Web からのデータ収集と探究事例の紹介

石原祥太郎
人工知能学会企画委員 / 日本経済新聞社
第 5 回 情報 II 全国指導力向上研修会
2025 年 1 月 30 日

自己紹介：石原祥太郎 https://upura.github.io/

- 大学新聞での記者経験を経て、日本経済新聞社へ
- 日経の研究開発部門で日々、情報と情報技術を活用した問題発見・解決の探究に従事
- 社内外での講演や技術書の出版など、培った知見を積極的に共有している

情報 II 第 1-4 章 (S) での学習内容

- S1: 情報社会と情報技術
- S2: コミュニケーションのための情報技術の活用
- S3: データを活用するための情報技術の活用
- S4: コンピュータや情報システムの基本的な仕組みと活用

教員用教材から引用

情報 II 第 5 章「情報と情報技術を活用した問題発見・解決の探究」

> 地域や学校の実態及び生徒の状況に応じて情報と情報技術を活用した問題発見・解決の探究を通して，情報の科学的な見方・考え方を働かせて，情報と情報技術を適切かつ効果的に活用するための知識及び技能の深化・総合化，思考力，判断力，表現力等の向上を図る。数学科など他教科とも積極的に連携を図る。

教員用教材から引用

第 5 章は、S1-4 の学習内容を踏まえた実践 (探究活動)

- S1: 情報社会と情報技術
- S2: コミュニケーションのための情報技術の活用
- S3: データを活用するための情報技術の活用
- S4: コンピュータや情報システムの基本的な仕組みと活用

教員用教材から引用

本発表の概要

- 新聞社での事例紹介
  - 具体例 1：ChatGPT の登場で、新聞記者・編集者の仕事はどう変わるか？
  - 具体例 2：画像を用いた記事推薦
  - 具体例 3：政治資金収支報告書からの情報抽出
- Web からのデータ収集の具体的な方法

**話題	学習内容**	S1: 情報社会と情報技術	S2: コミュニケーションのための情報技術の活用	S3: データを活用するための情報技術の活用	S4: コンピュータや情報システムの基本的な仕組みと活用
具体例 1：ChatGPT の登場で、新聞記者・編集者の仕事はどう変わるか？	✅ 可能性と課題は？			✅ 独自の生成 AI の構築
具体例 2：画像を用いた記事推薦		✅ 拡張現実との向き合い		✅ 生成 AI を用いたシステム
具体例 3：政治資金収支報告書からの情報抽出	✅ 新たなジャーナリズムの模索		✅ Web からのデータ収集や加工	
Web からのデータ収集の具体的な方法			✅ Web からのデータ収集や加工	

本発表の概要

- 新聞社での事例紹介
  - 具体例 1：ChatGPT の登場で、新聞記者・編集者の仕事はどう変わるか？
  - 具体例 2：画像を用いた記事推薦
  - 具体例 3：政治資金収支報告書からの情報抽出
- Web からのデータ収集の具体的な方法

話題	学習内容	S1: 情報社会と情報技術	S2: コミュニケーションのための情報技術の活用	S3: データを活用するための情報技術の活用	S4: コンピュータや情報システムの基本的な仕組みと活用
具体例 1：ChatGPT の登場で、新聞記者・編集者の仕事はどう変わるか？		✅ 可能性と課題は？			✅ 独自の生成 AI の構築
具体例 2：画像を用いた記事推薦			✅ 拡張現実との向き合い		✅ 生成 AI を用いたシステム
具体例 3：政治資金収支報告書からの情報抽出		✅ 新たなジャーナリズムの模索		✅ Web からのデータ収集や加工	
Web からのデータ収集の具体的な方法				✅ Web からのデータ収集や加工	

関連する情報 II の学習内容

教員用教材から引用

ア

情報社会と情報技術

1 現在使われている情報技術により情報社会が受ける効果や影響

情報システムにより個人情報が収集されること，
その利便性と危険性などについてまとめる

2 将来予測される情報技術により情報社会が受ける効果や影響

人工知能の発達，人間に求められる能力の変化，
社会で必要とされる新たな職業などについて提案する

エ

コンピュータや
情報システムの
基本的な仕組みと活用

1 問題の発見と解決

コンピュータの仕組みの活用，情報システムの活用
物理現象や数学的事象のシミュレーション

2 機能追加，ユーザビリティやアクセシビリティの向上

画像認識，音声認識，カメラやセンサなどの外部機器の活用
管理に必要なプログラムの作成，機械学習などの外部プログラムの活用

独自の生成 AI の構築プロジェクト

- **収集・編集・提供・計測**における**新機能開発**や**業務効率化**に繋げる目的
- ニュースメディアとしての責任ある使い方を模索
  - 自動化できる業務と、人間が注力すべき業務
  - 何がどこまで実現できるのか、何が課題となるのか？

言語モデルとは？

単語列の生成確率をモデル化したもの

P(吾輩は猫である): 単語列の生成確率

P(吾輩) * P(は | 吾輩) * P(猫 | 吾輩は) * P(で | 吾輩は猫) * P(ある | 吾輩は猫で)

事前学習 (自己教師あり学習)

