

地理空間情報と自然言語処理

「地球の歩き方旅行記データセット」の高付加価値化を通じて

奈良先端科学技術大学院大学 ヒューマンAIインタラクション研究室 准教授
サイバーエージェント AI Lab 行動理解チーム リサーチサイエンティスト

大内 啓樹

自己紹介

大内 啓樹 (おおうち ひろき)

- 奈良先端科学技術大学院大学 准教授
- 専門：自然言語処理

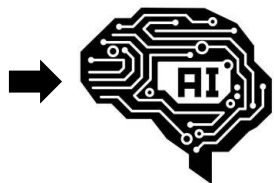


研究テーマ

言語知能と空間知能を兼ね備えたAI

【言語to空間】 文章中の人物の移動を地図上に再現

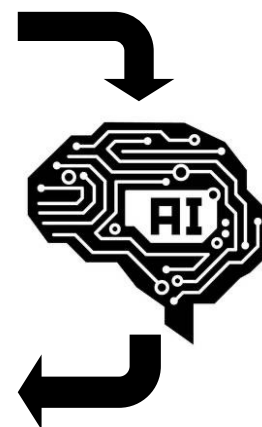
午前8時近鉄奈良駅到着。
ホテルニューわかさに
9時チェックイン予定。
時間に余裕があったので、
途中のスタバで一服。
30分過ぎてしてから店を出て、
ホテルにチェックイン。



【空間to言語】 衛星画像の「差」を説明



The desolate tree-lined street has been transformed into a bustling residential neighborhood with houses and parked cars.



2022年11月「地球の歩き方旅行記データセット」提供開始

一次データを基盤として研究開発を展開

ソフトウェア層

地名
抽出器

同一地点
認識器

地図DB
リンク器

訪問状態
予測器

地理的包含
関係予測器

訪問順序
予測器

旅行記
翻訳器

二次データ層

地理情報データ

訪問状態・順序データ

英語翻訳データ

一次データ層

地球の歩き方旅行記データセット

2022年11月「地球の歩き方旅行記データセット」提供開始 一次データを基盤として研究開発を展開

国際学会論文4件，国内学会論文12件，受賞3件



ソフトウェア層

地名
抽出器

同一地点
認識器

地図DB
リンク器

訪問状態
予測器

地理的包含
関係予測器

訪問順序
予測器

旅行記
翻訳器

二次データ層

地理情報データ

訪問状態・順序データ

英語翻訳データ

一次データ層

地球の歩き方旅行記データセット

本講演の目的

一次データ高付加価値化のヒントを提供

本講演で提供したい話題

データ提供者の視点: 一次データをもっと価値あるものにするには？

データ使用者の視点: 一次データを「工夫」して研究開発の幅を広げるには？

ソフトウェア層

地名
抽出器

同一地点
認識器

地図DB
リンク器

訪問状態
予測器

地理的包含
関係予測器

訪問順序
予測器

旅行記
翻訳器

二次データ層

地理情報データ

訪問状態・順序データ

英語翻訳データ

一次データ層

地球の歩き方旅行記データセット

本講演の目的

一次データ高付加価値化のヒントを提供

本講演で提供したい話題

データ提供者の視点: 一次データをもっと価値あるものにするには？

データ使用者の視点: 一次データを「工夫」して研究開発の幅を広げるには？

「地球の歩き方旅行記データセット」を
利用した実践例を通じて上記の話題を深掘り

主張したいこと

二次データ(アノテーションデータ)が鍵

※モデルや技術の詳しい説明はしません

私が「旅行記」に興味を持ったきっかけ

2021年8月の夜、科研費のテーマを考えていた

そのときにたまたまネット上で「**おくのほそ道**」見かけた

1689年、江戸を出発した松尾芭蕉が、奥州、北陸を巡った紀行文



https://www.kyohaku.go.jp/jp/exhibitions/feature/b/buson_2022/



<http://www.basho-ogaki.jp/hosomiti/about/>

「文章」を読み解き 「地図」とつなぐマルチメディア情報処理

2022年4月：研究プロジェクト発足

文章中の人物の移動軌跡を地図上に再現

入力: 文章

午前8時 近鉄奈良駅 到着。
ホテルニューわかさに
9時 チェックイン予定。
時間に余裕があったので、
途中のスタバで 一服。
30分過ぎしてから 店を出て、
ホテルにチェックインしました。

自然言語処理
地理情報処理

出力: 移動の軌跡



移動軌跡を再現するには 複合的な処理が必要

入力: 文章

午前8時 近鉄奈良駅 到着。

ホテルニューわかさに

9時 チェックイン予定。

時間に余裕があったので、

途中のスタバで 一服。

30分過ぎしてから 店を出て、

ホテルにチェックインしました。

① 場所を表す表現を抽出

移動軌跡を再現するには 複合的な処理が必要

入力: 文章

午前8時 近鉄奈良駅 到着。

ホテルニューわかさに

9時 チェックイン予定。

時間に余裕があったので、

途中のスタバで 一服。

30分過ぎしてから 店を出て、

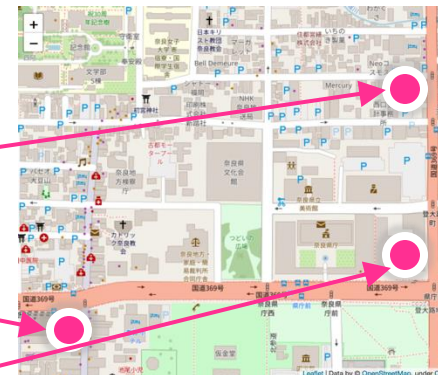
ホテルにチェックインしました。

- ① 場所を表す表現を抽出
- ② 同じ場所を指す表現の認識

移動軌跡を再現するには 複合的な処理が必要

入力: 文章

午前8時 **近鉄奈良駅** 到着。
ホテルニューわかに
9時 チェックイン予定。
時間に余裕があったので、
途中の**スタバ**で 一服。
30分過ぎしてから **店**を出て、
ホテルにチェックインしました。

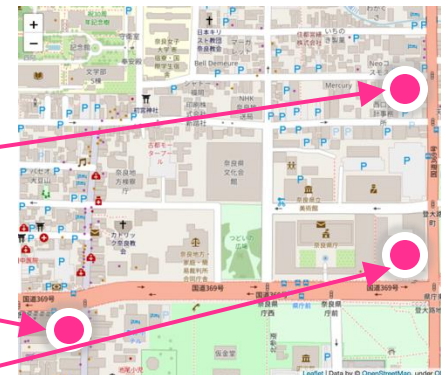
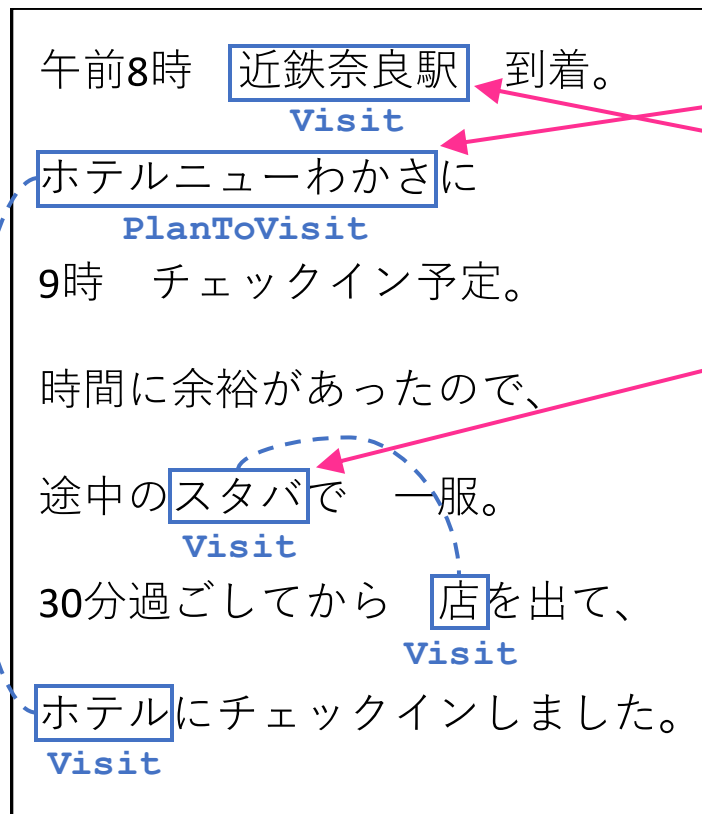


地図データベース
OpenStreetMap

- ① **場所を表す表現**を抽出
- ② 同じ場所を指す表現の認識
- ③ 各場所の表現と 地図上の対応箇所をつなぐ 

移動軌跡を再現するには 複合的な処理が必要

入力: 文章



地図データベース
OpenStreetMap

- ① **場所を表す表現**を抽出
- ② 同じ場所を指す表現の認識
- ③ 各場所の表現と 地図上の対応箇所をつなぐ
- ④ 各場所を訪問したか否かを予測 Visit/PlanToVisit/See...

