日本語自然言語処理リポジトリに対する研究分野マルチラベルの付与

池田大志 Quan HoangDanh 長野紘士朗 早田啓介
コニカミノルタ株式会社
{taishi.ikeda, quan.hoangdanh,
koshiro.nagano, keisuke.hayata}@konicaminolta.com

概要

本研究では，日本語自然言語処理に関連する 484 件の GitHub リポジトリを対象とし，合計 1537 件の研究分野マルチラベルを付与したデータセットを構築した．本データセットは，README ファイル，PDF ドキュメント，スクリーンショット画像など，複数形式のマルチモーダルデータを入力とすることでラベルを予測できるように設計されている．実験の結果，本データセットが日本語言語資源の効率的な検索に寄与することを示す．

1 はじめに

大規模言語モデルの普及を背景として，日本語言語資源の開発および公開が活発化している．例えば，「日本語言語資源の構築と利用性の向上ワークショップ[1)]」が継続的に開催されるなど，日本語言語資源の整備に向けた取り組みが進められている．しかし，日本語言語資源の体系的な整理は道半ばであり，特定の研究分野に関連する言語資源を効率的に検索することが難しいという課題がある [1]．また，GitHub 上には多種多様な日本語自然言語処理に関するリポジトリが存在するが，これらがどの研究分野に属し，どのような用途に活用できるかといった情報が明確に整理されていないのが現状である．そのため，研究者や開発者が必要な言語資源を見つけ出すまでに時間を要するケースが少なくない．

そこで本研究では，日本語自然言語処理に関連する 484 件の GitHub リポジトリを対象として，Schopf ら [2] が提案した自然言語処理に関する研究分類体系に基づいて，合計 1537 件の研究分野マルチラベルを付与したデータセットを構築した．先行研究 [1] では，「日本語の自然言語処理に関連するリポジトリを GitHub から抽出する」という二値分類を行っていたが，「機械翻訳」や「対話」など具体的な研究分野名を手がかりに言語資源を検索することは困難であった．そこで，Schopf ら [2] の分類体系を参考に，リポジトリにより細分化されたラベルを付与することで，日本語自然言語処理に関するリポジトリがどの研究分野に属するのかを明確化した．

実験では，キーワード検索とラベル予測の手法を比較し，本データセットを活用したラベル予測の結果が，日本語言語資源の検索性向上に寄与することを示す．本データセットは，Hugging Face[2)]で公開している．

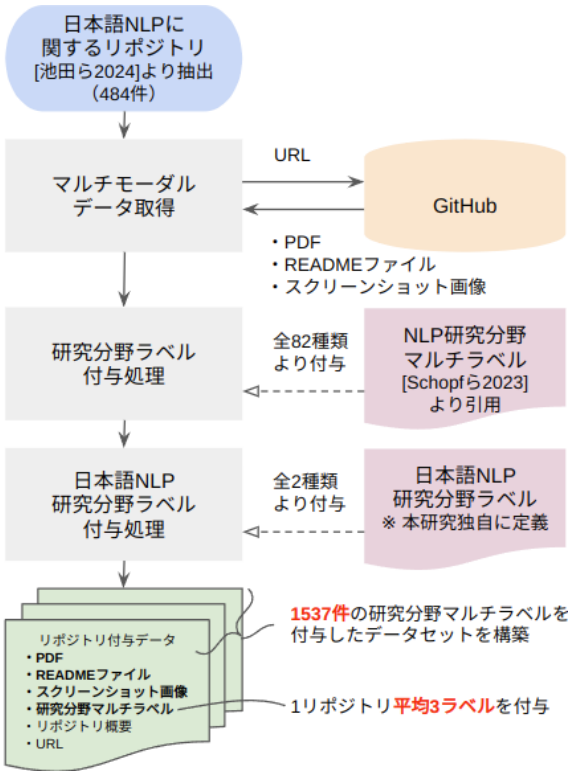

図 1 本データセット構築手順の概要．

1） https://jedworkshop.github.io/JLR2024/

2） https://huggingface.co/datasets/taishi-i/awesome-japanese-nlp-multilabel-dataset