大量の文から、入力と出力の対を自動生成して、
単語列の生成確率を推定する

日経電子版での学習

学習に使ったテキストの言い回しに近づく可能性
=> 記事の下書きや校正など、業務効率化に繋がる

日経電子版特有の言語表現を獲得したい

- 独自の表記規則を、全てプロンプトに記述するのは現実的でない
- 事前学習済み言語モデルの生成結果を、編集者が逐一修正していくのも手間

=> 研究課題：日経電子版の記事を用いた事前学習済みモデルで、表記規則を模倣できないか？

事前学習済みモデルの構築

- 日経電子版など、日経グループの記事の収集
- 権利面の確認や HTML タグの除去や重複排除などの前処理
- 計算資源の確保
- Transformers ライブラリを駆使したモデルの事前学習

日経電子版 T5 の構築と評価

- T5 をフルスクラッチで事前学習
- 事前学習済み T5 を、編集者作成の {本文, 見出し}, {本文, 3 行まとめ} の対でファインチューニング
- ファインチューニングに利用していないデータで、性能を他モデルと比較して評価
  - 一般的な T5 をファインチューニングしたモデル
  - gpt-3.5-turbo に少数の事例を提供

日経電子版 T5 で ROUGE が最良に

編集者の見出し・3 行まとめとの一致度合いで評価

ファインチューニングに用いた事前学習済みモデルと，外部 API	学習	評価	損失	ROUGE 1	ROUGE 2	ROUGE L
t5-base-japanese-nikkei	日経電子版	日経電子版	2.16	**0.47292**	**0.28639**	**0.37113**
sonoisa/t5-base-japanese			2.50	0.42441	0.25270	0.33836
megagonlabs/t5-base-japanese-web			2.63	0.41092	0.23583	0.31574
t5-base-japanese-nikkei	livedoor	日経電子版	2.91	0.39372	0.19094	0.27653
sonoisa/t5-base-japanese			2.68	0.37471	0.17136	0.26145
megagonlabs/t5-base-japanese-web			2.72	0.37938	0.17678	0.26378
gpt-3.5-turbo (ゼロショット)	-	日経電子版	-	0.38787	0.18152	0.26248
gpt-3.5-turbo (3 ショット)			-	0.39619	0.18594	0.27331
gpt-3.5-turbo (6 ショット)			-	0.39450	0.18448	0.27253

[石原ら24a] 表 7：3 行まとめの性能評価

編集支援ツール

- 複数候補を提示しユーザが選択・編集する (文字数や含める・含めない単語などを調整可能) [Ishihara21]
- 予測 CTR も提示

[石原ら24a] 図 2：スクリーンショット

議論：独自の生成 AI の構築プロジェクト

- 一般を上回る性能が出る活用場面を確認
- 社内共有を通じて、メリット・デメリットを考察
  - 誤りが生成される場合も [石原ら24a]
  - 時系列で性能が劣化する可能性も [石原ら24b]
  - 訓練データが暗記される現象も [Ishihara+24]

本発表の概要

- 新聞社での事例紹介
  - 具体例 1：ChatGPT の登場で、新聞記者・編集者の仕事はどう変わるか？
  - 具体例 2：画像を用いた記事推薦
  - 具体例 3：政治資金収支報告書からの情報抽出
- Web からのデータ収集の具体的な方法

話題	学習内容	S1: 情報社会と情報技術	S2: コミュニケーションのための情報技術の活用	S3: データを活用するための情報技術の活用	S4: コンピュータや情報システムの基本的な仕組みと活用
具体例 1：ChatGPT の登場で、新聞記者・編集者の仕事はどう変わるか？		✅ 可能性と課題は？			✅ 独自の生成 AI の構築
具体例 2：画像を用いた記事推薦			✅ 拡張現実との向き合い		✅ 生成 AI を用いたシステム
具体例 3：政治資金収支報告書からの情報抽出		✅ 新たなジャーナリズムの模索		✅ Web からのデータ収集や加工	
Web からのデータ収集の具体的な方法				✅ Web からのデータ収集や加工	

関連する情報 II の学習内容

教員用教材から引用

イ

コミュニケーションのための情報技術の活用

1 コンテンツの編集

文字，音・音声，静止画，動画など

2 新しい技術を含めたコンテンツの制作

仮想現実，拡張現実，複合現実，
仮想世界を探検する中で，様々な情報を提供する作品制作

エ

コンピュータや情報システムの基本的な仕組みと活用

1 問題の発見と解決

コンピュータの仕組みの活用，情報システムの活用

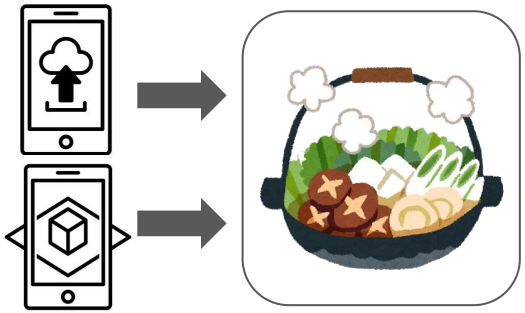
物理現象や数学的事象のシミュレーション

2 機能追加，ユーザビリティやアクセシビリティの向上

画像認識，音声認識，カメラやセンサなどの外部機器の活用

管理に必要なプログラムの作成，機械学習などの外部プログラムの活用

画像からのニュース記事推薦



マルチモーダル AI

"白菜",
"きのこ",
"豆腐",
"白滝",
"ねぎ",
"鍋"

