オランダ語 | 1,955,800 | 199,983 | ベトナム語 | 1,200,858 | 116,280 |
| ペルシア語 | 661,323 | 169,053 | ノルウェー語 | 501,699 | 135,935 | 中国語 | 1,041,899 | 267,107 |
| | | | | | | 合計 | 32,555,929 | 5,029,617 |

表 5. 森羅日本語属性値抽出データ（2022 属性）

| | |
|---|---|
| ページ数 | 19,711 |
| 属性種類数 | 1,671 |
| 属性値延数 | 910,567 |
| 1 属性当たりインスタンス数 | 544.92 |
| Wikipedia バージョン | Wikipedia2017(2018,2019,2020-JP 対象の 78 カテゴリー)<br>Wikipedia2019 (2022 に追加された 100 カテゴリー) |

システム出力結果として、森羅 2020-JP タスクに参加した 13 システムの出力結果を公開している。35 カテゴリーの 470,772 ページを対象としたデータであり、336 種類の属性に対して、インスタンス数として 6,089,547 の属性値のデータとなっている。アンサンブル学習などの実験に利用できる。

### 5.4 リンクデータ

リンクタスクは 2021、2022 で実施し、表 6 にある教師、開発データの 3 種類のデータを公開している。

## 6 森羅 2023 タスク

森羅 2023 では、2022 と同様に日本語を対象にした 3 つのタスクを実施する予定である。ただし、2022 で参加者の少なかった属性値抽出とエンティティーリンキングの参加のための敷居を低くすることを主な目的として下記の改良を加えることを考えている。

- 属性値抽出、リンクタスクでは、すべての Wikipedia 記事を対象とするのではなく、特定のカテゴリーなど限定された記事を対象に参加することを可能とする
- 機械学習がより精度高くなるように学習データを拡充する
- 前段階の間違えを含んだデータを対象に End-to-End タスクだけではなく、前段階の正解データに基づいて参加できる仕組みを作る
- 多くの開発データを配布し、参加者が詳細な分析をできるようにする
- データの一貫性、前処理の簡易化を検討する
- 森羅データを利用した応用システムのデモタスクを用意する

## 7 まとめ

Wikipedia の構造化データ「森羅」の作成を目指したプロジェクトを推進している。前述の通りこのプロジェクトは多くの方の協力なくしては進まない。これまでの森羅のタスクにご協力いただいた皆様、特に評価にご参加いただいた全ての団体にはここで感謝を申し上げたい。今後もより深い知識処理を実現するためにも、本プロジェクトに多くのご協力をいただけるようにお願い申し上げたい。

表 6. 森羅日本語リンクデータ

| データ名 | カテゴリー数 | ページ数 | 属性種類数 | リンク元数 | リンク先数 |
|---|---|---|---|---|---|
| 2021 リンク教師データ | 7 | 350 | 83 | 7284 | 7366 |
| 2021 リンク開発データ | 7 | 706 | 85 | 13887 | 13997 |
| 2022 リンク教師データ | 178 | 1397 | 958 | 59429 | 59715 |

