様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　６ 年　 ６ 月　 ７ 日現在

機関番号：８２４０１
研究種目：基盤研究(A)（一般）
研究期間：2020～2022
課題番号：２０Ｈ００６１７

研究課題名（和文）信頼される自然言語処理応用システムのための構造化世界知識の協働的貢献による構築

研究課題名（英文）Building structured Knowledge Base for trustable NLP application systems by Resource by Collaborative Construction scheme

研究代表者

関根　聡（SEKINE, SATOSHI）

国立研究開発法人理化学研究所・革新知能統合研究センター・チームリーダー

研究者番号：００８１３２５５

交付決定額（研究期間全体）：（直接経費）　　34,800,000円

研究成果の概要（和文）：森羅プロジェクトの主な成果物は以下の通りであり，森羅プロジェクトホームページにて公開している．
１）構築してきたWikipedia構造化データ．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データを含む森羅データ　２）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森羅ベースラインシステム　３）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アクセスAPI

研究成果の学術的意義や社会的意義
森羅データを含む、本プロジェクトの成果は、自然言語処理において必要不可欠なものであり、生成AIにおけるハルシネーション対応や信頼できる人工知能のための説明できる自然言語処理コンポーネントの中心的なデータとして利用できる。このようなデータは世界的にもユニークな内容であり、このデータの応用は社会において広く活用され、信頼できる人工知能の普及に役立つ。

研究成果の概要（英文）：The main results of the Shinra Project are as follows. These are published on the Shinra Project homepage.
1) Category classification data for all items in Japanese Wikipedia, attribute value extraction data for samples of all categories, entity linking data, and comprehensive data for category classification data in 30 languages. 2) A baseline system that semi-automatically performs each task of category classification, attribute value extraction and entity linking. 3) Access API for Shinra data from applications that use the data.

研究分野：自然言語処理

キーワード：知識グラフ　自然言語処理　情報抽出　固有表現　属性性抽出　テキスト分類　エンティティーリンキング

様　式　Ｃ－１９、Ｆ－１９－１（共通）

１．研究開始当初の背景

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。しかしながら、大規模な知識の作成は非常に膨大なコストがかかり、メンテナンスも非常に難しい問題である。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知百科事典の概念を一新した。しかし、この Wikipedia を自然言語処理のための知識として活用しようと考えると障壁は高い。Wikipedia は人が読んで理解できるように書かれており、計算機が利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、DBpedia、YAGO、Freebase、Wikidata などがあるが、それぞれに解決すべき課題があると考えている。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重要である。

２．研究の目的

上記の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 Wikipedia を構造化する「しんら」プロジェクトを開始した．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た．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．

３．研究の方法

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」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)」のページにリンクされた．

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．

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

4．研究成果

【概要】

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（http://shinra-project.info）にて公開している．

(1) **森羅データ**　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．

(2) **森羅ベースラインシステム**
以下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．

(3) **森羅データアクセス API**
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．

(4) **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において公開している．

【詳細】

(1) 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．

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．

公開データの一覧は(4)の付録に示す．

(1.1)　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．

(1.2)　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．

(1.3) 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．

(1.4) 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．

(2) 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 1. 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

(2.1) 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．

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

(2.2) 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．

(2.3) 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．

(3) 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．

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．

(4) 付録：森羅プロジェクトで公開しているデータの一覧

<u>本プロジェクトで構築・改変したデータ</u>

- 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

<u>本プロジェクトで再配布する Wikipedia データ</u>

- 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）

５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計14件（うち招待講演　0件／うち国際学会　8件）

１．発表者名 1.Satoshi Sekine, Kouta Nakayama, Masako Nomoto, Maya Ando, Asuka Sumida, Koji Matsuda
２．発表標題 Resource of Wikipedias in 31 Languages Categorized into Fine-Grained Named Entities
３．学会等名 COLING 2022（国際学会）
４．発表年 2022年

