Wikipedia構造化プロジェクト「森羅2019-JP」

小林 暁雄[1], 中山 功太[1,2], 安藤 まや[3], and 関根 聡[1]
[1] 理研 AIP
[2] 豊橋技科大 博士前期課程 情報・知能工学専攻
[3] ランゲージ・クラフト
{akio.kobayashi, kouta.nakayama, satoshi.sekine}@riken.jp
ando@languagecraft.com

1 森羅プロジェクトについて

自然言語理解の実現に向けて、言語的及び意味的な大規模知識ベースを構築する様々なプロジェクトが行われている。特に、クラウドソースによって日々更新・拡張が行われている大規模な百科事典であるWikipediaは、大規模知識ベース構築のための中心的なデータとして取り扱われている。Wikipediaから知識ベースを構築するプロジェクトとしては、DBpedia、YAGO、Freebase、Wikidataなどが行われているが、これらのタスクには、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」[1]（以降、ENE）にWikipedia記事を分類し、ENEに定義されている属性に対応する値を記事中から抽出することで構造化を目指す、森羅プロジェクト（以降、森羅）を実施している。森羅では、多くの異なる種類のシステムが協力することによって、単一のシステムでは実現できない、高精度かつ大規模なリソース構築を行うResource by Collabolative Contribution (RbCC)[2]アプローチを提案している。これを実現するため、森羅では評価型ワークショップを実施し、様々なシステムを募集している。森羅の評価型ワークショップでは、知識ベースの構築段階にあわせて2種類のタスク（分類タスク、構造化タスク）を行う（図1）。日本語Wikipedia記事を対象とした構造化タスクは2018年度から継続して行われており、本稿で解説する「森羅2019-JP」は、その第2回目のタスクである（日本語Wikipedia記事はすでに分類が完了している）。「森羅2019-JP」では、11チームに参加いただき、2018年度タスク「森羅2018-JP」を上回る数のチームに参加をいただいた。また、対象カテゴリを増加することで、日本語Wikipedia記事のうち、約半数を構造化対象とすることができた。

2 日本語Wikipedia構造化タスク

日本語Wikipedia構造化タスクは、Wikipedia項目中から、ENEカテゴリに設定された属性に対応する文字列を検出することを目的とする。森羅では、ENEカテゴリに分類済みの日本語Wikipedia項目のデータを公開している。[1]ENEは最大4階層の上位下位関係からなる階層構造を持ち、末端カテゴリには属性が設定されている。公開中の分類データは、この末端カテゴリに各Wikipedia項目が分類されている。ENEカテゴリ体系は、森羅を進めるに伴い現在段階的にアップデートを行っており[3]、公開中の最新版はバージョン8.0.0[2]となっている。

日本語構造化タスクは2018年度から行われており、森羅2019-JPで2回目の開催となる。森羅2018-JPでは、分類記事数の多いカテゴリを中心とした5カテゴリ（人名、市区町村名、企業名、化合物名、空港名）を対象として評価型ワークショップを開催した。森羅2019-JPでは、残りのカテゴリから、「地名」に関する14カテゴリと「組織名」に関する16カテゴリを新たに対象として評価型ワークショップを開催した。以下にタスク概要を示す。

2.1 森羅2019-JPタスク概要

森羅2019-JPでは、構造化システムを募るにあたり、以下のデータを配布している。

- 各ENEカテゴリに分類された記事のHTMLデータ及びテキストデータ
- 人手による属性値のアノテーションサンプル

参加者には、アノテーションサンプルをトレーニングデータとして使用し、自動で属性値のアノテーションを行うシステムの開発を行っていただいた。知財権の観点から、参加者にはシステムの出力結果のみを提

[1] http://shinra-project.info/download/#ene-jawiki ダウンロードにはアカウント作成が必要。
[2] http://ene-project.info/ene8/

