# 日本語自然言語処理リポジトリ分類データセットの構築

池田大志 樋本一晴 寺中駿人
コニカミノルタ株式会社
{taishi.ikeda, kazuharu.himoto, hayato.teranaka}@konicaminolta.com

## 概要

日本語の自然言語処理に関連する言語資源を集約し，「**日本語自然言語処理リポジトリ分類データセット**」を構築した．本研究では，6752 件の GitHub リポジトリに対して，日本語の自然言語処理に関連するか否かの二値の分類ラベルを付与した．本データセットのタスク難易度を評価するために，テキスト分類による実験を行った．本稿では，本データセットの構築方法，テキスト分類の実験結果，既存研究との差分，今後の展望を述べる．本データセットは，Hugging Face[1] で公開されている．

## 1 はじめに

大規模言語モデルの登場に伴い，自然言語処理研究の重要性が増している．この傾向は，日本語を解析対象とする言語処理の研究開発の活性化に寄与している．例えば，2023 年 5 月より国立情報学研究所が中心となり主催とする LLM 勉強会[2]が開催され，産学連携し日本語大規模言語モデルの構築と，そのモデルの公開[3]が行われている．さらに，企業間においても大規模言語モデルの開発競争が加熱している．例えば，大規模な日本語データセットで学習を行ったモデル [1, 2] や，LLaMA 2 [3] をベースとし日本語データセットで追加学習を行ったモデル [4, 5] が公開されている．

このように新たな大規模言語モデルが次々と登場する中，これらのモデルの言語理解能力を網羅的に評価し，効率的な性能向上を実現するためには，既存の言語資源の有効活用が不可欠である．しかしながら，これまで日本語の自然言語処理に関連する言語資源（ツール，コーパス，辞書，モデル）の情報が十分に集約されていない現状であり，情報の効率性に欠けるという問題が存在していた．

そこで本研究では，日本語の自然言語処理に関連する言語資源の情報集約を行い，日本語自然言語処理リポジトリ分類データセット（awesome-japanese-nlp-classification-dataset）を構築した．このデータセットは，6752 件の GitHub リポジトリに対して，日本語の自然言語処理に関連するか否かの二値の分類ラベルを付与したものである．データセット構築の初期段階では，日本語の自然言語処理に関連する言語資源を人手により収集，構造化した情報を GitHub リポジトリを通じて公開した．ここで公開したリポジトリ情報を正例と位置づけ，対照的な負例を収集することにより，テキスト分類のためのデータセットの構築を行った．

本データセットは，以下の特徴を有する．

- **不均衡なラベル分布**
- **日本語と英語のテキストが混在**
- **主観的な判断に基づくアノテーション**

これらの特性は，実世界おけるテキスト分類タスクが直面する一般的な課題である．そのため，本データセットを利用することで，現実的なテキスト分類タスクを想定し，モデル性能の評価を行うことが可能となる．さらに，本データセットを利用することで，日本語の自然言語処理に関連する言語資源を新たに検出するモデルの作成に活用でき，情報集約の効率化の観点で日本語の自然言語処理研究に貢献できると考える．

## 2 データセット構築

本データセットは，GitHub リポジトリ awesome-japanese-nlp-resources [4] に掲載された情報に基づいて構築されている．この GitHub リポジトリは 2022 年 6 月より情報の掲載を開始した．GitHub で公開されている日本語の自然言語処理に関連するツール，コーパス，辞書，モデルのリポジトリ情報を継続的

1） https://huggingface.co/datasets/taishi-i/awesome-japanese-nlp-classification-dataset
2） https://llm-jp.nii.ac.jp/
3） https://huggingface.co/llm-jp
4） https://github.com/taishi-i/awesome-japanese-nlp-resources

**表 1** 本データセットにおける正例と負例の例を示す．

| | 正例 | 負例 |
|---|---|---|
| ラベル | 1 | 0 |
| 概要 | JGLUE: Japanese General Language Understanding Evaluation for huggingface datasets | Official repository of FaceLit: Neural 3D Relightable Faces (CVPR 2023) |
| URL | https://github.com/shunk031/huggingface-datasets_JGLUE | https://github.com/apple/ml-facelit |
| 公開日 | 2023-02-25 | 2023-04-03 |

