森羅プロジェクト

関根 聡[1] 宇佐美 佑[2] 門脇 一真[3] 三浦 明波[4] 中山 功太[1] 安藤 まや[5]
[1]理化学研究所 AIP [2]Usami LLC [3]株式会社日本総合研究所
[4]株式会社アティード [5]フリー
{satoshi.sekine,kouta.nakayama}@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きた．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た．本稿では 2017 年から行ってきた森羅プロジェクトを総括し，プロジェクト終了時に残す成果を整理して報告する．

1. 背景

自然言語理解の実現に重要となる言語的及び意味的な知識は，その構築に膨大なコストがかかり，メンテナンスが難しい．計算機の利用を念頭においた知識ベースである CYC[1] はその構築手法からカバレッジが十分ではなかった．一方，これまでに Wikipedia をベースに構築されてきた DBpedia[2], YAGO[3], Freebase[4], Wikidata[5]などの知識ベースは，首尾一貫した知識体系に基づいておらず，その活用が容易ではないという問題があった．

このような課題を解決するため，私たちが 2017 年から行ってきた森羅プロジェクトでは，名前のオントロジーである「拡張固有表現」（ENE）[6][7][8]を用いている．この拡張固有表現に基づいて大規模な知識を含む Wikipedia 記事を分類し，属性情報を抽出し，Wikipedia のエンティティーにリンクすることで，計算機に利用可能とするような構造化を進めてきた[9][10][11][12][13][14][15].

また森羅プロジェクトでは，日々更新される Wikipedia を対象にこのような構造化を継続できるよう，半自動化の仕組みの構築も目指してきた．Resource by Collaborative Contribution （RbCC:協働による知識構築）と名付けたこのスキームでは，「評価型ワークショップ」を開催し多くの機械学習システムを募ることで，より精度の高い構造化知識の構築モデルの開発を行ってきた．

本稿ではこれまでの森羅プロジェクトを総括し，プロジェクトの成果を報告する．

2. 成果物

森羅プロジェクトの主な成果物は以下のとおりであり，これらは森羅プロジェクトホームページ[9]にて公開する．

森羅データ（4 章） これまで構築してきた Wikipedia 構造化データである．これには日本語 Wikipedia の全項目を対象にしたカテゴリー分類データ，全カテゴリーのサンプルを対象にした属性値抽出データ，エンティティーリンキングデータ，および 30 言語のカテゴリー分類タスクのデータが含まれる．詳細は 4 章で述べる．

森羅ベースラインシステム（5 章） 3 章で説明するカテゴリー分類，属性値抽出，エンティティーリンキングの各タスクを半自動的に実施するシステムである．これまでに開催した評価型ワークショップにおいて参加者から提出された精度の高いシステムの実装を参考にしたものである．詳細は 5 章で述べる．

森羅データアクセス API（6 章） 森羅データを利用するアプリケーションから，上記の森羅データへアクセスするための API である．ただし，属性値とエンティティーリンキングについては，森羅ベースラインシステムの出力を用いているため，API の返す値の一部には誤りも含まれる．詳細は 6 章で述べる．

その他 アノテーションマニュアル，アノテーションツールについてもプロジェクトの中で構築した．アノテーションマニュアルのうち重要な部分は，拡張固有表現ホームページ[6]において公開する．

3. 森羅タスク

3.1 Wikipedia 構造化のプロセス

森羅プロジェクトでは，以下の 3 プロセスを順に行うことで，固有表現に関する定義である「拡張固有表現」[6][7][8]に基づいた Wikipedia の構造化を行った．

カテゴリー分類　拡張固有表現では，固有表現のカテゴリーとして「人名」，「地名」，「イベント名」といった上位カテゴリーと，例えば「地名」における「河川名」「湖沼名」等の下位カテゴリーが定義される．Wikipedia 構造化の最初のプロセスでは，Wikipedia の各ページを最下位のカテゴリー（246 カテゴリーのいずれか）に分類した．例えば「島崎藤村」は「名前>人名」，彼の作品の「嵐」は「名前>プロダクト名>出版物名>書物名」に分類された．