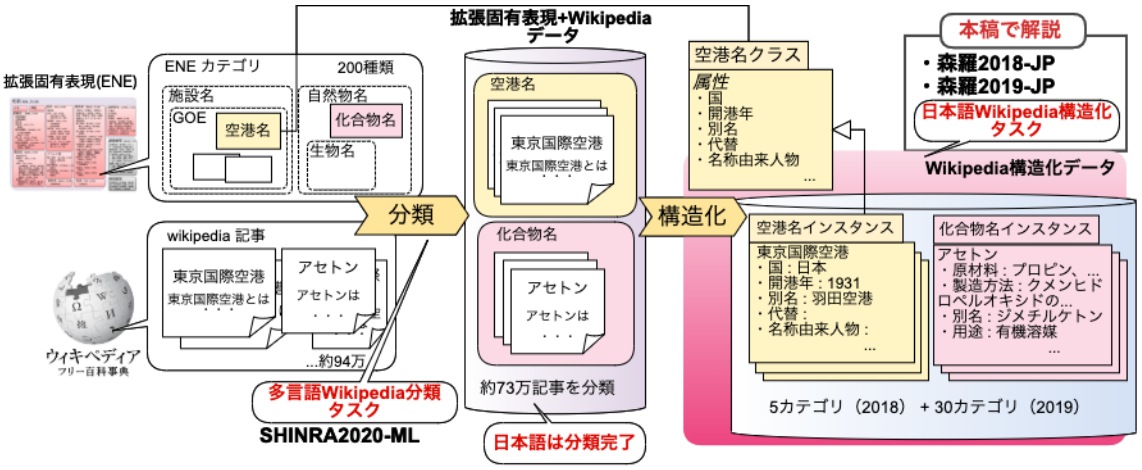

図 1: 森羅全体の流れ

出してもらい、その結果をアンサンブルすることで、RbCC による大規模知識ベースの実現を目指す。

2.2 森羅 2018-JP タスクとの違い

森羅 2019-JP を開催するに当たり、森羅 2018-JP の結果 [4] から、以下の 2 点について主に変更を行った。

1. 属性値となる文字列だけでなく、その記述位置（オフセット）についても出力することを課題とする
2. 対象カテゴリの追加

1. について、森羅 2018-JP では属性値の記述位置については考慮しておらず、記事中に同一表記がある場合に参加者のシステムが正確に学習を行えないという問題があった。例えば「人名」の属性「国籍」に該当する国名は記事中に何度も出現することがある（具体例として、「安倍晋三」の記事では国籍と同じ「日本」という文字列は 244 回出現する (2019/1/12 現在) が、その中で国籍を表すものは 8 件のみである）。これを解消するため、森羅 2019-JP では、タスクを記事中の属性値に該当する文字列の位置情報も取得するタスクに変更を行った。これに伴いトレーニングデータを新たに構築した。

2. について、森羅の最終的な目的は、Wikipedia のすべての記事を構造化するシステムの実現である。しかしながら、森羅 2018-JP の段階では、全体の約 40% の記事しか対象にできていない。前述の通り、ENE は 4 階層の上位下位関係からなる意味体系を持っており、Wikipedia 項目を直接分類している末端カテゴリの多くは何らかの抽象カテゴリの下位概念となっている。これを利用することで、参加者が複数のカテゴリに参加しやすくなるよう、森羅 2019-JP では、抽象カテゴリ下位の兄弟関係にある末端カテゴリをまとめてタスクの対象として追加した。具体的には、「地名」と「組織名」より、以下の 30 カテゴリを対象とした。これらのカテゴリを JP-30 と呼び、森羅 2018-JP と同じカテゴリを JP-5 と呼ぶこととした。

表 1: 日本語 Wikipedia 構造化タスク対象カテゴリの現状

カテゴリ	記事数	比率
人名	269,688	30.76%
市区町村名	49,028	5.73%
企業名	30,120	3.52%
化合物名	5,087	0.59%
空港名	1,562	0.18%
地名（14 カテゴリ）	34,027	4.00%
組織名（16 カテゴリ）	46,133	5.43%
残カテゴリ	420,419	49.79%

地名 GPE__その他, 都道府県州郡名, 国名, 大陸地域名, 国内地域名, 地名__その他, 温泉名, 地形名__その他, 山地名, 島名, 河川名, 湖沼名, 海洋名, 湾名

組織名 組織名その他, 国際組織名, 公演組織名, 家系名, 民族名__その他, 国籍名, 競技団体名, 競技リーグ名, 競技連盟名, 非営利団体名, 企業グループ名, 政治的組織名__その他, 政府組織名, 政党名, 内閣名, 軍隊名