## 謝辞

本研究は JSPS 科研費 JP20269633 の助成を受けたものです。

## 参考文献

1. SHINRA-HP. 森羅プロジェクトホームページ: http://shinra-project.info.
2. ENE-HP. 拡張固有表現ホームページ: http://ene-project.info.

3. 関根聡, 小林暁雄, 安藤まや, 馬場雪乃, 乾健太郎. Wikipedia 構造化データ「森羅」構築に向けて.言語処理学会第 24 回年次大会(2018)
4. Satoshi Sekine, Akio Kobayashi, and Kouta Nakayama. 2019. SHINRA: Structuring Wikipedia by Collaborative Contribution. In *Proceedings of the 1st conference on the Automatic Knowledge Base Construction AKBC-2019*.
5. 小林暁雄, 関根聡, 安藤まや. Wikipedia 構造化プロジェクト「森羅 2018」言語処理学会第 25 回年次大会(2019)
6. 小林暁雄, 中山功太, 安藤まや, 関根聡. Wikipedia 構造化プロジェクト「森羅 2019」. 言語処理学会第 26 回年次大会(2020)
7. 関根聡, 安藤まや, 小林暁雄, 松田耕史, Duc Nguyen, 鈴木正敏, 乾健太郎 「拡張固有表表現+Wikipedia」データ（2015 年 11 月版 Wikipedia 分類作業完成版）.言語処理学会第 24 回年次大会(2018)
8. 関根聡, 竹内康介. 拡張固有表現オントロジー. 言語処理学会第13回年次大会ワークショップ「言語的オントロジーの構築・連携・利用」(2007)
9. Satoshi Sekine, Kiyoshi Sudo, and Chikashi Nobata (2002). Extended named entity hierarchy. In the Third International Conference on Language Resources and Evaluation (LREC'02).
10. Satoshi Sekine. 2008. Extended Named Entity Ontology with Attribute Information. In *Proceedings of the Sixth Internationa Conference on Language Resource and Evaluation (LREC08).*
11. Douglas B. Lenat. CYC: a large-scale investment in knowledge infrastructure. ACM 38, pp. 32–38.
12. Lehmann, J., Isele, R., Jakob, M., Jentzch, M., Kontokostas, D., Mendes, P.N., Hellman, S., Morsey M., Kleef, P., Auer, S. and Bizer, C. DBpedia – A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. Semantic Web Journal, 6(2) :167—195
13. Farzaneh Mahdisoltani, Joanna Biega, Fabian M. Suchanek. YAGO3: A Knowledge Base from Multilingual Wikipedias. Proceedings of the Conference on Innovative Data Systems Research (CIDR 2015).
14. Bollacker, K., Evans, C., Paritosh, P., Sturge, T. and Taylor, J. Freebase: a collaboratively created graph database for structuring human knowledge. Proc. International conference on Management of data (SIGMOD '08). ACM, pp.1247-1250.
15. Vrandečić, D. and Krötzsch, M. Wikidata: a free collaborative knowledgebase. Commun. ACM57, pp. 78-85

# A 付録


図 1. 拡張固有表現定義（バージョン 9.0）における階層定義

表 1. 拡張固有表現定義（バージョン 9.0）における属性定義（一部抜粋）

| カテゴリ | 属性 |
|---|---|
| 1.1<br>人名 | 異表記，本名，別名・旧称，国籍，地位職業，生年月日，没年月日，時代，所属組織，学歴，誕生地，居住地，没地，死因，作品，受賞歴，参加イベント，師匠，父母，家族，読み |
| 1.2<br>神名 | 読み，正式名称，別名，旧称，信仰大陸地域，信仰国・GPE，信仰地域，宗教・信仰，性別，住処，姿，位，性質(神)，同一神，配偶者，父母，子，兄弟姉妹，眷属，祭祀施設，祭礼，武器(神)，シンボル，聖獣，乗り物(神)，登場作品1，登場作品2，構成する神，名前の謂れ |
| … | |
| 1.6.5.1<br>停車場名 | 読み，正式名称，別名，旧称・前身，種類，国，所在地，開設年，路線(交通施設)，設計者・組織，管理・運営，座標・緯度，座標・経度，敷地面積，名前の謂れ |
| 1.6.5.2<br>鉄道駅名 | 読み，正式名称，別名，旧称・前身，種類，国，所在地，路線(交通施設)，連絡路線，連絡駅，開設年，閉鎖年，設計者・組織，建築様式，所属事業者，座標・緯度，座標・経度，敷地面積，ホーム，1日の平均乗降人員数，1日の平均乗降人員数データの年，1日の平均乗車人員数，1日の平均乗車人員数データの年，年間利用者数，年間利用者数データの年，地位・規模・タイトル，名前の謂れ |
| 1.6.5.3<br>空港名 | 読み，別名，旧称・前身，IATA(空港コード)，ICAO(空港コード)，国，所在地，母都市，近隣空港，開設年，滑走路数，滑走路の長さ，敷地面積，標高(施設)，年間発着数データの年，年間発着数，年間利用者数データの年，年間利用者数，管理・運営，運用時間，座標・緯度，座標・経度，名前の謂れ，名称由来人物の地位職業名，正式名称，種類，航空会社，就航地，設計者・組織，地位・規模 |
| 1.6.5.4<br>港名 | 読み，正式名称，別名，旧称・前身，種類，国，所在地，河川(施設)，湖沼(施設)，湾(施設)，海洋(施設)，開設年，取扱品，水揚げされる魚介類，埠頭，見出し語内の施設，橋・トンネル，航路(施設)，姉妹港，設計者・組織，管理・運営，座標・緯度，座標・経度，泊地面積，陸地面積，年間発着数，年間発着数データの年，年間貨物取扱量，年間貨物取扱量データの年，年間コンテナ数，年間コンテナ数データの年，年間利用者数，年間利用者数データの年，年間総陸揚量，年間総陸量データの年，年間陸揚金額，年間陸揚金額データの年，地位・規模，名前の謂れ，就航船 |