	テキストデータ									画像データ					
	リポジトリ概要			README ファイル			PDF ドキュメント			幅 (ピクセル)			高さ (ピクセル)		
	最小	最大	平均	最小	最大	平均	最小	最大	平均	最小	最大	平均	最小	最大	平均
学習	5	348	57.8	0	81292	4389.4	98	67107	4046.7	1280	1280	1280.0	904	44851	4084.0
開発	12	226	59.5	683	10809	3424.0	884	5164	3006.6	1280	1280	1280.0	1528	7258	3648.9
評価	14	233	66.3	0	94977	7614.4	98	39875	5789.9	1280	1280	1280.0	1123	39111	5850.4

表 1　本データセットに含まれるマルチモーダルデータの統計情報示す．テキストデータ（リポジトリ概要，README ファイル，PDF ドキュメント）の文字数と，画像データ（幅，高さ）のピクセル数について，最小値，最大値，および平均値を示す．

	最小	最大	平均	合計
学習	1	12	3.12	1270
開発	2	5	3.24	55
評価	1	11	3.53	212

表 2　本データセットに付与された 1 リポジトリあたりの研究分野マルチラベル数の最小値，最大値および平均値を示す．合計値は，研究分野マルチラベルの総数を示す．

2　関連研究

日本語言語資源を研究分野ごとに整理・分類する例として，「日本の言語資源・ツールのカタログ」[3] や「自然言語処理の餅屋」[4] が存在する．前者は年次大会の論文を情報源とする自動的な情報抽出が特徴だが，頻繁な更新が難しいという課題がある．後者はウェブ上の言語資源を手動で収集・更新しており，作業が人手に大きく依存するため，未整備の領域が生じやすい可能性がある．また，日本語の大規模言語モデルに特化した「awesome-japanese-llm」[5, 6] も存在するが，大規模言語モデルに焦点を当てているため，汎用的な日本語言語資源を包括的に網羅しているわけではない．これらの例はいずれも有益な情報を提供している一方，GitHub 上に数多く存在する日本語言語資源を包括的に収集・分類し，研究分野ごとに細分化する仕組みは十分に整備されていないのが現状である．そこで本研究では，GitHub 上のリポジトリを対象に，自動的な研究分野マルチラベルを付与する手法を検討し，この課題を解決することを目指す．

3　データセット構築

3.1　日本語言語処理特有の研究分野調査

本研究では，Schopf ら [2] が提案した研究分類体系を参考にデータセットを構築している．Schopf ら [2] は，ACL などの自然言語処理に関する会議の投稿トピックをもとに，82 種類の研究分類体系を作成している（詳細は付録 A.1 参照）．しかし，Schopf ら [2] は英語以外の言語を含む論文を事前に除外しており，日本語特有の研究分野が十分に反映されていない可能性がある．そこで本研究では，言語処理学会の年次大会（1995 年〜2024 年）における口頭発表のセッション名を収集し，これらを Schopf ら [2] の分類体系と比較することで，日本語特有の研究分野を調査した（詳細は付録表 4 参照）．その結果，以下のセッション名に対応する研究分野ラベルが不足していることが判明した．

- **言語資源・コーパス・アノテーション**
- **語彙・辞書**

これら 2 つのセッションは，日本語自然言語処理において研究者や開発者の需要が高く，関連する GitHub リポジトリも多数存在している．そこで，Schopf ら [2] による 82 種類のラベルに加え，本研究独自の 2 種類のラベルとして「**Annotation and Dataset Development**」および「**Vocabulary, Dictionary, and Language Input Method**」を追加し，日本語特有の研究分野に対応した．

3.2　データセット構築手順

本研究では，先行研究 [1] によって「日本語の自然言語処理に関連する」と判定された 484 件の GitHub リポジトリをアノテーション対象とした．そして前節で示した 84 種類のラベル（Schopf らの 82 種類 + 本研究独自の 2 種類）をマルチラベル形式で付与している．データセット構築手順の概要を図 1 に示す．また具体的な構築手順は以下のとおりである．

1. **マルチモーダルデータの取得**：アノテーション対象のリポジトリが属する研究分野ラベルをデータから判断するため，README ファイル，スクリーンショット画像，PDF ドキュメントを取得する（詳細は付録 A.2 参照）．
2. **論文実装の判定**：対象リポジトリが論文の公式実装または再実装であるかを確認する．公式