移動軌跡を再現するには 複合的な処理が必要

入力: 文章

午前8時 **近鉄奈良駅** 到着。

Visit ①

ホテルニューわかに

PlanToVisit

9時 チェックイン予定。

時間に余裕があったので、

途中の**スタバ**で 一服。

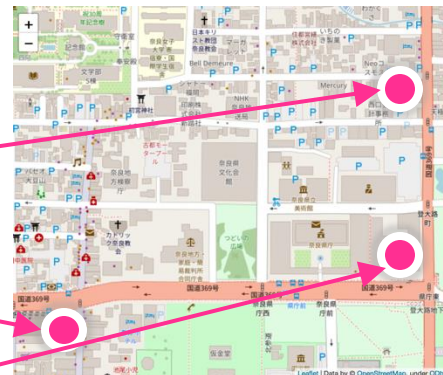
Visit ②

30分過ぎしてから **店**を出て、

Visit

ホテルにチェックインしました。

Visit ③



地図データベース
OpenStreetMap

- ① **場所を表す表現**を抽出
- ② 同じ場所を指す表現の認識
- ③ 各場所の表現と 地図上の対応箇所をつなぐ 
- ④ 各場所を訪問したか否かを予測 **Visit/PlanToVisit/See...**
- ⑤ 訪問場所を時系列順に整序 **①, ②, ③...**

移動軌跡を再現するには 複合的な処理が必要

入力: 文章

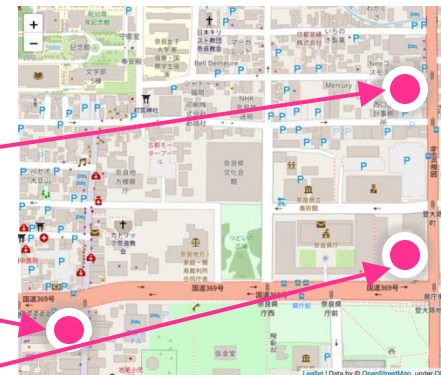
午前8時 **近鉄奈良駅** 到着。
Visit①

ホテルニューわかに
PlanToVisit
9時 チェックイン予定。

時間に余裕があったので、
途中の**スタバ**で一服。
Visit②

30分過ぎたので、
ホテルにチェックインした。
Visit③

本講演で
フォーカス



地図データベース
OpenStreetMap

- ① **場所を表す表現**を抽出
- ② 同じ場所を指す表現の認識
- ③ 各場所の表現と 地図上の対応箇所をつなぐ 
- ④ 各場所を訪問したか否かを予測 **Visit/PlanToVisit/See...**
- ⑤ 訪問場所を時系列順に整序 **①, ②, ③...**

最初の関門

「旅行記の一次データをどうするか？」

我々は「地球の歩き方」に問い合わせしてみた。
議論を重ね、旅行記をデータとして公開できることに。

※機械学習モデルの構築を前提としているため、データは必須であると考えていた

2022年11月24日

「地球の歩き方旅行記データセット」公開

国立情報学研究所情報学研究データリポジトリ(IDR)から学術目的で提供開始

The screenshot shows the NII IDR website. The header includes the NII logo and the IDR logo. The main navigation bar has links for HOME, データー一覧, 研究成果一覧, ユーザーフォーラム, 組織, 関連リンク, and お問い合わせ. Below the navigation bar, there is a breadcrumb trail: > HOME > データー一覧 > 地球の歩き方旅行記データセット. The main content area is titled '地球の歩き方旅行記データセット'. It includes a description: '株式会社地球の歩き方が国立情報学研究所を通じて研究者に提供しているデータセットです。' and the date '2022/11/24 作成'. There is a section for 'データ概要' (Data Overview) with a sub-section '1. 投稿旅行記データ' (Submitted Travelogue Data). This section describes the data as travelogue text from the 'Earth's Way Web' website, covering the period from November 2007 to February 2022, with approximately 14,000 travelogues and 9,500 text pieces. It also mentions the research project 'JSPS 科学研究費助成事業 基盤研究(B) 22H03648' led by Professor Kenji Ooyama.

ニュースリリース

2022/11/24

「地球の歩き方」の利用者投稿旅行記データを学術研究用に無償で提供開始

株式会社地球の歩き方(地球の歩き方、東京都品川区、代表取締役社長:新井 邦弘)と情報・システム研究機構 国立情報学研究所(NII 、東京都千代田区、所長:喜連川 まさる)は学術研究分野でのデータ活用促進に向けて提携し、地球の歩き方が運営するウェブサイトの「旅スケ」サービスに投稿された旅行記のテキストデータを、研究用に「地球の歩き方旅行記データセット」として11月24日(木)より無償提供を開始しました。

このデータセットは、NIIのデータセット共同利用研究開発センター(センター長:コンテンツ科学研究系教授 大山 敬三)が運営する情報学研究データリポジトリ(IDR)を通じて提供されます。

申込サイト: <https://www.nii.ac.jp/dsc/idr/arukikata/>

データセットDOI: <https://doi.org/10.32130/idr.18.1>

まずはデータセットの実例を見てみよう

「旅行記」と「旅スケジュール」

「旅行記」の実例

会津若松へ向かう磐越西線の接続を考慮して選んだやまびこ203号はE5系での運転でした。何度も乗っているE2系よりも座席が広く感じ、快適な移動でした。

...

会津若松駅から快速あいつ4号に乗車し、郡山へ向かいました。

会津若松は晴れ間がありましたが、山を上につれて雲が増えて行き、途中から雨が降り出しました。天気予報通りでしたが、今回の旅行は暖かい2日間で移動時間を除いて雨に降られることがなかったのはラッキーでした。

「旅スケジュール」の実例

1日目 2021年10月15日(金)

05:40 - 05:50 自宅

05:50 - 05:53 最寄駅

...

15:33 - 23:59 庄助の宿 瀧の湯

2日目 2021年10月16日(土)

00:00 - 09:25 庄助の宿 瀧の湯

09:25 - 09:32 東山温泉入口
(瀧の湯前)バス停

...

17:33 - 17:38 最寄駅

17:38 - 17:45 自宅

本データセットの特徴

国内旅行記4500記事と海外旅行記9500記事

	国内旅行記		海外旅行記		全旅行記
	旅スケ付き	旅スケなし	旅スケ付き	旅スケなし	
記事数	3,153	1,519	6,419	3,188	14,279
段落数	76,307	16,412	188,908	58,700	340,327
文字数	5,878,704	1,541,124	19,273,201	4,870,061	31,563,090
単語数	3,568,354	928,936	10,950,950	2,785,494	18,233,734
固有表現数	304,606	71,598	928,487	227,191	1,531,882
地名・施設名数	95,282	25,455	268,417	71,830	460,984
段落数/記事	24.2	10.8	29.4	18.4	23.8
文字数/記事	1864.4	1014.5	3002.5	1527.6	2210.4
単語数/記事	1131.7	611.5	1706.0	873.7	1276.9
固有表現数/記事	96.6	47.1	144.6	71.2	107.2
地名・施設名数/記事	30.2	16.7	41.8	22.5	32.2

3100万文字規模

Xの投稿の
10倍以上の長さ

豊富な
地名・施設名

なぜ「文章」に着目するのか？

「文章」は地理空間情報の鉱脈

- GNSS（GPSなどの衛星測位システム）で十分では？
 - 人間の位置を把握するだけならGPSデータで事足りる
 - しかし「人間」と「場所」の相互関係を捉えるのはむずかしい

場所における行為，場所に対する価値付与，場所から受ける印象や感覚

場所の認識 + 情報抽出

この週末、東福寺へ紅葉を見に行った。
地面いっぱいに むした 苔の緑と 真っ赤な
紅葉との コントラストが 美しかった。

目的語 述語
主語 述語



なぜ「文章」に着目するのか？

「文章」は地理空間情報の鉱脈

- GNSS（GPSなどの衛星測位システム）で十分では？
 - 人間の位置を把握するだけならGPSデータで事足りる
 - しかし「人間」と「場所」の相互関係を捉えるのはむずかしい

場所における行為，場所に対する価値付与，場所から受ける印象や感覚

「場所を表す言語表現」を「地図上の位置・領域」に
接地(グラウンディング)する技術が鍵



ジオパーズング

「文章の世界」と「地図の世界」の架け橋

ジオパージング

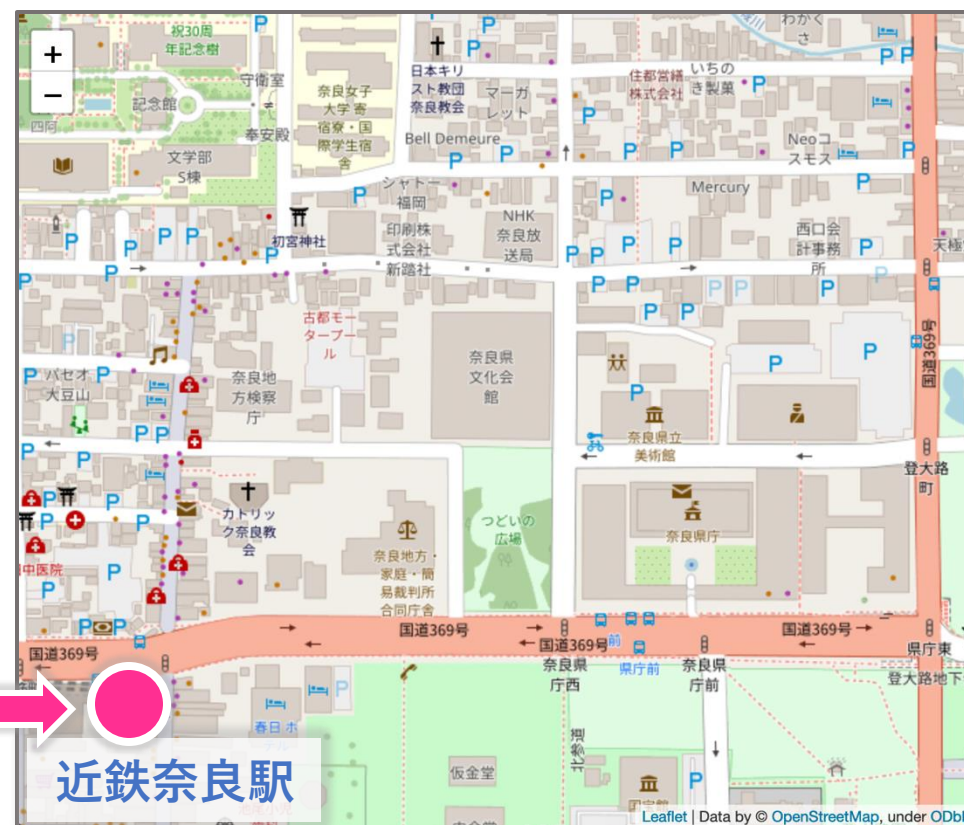
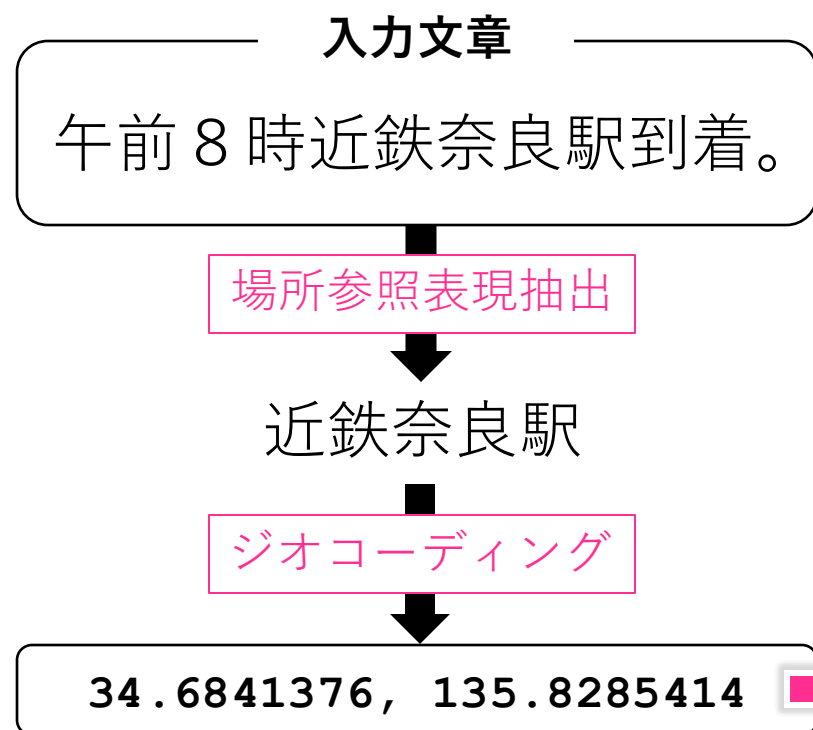
詳しい解説は以下の論文を参照