に収集し，構造化した情報を GitHub リポジトリを通じて公開している．2024 年 1 月時点で 528 件のリポジトリ情報が掲載されている．

構造化された掲載情報としては，リポジトリ名，URL，リポジトリ概要（Description）が含まれる．ここでのリポジトリ概要とは，リポジトリの目的を簡単に説明するテキストであり，各リポジトリに対して設定されている．実際に掲載されている情報は付録 1 に示す．リポジトリ概要は，英語，日本語，中国語に翻訳されており，多言語で情報を掲載している．さらに，ツールに関しては，ダウンロード数や GitHub スター数の統計情報も掲載している．実際に掲載されている統計情報は付録 2 に示す．

## 2.1 アノテーション手順

本節では，具体的なリポジトリ情報の収集方法と，そのリポジトリに対する分類ラベルのアノテーション手順について説明する．GitHub リポジトリ awesome-japanese-nlp-resources では，Twitter API[5] と GitHub API[6] を利用し，リポジトリ情報を収集した．

Twitter API では，GitHub の URL（https://github.com）を含むツイートを収集した．ここでは，日々のツイートから出現頻度の上位 25 件のリポジトリを抽出し，その中から日本語の自然言語処理に関連するリポジトリの選定を行った．

GitHub API では，Search API を用いて，日本語の自然言語処理に関連するリポジトリを検索し，リポジトリの選定を行った．検索キーワードとして，1995 年から 2015 年にかけて発表された言語処理学会年次大会の論文集のタイトルから，出現頻度の高い単語を取得し，検索に利用した．例えば，「日本語」「言語処理」「形態素解析」といったキーワードを検索に利用した．論文集のタイトルは，anlp-jp-history[7] から取得した．選定したリポジトリとしては，日本語を解析対象とするツール，日本語テキストに対してアノテーションされたコーパス，日本語コーパスで学習されたモデルなどが存在する．

このように『日本語の自然言語処理に関連する』という定義が広く曖昧であるため，明確なアノテーション基準はなく，主観的な判断に基づくリポジトリの選定を行った．選定されたリポジトリは，本データセットの正例とした．負例は，Twitter API で取得した出現頻度が高いリポジトリから，日本語の自然言語処理に関連しないポジトリを選択した．

各リポジトリには，アノテーション情報として，**日本語の自然言語処理に関連するか否かを示すラベル**を付与している．関連する場合はラベル 1，関連しない場合はラベル 0 を付与した．また，**リポジトリ概要**，**URL**，**リポジトリ公開日**も付与されており，これらの情報は，GitHub API を通じて取得した．表 1 は，本データセットの正例と負例の例を示している．ここでの正例は，日本語の言語理解能力を評価するためのデータセットのリポジトリであるため，ラベル 1 を付与している．また，負例は，画像処理関連のリポジトリであるため，ラベル 0 を付与している．

## 2.2 データセットの特徴

表 2 は，本データセットの期間とラベル分布を示している．学習データは，ラベル 0 が 5089 件，ラベル 1 が 407 件の合計 5496 件のリポジトリで構成されている．これらは，2008 年 2 月から 2022 年 9 月に公開されたものである．開発データは，ラベル 0 が 383 件，ラベル 1 が 17 件の合計 400 件のリポジトリで構成されている．これらは，2022 年 10 月から 2022 年 12 月に公開されたものである．評価データは，ラベル 0 が 796 件，ラベル 1 が 60 件の合計 856 件のリポジトリで構成されている．これらは，2023 年 1 月から 2023 年 8 月に公開されたものである．このように，過去のリポジトリ情報から学習する特徴を用いて，新たに公開されるリポジトリに対

5） リポジトリ情報取得時には，Twitte API v1.1 （https://developer.twitter.com/en/docs/twitter-api/v1）を利用した．
6） https://docs.github.com/ja/rest/search/search
7） https://github.com/whym/anlp-jp-history

**表 2** 本データセットにおける GitHub リポジトリの公開日の期間とラベル分布を示す．

| | 公開日の期間 | 0 | 1 | 合計 |
|---|---|---|---|---|
| 学習 | 2008/02 - 2022/09 | 5,089 | 407 | 5,496 |
| 開発 | 2022/10 - 2022/12 | 383 | 17 | 400 |
| 評価 | 2023/01 - 2023/08 | 796 | 60 | 856 |

**表 3** 本データセットにおける言語の割合を示す．

| | 日本語 | 英語 |
|---|---|---|
| 学習 | 10.66% | 89.34% |
| 開発 | 18.00% | 82.00% |
| 評価 | 16.82% | 83.18% |