属性値抽出　分類されたページから拡張固有表現定義でカテゴリーごとに規定された属性（計 1,712 種類）の値を抽出した．例えば拡張固有表現の「人名」カテゴリーでは異表記，本名，別名・旧称，国籍などの 21 属性が定義されるため，「島崎藤村」の「本名」として「島崎春樹」が，また「作品」として「嵐」などが抽出された．

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 のページをリンクした．例えば，上記の「島崎藤村」の「作品」として抽出された「嵐」という文字列は，Wikipedia の「嵐 (小説)」のページ[i]にリンクされた．

3.2 評価型ワークショップの実施

森羅の評価型ワークショップは 2018 年に始まり，2023 年までに表 1 に示す 12 タスクを実施した．

4. 森羅データ

森羅プロジェクトではこれまで日本語を対象にカテゴリー分類，属性値抽出，エンティティーリンキングの 3 タスクを，また日本語以外の 30 言語を対象にカテゴリー分類タスクを実施してきた．表 2 に含まれる言語・タスクが同一のタスクは，データセットの追加や修正を行ったものであり，Wikipedia の構造化を目的とする場合には，それぞれ最新のデータセットを用いれば十分である．すなわち日本語の 3 タスクにおいては森羅 2022/2023 のデータが，また 30 言語のカテゴリー分類タスクにおいては SHINRA 2021-ML のデータがそれぞれ森羅プロジェクトの成果となる．

表 1．6 年間に行なった全 12 タスク

タスク名	タスク	言語	詳細
2018[12]	属性値抽出	日本語	5 カテゴリー
2019[14]	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[15], 2023[9]	分類 属性値抽出 リンク	日本語	全(178)カテゴリー

各タスクのために森羅プロジェクトが構築したデータセットとしては，教師データ，開発データ，テストデータ等があるが，各タスクを実装する上ではこの他に拡張固有表現の定義書や Wikipedia のダンプデータ等も必要となる．Wikipedia のデータは日々更新されているため，森羅プロジェクトでは，データを構築した時点でのダンプデータを XML 形式（WikiDump）および JSON 形式（CirrusSearchDump）で再配布するとともに，各ページを HTML 形式やプレーンテキスト形式に変換したものも配布している．

また，森羅プロジェクトでは RbCC の考え方に基づいたアンサンブルの手法も提案した[16]．RbCC やアンサンブル手法の研究・評価のために，SHINRA 2020-ML のタスク参加者より提出された各システムの出力も森羅プロジェクトの成果として公開している．

公開データの一覧については，付録 A.1 に示す．

4.1 カテゴリー分類タスク（日本語）

日本語のカテゴリー分類タスクでは，20190120 版の Wikipedia 全ページ（920,444 ページ）に拡張固有表現 9.0 のカテゴリー番号を付与し，教師データとして配布した．この分類は，まず 20151123 版の Wikipedia の 22,667 件を対象に人手で実施して予測

[i] Wikipedia には他に自然現象・気象の「嵐」，「嵐 (駆逐艦)」，「嵐 (グループ)」などのページが含まれる。

モデルを構築し，そのモデルの出力全件を人手でチェックしたものである[11]. 開発・テストデータには，20210820 版の Wikipedia ダンプデータを使用し，それぞれ 1,216 件，2,389 件が含まれる．開発データは全件ラベル付きで公開している．テストデータの正解ラベルは非公開としているが，後続の属性値抽出タスクのみに取り組む参加者のために，ベースラインシステム（Micro-F1: 92.6849）の出力を公開している．

4.2 属性値抽出タスク

属性値抽出タスクでは，20171103 版および 20190120 版の Wikipedia のうち計 19,717 ページから，拡張固有表現 9.0 で定義された属性の値を人手で抽出し，教師データとして配布した（ここには延べ 910,567 件の属性の値が含まれる）．教師データの各ページには，正解となる拡張固有表現のカテゴリー番号も付与されている．テストデータには，20210820 版の Wikipedia ダンプデータのうち 2,389 ページを使用したが，リーダーボードではこのうちの 469 ページで評価したスコアを表示した．テストデータの正解ラベルは非公開としているが，後続のエンティティーリンキングタスクのみに取り組む参加者のために，ベースラインシステム（Macro-F1: 44.9441, Micro-F1: 51.5130）の出力を公開している．