[Higashiyama+ EACL2024, Arukikata Travelogue Dataset with Geographic Entity Mention, Coreference, and Link Annotation](#)

文章のなかの「移動軌跡」解析の中核技術 「ジオパーズング」とは？

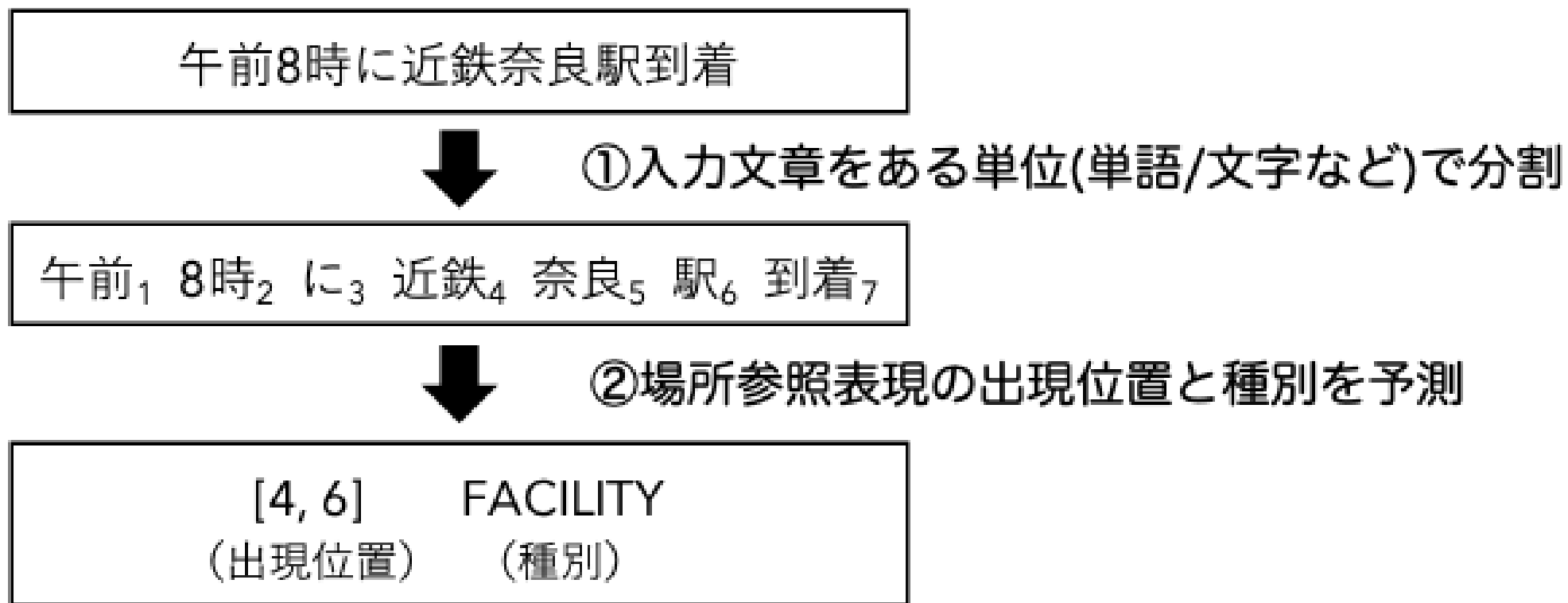
- 場所を表す表現(**場所参照表現**)を抽出し、緯度・経度を推定する



場所参照表現抽出とジオコーディングの二段階から構成される

場所を表す言語表現の出現位置と種別を同定

- 場所を表す言語表現 = 「**場所参照表現**」と呼ぶ



場所参照表現抽出 = 固有表現抽出のひとつの特殊なタスク

2つの主流アプローチ

直説法

緯度・経度を直接出力するアプローチ

近鉄奈良駅

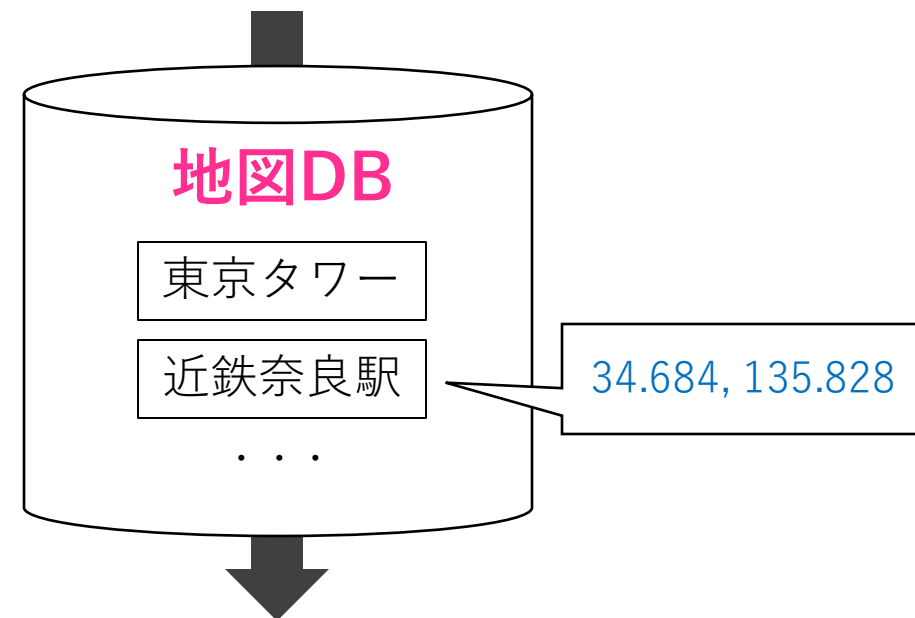


34.684, 135.828

間接法

地図DBのIDへの紐付けを介して
緯度・経度を出力するアプローチ

近鉄奈良駅



34.684, 135.828

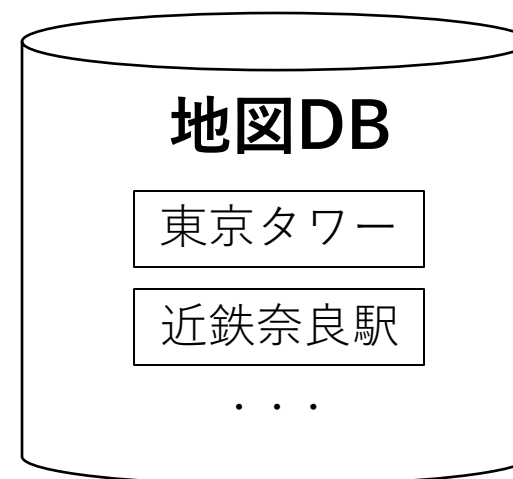
2つのアプローチの長所・短所

直説法

- 長所: 地図DBを必要としない
- 短所: 予測結果の少しのズレが実用上許容できるものではない

間接法

- 長所: 緯度・経度以外の情報も出力可能
- 短所: 地図DBのカバー率が性能の上限を決定する(DBにない場所は扱えない)



もし「JR奈良駅」が
収録されていない場合、
「JR奈良駅」は絶対に
正解できない

2つのアプローチの長所・短所

直説法

- 長所: 地図DBを必要としない
- 短所: 予測結果の少しのズレが実用上許容できるものではない

本研究プロジェクトでは

「間接法」を採用

間接法

- 長所: 緯度・経度以外の情報も出力可能
- 短所: 地図DBのカバー率が性能の上限を決定する(DBにない場所は扱えない)

地図DB

東京タワー

近鉄奈良駅

...

もし「JR奈良駅」が収録されていない場合,
「JR奈良駅」は絶対に正解できない

ジオパージング

アノテーションデータ構築

機械学習モデルで各処理を実現するために
必要な情報を生データに付与する

各処理を機械学習モデルで実現するには？

アノテーション情報が必要

- 「場所参照表現抽出」に必要な情報とは？

途₁ 中₂ の₃ ス₄ タ₅ バ₆ で₇ ー₈ 服₉ 。 ₁₀

3₁₁ 0₁₂ 分₁₃ 過₁₄ ご₁₅ し₁₆ て₁₇ か₁₈ ら₁₉ 店₂₀ を₂₁ 出₂₂ て₂₃ 、 ₂₄

ホ₂₅ テ₂₆ ル₂₇ に₂₈ チ₂₉ エ₃₀ ツ₃₁ ク₃₂ イ₃₃ ン₃₄ し₃₅ ま₃₆ し₃₇ た₃₈ 。 ₃₉

[4, 6]

[20, 20]

[25, 27]

- 場所参照表現の文章中の「位置(文字インデックス)」の情報を生データに追加

各処理を機械学習モデルで実現するには？

アノテーション情報が必要

- 「ジオコーディング」に必要な情報とは？

途₁ 中₂ の₃ ス₄ タ₅ バ₆ で₇ ー₈ 服₉ 。 ₁₀

3₁₁ 0₁₂ 分₁₃ 過₁₄ ご₁₅ し₁₆ て₁₇ か₁₈ ら₁₉ 店₂₀ を₂₁ 出₂₂ て₂₃ 、 ₂₄

ホ₂₅ テ₂₆ ル₂₇ に₂₈ チ₂₉ エ₃₀ ツ₃₁ ク₃₂ イ₃₃ ン₃₄ し₃₅ ま₃₆ し₃₇ た₃₈ 。 ₃₉

[4, 6]