１．発表者名 2.Satoshi Sekine, Kouta Nakayama, Koji Matsuda, Asuka Sumida, Maya Ando, Yu Usami, Masako Nomoto
２．発表標題 SHINRA2020-ML: Categorizing 30-language Wikipedia into fine-grained NE based on “Resource by Collaborative Contribution” scheme”
３．学会等名 Automated Knowledge Base Construction（国際学会）
４．発表年 2021年

１．発表者名 3.Kouta Nakayama, Yukino Baba, Satoshi Sekine
２．発表標題 Co-Teaching Student-Model through Submission Results of Shared Task
３．学会等名 EMNLP 2021（国際学会）
４．発表年 2021年

１．発表者名 関根聡（理研），中山功太（理研/筑波大），隅田飛鳥（理研），渋木英潔（BESNA），門脇一真（日本総研），三浦明波（アティード），宇佐美佑（Usami LLC），安藤まや（フリー）
２．発表標題 森羅タスクと森羅公開データ
３．学会等名 言語処理学会年次大会
４．発表年 2023年

１．発表者名
関根聡，中山功太，野本昌子（理研），安藤まや（フリー），隅田飛鳥，松田耕史（理研）

２．発表標題
拡張固有表現に分類された31言語のWikipedia知識ベース

３．学会等名
言語処理学会年次大会

４．発表年
2022年

１．発表者名
Sosuke Nishikawa and Ikuya Yamada

２．発表標題
Studio Ousia at the NTCIR-15 SHINRA2020-ML Task

３．学会等名
In Proceedings of the NTCIR-15 Conference（国際学会）

４．発表年
2020年

１．発表者名
Masaharu Yoshioka and Yoshiaki Koitabashi

２．発表標題
HUKB at SHINRA2020-ML task

３．学会等名
In Proceedings of the NTCIR-15 Conference（国際学会）

４．発表年
2020年

１．発表者名
Kouta Nakayama and Satoshi Sekine

２．発表標題
LIAT Team's Wikipedia Classifier at NTCIR-15 SHINRA2020-ML: Classification Task

３．学会等名
In Proceedings of the NTCIR-15 Conference（国際学会）

４．発表年
2020年

１．発表者名
Satoshi Sekine, Masako Nomoto, Kouta Nakayama, Asuka Sumida, Koji Matsuda, and Maya Ando

２．発表標題
Overview of SHINRA2020-ML Task

３．学会等名
In Proceedings of the NTCIR-15 Conference（国際学会）

４．発表年
2020年

１．発表者名
関根聡，野本昌子，中山功太，隅田飛鳥，松田耕史，安藤まや

２．発表標題
SHINRA2020-ML:30 言語の Wikipedia ページの分類

３．学会等名
言語処理学会第27回年次大会

４．発表年
2021年

１．発表者名
中山功太，栗田修平，馬場雪乃，関根聡

２．発表標題
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

３．学会等名
言語処理学会第27回年次大会

４．発表年
2021年

１．発表者名
Satoshi Sekine, Kouta Nakayama, Maya Ando, Yu Usami, Masako Nomoto and Koji Matsuda

２．発表標題
SHINRA2020-ML: Categorizing 30-language Wikipedia into fine-grained NE based on “Resource by Collaborative Contribution” scheme

３．学会等名
3rd conference on the Automated Knowledge Base Construction (AKBC 2021)（国際学会）

４．発表年
2021年

１．発表者名 野本昌子，宇佐美佑，安藤まや，中山功太，関根聡
２．発表標題 森羅2021-LinkJP結果の分析:BERTとルールベースの比較
３．学会等名 言語処理学会第28回年次大会
４．発表年 2022年

１．発表者名 関根聡，中山功太，野本昌子，安藤まや，隅田飛鳥，松田耕史
２．発表標題 拡張固有表現に分類された31言語のWikipedia知識ベース
３．学会等名 言語処理学会第28回年次大会
４．発表年 2022年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関