記事検索

"白菜": "手作り漬物の販売ピンチ？衛生管理、国際規格で厳しく"
https://www.nikkei.com/article/DGXZQOUE117TN0R10C24A4000000/

ニュース配信サービスの新規ユーザ向けに、新たな興味関心の発見に繋がる記事を推薦するため、ユーザの身近な画像を入力とする手法を検証したい

ニュース配信サービスの新規ユーザ

- 情報収集のために、登録してみよう
- たくさんのニュースから何を読めば良いか……
- 閲覧傾向に基づくニュース推薦も、興味関心に刺さるものが少ない……
- 閲覧習慣が定着せず、サービス離脱に……

システム実装の詳細

1. **画像の入力**：画像アップロード機能
2. **物体名の抽出**：視覚言語モデル (Gemini 1.0 Pro Vision) を利用
3. **ニュース記事の検索**：「日経電子版」を題材に、全文検索システム (Elasticsearch) を利用

実装したシステム

[田邉ら24] 図 2

画像記事検索

1.画像をアップロードしてください

Drag and drop file here

Limit 200MB per file • PNG, JPG, JPEG

Browse files

VA-02-05-1…

66.8KB

2.物体検出を利用しますか？

○ はい ● いいえ

3.画像についてのプロンプトを入力してください

この画像に含まれている物体をすべて検出し、jsonにブランド名または商

アップロードした画像

処理開始

Gemini処理結果


```
{
  "objectname": [
    "キッチン",
    "キャビネット",
    "換気扇",
    "コンロ",
    "まな板",
    "包丁",
    "フォーク",
    "スプーン"
  ]
}
```


JSONフォーマットは適切です。

記事検索結果

検索クエリ: キッチン

	表示時刻	記事タイトル
0	2024-07-05	中国、キッチン用エアコン　炒め物中も涼しさキープ
1	2024-07-02	輪島に復興の芽吹きを　被災料理人ら出店、炊き出しで絆
2	2024-07-02	人型ロボットにAIのワザ「頼れる機械」を社会に
3	2024-07-01	山形県西川町、住民や企業の交流拠点　5.6億円かけ開所
4	2024-07-01	夏のお弁当、食中毒防ぐには　冷凍食品に曲げわっぱ

検索クエリ: 食器棚

	表示時刻	記事タイトル
0	2024-05-20	ポラスグループ、大宮駅近くにマンション　共働きに照準
1	2024-04-18	深夜に「経験ない揺れ」住民に広がる不安　四国で震度6弱
2	2024-01-22	勉強がしたくなる部屋の作り方　ロジカル片付け術
3	2024-01-01	石川・能登地震、家屋倒壊　住民ら避難「経験ない揺れ」
4	2024-01-01	余震に備えを　グラッとくる前に、防災対策の再確認

検索クエリ: 換気扇

	表示時刻	記事タイトル
0	2024-05-12	「円安にもほどがある」 暮らしも企業活動も影響多岐に
1	2024-05-10	蒸し暑いカフェ増える？　電気代高騰再び、円安で拍車

実験設定

- 日常画像データセットから、オフィス・寝室・パン屋・キッチン・クローゼットの 5 カテゴリで 1 枚ずつ画像を利用
- 5 人の参加者が 5 枚の画像をシステムに入力し、推薦された合計 115 記事をそれぞれが評価

評価観点：セレンディピティ

全てを満たす場合に「セレンディピティがある」

- 関連性「提示された物体名やニュース記事が、妥当であると感じる」
- 新規性「提示された物体名やニュース記事を、知らなかった」
- 意外性「提示された物体名やニュース記事を、システムのおかげで発見できたと感じる」

評価の具体例（寝室カテゴリ）

- 物体名の抽出：
  デスク、ベッド、
  ランプ、椅子、窓
- それぞれに対し、5
  件のニュース記事
  を検索し推薦

画像はデータセットから

5 記事と評価の平均値

[田邉ら24] 表 3

記事番号	記事タイトル	関連性	新規性	意外性	全て 1
ベッド-1	中小型株、地味にスゴい コメ兵やフランスベッド…	1.00	1.00	1.00	1.00
ベッド-2	乳幼児用バウンサーの安全基準改正 米国で窒息死多発	1.00	1.00	0.80	0.80
ベッド-3	ユーラシア最南端に中国バブル崩壊の爪痕 「鬼城」は今	0.00	0.80	0.60	0.00
ベッド-4	静岡の SUS、仮眠用個室生む 2 段ベッド JR 東と開発	1.00	0.40	0.40	0.40
ベッド-5	星野リゾート、ディズニー近くにホテル 1 泊 9000 円から	0.80	0.00	0.00	0.00