OSM_ID= “node/6976944686”

[20, 20]

OSM_ID= “node/6976944686”

[25, 27]

OSM_ID= “way/229252714”

- 各場所エンティティに「OSM_ID (OpenStreetMapのID)」を追加

※ **OpenStreetMap (OSM)** = 誰でも自由に地図を使えるよう、みんなでオープンデータの地理情報を作るプロジェクト、および、その地図データベース

アノテーションデータ構築 場所参照表現を付与

- アノテーションツール [Brat](https://brat.nlplab.org/) を使ってアノテーション作業

← → /sec054_analysis/30379_concatd

1 朝、OTM_LOC_NAME
IDE_LOC_NAME
HIG_LOC_NAME 京都 OTM_FAC_NAME
IDE_LOC_NAME へ向かう OTM_FAC_NAME
IDE_LOC_NAME 東海道新幹線 の車窓から、OTM_LOC_GENERAL
IDE_LOC_GENERAL 頂上 付近に雲をまとった OTM_LOC_NAME
IDE_LOC_NAME
HIG_LOC_NAME 富士 が良く見えました。

2 OTM_FAC_GENERAL
IDE_FAC_GENERAL 新幹線 利用時の楽しみの一つが車窓からの OTM_LOC_NAME
IDE_LOC_NAME
HIG_LOC_NAME 富士 の眺めですが、今回も良く見えました。

4 OTM_FAC_NAME
IDE_FAC_NAME
HIG_FAC_NAME 近鉄奈良駅 に到着後、OTM_FAC_NAME
IDE_FAC_NAME
HIG_FAC_NAME 興福寺 OTM_FAC_NAME
IDE_FAC_NAME
HIG_FAC_NAME へ向かうために表示に従って OTM_FAC_NAME
IDE_FAC_NAME
HIG_FAC_NAME 東向商店街 へ。

<https://brat.nlplab.org/>

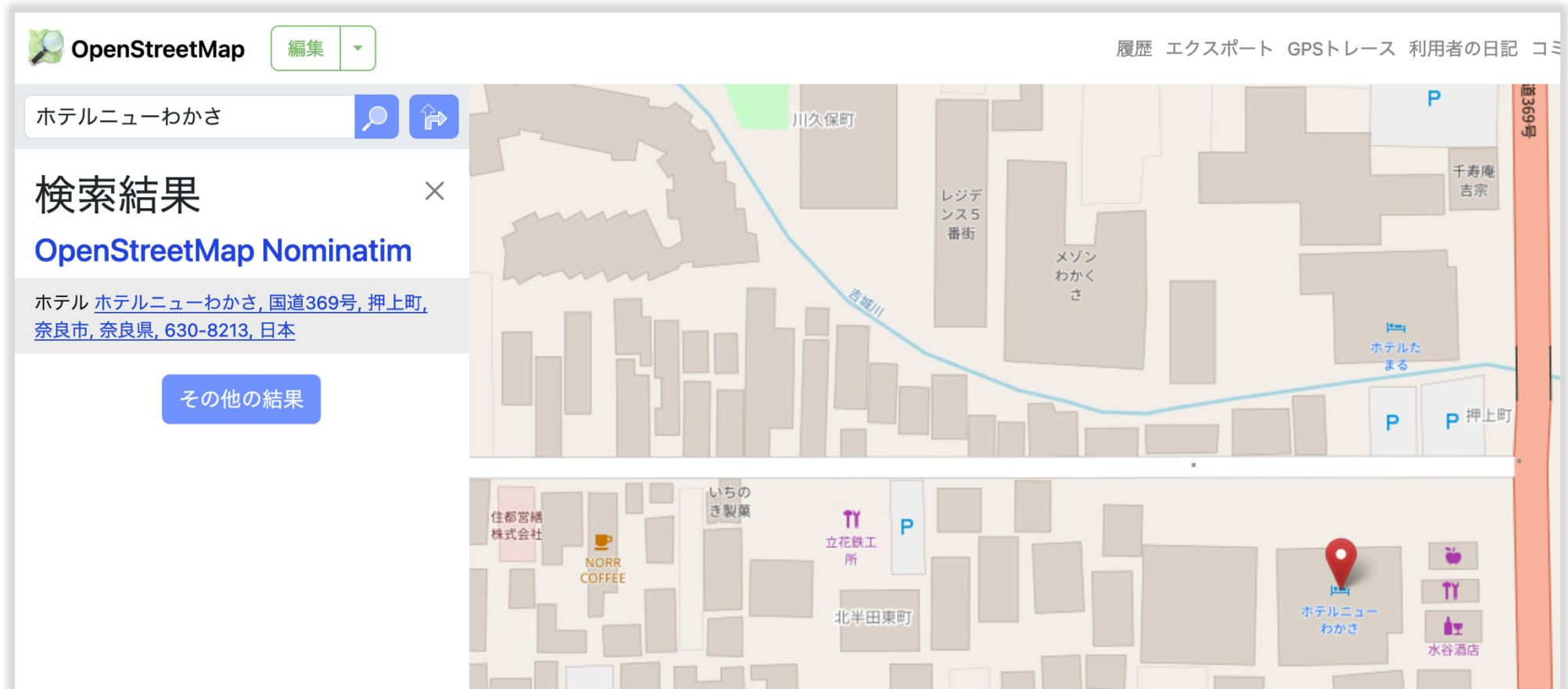
アノテーションデータ構築

OSM IDを付与

■ OpenStreetMap



を使ってアノテーション作業



<https://www.openstreetmap.org>

各処理を機械学習モデルで実現するには？

アノテーションデータの記述統計

	#記事	#文	#単語	#場所参照表現
Set-A	100	5,949	85,741	6,052
Set-B	100	6,324	87,074	6,119
Total	200	12,273	172,815	12,171

Set-A: 場所参照表現のみ付与

Set-B: OSM_ID も付与

※Set-AにOSM_IDを付与できなかったのは予算の関係

既存研究におけるジオパーキングの課題

場所参照表現の網羅性が低い

■ 既存のジオパーキングデータセットのまとめ

Dataset Name		Lang	Text Genre	Geo-DB	#Men	Facility	Nominal
LGL Corpus	(Lieberman et al., 2010)	en	News	GeoNames	4.8K	✗	✗
TR-News	(Kamalloo and Rafiei, 2018)	en	News	GeoNames	1.3K	✗	✗
GeoVirus	(Gritta et al., 2018)	en	News	Wikipedia	2.2K	✗	✗
GeoWebNews	(Gritta et al., 2020)	en	News	GeoNames	2.7K	△	✓
SemEval-2019 T12	(Weissenbacher et al., 2019)	en	Science	GeoNames	8.4K	✗	✗
CLDW	(Rayson et al., 2017)	en	Historical	Unlock	3.7K	△	✗
GeoCorpora	(Wallgrün et al., 2018)	en	Microblog	GeoNames	3.0K	△	✗
LRE Corpus	(Matsuda et al., 2017)	ja	Microblog	ISJ & Orig.	1.0K	△	✓
ATD-MCL	(Ours)	ja	Travelogue	OSM	12.3K	✓	✓

Table 10 in [Higashiyama+ 2024]

既存研究におけるジオパーキングの課題

場所参照表現の網羅性が低い

- 我々のデータセットでは多くの施設名をカバーしている

これまでは
対象の施設名が
少なかった

Dataset Name		Lang	Text Genre	Geo-DB	#Men	Facility	Nominal
LGL Corpus	(Lieberman et al., 2010)	en	News	GeoNames	4.8K	✗	✗
TR-News	(Kamalloo and Rafiei, 2018)	en	News	GeoNames	1.3K	✗	✗
GeoVirus	(Gritta et al., 2018)	en	News	Wikipedia	2.2K	✗	✗
GeoWebNews	(Gritta et al., 2020)	en	News	GeoNames	2.7K	△	✓
SemEval-2019 T12	(Weissenbacher et al., 2019)	en	Science	GeoNames	8.4K	✗	✗
CLDW	(Rayson et al., 2017)	en	Historical	Unlock	3.7K	△	✗
GeoCorpora	(Wallgrün et al., 2018)	en	Microblog	GeoNames	3.0K	△	✗
LRE Corpus	(Matsuda et al., 2017)	ja	Microblog	ISJ & Orig.	1.0K	△	✓
ATD-MCL	(Ours)	ja	Travelogue	OSM	12.3K	✓	✓

Table 10 in [Higashiyama+ 2024]

OpenStreetMap (OSM) を利用することで、旅行記に登場する施設名の3/4をカバーできるようになった

対象の施設名が
多くなった

場所参照表現の網羅性が低い

一般名詞語句や
指示語が
対象外だった

- 我々のデータセットでは固有名以外も対象としている

Dataset Name		Lang	Text Genre	Geo-DB	#Men	Facility	Nominal
LGL Corpus	(Lieberman et al., 2010)	en	News	GeoNames	4.8K	✗	✗
TR-News	(Kamalloo and Rafiei, 2018)	en	News	GeoNames	1.3K	✗	✗
GeoVirus	(Gritta et al., 2018)	en	News	Wikipedia	2.2K	✗	✗
GeoWebNews	(Gritta et al., 2020)	en	News	GeoNames	2.7K	△	✓
SemEval-2019 T12	(Weissenbacher et al., 2019)	en	Science	GeoNames	8.4K	✗	✗
CLDW	(Rayson et al., 2017)	en	Historical	Unlock	3.7K	△	✗
GeoCorpora	(Wallgrün et al., 2018)	en	Microblog	GeoNames	3.0K	△	✗
LRE Corpus	(Matsuda et al., 2017)	ja	Microblog	ISJ & Orig.	1.0K	△	✓
ATD-MCL	(Ours)	ja	Travelogue	OSM	12.3K	✓	✓

Table 10 in [Higashiyama+ 2024]