実装または再実装であり，Schopf らが提供する nlp_taxonomy_data[3] に該当する論文が存在し，研究分野マルチラベルが付与されている場合は，そのラベルを付与して 5 に進む．該当する論文が見つからない場合は 3 に進む．

3. **類似リポジトリとの照合**：すでにアノテーション済みのリポジトリと機能，手法や目的が類似している，あるいは同種のタスクである場合は，同様のマルチラベルを付与して 5 に進む．類似するリポジトリがない場合は 4 に進む．
4. **ラベルの手動付与**：対象リポジトリの概要からキーワードを抽出し，nlp_taxonomy_data に対してキーワード検索を行う．その後，README ファイルやスクリーンショット画像などを参照し，タスク，機能，手法や目的が最も類似している論文のマルチラベルを付与して 5 に進む．類似する論文がない場合も 5 に進む．
5. **日本語言語処理独自ラベルの付与**：対象リポジトリがアノテーション済みのコーパスやテキストデータセットを含む場合は「**Annotation and Dataset Development**」のラベルを追加する．また，形態素解析用の辞書や IME 用の語彙拡充などに関連する場合は「**Vocabulary, Dictionary, and Language Input Method**」のラベルを追加してアノテーションを完了する．

本研究におけるアノテーション作業はクラウドソーシングなどの外部サービスを利用せず，著者の 1 名が実施した．そのため，Schopf ら [2] によるラベルをできる限り再現するよう努めた一方，一部に主観的な判断が含まれている可能性がある点には留意が必要である．

3.3 データセット特徴

表 2 に，本データセットに付与された研究分野マルチラベルの統計を示す．合計で 1537 件の研究分野マルチラベルが付与され，各リポジトリには平均 3 つのラベルが付与されている．付録表 5 では，マルチラベルの頻度分布に関する詳細を示す．さらに，表 1 では，本データセットのマルチモーダルデータに関する統計値を示している．実際に付与されるマルチモーダルデータと研究分野マルチラベルの具体例については，付録 A.3 で説明する．

3） Schopf らが提供するデータセット．論文のタイトル・概要と研究分野マルチラベルが付与されている．https://huggingface.co/datasets/TimSchopf/nlp_taxonomy_data

4 実験

4.1 タスク設定

本研究で構築したデータセットが日本語言語資源の効率的な検索に寄与するかを評価するため，マルチラベル分類タスクを用いた実験を行った．本タスクでは，リポジトリ概要，README ファイル，PDF ドキュメント，スクリーンショット画像の各マルチモーダルデータを入力とし，リポジトリに関連する研究分野マルチラベルを正しく予測できるかを検証する．評価指標としては，マルチラベル分類タスクで一般的に用いられる適合率（Precision），再現率（Recall），F1 スコア（Micro F1）を採用する．これらの指標を用いて以下を観点を比較し，本データセットの有効性を示す．

- キーワード検索とラベル予測に基づく検索手法
- 本研究で構築したデータセットを用いた追加学習（ファインチューニング）の有無
- マルチモーダルデータの活用効果

4.2 分類手法

本研究が構築したデータセットを用いて研究分野マルチラベルを予測するにあたり，以下の手法を比較・評価する．

ランダム予測　精度比較の基準として，ランダムにラベルを付与する手法を導入する．具体的には，各研究分野のラベルを 50% の確率で付与する．

キーワード検索　研究分野マルチラベルのラベル名（例：Text Generation）をクエリとして入力し，そのクエリがテキストデータ（リポジトリ概要，README ファイル，PDF ドキュメント）に含まれている場合に該当ラベルを付与する．ラベル名とテキストデータはいずれも小文字に変換してから検索を行う．

ベースラインモデル　Schopf ら [2] が提案した TimSchopf/nl_taxonomy_classifier[4] をベースラインモデルとする．このモデルは，学術論文の表現に特化した Transformer ベースの allenai/specter2_base[5] [7] を事前学習モデルとして用いたテキスト分類モデルであり，178521 件の論文データセットを用いて学習されている．Schopf