- 1 記事目は全員が 3 観点が妥当で、セレンディピティがあると判断した
- 関連性・新規性・意外性のいずれかが欠けても、セレンディピティがないと見なす

有用性の評価

[田邉ら24] 表 4

0.12 の割合で、セレンディピティがある推薦を実現

		ニュース記事（0-1 点）			
カテゴリ名	数	関連性	新規性	意外性	全て 1
オフィス	25	0.47 (0.07)	0.68 (0.06)	0.41 (0.14)	0.14
寝室	25	0.41 (0.02)	0.67 (0.05)	0.46 (0.12)	0.13
パン屋	15	0.33 (0.00)	0.63 (0.15)	0.25 (0.18)	0.08
キッチン	25	0.41 (0.05)	0.60 (0.10)	0.26 (0.15)	0.10
クローゼット	25	0.46 (0.04)	0.52 (0.12)	0.28 (0.10)	0.12
全体	115	0.42	0.62	0.34	0.12

実装したシステムの改善点

- 全体的に関連性が十分に高くない
  - 「ランプ（照明）」で高速道路の文脈の「ランプ（相互を連結する道）」が検索される
  - 必ずしも主題ではない記事が検索される

ユーザ実験の改善点

- 新規性と意外性の定義が不明瞭で，関連性と比べて，標準偏差が大きい
- 「戦争や政治など，意図的に避けている話題が出てきた際の評価が苦痛」

今後の展望

- システム実装の改善（特にニュース記事の検索における関連性の向上）
- 大規模なユーザ実験（サービス実装も視野に）
- AR/VR システムとの繋ぎ込み

議論：画像を用いた記事推薦

- 読者との新しいコミュニケーション方法を提案し実際にシステムを実装
- 定量・定性評価を通じて提案内容を考察
  - 一定割合で目的に合致する推薦を実現
  - システム実装などの課題と今後の展望を確認

本発表の概要

- 新聞社での事例紹介
  - 具体例 1：ChatGPT の登場で、新聞記者・編集者の仕事はどう変わるか？
  - 具体例 2：画像を用いた記事推薦
  - 具体例 3：政治資金収支報告書からの情報抽出
- Web からのデータ収集の具体的な方法

話題	学習内容	S1: 情報社会と情報技術	S2: コミュニケーションのための情報技術の活用	S3: データを活用するための情報技術の活用	S4: コンピュータや情報システムの基本的な仕組みと活用
具体例 1：ChatGPT の登場で、新聞記者・編集者の仕事はどう変わるか？		✅ 可能性と課題は？			✅ 独自の生成 AI の構築
具体例 2：画像を用いた記事推薦			✅ 拡張現実との向き合い		✅ 生成 AI を用いたシステム
具体例 3：政治資金収支報告書からの情報抽出		✅ 新たなジャーナリズムの模索		✅ Web からのデータ収集や加工	
Web からのデータ収集の具体的な方法				✅ Web からのデータ収集や加工	

関連する情報 II の学習内容

教員用教材から引用

ア

情報社会と情報技術

1 現在使われている情報技術により情報社会が受ける効果や影響

情報システムにより個人情報が収集されること，
その利便性と危険性などについてまとめる

2 将来予測される情報技術により情報社会が受ける効果や影響

人工知能の発達，人間に求められる能力の変化，
社会で必要とされる新たな職業などについて提案する

ウ

データを活用するための情報技術の活用

1 問題の発見や解決

インターネット上で公開されたデータなどの活用

2 今後の方向性の予測

データマイニング，ビッグデータを含むデータの解析

政治資金収支報告書からの情報抽出

- 政治団体の 1 年間の収支を記した報告書から、情報を抽出したい
  - 入力：紙媒体の書類がスキャンされた画像
  - 特徴：手書き、修正印あり、雛形は非統一
  - 出力：表形式
- 情報抽出手法や結果の活用可能性を検証したい

[山田ら24] 図 1

【エラーチェック済】

（その７）

行番号	(7) 寄附の内訳 寄附者の氏名（又は名称）	金額	年月日	寄附者の区分 住所（又は所在地）	1. 個人 職業（又は代表者の氏名）	備考
1	石田 晃男	500,000	H24/1/14	岡山県苫田郡鏡野町古川383-3	農業	
2						
3						
4						
5						
6						
7						
8						
9						
10						
11						
12						
13						
14						
15						
	その他の寄附	0				
	合計	500,000				

（その7）

(7) 寄附の内訳　政治団体からの寄附　寄附者の区分　政治団体からの寄附

寄附者の氏名	金額	年月日	主たる事務所の所在地	代表者の氏名	備考
自由民主党北海道参議院選挙区第一支部	500000	H24.6.30	北海道札幌市中央区北3条西3丁目 札幌ビル9F	伊達忠一	
〃	598548	H24.12.30	〃	〃	
この頁の小計	1098548				
その他の寄附	0				
合計	1098548				

（その7）

(7) 寄附の内訳　寄附者の区分　政治団体

寄附者の氏名（団体にあっては、その名称）	金額	年月日	住所（団体にあっては、主たる事務所の所在地）	職業（団体にあっては、代表者の氏名）	備考
明風会	8000000	2.1.24	[illegible]	伊吹文明	
〃	9000000	2.8.7	〃	〃	
〃	4000000	2.9.11	〃	〃	
この頁の小計	21000000				
その他の寄附	0				
合計	21000000				