するラベル予測性能を評価するタスク設定としている．ラベル分布は，ラベル 1 の割合が少ない不均衡な分布であることが特徴である．

表 3 は，本データセットの言語の割合を示している．オープンソースライブラリ langdetect[8] を用いて，概要に対する言語ラベルを付与し，言語判定を行った．この結果から日本語と英語が混在していることがわかる．表 4 は，本データセットの概要に対する文字数の統計情報を示している．この表では，最小文字数，平均文字数，最大文字数を示している．これらの特性は実世界のテキスト分類タスクにおける一般的な課題であり，現実的なモデル性能の評価に利用可能である．

# 3 実験

## 3.1 タスク設定

本データセットのタスク難易度を評価するためにテキスト分類タスクによる実験を行う．テキスト分類のタスク設定としては，リポジトリ概要を入力とし，そのリポジトリが日本語の自然言語処理に関連するか否かのラベルを正しく予測できるか評価する．評価指標として，テキスト分類タスクの一般的な指標である精度（Acc.），適合率（Prec.），再現率（Rec.），F1 スコア（F1）を用いた．これらの指標を用いて，各モデルの性能を比較し，データセットの難易度を示す．

## 3.2 モデル

本節では，実験で採用したモデルについて述べる．学習データを用いて，ファインチューニングを行うモデルとして，BERT [6] を採用する．データセットに英語と日本語の両方が含まれているため，bert-base-multilingual-cased[9] を事前学習モデルに利用した．

次に，大規模言語モデルを用いてゼロショットおよび少数ショットを行うモデルとして，Vicuna [7] を採用した．Vicuna は，ShareGPT [10] から収集された会話データに基づいて，LLaMA 2 [3] をファインチューニングしたモデルである．ここでは，Vicuna の 70 億と 130 億という異なるパラメータサイズによる性能比較を行う．Vicuna を利用する場合のラベル予測は，プロンプトに続く「Yes」と「No」の各トークンの対数尤度を計算し，対数尤度が高いものを予測ラベルとする．プロンプトに利用される少数事例は，学習データから選出され，開発データの F1 スコアを参考に選定した．付録 A.1 には，実験で使用したプロンプトを示す．

最後に，OpenAI API [8] を用いる場合として，gpt-3.5-turbo と gpt-4 を採用した．モデルの入力とするプロンプトは，Vicuna と同様のものを利用した．OpenAI API を利用する場合，「Yes」または「No」を直接出力させ，それを予測ラベルとする．

**表 4** 本データセットにおける概要の文字数の統計情報を示す．

| | 最小 | 平均 | 最大 |
|---|---|---|---|
| 学習 | 2 | 58.05 | 609 |
| 開発 | 8 | 54.33 | 226 |
| 評価 | 3 | 58.85 | 341 |

## 3.3 結果

表 5 に実験結果を示す．ここでは，F1 スコアに関する結果について述べる．参考として，ラベルをランダムに選択した場合，全てのラベルを 0 または 1 と仮定した場合の結果について説明する．ランダムの場合，F1 スコアは 0.12 であった．ラベルを 0 と仮定した場合，F1 スコアは 0.0 であった．ラベルを 1 と仮定した場合，F1 スコアは 0.13 であった．この結果を基準に，各モデルの性能を比較し，データセットの難易度を示す．

最も高い F1 スコアは，BERT [6] をファインチューニングしたモデルと，パラメータサイズが 130 億のモデルで少数ショットで予測を行った Vicuna [7] で，0.74 であった．Vicuna は，2 件の少数事例にもかかわらず，BERT をファインチューニングしたモ

8） https://github.com/Mimino666/langdetect

9） https://huggingface.co/bert-base-multilingual-cased

10） https://sharegpt.com/

**表 5**　本データセットにおける各モデルの実験結果を示す．

| モデル | 設定 | 学習 | Acc. | Prec. | Rec. | F1 |
|---|---|---|---|---|---|---|
| - | ランダム | - | 0.50 | 0.07 | 0.50 | 0.12 |
| - | ラベル 0 | - | 0.93 | 0.00 | 0.00 | 0.00 |
| - | ラベル 1 | - | 0.07 | 0.07 | 1.00 | 0.13 |
| bert-base-multilingual-cased [6] | ファインチューニング | ✓ | 0.97 | 0.79 | 0.70 | **0.74** |
| lmsys/vicuna-7b-v1.5 [7] | ゼロショット | - | 0.29 | 0.09 | 0.98 | 0.16 |
| lmsys/vicuna-7b-v1.5 [7] | 少数ショット（2 件） | - | 0.89 | 0.36 | 0.87 | 0.50 |
| lmsys/vicuna-13b-v1.5 [7] | ゼロショット | - | 0.67 | 0.13 | 0.63 | 0.21 |
| lmsys/vicuna-13b-v1.5 [7] | 少数ショット（2 件） | - | 0.97 | 0.80 | 0.68 | **0.74** |
| gpt-3.5-turbo [8] | ゼロショット | - | 0.88 | 0.31 | 0.47 | 0.40 |
| gpt-3.5-turbo [8] | 少数ショット（2 件） | - | 0.86 | 0.31 | 0.83 | 0.46 |
| gpt-4 [8] | ゼロショット | - | 0.95 | 0.62 | 0.82 | 0.71 |
| gpt-4 [8] | 少数ショット（2 件） | - | 0.94 | 0.57 | 0.82 | 0.67 |