4.3 エンティティーリンキングタスク

エンティティーリンキングタスクでは，20171103 版および 20190120 版の Wikipedia のうち計 1,397 ページに含まれる各属性（延べ 59,715 属性）の値について，リンク先のページ ID ならびにリンク種別を人手で付与し，教師データとして配布した．教師データに含まれるページはすべて属性値抽出の教師データにも含まれ，正解となる拡張固有表現のカテゴリーならびに各属性の値も付与されている．テストデータには，20210820 版の Wikipedia ダンプデータを使用し，属性値抽出タスクで用いた 1,994 ページが含まれている．リーダーボードではこのうちの 126 ページで評価したスコアを表示した．テストデータの正解ラベルは非公開としている．

4.4 カテゴリー分類タスク（30 言語）

30 言語のカテゴリー分類タスクでは, 20190120 版の日本語 Wikipedia 全ページを拡張固有表現 8.0 のカテゴリーに分類した結果と, Wikipedia に含まれる言語間リンクの情報を用いた．日本語からリンクのあるページ数は 30 言語の合計で 5,029,617（うち最多の英語は 439,352）であり，これが各言語における教師データとなる. Wikipedia ダンプ全体に含まれる記事数は 30 言語の合計で 32,555,929（うち最多の英語は 5,793,197）であり，このうち教師データに含まれない全ページを予測対象とする．予測対象のうち 6,298 ページは開発データとして正解ラベルを公開している．またリーダーボードでは予測対象の一部で評価したスコアを表示したが，その正解ラベルについては非公開としている．

また，本タスクにおいては SHINRA 2020-ML で提出された一部のシステムの出力も公開している．対象は 7 チームによる 12 システムであるが，システムによっては 30 言語すべてではなく，一部の言語の予測結果のみを提出している場合がある．

5. 森羅ベースラインシステム

Wikipedia の構造化のための 3 つのサブタスクについて，これまでの評価型ワークショップで高い精度が得られたシステムを用いて，森羅ベース LINE システムを構築し公開した．そして，最新版である森羅 2023 の訓練用ラベル付きデータを元に，前段タスクでの正解ラベルが与えられた条件での個別タスク評価と，正解ラベルが全て不明な条件での End-to-End タスク評価を実施した．表 2 に各タスクにおけるベースラインシステムをカテゴリ横断の Micro-F1 スコアとして評価した結果を掲載する．また，以降の小節において，各ベースラインシステムに関する採用手法の基本的な説明と公開情報について記載する．

ベースラインシステムの実装は全てオープンソースソフトウェアとして公開しており，森羅プロジェクトのホームページ[9]にリンクをまとめてある．

表 2．各タスクにおけるベースライン評価

タスク	Micro-F1 スコア [%]	
	個別タスク	End-to-End
分類	95.9283	
属性値抽出	71.2601	68.3038
リンク	76.8072	49.4401

5.1 カテゴリー分類システム

本システムではまず，Word2Vec における単語分

散表現を，Wikipedia のページ名に拡張させた Wiki Entity Vectors [11]によって Wikipedia 日本語全記事データを用いた教師なし学習を行うことで，ページ名の分散表現を獲得し，エンティティベクトルとする．このベクトルを主要な素性とし，各記事に付与されたメタ情報も離散値的素性に変換し，結合して多層パーセプトロンで学習・推論を行う．