これにより、日本語 Wikipedia 分類データのうち、50% の Wikipedia 記事が構造化の対象となった。それぞれのカテゴリについての記事数と Wikipedia 全体にしめる割合を表 1 に示す。

タスク参加者には、JP-5 については各カテゴリ約 1000 記事、JP-30 については各カテゴリ約 200

記事をトレーニングデータとして配布した。タスクは 2019/4/19 のキックオフミーティングにて公開し、2019/9/10 に結果提出締め切り、2019/10/23 に最終報告会を行い、提出結果全体の報告、及び、参加者からのシステム紹介を行っていただいた。

3 提出結果

森羅 2019-JP では、11 チームに参加頂いた。参加者別の参加カテゴリの一覧を表 2 に示す（2 チームについては、非公開希望などにより結果を割愛）。表中の各スコアはカテゴリごとに全属性について Micro F 値を求めたものである。紙面の都合上、JP-30 は地名、組織名でカテゴリ間の平均スコアを示す（太字は最高スコアのもの）。

参加者数が最大となったカテゴリは 8 チーム参加の JP-5「化合物名」、最小だったカテゴリは 6 チーム参加の JP-5「市区町村名」と JP-30 の各カテゴリとなった。スコアが最大となったカテゴリは JP-5「空港名」の 87.5、最低だったのは JP-30 の「政治組織名__その他」の 47.4 となった。各システムの採用したアプローチを以下に示す。

CRF NRI-UDI, Ricoh, Tanaka, NUT

BERT[5, 6] Toppan, AIP

DrQA[7] Nihon Unisys, OPU

ルールベース TUT

結果としては、BERT ベースの手法と DrQA による手法が多くのカテゴリで最高性能となっていたが、一部カテゴリについては CRF ベースの手法が最高となるなど、手法によって有効なカテゴリが異なる結果となった。BERT ベース、DrQA ベースの手法については、再現率が高い傾向にあった一方、ルールベース、CRF ベースの手法は精度が高い傾向にあった。特に、「座標」などの、地名などのカテゴリの多くに設定されており、かつ記事中でテンプレートを使用して記載されることが多い属性については、CRF ベース手法が最も高い性能となっていた。

4 結果考察

4.1 JP-5 について

森羅 2018-JP の結果との比較を表 3 に示す。全てのカテゴリについて森羅 2018-JP の結果よりスコアが向上していた。個別の属性についてはスコアがあまり改善されていないものも存在しているが、森羅 2018-JP 結果より明らかになった課題である、値が並列表記されているもの [4] については、改善が見られる。これは、オフセット位置の推定も課題に設定したことにより、提案システム側にそれに対応する工夫が導入された（文だけでなく段落まで見るなど）ことによる改善と考えられる。スコアの向上については、特に「人名」カテゴリについて向上が著しい。これについては、ENE を更新するに当たり、トレーニングデータ構築、及び森羅 2018-JP タスクにて問題になっていた属性について設計を見直したことが大きく影響している。具体的には、「代表作」属性を「作品」属性に変更し、記事の人物の代表的作品だけでなく、記事中の全ての作品を対象とすることに変更したことにより、値抽出の判断基準が明確になった影響が大きい。他の属性についても概ね値が向上する傾向にあった。これは、トレーニングデータの増加が影響していると考えられる。森羅 2018-JP からあまりスコアが改善されなかった「化合物名」については、森羅 2018-JP の結果 [4] において、属性値が単語ではなく表現で記載されやすい属性である「製造方法」や「特性」については抽出が難しいことが判明していた。森羅 2019-JP では、各システムがこれに対応できるよう工夫していたにも関わらず、いずれのスコアも森羅 2018 より低下する結果となった。これは、「化合物名」は専門家の協力のもとでトレーニングデータの見直しを行ったため、大幅に属性値が変更されたことが要因と考えられる。