松島に行くには、空港から電車で仙台駅に向かい、
小牛田行きの電車に乗り換え、松島駅で降ります。

場所に関する
言語表現全般を
対象に含める

💡 「松島」までの経路を特定したい場合、一般名詞句「空港」も抽出する必要がある

ジオパージング

機械学習モデルの構築

アノテーションデータをモデルの学習と評価に使用

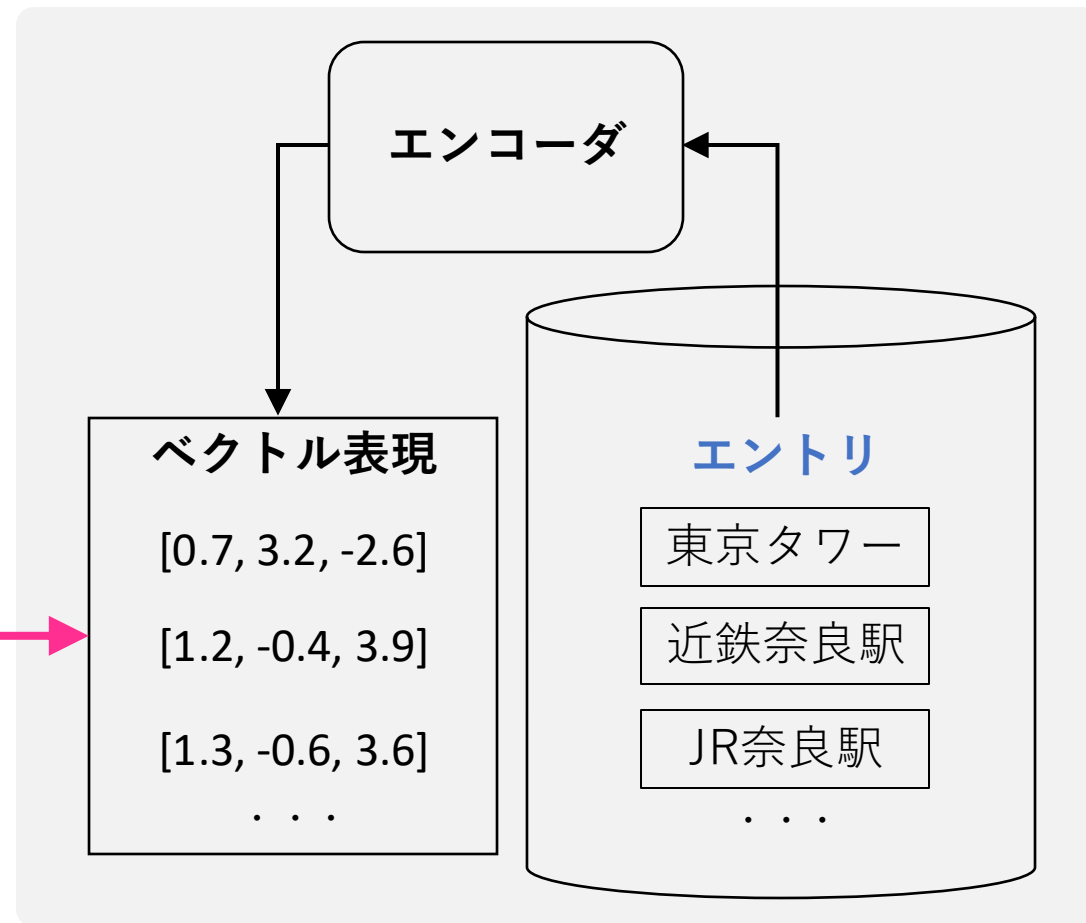
ベクトル検索の処理過程

②場所参照表現をベクトル化



③ベクトル間の類似度計算 → 類似度でランク付け

①地図DBの各エントリをベクトル化



※エンコーダとして我々の研究では BERT を使用

性能評価実験

- 評価指標 $R@N$: トップN個の出力に正解エントリが含まれる割合

手法	$R@1$	$R@5$	$R@10$
文字列マッチ	.338	.618	.700
ベクトル検索	.607	.842	.877

- ベクトル検索のほうが、正解エントリを上位にランク付けできている
- 上位予測エントリのより精緻な予測（リランキング）が今後の課題

ジオコーディング 構築したモデルの解析例

解析をしたい文章を入力してください

東京プリンスホテルから東京タワーに行きました。

実行

解析結果

1. 東京プリンスホテル
2. 東京タワー

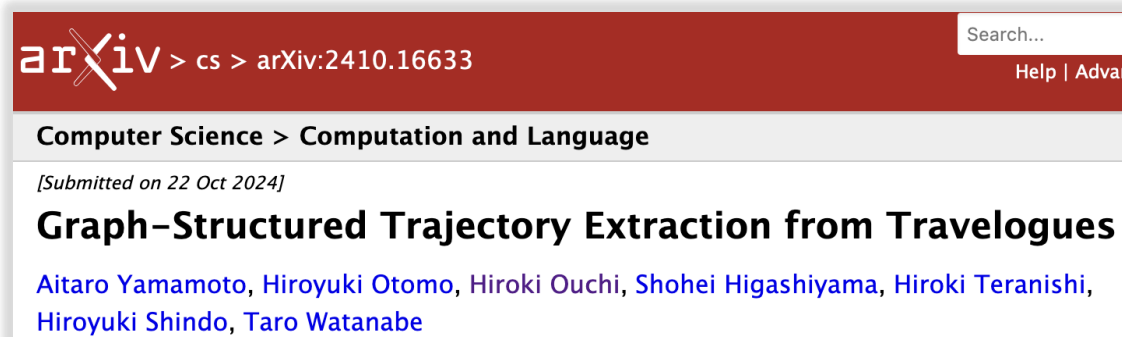


📍 は場所を表すマーカであり、📍 はエリアを表すマーカである

文章中の人物の地理的な移動をとらえる

文章のための移動軌跡解析

詳しい解説は以下の論文を参照



[Yamamoto+ arXiv 2024, *Graph-Structured Trajectory Extraction from Travelogues*](#)

これまでの研究

移動軌跡を場所の「系列」として表現

■ 移動軌跡の「発着地」の抽出

[Ishino+ 2012]

「出発地」「到着地」と「交通手段」をSNS投稿から自動抽出

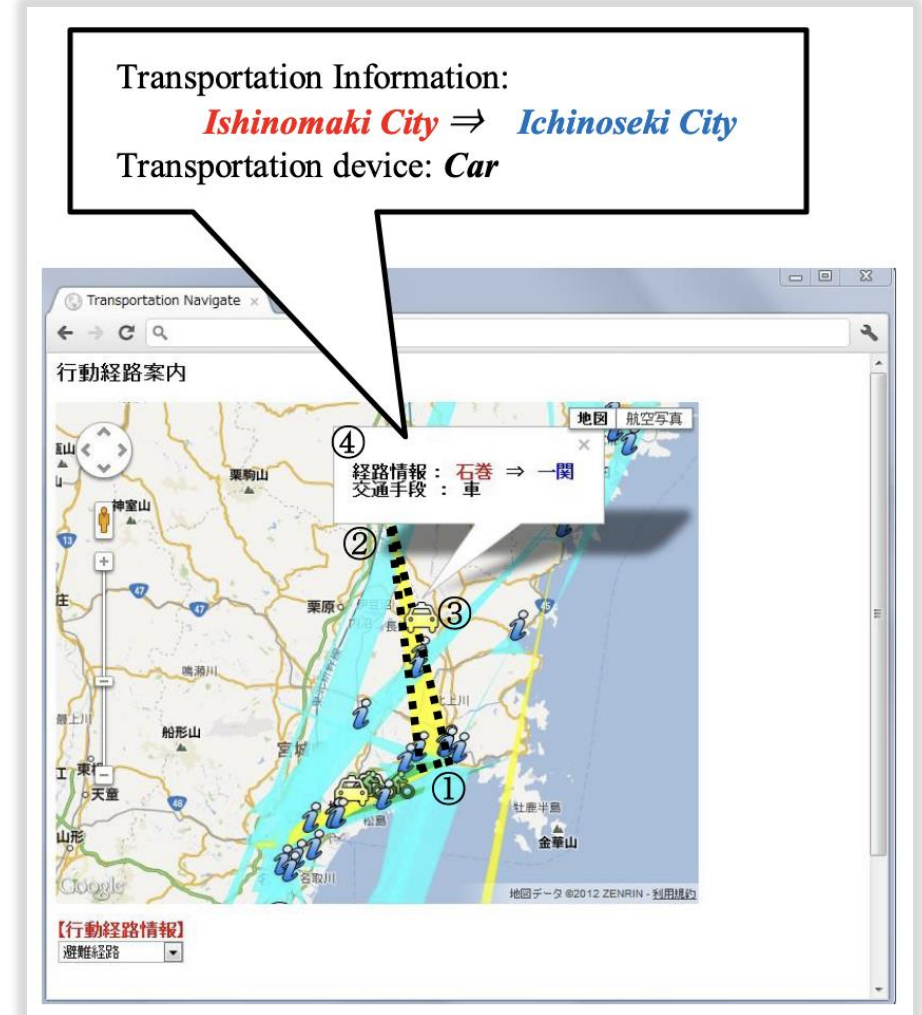
■ 移動軌跡全体を抽出

[Wagner+ 2023]

ホロコースト生存者の証言からあるイベントが起こった場所の系列を抽出

[Kori+ 2023]

旅行ブログからユーザの行動経路(移動軌跡)を抽出



[Ishino+ 2012]のFigure 1からの引用

これまでの研究

移動軌跡を場所の「系列」として表現

■ 移動軌跡の「発着地」の抽出

[\[Ishino+ 2012\]](#)

「出発地」「到着地」と「交通手段」をSNS投稿から自動抽出

■ 移動軌跡全体を抽出

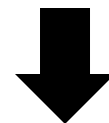
[\[Wagner+ 2023\]](#)

ホロコースト生存者の証言からあるイベントが起こった場所の系列を抽出

[\[Kori+ 2023\]](#)