4） https://huggingface.co/TimSchopf/nlp_taxonomy_classifier

5） https://huggingface.co/allenai/specter2_base

分類手法	モデル	入力データ				開発			評価		
		リポジトリ概要	README	PDF	スクリーンショット	Prec.	Rec.	F1	Prec.	Rec.	F1
ランダム予測	-	-	-	-	-	0.034	0.455	0.064	0.042	0.513	0.078
キーワード検索	-	✓	-	-	-	0.000	0.000	0.000	1.000	0.015	0.030
キーワード検索	-	-	✓	-	-	0.600	0.055	0.100	0.414	0.045	0.081
キーワード検索	-	-	-	✓	-	0.500	0.036	0.068	0.343	0.045	0.079
ベースライン	TimSchopf/nlp_taxonomy_classifier	✓	-	-	-	0.538	0.382	0.447	0.630	0.354	0.453
ベースライン	TimSchopf/nlp_taxonomy_classifier	-	✓	-	-	0.538	0.382	0.447	0.503	0.354	0.416
ベースライン	TimSchopf/nlp_taxonomy_classifier	-	-	✓	-	0.532	0.455	0.490	0.414	0.274	0.330
ベースライン	TimSchopf/nlp_taxonomy_classifier	✓	✓	-	-	0.500	0.418	0.455	0.620	0.476	0.539
ベースライン	TimSchopf/nlp_taxonomy_classifier	✓	-	✓	-	0.571	0.509	0.538	0.506	0.387	0.439
ファインチューニング	TimSchopf/nlp_taxonomy_classifier	✓	-	-	-	0.611	0.400	0.484	0.663	0.449	0.536
ファインチューニング	TimSchopf/nlp_taxonomy_classifier	-	✓	-	-	0.538	0.509	0.523	0.566	0.483	0.521
ファインチューニング	TimSchopf/nlp_taxonomy_classifier	-	-	✓	-	0.516	**0.582**	0.547	0.579	0.506	0.540
ファインチューニング	TimSchopf/nlp_taxonomy_classifier	✓	✓	-	-	0.585	0.564	**0.574**	**0.681**	**0.592**	**0.633**
ファインチューニング	TimSchopf/nlp_taxonomy_classifier	✓	-	✓	-	0.564	0.564	0.564	0.591	0.536	0.562
ゼロショット	gpt-4o-2024-08-06	✓	-	-	-	0.560	0.255	0.350	0.476	0.184	0.265
ゼロショット	gpt-4o-2024-08-06	-	✓	-	-	**0.632**	0.218	0.324	0.429	0.184	0.257
ゼロショット	gpt-4o-2024-08-06	-	-	✓	-	0.500	0.182	0.267	0.481	0.184	0.266
ゼロショット	gpt-4o-2024-08-06	-	-	-	✓	0.500	0.200	0.286	0.511	0.222	0.309
少数ショット（5 件）	gpt-4o-2024-08-06	✓	✓	-	-	0.487	0.345	0.404	0.555	0.311	0.399

表 3　開発および評価データセットに対する各分類手法（モデルと入力データの組み合わせ）の精度を示す．

ら [2] の 82 種類の研究分野ラベルを出力するよう設計されているため，リポジトリ概要，README ファイル，PDF ドキュメントを入力し，出力スコアが 0.5 を超えたラベルを予測ラベルとする．入力テキストの長さはモデルの制約により 512 文字に制限した．ベースラインモデルの評価では，学習済みのモデルをそのまま用い，本研究のデータセットによる追加学習は行わない．

ファインチューニング　ベースラインモデル（`TimSchopf/nl_taxonomy_classifier`）を初期モデルとし，本研究で構築したデータセットを用いてファインチューニングを行う．その他の設定はベースラインモデルと同様である．

ゼロショット・少数ショット　大規模言語モデルを用いたゼロショットおよび少数ショット手法について説明する．本研究では OpenAI API [8] を利用し，画像入力に対応する `gpt-4o-2024-08-06` を採用した．入力テキストの長さはベースラインと同じく 512 文字に制限し，画像は前処理なしにそのままモデルに入力する．モデルの出力テキストをそのまま予測ラベルとする．少数ショット (5 件) では追加の事例をモデルに提示することで精度向上を図る．実際に用いたプロンプト詳細は付録 A.4 に示す．