4.2 JP-30 について

スコアは平均がどちらも 58 前後であり、「地名」については 47.9〜73.6、「組織名」については 47.4〜71.0 の間の値となった。トレーニングデータ件数が 200 件程度と少ないため、トレーニングデータ中で頻度の低い属性については、カテゴリで最高スコアを出した手法であっても、一つも値をアノテーションできなかったケースも散見された（「競技連盟名」カテゴリの属性「スローガン」、「家族名」の「解散年」など）。JP-30 のカテゴリは、表 1 にも示したように、各カテゴリの記事数は少なく、これ以上トレーニングデータを増やすことは難しい。このため、これらの低頻度の属性については、機械学習以外の手法が必要と考えられる。ただし、対象カテゴリはそれぞれ兄弟にあるため、カテゴリ間で同名・類似した属性が数多く設定されている（例えば、「港湾名」「湖沼名」の 2 カテゴリでは、それぞれ 32 種類、28 種類設定されている属性のうち、19 種類が同名の属性となっている）。このことから、カテゴリ横断で抽出可能な属性があると考えられる。実際に、「地名」以下のカテゴリの全てに、「地名の謂れ」または「名前の謂れ」という類似した属性が設定

表 2: 提出システム評価結果一覧

カテゴリ	AIP	NUT	Toppan	TUT	NRI-UDI	NihonUnisys	Ricoh	OkaPU	Tanaka
人名	71.8			3.4	54.7	**75.5**		34.7	68.9
企業名	**64.7**		60.8	12.1	38.4	59.5	10.1		52.5
市区町村名	**64.0**		58.5	8.2	54.0	63.3			60.1
化合物名	**47.6**	47.5	45.6		47.0	43.7			46.6
空港名	85.6		77.5	44.5	84.1	**87.5**			82.5
JP-30 地名	57.4		57.4	3.4	48.0	**58.3**			52.9
JP-30 組織名	**57.3**		53.1	2.3	36.2	50.4			48.1

表 3: JP-5 の森羅 2018-JP との最高スコア比較

タスク	人名	企業名	市区町村名	空港名	化合物名
2018	44	53	46	72	47
2019	76	65	64	88	48

されているが、その殆ど（14 カテゴリ中 10 カテゴリ）について、「AIP」チームのシステムが最高スコアとなっていた。これらの属性は、本文中に「〜の由来は」や「〜の名称で」などの表現が付近に出現するといった類似性を持つ一方で、どのカテゴリにおいても 10% 前後の記事にしか記載されていない、スパースな属性であるという特徴がある。このような属性に対し、カテゴリ横断で学習を行う手法が有効であったと考えられる。トップノードである「地名」などの下位全てで共通する属性はあまり種類数はないが、中間ノード以下にしぼって注目すると、末端カテゴリ間で共通する属性の種類数は多くなる。このため、階層毎に手法を切り替えるシステムなどがあれば、より良い結果がえられる可能性がある。あるいは、次年度以降のタスクにて、中間層ごとに共通の属性を対象としたサブタスクなどを行うことで、RbCC にとって有用なデータが構築されると期待される。

5 まとめ

Wikipedia の構造化を目指したプロジェクト「森羅」を推進している。2019 年度はアノテーションタスクへの変更、及び対象カテゴリの追加を行った。ご協力いただいた皆様、特に本年度タスクにご参加頂いた 11 団体にはここで感謝を述べたい。2020 年度は、更に新たなカテゴリを対象として追加した日本語 Wikipedia 構造化タスク「森羅 2020-JP」と、多言語の Wikipedia 分類タスク「SHINRA2020-ML」[3]を実施する。本プロジェクトへのご協力をいただけるようお願いしたい。

[3]NTCIR-15 にて実施（http://http://research.nii.ac.jp/ntcir/ntcir-15/index-ja.html）

参考文献

[1] Satoshi Sekine, Kiyoshi Sudo, and Chikashi Nobata. Extended named entity hierarchy. In *LREC*, 2002.

[2] Kouta Nakayama Satoshi Sekine, Akio Kobayashi. Shinra: Structuring wikipedia by collaborative contribution. In *AKBC*, 2019.

[3] Satoshi Sekine, Maya Ando, Akio Kobayashi, and Asuka Sumida. Extended named entity meets wikipedia: Updated definition of ene and wikipedia categorized by ene. In *LREC*, 2020 (to appear).

[4] 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 発表論文集, pp. 538–541, 2019.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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*. Association for Computational Linguistics, 2019.

[6] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, 2019.

[7] 石井愛. 機械読解による wikipedia からの情報抽出. 言語処理学会 第 25 回年次大会 発表論文集, pp. 438–441, 2019.