旅行ブログからユーザの行動経路(移動軌跡)を抽出

移動軌跡を
1本の系列として表現



それでいいのか？

文章の移動軌跡は系列で表現しづらい

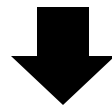
入力文章

その日は京都市から奈良市に向かいました。京都駅からは京都タワーが見えました。近鉄奈良駅で降り、東大寺へ少し歩いて行きました。大仏堂はとても雄大でした。

■ 移動軌跡は一本の系列として描けない場合も多い

➤ 地理的な包含関係を適切に表せない

例) 「京都市」と「京都駅」, 「東大寺」と「大仏堂」など



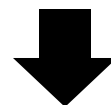
提案

地理的包含関係を考慮した**グラフ**として定義

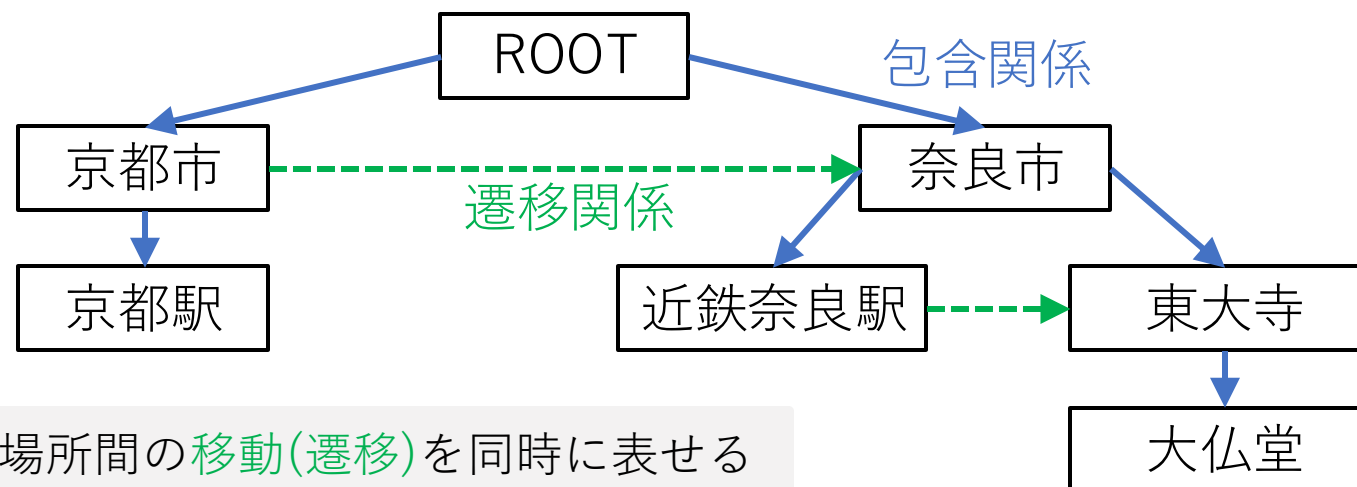
移動軌跡を表現するグラフ構造

入力文章

その日は京都市から奈良市に向かいました。京都駅からは京都タワーが見えました。近鉄奈良駅で降り、東大寺へ少し歩いて行きました。大仏堂はとても雄大でした。



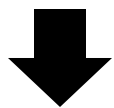
提案：訪問順序グラフ



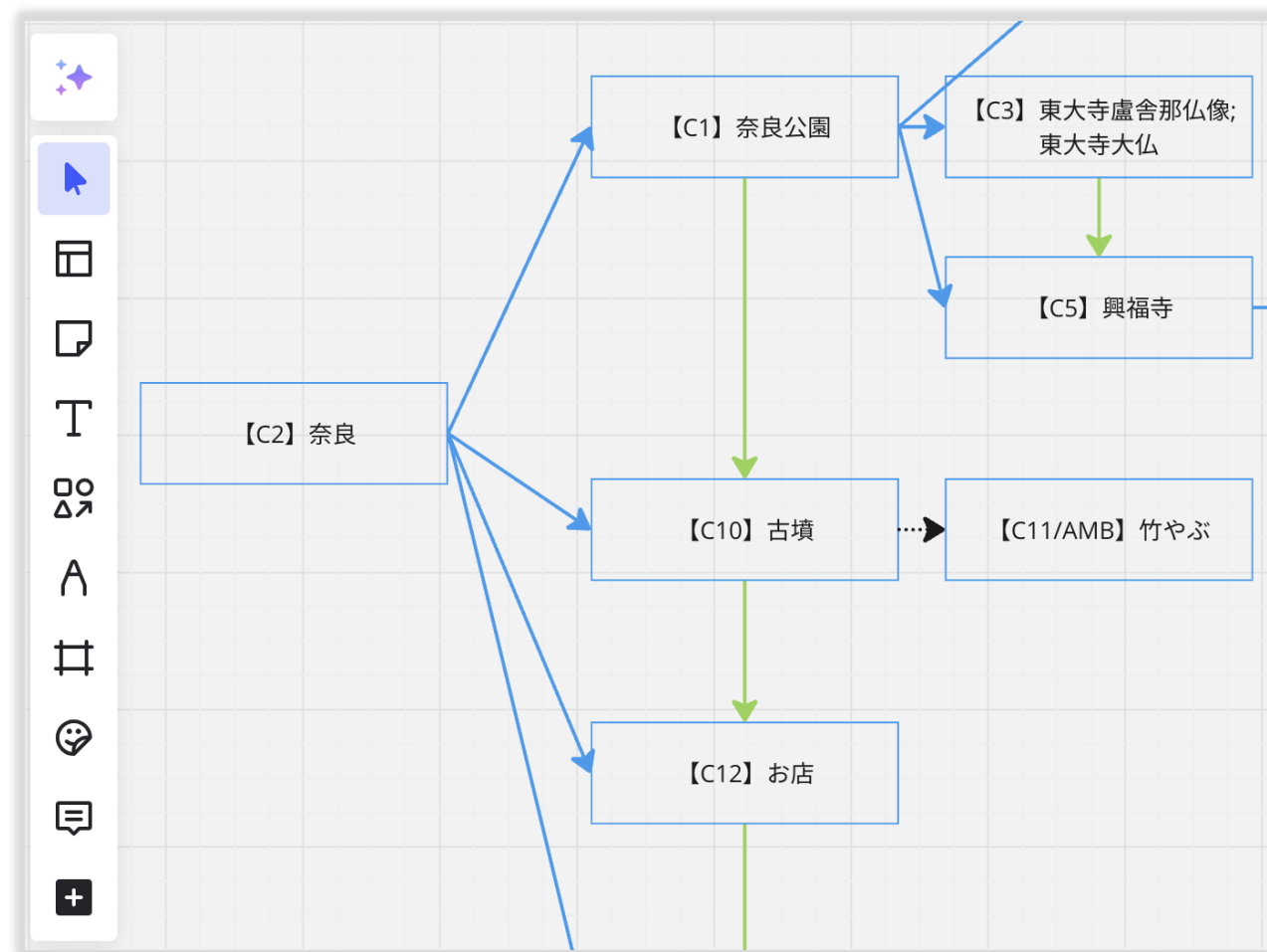
地理的な包含関係と場所間の移動(遷移)を同時に表せる

グラフ構造をアノテーション

-  **miro** を使って
アノテーション作業
 - 場所参照表現などは
アノテーション済み
 - 各場所をノードとして、
ノード間に**包含関係**や
遷移関係をエッジで結ぶ



#包含関係	#遷移関係
1,863	1,506



作業画面例

性能評価実験

- Masked Language Models の LUKE をファインチューニング
- Causal Language Models の ELYZA と Swallow は 0-shot

モデル	#包含関係	#遷移関係
LUKE	0.355	0.748
Llama3-ELYZA	0.115	0.456
Llama3-Swallow	0.132	0.388

- ファインチューニングした LUKE が最も良い結果
- 入力文章だけを手がかりに包含関係を予測することは難しい
 - ジオコーディングと組み合わせる必要性が示唆された

まとめ

一次データの高付加価値化のためには

「二次データ(アノテーションデータ)」が鍵

ソフトウェア層

地名
抽出器

同一地点
認識器

地図DB
リンク器

訪問状態
予測器

地理的包含
関係予測器

訪問順序
予測器

旅行記
翻訳器

二次データ層

地理情報データ

訪問状態・順序データ

英語翻訳データ

一次データ層

地球の歩き方旅行記データセット

一次データの高付加価値化のためには

「二次データ(アノテーションデータ)」が鍵

本講演で提供したい話題

データ提供者の視点: 一次データをもっと価値あるものにするには？

データ使用者の視点: 一次データを「工夫」して研究開発の幅を広げるには？

「地球の歩き方旅行記データセット」の実践例を通じて

Take-Home Message

- (1) 解きたい重要な問題を考え、(2) その解き方を考え、
- (3) それを実現するための**アノテーションデータ**を構築しよう！

「地球の歩き方旅行記データセット」の研究プロジェクト リンク集

■ 一次データ

- 地球の歩き方旅行記データセット

<https://www.nii.ac.jp/dsc/idr/arukikata/>

■ 二次データ

- ATD-MCL (ジオパーキング)
- ATD-VSO (訪問グラフ)
- ATD-MCL-Overseas (ジオパーキング海外旅行記)
- ATD-Para(仮) (英語翻訳データ)

<https://github.com/naist-nlp/atd-mcl>

Coming soon!

<https://github.com/naist-nlp/atd-mcl-overseas>

Coming soon!

■ ソフトウェア

- 場所参照表現抽出モデル
 - HuggingFace
 - Google Colab
- ジオコーディングモデル
- 訪問グラフ予測モデル

<https://huggingface.co/naist-nlp/mluke-large-lite-finetuned-atd>

<https://colab.research.google.com/drive/1BI2hxfOzxyT7mET8OzJ7RbF7zFCvqSEo?usp=sharing>

Coming soon!

Coming soon!