本手法における評価データにおける精度は Micro-F1 スコアにおいて 95.93 ポイントとなった．

5.2 属性値抽出システム

本手法は事前学習済みモデルとして日本語データで学習された BERT [17][18]によって，ページのトークナイズ済み HTML データを入力して得られる各トークンに対する文脈情報付き分散表現を取得し，追加の線形変換層で BIO タグ[19]への分類の学習と推論を行う．ただし属性値抽出の場合，固有表現抽出とは異なり同一出現箇所の文字列に対して複数の異なる種類（属性名）の属性が抽出対象となりうる性質があるため，BIO タグの 3 候補に対する分類ではなく，全属性名候補に対して BIO タグの分類確率が得られるよう線形層の次元数を追加している．これによって単一モデルの一度のフィードフォワード計算により，全属性名候補の一括抽出が可能である．

本手法における評価データにおける精度は Micro-F1 スコアにおいて個別タスク設定で 71.26 ポイント，End-to-End 設定で 68.30 ポイントとなった．

5.3 エンティティーリンキングシステム

本システムは現実的な実行速度を考慮して，ヒューリスティックに基づく幾つかのルールによって属性値に対するリンクの有無とリンクページ推定を行う．採用しているルールの例を以下に示す．

- 属性値に対して異なる記事への URL がある場合，そのページをリンクページとする
- 属性値文字列とタイトルが完全一致するページが存在する場合，そのページをリンクページとする
- 全記事に対してリンク文字列とリンク先のペアを取得して集計し，属性値文字列に完全一致するリンク文字列が存在する場合，最も件数の多いリンク先をリンクページとする
- ラベル付きデータに基づき，カテゴリー名と属性名のペアに対する正解リンク先が自己リンクである割合が一定値以上である場合，自己リンクをページリンクとする

本手法における評価データにおける精度は Micro-F1 スコアにおいて個別タスク設定で 76.81 ポイント，End-to-End 設定で 49.44 ポイントとなった．

6. 森羅 API

本プロジェクトにおいて作成した Wikipedia 構造化データを広く利用可能とすることは，自然言語処理の研究及び自然言語処理技術を用いた広範な技術開発を後押しし，本プロジェクトの目指す高度な言語理解システムの普及を促すことに繋がると考えている．そこで，プロジェクトで作成した構造化データを任意に取得可能な API サーバを実装し，公開することとした．現在，作成した API サーバは森羅パブリック API として公開されており，2025 年 3 月まで公開を継続する見込みである．森羅パブリック API で取得可能なデータとしては，2023 年の森羅タスク実施にあたり開発したベースモデル (5 章) を，日本語の Wikipedia を対象としてカテゴリー分類，属性値抽出，エンティティーリンキングを行い作成した構造化データを用いている．

API の利用のためには森羅パブリック API 公開ページ[20]においてアカウントの作成が必要となる．アカウントを作成後，ログインをすることで API を利用するためのアクセストークンが発行される．森羅パブリック API はアクセストークンを用いて，シェルやプログラムからアクセスすることができるようになっている．利用例を付録 A.2 に示す．

森羅パブリック API の利用方法や API の詳細については，森羅パブリック API のドキュメント[20]を参照されたい．

7. まとめ

世界知識の構造化を目指し、Wikipedia の構造化データ作成を行う「森羅」プロジェクトを推進してきた．森羅データ，構造化のためのツール，森羅データアクセス API などを構築し，公開した．これまでの森羅のタスクにご協力いただいた，また評価タスクにご参加いただいた全ての団体，個人の皆様に心から感謝を申し上げたい．今後は，より深い自然言語処理および知識処理を実現するために，本プロジェクトの成果を活用し，この分野の研究がより深化していくことを切に望んでいる．