デルと同様の性能を示した．しかし，Vicuna のゼロショット設定で予測を行った場合，ラベルをランダムに選択した場合との差がほとんどなく，正しい予測が行われていないことがわかる．また，その他の結果から，パラメータの大きさやモデルの種類により性能の変化があるため，データセットの難易度はバランスのとれた設定となっていると言える．

この結果から，本データセットはモデルの性能評価に利用可能であると考える．また，日本語の自然言語処理に関連するリポジトリの検出性能（F1 スコア）は 7 割を超えているため，新たな言語資源の検出にも利用可能である．

### 3.4　エラー分析

本節では，BERT [6] をファインチューニングしたモデルのエラー分析結果について述べる．評価データには，ラベル 1 である「日本語の自然言語処理に関連するリポジトリ」が 60 件含まれている．このうち 18 件が予測誤り（偽陰性）であった．例えば，モデルの入力となる概要が「ChatGPT を使用して VRChat 上で会話を可能にするプログラム」である場合，「ChatGPT」という単語が正しく判断されず，予測誤りとなった事例が存在した．さらに，ラベル 0 であるリポジトリ 796 件のうち，11 件が予測誤り（偽陽性）であった．例えば，モデルの入力となる概要が「Exercises on Riemannian geometry (in Japanese)」である場合，「Japanese」という単語に反応し，予測誤りとなった事例が存在した．今回ファインチューニングしたモデル[11]は，Hugging Face で公開しているため，同様のエラー分析を再現することも可能である．

## 4　関連研究

日本語の自然言語処理分野における研究動向に関して，言語処理学会年次大会予稿集を分析対象にした研究が存在する．Murata ら [9] は，年次大会への投稿を行う研究機関や論文タイトルの単語を分析し，研究動向の調査を実施している．また，増田ら [10] はテキストマイニングシステムを構築し，研究動向の調査を実施している．さらに，「日本の言語資源・ツールのカタログ[12]」では，複数の学会の論文集から自動抽出した言語資源の情報が公開されている．awesome-japanese-llm [11] では，GitHub リポジトリを通じて，日本語に関連する大規模言語モデルについての情報が公開されている．

これらの研究は，日本語の自然言語処理における動向を調査し，情報を提供している点で本研究と類似している．しかしながら，本研究は GitHub リポジトリから情報を収集し，これを構造化し，さらに負例を含む分類ラベルを付与してデータセットを構築した点において，既存研究とは異なる新規性を有していると考えられる．

## 5　おわりに

本研究では，日本語の自然言語処理に関連する言語資源の情報を集約し，日本語自然言語処理のためのリポジトリ分類データセットを構築した．今後の展望として，大学の各研究室で公開されている言語資源や，Hugging Face に公開されているリポジトリを収集し，データセットのさらなる充実を図る予定である．付録には，本データセットの利用方法や検索ツールに関する内容も掲載しているので，そちらも参考にしていただけたら幸いである．

11）https://huggingface.co/taishi-i/awesome-japanese-nlp-classification-model

12）https://www.jaist.ac.jp/project/NLP_Portal/doc/LR/lr-cat-j.html

# 参考文献


[1] Meng Lee, Fujiki Nakamura, Makoto Shing, Paul McCann, Takuya Akiba, and Naoki Orii. Japanese stablelm base alpha 7b. https://huggingface.co/stabilityai/japanese-stablelm-base-alpha-7b.

[2] Inc Preferred Networks. Plamo-13b, 2023. https://huggingface.co/pfnet/plamo-13b.

[3] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. **arXiv preprint arXiv:2307.09288**, 2023.

[4] Tianyu Zhao, Akio Kaga, and Kei Sawada. rinna/youri-7b. https://huggingface.co/rinna/youri-7b.

[5] Akira Sasaki, Masato Hirakawa, Shintaro Horie, and Tomoaki Nakamura. Elyza-japanese-llama-2-7b, 2023. https://huggingface.co/elyza/ELYZA-japanese-Llama-2-7b.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. **arXiv preprint arXiv:1810.04805**, 2018.