付録

地理空間情報

「地理空間情報」 とは？

平成十九年法律第六十三号

地理空間情報活用推進基本法

(定義)

第二条 この法律において「地理空間情報」とは、第一号の情報又は同号及び第二号の情報からなる情報をいう。

- 一 空間上の特定の地点又は区域の位置を示す情報（当該情報に係る時点に関する情報を含む。以下「位置情報」という。）
- 二 前号の情報に関連付けられた情報

「地理空間情報」とは？



地理空間情報ライブラリー

Geospatial Information Library

ホーム

お知らせ

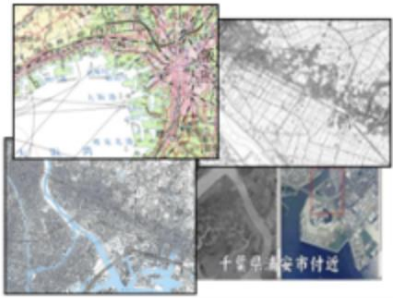
ライブラリーについて

地図・空中写真の総合ライブラリー

地理空間情報ライブラリーには、国土地理院の地図・空中写真及び国・地方公共団体が整備した図面等が登録されており、インターネットを通じて検索・閲覧・入手が行えます。

研究プロジェクト紹介

「地理空間情報」とは？



地図・空中写真閲覧サービス

地図や空中写真を検索して閲覧するサービスです。

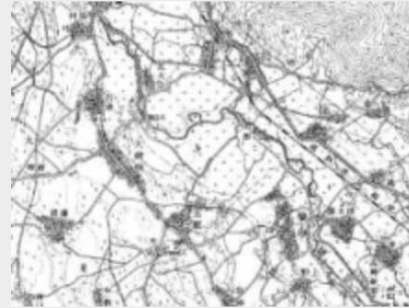
もっと見る



地理院地図（電子国土Web）

インターネットで地図や空中写真等を閲覧することができます。

もっと見る



図歴（旧版地図）

国土地理院で図・地勢図の図歴を閲覧することができます。



ハザードマップポータルサイト

防災に役立つ様々な情報を閲覧することができます。



自然災害伝承碑

過去に発生した自然災害の被害等を伝える石碑やモニュメントです。

研究プロジェクト紹介

「地理空間情報」 とは？



https://commons.wikimedia.org/wiki/File:Hasedera,%E9%95%B7%E8%B0%B7%E5%AF%BA_%E6%9C%AC%E5%A0%82_-_panoramio_-_z_tanuki.jpg?uselang=ja

研究プロジェクト紹介

「地理空間情報」 とは？

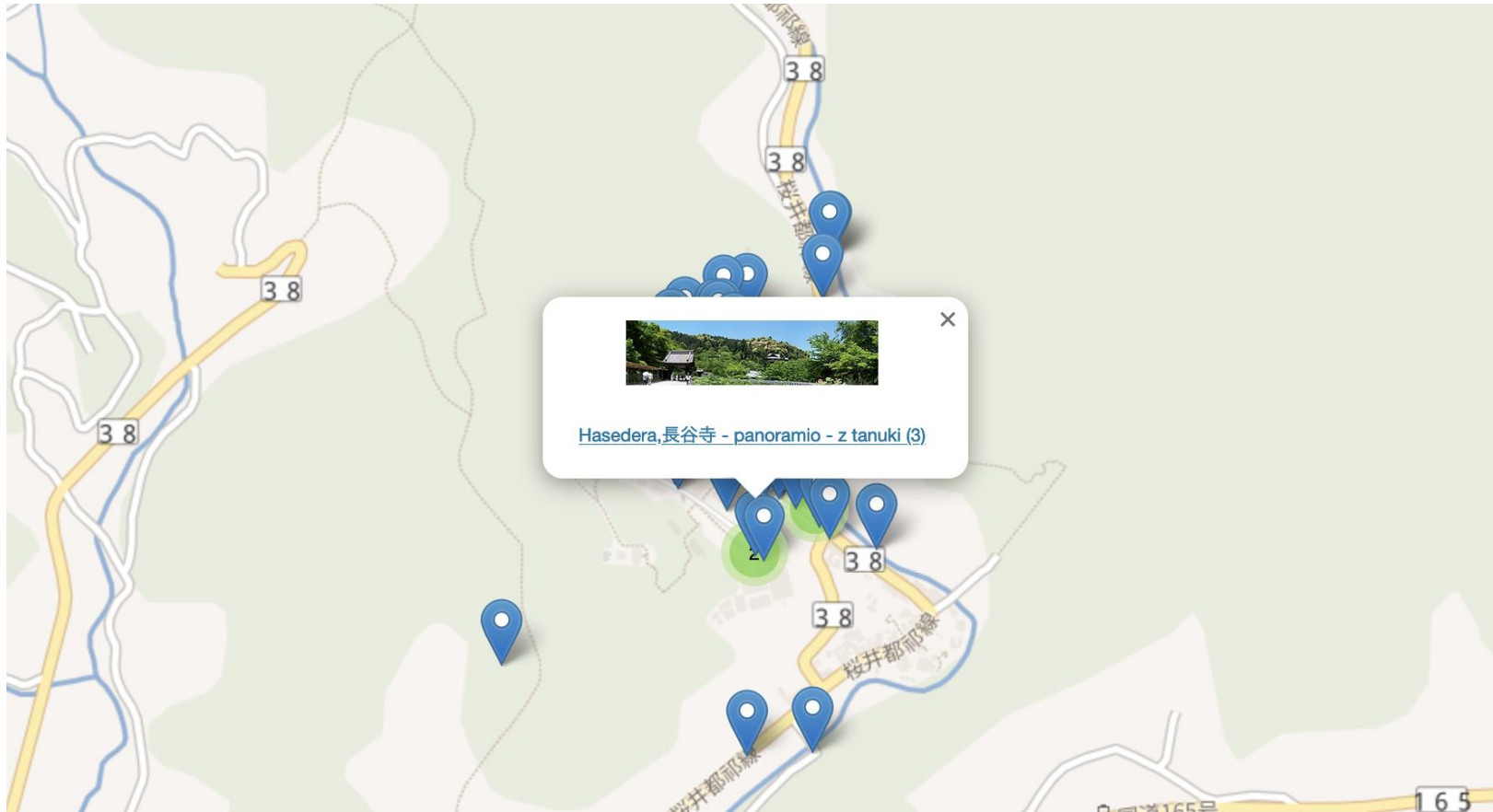


https://commons.wikimedia.org/wiki/File:Hasedera,%E9%95%B7%E8%B0%B7%E5%AF%BA_%E6%9C%AC%E5%A0%82_-_panoramio_-_z_tanuki.jpg?uselang=ja

カメラの位置

 34° 32' 01.96" 北, 135° 54' 27.24" 東 

「地理空間情報」 とは？



研究プロジェクト紹介

「地理空間情報」 とは？



<https://commons.wikimedia.org/wiki/File:Nachikatsuura-kinomatsushima85st2040.jpg?uselang=ja>

解説 English: Ki-no-Matsushima, Nachikatsuura, Wakayama prefecture, Japan
日本語: 紀の松島, 和歌山県那智勝浦町

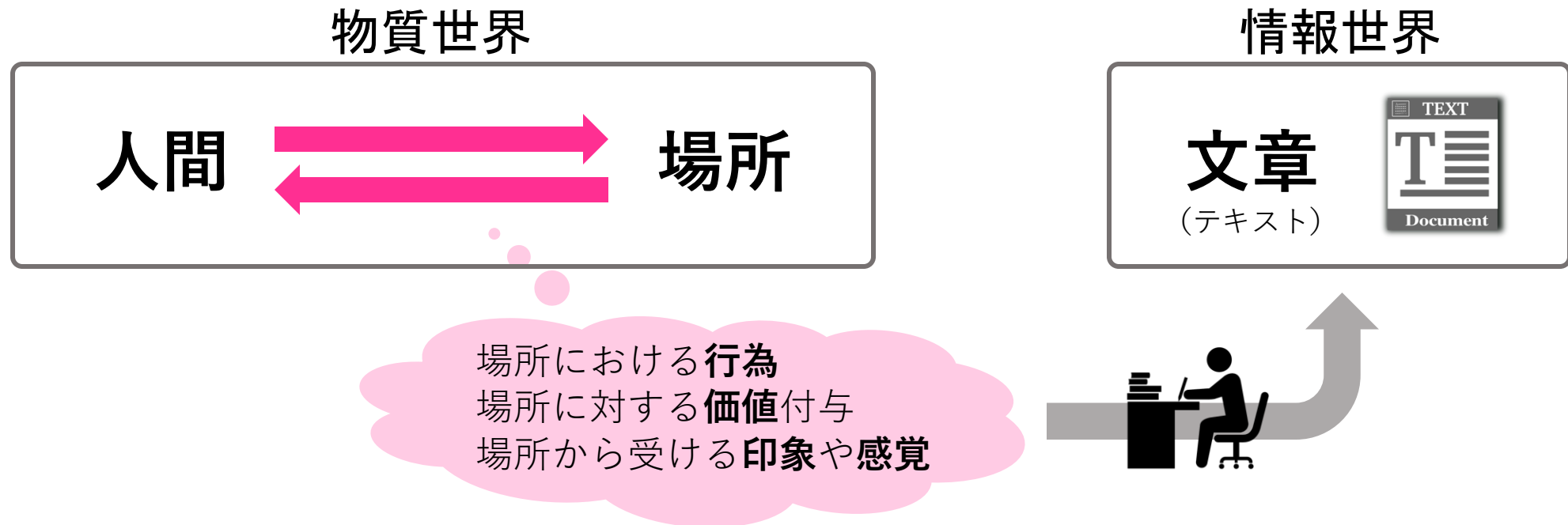
地球の歩き方旅行記データセット

なぜ「文章」に着目するのか？

「文章」は地理空間情報の鉱脈

■ GPSデータで十分では？

- 人間の位置を把握するだけならGPSデータで事足りる
- しかし「人間」と「場所」の相互関係を捉えるのはむずかしい
- 「文章」にはそういった情報も書かれる



共通利用可能な世界初の旅行記データセット

- これまでの課題：共通利用可能な旅行記データがなかった
 - それぞれの研究者がウェブ上の旅行記投稿サイトなどから独自に取得した旅行記データを用いることが多かった
 - そのため、研究の再現や実験結果の公平な比較分析が困難であった
- 本データセット：一定条件下での利用機会について**オープン化**
 - 研究に使用したデータが特定可能になる → 研究の**透明性・再現性の担保**
 - 他の研究との比較も格段に容易となる → 研究の**知見の蓄積が加速**

[大須賀ら 2021 <https://www.ipsj.or.jp/dp/contents/publication/46/TR0202-05.html>]

オープンサイエンスの促進

本データセットの特徴

旅行記の言及する地域

国内旅行記：全都道府県をカバー

都道府県名	記事数	割合
東京	1,206	25.81%
大阪	492	10.53%
沖縄	353	7.56%
北海道	336	7.19%
神奈川	313	6.70%
京都	299	6.40%
長野	295	6.31%
兵庫	242	5.18%
愛知	231	4.94%
千葉	219	4.69%