(注) 同項目の記載が数ページにわたる場合は、最終ページのみに記載してください。
(注) 様式(その2)の寄附額と合算します。
(注) 個人、法人その他の団体又は政治団体の区分を記載し、区分ごとに別葉としてください。
(注) ＊特定寄附については、氏名の前に㊕と記載し、他の寄附と区別してください。（資金管理団体のみが対象となる寄附です。詳しくは記載要領7ページを参照ください。）
＊同一の者からの寄附で明細を記載する場合は、寄附者ごとにまとめて年月日順に記載してください。
＊遺贈によってする寄附については、備考欄に「遺贈」と記載してください。

（その７）

行番号	(7) 寄附の内訳 寄附者の氏名（団体にあっては、その名称）	金額	年月日	寄附者の区分 住所（団体にあっては、主たる事務所の所在地）	1. 個人 職業（団体にあっては、代表者の氏名）	備考
1	山内 俊文	30,000	H31/1/17	神戸市東灘区住吉町3-9-28-606	会社員	
2	武藤 雄一郎	12,000	H31/1/17	千代[illegible]町7-1-1-1206	会社員	
3	武藤 雄一郎	12,000	R1/8/5	千代[illegible]町7-1-1-1206	会社員	
4	（小計）	24,000				
5	西田 一也	10,000	H31/1/23	山口県山口市阿知須3745-1	医療法人理事長	
6	西田 一也	10,000	H31/2/27	山口県山口市阿知須3745-1	医療法人理事長	
7	西田 一也	10,000	H31/3/27	山口県山口市阿知須3745-1	医療法人理事長	
8	西田 一也	10,000	H31/4/26	山口県山口市阿知須3745-1	医療法人理事長	
9	西田 一也	10,000	R1/5/29	山口県山口市阿知須3745-1	医療法人理事長	
10	西田 一也	10,000	R1/6/24	山口県山口市阿知須3745-1	医療法人理事長	
11	西田 一也	10,000	R1/7/31	山口県山口市阿知須3745-1	医療法人理事長	
12	西田 一也	10,000	R1/8/28	山口県山口市阿知須3745-1	医療法人理事長	
13	西田 一也	10,000	R1/9/30	山口県山口市阿知須3745-1	医療法人理事長	
14	西田 一也	10,000	R1/10/30	山口県山口市阿知須3745-1	医療法人理事長	
15	西田 一也	10,000	R1/11/29	山口県山口市阿知須3745-1	医療法人理事長	
16	西田 一也	10,000	R1/12/27	山口県山口市阿知須3745-1	医療法人理事長	
17	（小計）	120,000				
18	土肥 幸夫	120,000	H31/1/29	鎌倉市笹目町7-24	会社役員	

マルチモーダル AI

- 言語のみならず、画像・音声・動画など複数の種類 (モーダル) のデータを統合的に処理
- モデルの例
  - Gemini (by Google Deepmind)
  - Claude (by Anthropic)
  - GPT-4V (by OpenAI)

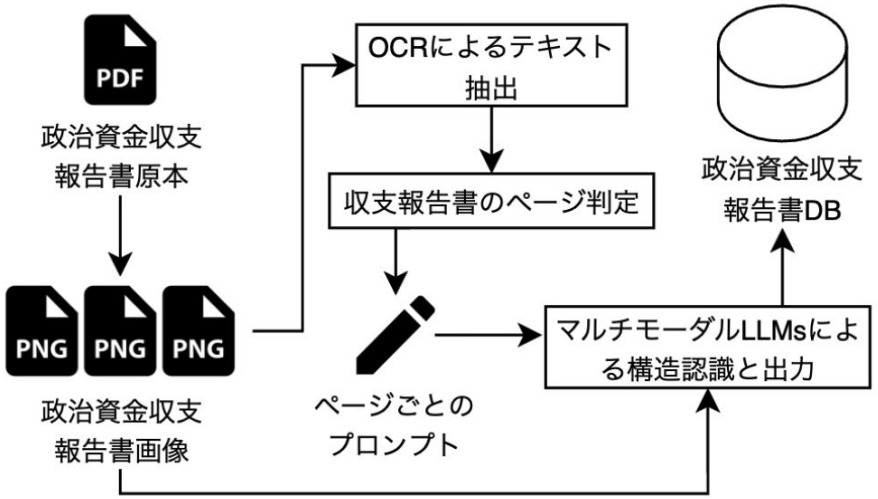
マルチモーダル AI への指示

モード	対象ページ	Prompt
USER	全ページ共通	与えられた政治資金収支報告書から情報を抽出し，必ずヘッダ付きのカンマ区切りの csv 形式の文字列を出力してください．金額は数値として出力してください．csv の形式を守ってください．”小計”や”合計”，それに類する行は出力しないでください．csv のヘッダーとなる列名を後述します．列名は csv のヘッダーに対応します．ヘッダーも出力してください．後述する列名以外は表示しないでください．
ASSISTANT		
	機関紙誌の発行その他の事業による収入	事業の種類，金額，備考
	借入金	借入先，金額，備考
	本部または支部から供与された交付金に係る収入	交付金を供与した本部又は支部，金額，年月日，主たる事務所の所在地，備考
	その他の収入	摘要，金額，備考
	寄附の内訳	寄附者の氏名，金額，年月日，住所，職業，備考
	寄附のうち寄附の斡旋によるものの内訳	寄附のあっせん者の氏名，金額，提供年月日，集めた期間，住所，職業，備考
	政党匿名寄附の内訳	政党匿名寄附を受けた場所，金額，年月日，備考
	機関紙誌の発行その他事業による収入のうち特定パーティーの対価に係る収入の内訳	特定パーティーの名称，対価に係る収入の金額，対価の支払いをした者の数，開催年月日，開催場所，備考
	政治資金パーティーの対価に係る収入の内訳	対価の支払いをした者の氏名，金額，年月日，住所，職業，備考
	政治資金パーティーの対価に係る収入のうち対価の支払のあっせんによるものの内訳	対価の支払いをした者の氏名，金額，提供年月日，集めた期間，住所，職業，備考