[7] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. https://lmsys.org/blog/2023-03-30-vicuna/.

[8] OpenAI. Gpt-4 technical report, 2023.

[9] Masaki Murata, Koji Ichii, Qing Ma, Tamotsu Shirado, Toshiyuki Kanamaru, and Hitoshi Isahara. Trend survey on japanese natural language processing studies over the last decade. In **Companion Volume to the Proceedings of Conference including Posters/Demos and tutorial abstracts**, 2005.

[10] 増田勝也, 丹治信, 植松すみれ, 美馬秀樹. 研究動向分析のための論文のデジタルテキスト化とマイニングシステム. 言語処理学会第 20 回年次大会発表論文集, pp. 792–795, 2014.

[11] LLM-jp. Overview of Japanese LLMs, July 2023. https://github.com/llm-jp/awesome-japanese-llm.

# A　付録

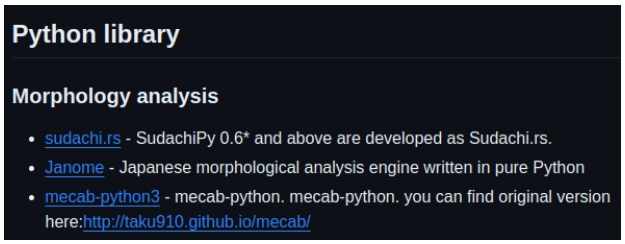


**図 1**　GitHub リポジトリ awesome-japanese-nlp-resources にて掲載されているリポジトリ情報をスクリーンショットした画像を示す．

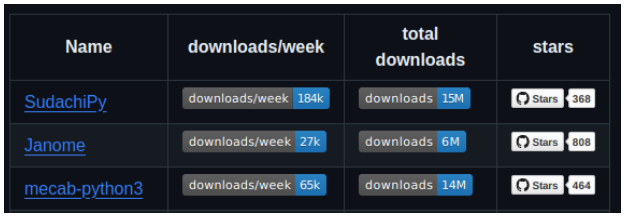


**図 2**　GitHub リポジトリ awesome-japanese-nlp-resources にて掲載されている統計情報をスクリーンショットした画像を示す．


**図 3**　Hugging Face にて日本語の自然言語処理に関連するリポジトリを検索するためのツールを公開している．ここでは，ツールの画面をスクリーンショットした画像を示す．

図 3 の検索ツールは，こちらのページ[13]から利用可能である．

## A.1　プロンプト

ここでは，少数ショットの設定で利用したプロンプトを示す．少数事例は，各ラベルから 1 件ずつ正例と負例を選び，2 件の事例を利用した．予測対象となるリポジトリ概要は，プロンプトの {description} 部分に代入する．ゼロショットの設定の場合は，少数事例の部分を取り除いたプロンプトを利用する．

```
"""
### Question:\nIs this description relevant to
    Japanese natural language processing? Please
     answer Yes or No.\n\n### Description:\nThe
    Symfony PHP framework\n\n### Answer:\nNo\n\n
    ### Question:\nIs this description relevant
    to Japanese natural language processing?
    Please answer Yes or No.\n\n### Description
    :\nGLUE: Japanese General Language
    Understanding Evaluation\n\n### Answer:\nYes
    \n\n### Question:\nIs this description
    relevant to Japanese natural language
    processing? Please answer Yes or No.\n\n###
    Description:\n{description}\n\n### Answer:\n
"""
```

## A.2　データセットの利用方法

本データセットは，Hugging Face 社によって開発された Python ライブラリ datasets[14]から利用可能である．ここでは，データセットを読み込む際のサンプルコードを示す．

```
from datasets import load_dataset

dataset = load_dataset(
    "taishi-i/awesome-japanese-nlp-
    classification-dataset"
)

# Negative example
print(dataset["train"][0])
{'label': 0, 'text': 'HTML Abstraction Markup
    Language - A Markup Haiku', 'url': 'https://
    github.com/haml/haml', 'created_at': '
    2008-02-11T22:55:26Z'}

# Positive example
print(dataset["train"][105])
{'label': 1, 'text': 'The Kyoto Text Analysis
    Toolkit for word segmentation and
    pronunciation estimation, etc.', 'url': '
    https://github.com/neubig/kytea', '
    created_at': '2010-12-22T02:56:51Z'}
```

13）https://huggingface.co/spaces/taishi-i/awesome-japanese-nlp-resources-search

14）https://github.com/huggingface/datasets．*pip install datasets* でインストール可能である．