4.3　結果

表 3 に実験結果を示す．まず，キーワード検索の結果について見ると，テキストデータに若干の差異はあるものの，ランダム予測と大差ない低い F1 スコアに留まった．これは，単に研究分野名をキーワードとして検索するだけでは，該当分野のリポジトリを十分に抽出できていないことを示唆する．

次に，ベースラインモデルでは，キーワード検索を上回る精度が得られた．特に複数のテキストを組み合わせることで F1 スコアが向上している．さらに，本研究で構築したデータセットを用いてファインチューニングを行うと，ベースラインを大幅に上回る結果が得られ，リポジトリ概要と README ファイルを併用した場合に最良の F1 スコアを示した．一方，ゼロショット・少数ショット手法は，ベースラインモデルやファインチューニングモデルと比較して精度が劣り，少数ショット（5 件）を付与してもわずかな改善にとどまった．

以上の結果から，本研究が構築したデータセットを用いてベースラインモデルをファインチューニングすることにより，研究分野マルチラベル分類の性能が大きく向上することが示された．これは，単純なキーワード検索では捉えきれない研究分野マルチラベルを，モデル出力で補うことにより，日本語言語資源の検索性向上に寄与することを示唆している．

5　おわりに

本研究では，日本語自然言語処理に関連する GitHub リポジトリを対象に，研究分野マルチラベルを付与したデータセットを構築した．本データセットを用いたラベル付与は，キーワード検索と比較して高い精度を示し，日本語言語資源の検索性向上に寄与することが明らかになった．しかし，さらなるラベル予測性能の向上余地は依然として残されている考え，今後はマルチモーダル情報を総合的に活用する手法や，大規模言語モデルの性能を活かしたマルチラベル分類手法などを検討する予定である．

参考文献

[1] 池田大志, 樋本一晴, 寺中駿人. 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. 言語処理学会第 30 回年次大会発表論文集, pp. 851–856, 2024.

[2] Tim Schopf, Karim Arabi, and Florian Matthes. Exploring the landscape of natural language processing research. In Ruslan Mitkov and Galia Angelova, editors, **Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing**, pp. 1034–1045, Varna, Bulgaria, September 2023. INCOMA Ltd., Shoumen, Bulgaria.

[3] 言語処理学会ポータル. 日本の言語資源・ツールのカタログ, (2025-1 閲覧). https://www.jaist.ac.jp/project/NLP_Portal/doc/LR/lr-cat-j.html.

[4] 山本和英. 自然言語処理の餅屋, (2025-1 閲覧). https://www.jnlp.org/nlp/.

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

[6] Kaito Sugimoto. Exploring Open Large Language Models for the Japanese Language: A Practical Guide. **Jxiv preprint**, 2024.

[7] Amanpreet Singh, Mike D'Arcy, Arman Cohan, Doug Downey, and Sergey Feldman. SciRepEval: A multi-format benchmark for scientific document representations. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, pp. 5548–5566, Singapore, December 2023. Association for Computational Linguistics.

[8] OpenAI. Gpt-4o system card, 2024.

[9] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In Tal Linzen, Grzegorz Chrupa la, and Afra Alishahi, editors, **Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP**, pp. 353–355, Brussels, Belgium, November 2018. Association for Computational Linguistics.

A 付録

言語処理学会セッション名	Schopf による研究分野	頻度
対応関係あり		
機械翻訳	Machine Translation	56
対話	Dialogue Systems & Conversational Agents	31
意味解析	Semantic Text Processing	28
知識獲得・情報抽出	Knowledge Representation, Information Extraction & Text Mining	18
生成	Text Generation	16
要約	Summarization	15
質問応答	Question Answering	13
情報抽出	Information Extraction & Text Mining	13
解析	Syntactic Text Processing	12
言語学	Linguistics & Cognitive NLP	12
構文解析	Syntactic Parsing	11
埋め込み表現	Representation Learning	11
抽出	Term Extraction	10
形態素解析	Morphology	8
対応関係なし		
機械学習	該当なし	20
教育応用	該当なし	12
言語資源・アノテーション	該当なし	12
言語教育と言語処理の接点	該当なし	11
言語資源・コーパス	該当なし	10
語彙・辞書	該当なし	9