[山田ら24] 表 2

実験

- 複数のモデルを用いて、性能を検証
  - OCR との組み合わせも調査



[山田ら24] 図 3

- 評価指標 Tree-Edit-Distance-Similarity (TEDS)
  - 比較対象の表の構造を HTML 形式にし、類似度を計算 (大きいほど良い)

実験結果

- OCR との組み合わせで、性能が改善
- 2024 年 5 月に論文を公開した時点の実験結果
- [山田ら24] 表 3

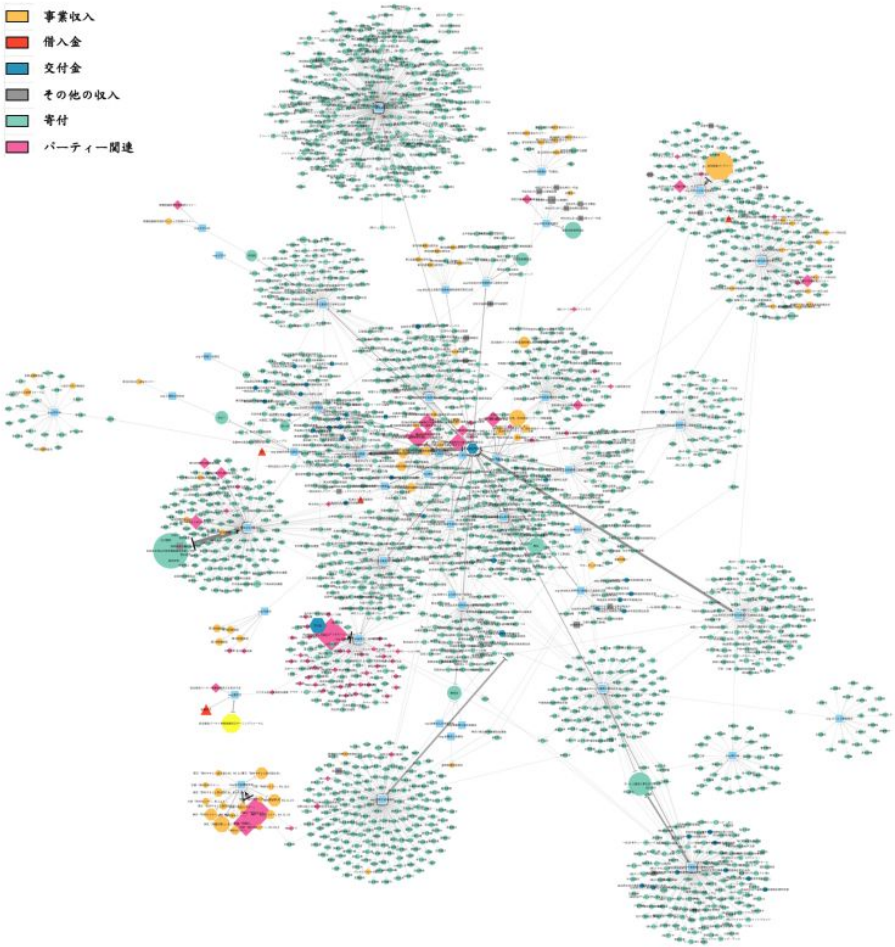
項目名	TEDS
Gemini Pro 1.0	0.6643
Gemini Pro 1.5	0.6487
GPT-4	0.6573
Claude 3 Haiku	0.7876
Claude 3 Opus	0.8411
Gemini Pro 1.0+OCR	0.7987
Gemini Pro 1.5+OCR	0.8314
GPT-4+OCR	0.8166
Claude 3 Haiku+OCR	0.8500
Claude 3 Opus+OCR	**0.9048**

抽出結果の活用 1

各種政治団体における
収入グラフネットワーク

- 情報抽出を半機械的に処理することで、分析や活用可能性の議論に注力できる

[山田ら24] 図 7



抽出結果の活用 2

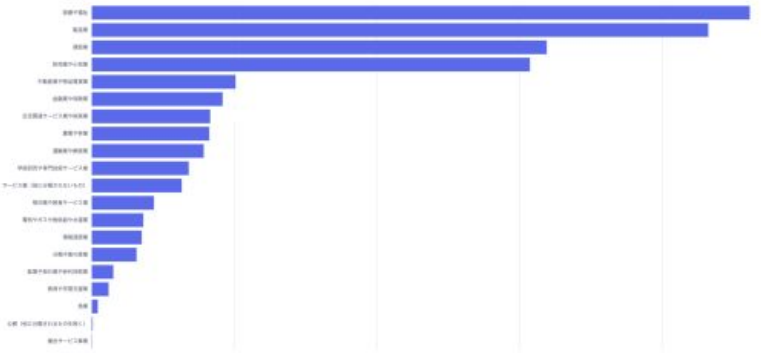


図 4: 閣僚らに関連する団体への日本標準産業分類ごとの企業・団体の献金額の合計 (5 年、集計)