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. Douglas B. Lenat. CYC: a large-scale investment in knowledge infrastructure. ACM 38, pp. 32–38.
2. 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
3. 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).
4. 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.
5. Vrandečić, D. and Krötzsch, M. Wikidata: a free collaborative knowledgebase. Commun. ACM57, pp. 78-85
6. 拡張固有表現ホームページ: Extended Named Entity - Fine-grained Named Entity Ontology. http://ene-project.info.
7. Satoshi Sekine. 2008. Extended Named Entity Ontology with Attribute Information. In Proceedings of the Sixth Internationa Conference on Language Resource and Evaluation (LREC08).
8. Satoshi Sekine, Kiyoshi Sudo, and Chikashi Nobata (2002). Extended named entity hierarchy. In the Third International Conference on Language Resources and Evaluation (LREC'02).
9. 森羅プロジェクトホームページ: 森羅 SHINRA - Wikipedia 構造化プロジェクト http://shinra-project.info.
10. 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)
11. Masatoshi Suzuki, Koji Matsuda, Satoshi Sekine, Naoki Okazaki, and Kentaro Inui. 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)
12. 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)
13. 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.
14. 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)
15. 関根聡, 中山功太, 隅田飛鳥, 渋木英潔, 門脇一真, 三浦明波, 宇佐美佑, 安藤まや. 森羅タスクと森羅公開データ. 言語処理学会第 29 回年次大会. (2023)
16. 中山功太, 栗田修平, 馬場雪乃, 関根聡. 能動的サンプリングを用いたリソース構築共有タスクにおける予測対象データ削減. 言語処理学会第27回年次大会発表論文集, pp.1187-1192. (2021)
17. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
18. 東北大学 自然言語処理研究グループ, Pretrained Japanese BERT models, https://github.com/cl-tohoku/bert-japanese. (参照 2024-01-09)
19. Erik F. Tjong Kim Sang and Fien De Meulder. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition (CoNLL '03)
20. 森羅パブリック API https://api.shinra-project.info.

A 付録

A.1 森羅プロジェクトで公開するデータの一覧

本プロジェクトで構築・改変したデータ

- 拡張固有表現 ver9.0 定義書（JSON 形式）
- 森羅 2023 分類タスク 教師データ，開発データ，テストデータ
- 森羅 2023 属性値抽出タスク 教師データ, テストデータ (分類タスク ベースラインシステムの出力)
- 森羅 2023 リンキングタスク 教師データ，テストデータ（属性値抽出タスク ベースラインシステムの出力）
- SHINRA 2021-ML 教師データ，開発データ
- SHINRA 2020-ML Wikipedia 言語間リンク情報
- SHINRA 2020-ML 各システムの出力

本プロジェクトで再配布する Wikipedia データ

- 20190120 版 日本語・30 言語 Wikipedia ダンプデータ（XML 形式 : WikiDump，HTML 形式，プレーンテキスト形式）
- 20190121 版 日本語・30 言語 Wikipedia ダンプデータ（JSON 形式 : CirrusSearchDump）
- 20190120 版 日本語 Wikipeida（HTML 形式，プレーンテキスト形式）
- 20210820 版 日本語・30 言語 Wikipedia ダンプデータ（XML 形式 : WikiDump，HTML 形式，プレーンテキスト形式）
- 20210823 版 日本語・30 言語 Wikipedia ダンプデータ（JSON 形式 : CirrusSearchDump）

A.2 森羅パブリック API の利用例

```
$ curl -s -H "Authorization: Bearer $TOKEN" "https://api.shinra-
project.info/pages?wikipedia_page_id=1732163" | jq '.[0] | "title: ¥(.title),
id: ¥(.id)"'
"title: 伊藤計劃, id: b35d70cc-74b9-4a91-8347-f01fdcbdc1d8"
$ curl -s -H "Authorization: Bearer $TOKEN" "https://api.shinra-
project.info/entities?page_id=b35d70cc-74b9-4a91-8347-f01fdcbdc1d8" | jq '.[]
| "¥(.attribute.name): ¥(.html_offset.text)"' | sort -f | uniq -i | head
"作品: Automatic Death:episode 0:No Distance, But Interface."
"作品: From the Nothing, with Love"
"作品: From the Nothing, With Love."
"作品: Heavenscape"
"作品: METAL GEAR SOLID GUNS OF THE PATRIOTS"
"作品: PEACE WALKER"
"作品: The Empire of Corpses"
"作品: The Indifference Engine"
"作品: つぎはぎの王国から"
"作品: セカイ、蛮族、ぼく。"
```

図 1. 森羅パブリック API の利用例