表 4 言語処理学会年次大会（1995 年〜2024 年）における口頭発表セッション名と，Schopf らによる研究分野ラベルを比較した結果を示す．頻度は年次大会の口頭発表セッション名に対する開催回数を表す．

A.1 Schopf らによる研究分類体系

Schopf ら [2] は，自然言語処理分野の研究論文を体系的に分類することを目的とし，主要な研究領域を整理している．具体的には，ACL，EMNLP，COLING などの国際会議で扱われるトピックを手作業で精査し，20 名以上の専門家から意見を募ったうえで最終的な研究分類体系を構築した．この分類体系は OWL 形式で公開されており[6]，本研究ではこれを参考にデータセットを構築した．

A.2 マルチモーダルデータの取得方法

スクリーンショット画像および PDF ドキュメントの取得には，`playwright`[7] を用いた．スクリーンショット画像はフルページをキャプチャし，PDF ドキュメントは A4 サイズで取得し，そのほかの設定はデフォルト値を使用している．README ファイルは，対象リポジトリを `git clone` した後，拡張子を手がかりに抽出した．そのため，README ファイルが存在しないリポジトリも一部含まれる．これらのデータは，GitHub の「情報使用の制限」[8] に従い，本研究成果をオープンアクセスとすることを前提に利用した．

A.3 マルチモーダルデータと研究分野マルチラベルの具体例

本研究で構築したデータセットの具体的なサンプルを JSON 形式で示す．この例では，GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding [9] と同様のマルチラベルを付与し，データセットを含むため「Annotation and Dataset Development」を追加ラベルとして付与している．

```
{
    "URL": "https://github.com/shunk031/huggingface-datasets_JGLUE",
    "labels": [
        "Language Models",
        "Low-Resource NLP",
        "Semantic Text Processing",
        "Responsible & Trustworthy NLP",
        "Annotation and Dataset Development",
        "Explainability & Interpretability in NLP"
    ],
    "PDF": "shunk031 / huggingface-datasets_JGLUE\n...",
    "README": "# JunoDB - A secure, consistent and ...",
    "Image": "huggingface-datasets_JGLUE.png",
    "Description": "JGLUE: Japanese General Language Understanding Evaluation for huggingface datasets"
}
```

学習 (2008/02 – 2022/09)	
研究分野ラベル	頻度
Syntactic Text Processing	161
Annotation and Dataset Development	93
Semantic Text Processing	86
Language Models	79
Text Segmentation	77
Tagging	62
Morphology	51
Information Extraction & Text Mining	51
Vocabulary, Dictionary, and Language Input Method	48
Multilinguality	44
開発 (2022/10 – 2022/12)	
研究分野ラベル	頻度
Text Generation	7
Language Models	5
Multimodality	4
Annotation and Dataset Development	4
Natural Language Interfaces	4
Semantic Text Processing	3
Captioning	2
Speech & Audio in NLP	2
Dialogue Systems & Conversational Agents	2
Dialogue Response Generation	2
評価 (2023/01 – 2023/08)	
研究分野ラベル	頻度
Text Generation	28
Language Models	24
Annotation and Dataset Development	18
Dialogue Systems & Conversational Agents	14
Natural Language Interface	13
Dialogue Response Generation	11
Semantic Text Processing	10
Syntactic Text Processing	8
Multimodality	8
Responsible & Trustworthy NLP	6

表 5 各データセットにおける研究分野マルチラベルの頻度分布を示す．

A.4 ゼロショット・少数ショットのプロンプト

本研究の実験で用いたプロンプトを以下に示す．

```
You are an AI assistant specialized in multi-label classification for GitHub repository descriptions. You are given a set of possible labels (label IDs and names) and a single repository description. Your task is to output **only** the numeric label IDs and names that are relevant to the repository, and nothing else.
Important guidelines:
1. Read the repository description carefully.
2. Identify **all** applicable labels from the list.
3. Output **only** label IDs and names, separated by new lines (or commas), with no additional explanation or text.
4. If no labels apply, output nothing (blank).
Here is the complete set of labels: {labels_as_text}
```

6) https://github.com/sebischair/Exploring-NLP-Research/blob/main/NLP-Taxonomy.owl
7) https://github.com/microsoft/playwright
8) https://docs.github.com/ja/site-policy/acceptable-use-policies/github-acceptable-use-policies