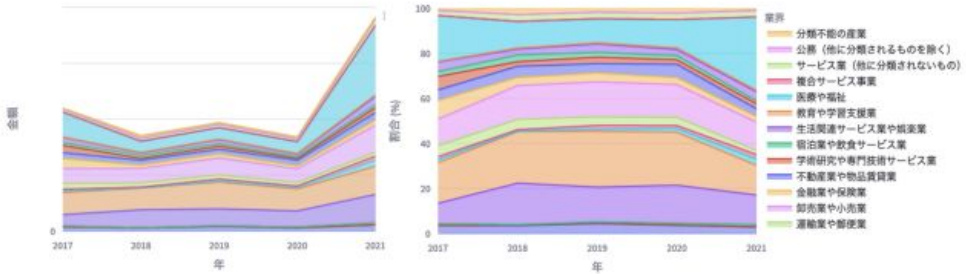
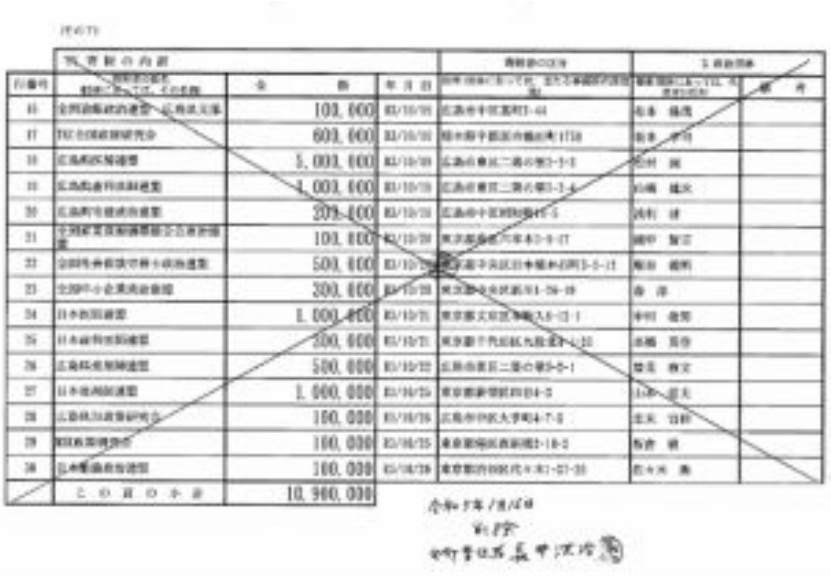


図 5: 閣僚らに関連する団体への日本標準産業分類ごとの企業・団体の献金額 (5 年、積み上げ)

[山田ら24] 図 4, 5

抽出が不十分だった例

[山田ら24] 図 6



	寄附者の区分		
	年 月 日	住所(団体にあっては、主たる事務所の所在地)	職業(団体にあっては、その代表者の
0	H31/1/17	神戸市東灘区江北町3-9-28-606	会社員
0	H31/1/17	千代田区[illegible]番町7-1-1-1206	会社員
0	R1/8/5	千代田区[illegible]番町7-1-1-1206	会社員
0			
0	H31/1/23	山口県山口市阿知須3745-1	医療法人翔
0	H31/2/27	山口県山口市阿知須3745-1	医療法人翔

図 6: 抽出された献金のうち，重複や削除を考慮できなかった例

議論：政治資金収支報告書からの情報抽出

- Web に公開されている (が乱雑な) 情報を収集し統一的な形式に加工・分析
- データ収集の方法や活用方法を考察
  - マルチモーダル AI を用いた収集の可能性
  - 収集の半自動化による分析への注力

本発表の概要

- 新聞社での事例紹介
- Web からのデータ収集の具体的な方法
  - 公開データセットや API の利用
  - データが公開されていない場合

話題	学習内容	S1: 情報社会と情報技術	S2: コミュニケーションのための情報技術の活用	S3: データを活用するための情報技術の活用	S4: コンピュータや情報システムの基本的な仕組みと活用
具体例 1：ChatGPT の登場で、新聞記者・編集者の仕事はどう変わるか？		✅ 可能性と課題は？			✅ 独自の生成 AI の構築
具体例 2：画像を用いた記事推薦			✅ 拡張現実との向き合い		✅ 生成 AI を用いたシステム
具体例 3：政治資金収支報告書からの情報抽出		✅ 新たなジャーナリズムの模索		✅ Web からのデータ収集や加工	
Web からのデータ収集の具体的な方法		☑	☑	✅ Web からのデータ収集や加工	☑

話題	学習内容	S1: 情報社会と情報技術	S2: コミュニケーションのための情報技術の活用	S3: データを活用するための情報技術の活用	S4: コンピュータや情報システムの基本的な仕組みと活用
具体例 1：ChatGPT の登場で、新聞記者・編集者の仕事はどう変わるか？		✅ 可能性と課題は？		(社内データの活用)	✅ 独自の生成 AI の構築
具体例 2：画像を用いた記事推薦			✅ 拡張現実との向き合い	(社内データの活用)	✅ 生成 AI を用いたシステム
具体例 3：政治資金収支報告書からの情報抽出		✅ 新たなジャーナリズムの模索		✅ Web からのデータ収集や加工	
Web からのデータ収集の具体的な方法		☑	☑	✅ Web からのデータ収集や加工	☑

「データ」という新たな武器

- データを用いることで、新たな視点で物事を捉えたり、新たな体験を生み出せたりする
- 「KKD (勘と経験と度胸)」の否定ではなく、データの得意分野を見つけるのが大事
- 日常の気づきや課題感を起点に、仮説を立てた上でデータを探しに行くと良い (「このデータから何か面白いことを発見して」は難しい)

私の場合は、大学新聞での経験が糧に

- 具体例 1：ChatGPT の登場で、新聞記者・編集者の仕事はどう変わるか？
  - 多忙な同僚のための業務効率化？
- 具体例 2：画像を用いた記事推薦
  - 若者にもニュースを届けるには？
- 具体例 3：政治資金収支報告書からの情報抽出
  - 「政治とカネ」への向き合い方？

本発表の概要

- 新聞社での事例紹介
- Web からのデータ収集の具体的な方法
  - 公開データセットや API の利用
  - データが公開されていない場合

正攻法は、公式提供の利用

- 公式で提供されているデータセットや API がないか確認する
  - 大抵は利用規約やライセンスなどが確認できる
- 「スクレイピング」は最終手段 (後述)
  - 利用規約に注意 (学術・教育的利用は許諾されている場合も)

公式提供の見つけ方

- インターネット検索
  - Google 検索やデータセット検索サイト
  - 日本語だけでなく英語でも
  - 「転載」の場合に注意
- 有識者に聞く

データセット検索

- Google Dataset Search
  https://toolbox.google.com/datasetsearch
- Kaggle Datasets
  https://www.kaggle.com/datasets
- Harvard Dataverse
  https://dataverse.harvard.edu/
- e-Stat https://www.e-stat.go.jp/

本発表の概要

- 新聞社での事例紹介
- Web からのデータ収集の具体的な方法
  - 公開データセットや API の利用
  - データが公開されていない場合

データが公開されていない場合

- (情報開示請求)
- Web スクレイピング
- データを作る

Web スクレイピング

- Web サイトから特定の情報を抽出・取集する技術
- ページにアクセス、情報を探す、参照などの処理をプログラミング言語で記述し、自動化
- UTokyo OpenCourseWare「メディアプログラミング入門」の「7. WebスクレイピングとWebAPI」
  - https://ocw.u-tokyo.ac.jp/course_11472/

Web スクレイピングは何をしているか

- (人間が見る) Web ページを構成する「情報ソース」に対して、機械的にアクセス
- (人間が見てメモする代わりに) HTML などを解析し情報を参照・保存
- 🔍 日経電子版 (https://www.nikkei.com/) の情報ソースを確認してみましょう

Web クローリングのマナー

- 利用規約の確認
- 法律面 (著作権など) の確認
- アクセス頻度の調整

データを作る

- 協力者を募ってアンケート評価 (具体例 1 や 2 でのユーザ評価)
- 生成 AI を用いた情報抽出や生成 (具体例 3 での政治資金収支報告書の解析)
- 人力でのラベル付け

本発表のまとめ

- 新聞社での事例紹介
  - 情報と情報技術を活用した、問題発見・解決の具体例 3 つを紹介
- Web からのデータ収集の具体的な方法
  - 公開データセットや API の利用方法や、データが公開されていない場合の対応策を紹介

具体例の参考文献

- [石原ら24a] 石原祥太郎, 村田栄樹, 中間康文, 高橋寛武 (2024). 日本語ニュース記事要約支援に向けたドメイン特化事前学習済みモデルの構築と活用. 自然言語処理, 2024, 31 巻, 4 号.
- [Ishihara+24] Shotaro Ishihara and Hiromu Takahashi (2024). Quantifying Memorization and Detecting Training Data of Pre-trained Language Models using Japanese Newspaper. Proceedings of INLG 2024.
- [石原ら24b] 石原祥太郎, 高橋寛武, 白井穂乃 (2024). Semantic Shift Stability: 学習コーパス内の単語の意味変化を用いた事前学習済みモデルの時系列性能劣化の監査. 自然言語処理, 31巻, 4号.
- [山田ら24] 山田健太, 青田雅輝 (2024). マルチモーダルな深層学習手法を用いた政治資金収支報告書の判読の試み. 2024年度日本選挙学会総会・研究会.
- [田邉ら24] 田邉耕太, 石原祥太郎, 山田健太, 青田雅輝, 又吉康綱 (2024). ニュースを身近に：日常風景からのニュース推薦. 第 210 回情報処理学会ヒューマンコンピュータインタラクション研究会.