ベスト10

海外旅行記：150以上の国・地域をカバー

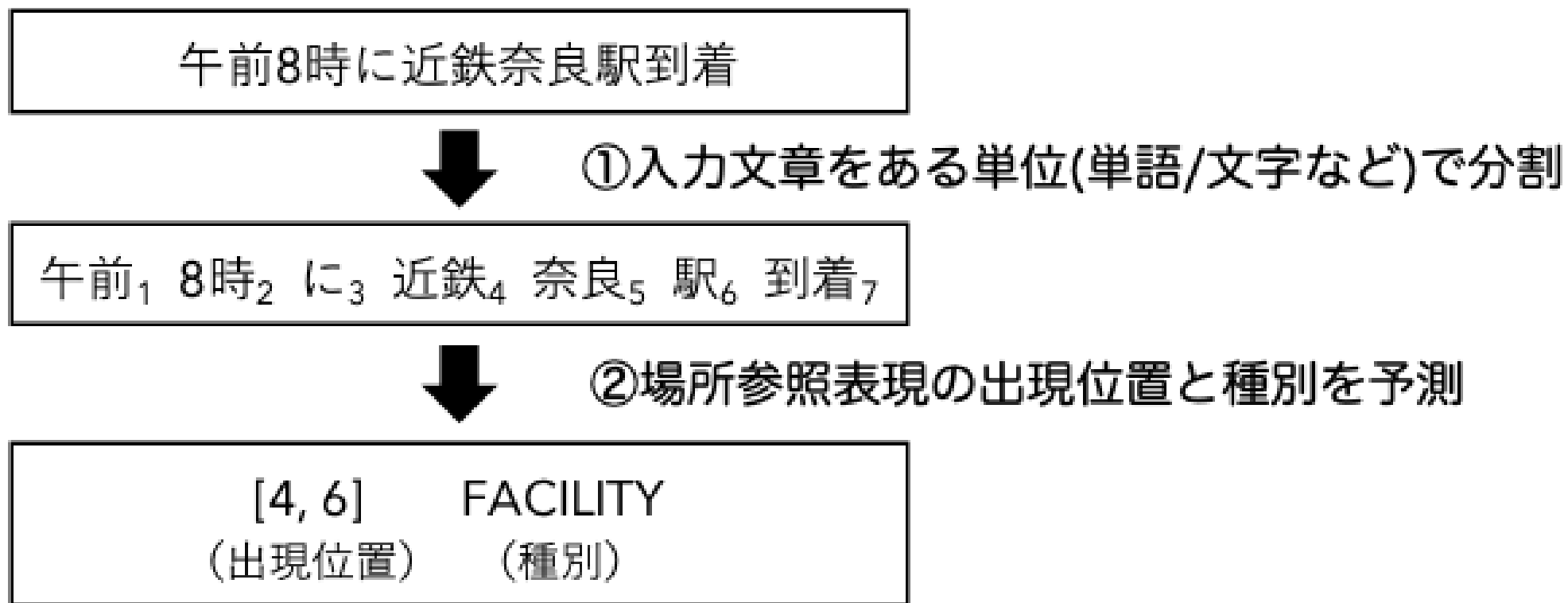
国・地域名	記事数	割合
アメリカ	732	7.62%
韓国	708	7.37%
フランス	697	7.26%
中国	692	7.20%
台湾	611	6.36%
ドイツ	577	6.01%
タイ	554	5.77%
イタリア	548	5.70%
スペイン	413	4.30%
スイス	411	4.28%

ベスト10

場所参照表現抽出

場所を表す言語表現の出現位置と種別を同定

- 場所を表す言語表現 = 「**場所参照表現**」と呼ぶ



場所参照表現抽出 = 固有表現抽出のひとつの特殊なタスク

性能評価実験

一致数 / システム予測事例数

一致数 / 正解事例数

種別ラベル	正解事例数	システム予測事例数	一致数	精度	再現率
LOC_NAME	881	916	760	0.830	0.863
LOC_NOM	349	377	282	0.748	0.808
FAC_NAME	1285	1230	1037	0.843	0.807
FAC_NOM	1135	1086	929	0.855	0.819
LINE_NAME	195	194	156	0.804	0.800
LINE_NOM	236	223	193	0.865	0.818
TRANS_NAME	148	123	87	0.707	0.588
TRANS_NOM	334	353	293	0.830	0.877

体感としてはほとんど正解できている
(まぎらわしい事例の種別ラベルを間違っている)

構築したモデルの解析例

以前から出雲に行きたいと思っていましたが、

LOC_NAME

中国山地を列車で越えて見たいとの思いもあり、

LOC_NAME TRANS_NOM

岡山経由で山陰に入ることになりました。

予測：FAC_NAME 予測：LOC_NOM

正解：LOC_NAME 正解：LOC_NAME

岡山空港から岡山駅に出て、

FAC_NAME

FAC_NAME

ここから特急やくもで出雲に向かいます。

DEICTIC

TRANS_NAME

LOC_NAME

「地球の歩き方旅行記データセット」の記事 (ID=23251)

正解箇所

不正解箇所

種別ラベル	例
LOC_NAME	奈良県; 生駒山
LOC_NOM	街; 島; 山
FAC_NAME	大神神社; 東京駅
FAC_NOM	駅; 公園; お店
LINE_NAME	近鉄奈良線
LINE_NOM	国道; 川; トンネル
TRANS_NAME	特急ひのとり
TRANS_NOM	バス; フェリー

8割以上の正確性で解析可能になった

解析誤りの例

■ 地名(**LOC_NAME**)と駅名 (**FAC_NAME**) の混同

特急"きりしま"は、鹿児島中央~宮崎を結ぶ、...

予測 : **LOC_NAME**

正解 : **FAC_NAME**

■ 字面が固有名(***_NAME**)っぽい一般名詞句(***_NOM**)を混同

輪王寺の金堂・三仏堂。

予測 : **FAC_NAME**

正解 : **FAC_NOM**

今後、学習データを増やすなどして対応していきたい

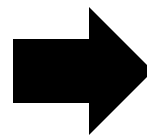
共参照解析

同じ場所を指す表現のグルーピング

- 共参照解析 = 同じ事物や概念を指す異なる言語表現をグルーピング
- 共参照関係(グループ) = 同じ事物や概念を指す表現のグループ

入力：文章と場所参照表現

午前8時に近鉄奈良駅到着。
ホテルニューわかさに9時チェックイン
予定。時間に余裕があったので、途中の
スタバで一服。
30分過ごしてから店を出て、ホテルに
チェックインしました。



出力：共参照グループ

Group ID: 1

近鉄奈良駅

Group ID: 2

ホテルニューわかさ
ホテル

Group ID: 3

スタバ
店

同じ場所を指す表現のグルーピング

■ 共参照解析の利点は？

- その場所に関する情報を**より広範な文脈**から取り出しやすくなる

入力：文章と場所参照表現

午前8時に近鉄奈良駅到着。

ホテルニューわかさに9時チェックイン
予定。時間に余裕があったので、途中の
スタバで一服。

30分過ごしてから店を出て、ホテルに
チェックインしました。

例) 「ホテルニューわかさ」にチェックインした
事実やタイミングを抽出したい場合

- 「ホテルニューわかさ」が登場する文を見ただけでは特定できないが、「ホテル」が登場する文を根拠として特定可能
- もし「ホテルニューわかさ」と「ホテル」をグループ化できていない場合、「ホテル」に関する情報を「ホテルニューわかさ」に関する情報として取り出すことは困難である

各処理を機械学習モデルで実現するには？

アノテーション(付加)情報が必要

- 「同じ場所(同一エンティティ)を指す表現の認識」に必要な情報とは？

途₁ 中₂ の₃ ス₄ タ₅ バ₆ で₇ ー₈ 服₉ 。 ₁₀

3₁₁ 0₁₂ 分₁₃ 過₁₄ ご₁₅ し₁₆ て₁₇ か₁₈ ら₁₉ 店₂₀ を₂₁ 出₂₂ て₂₃ 、 ₂₄

ホ₂₅ テ₂₆ ル₂₇ に₂₈ チ₂₉ エ₃₀ ツ₃₁ ク₃₂ イ₃₃ ン₃₄ し₃₅ ま₃₆ し₃₇ た₃₈ 。 ₃₉

[4, 6] ENT_ID=1

[20, 20] ENT_ID=1

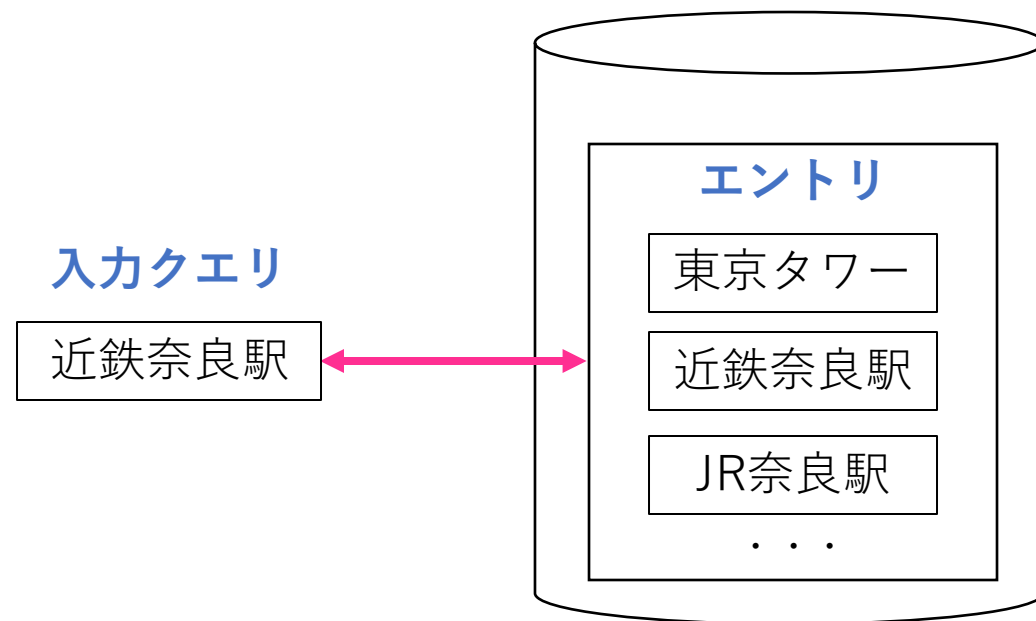
[25, 27] ENT_ID=2

- 各場所参照表現に「ENT_ID (エンティティID)」を追加
(同じ場所を指す場合は同じIDを付与)

ジオコーディング

地図DBの検索問題として取り組む

- 地図DBの各エントリには経緯度情報が付与されている
 - 入力クエリに対して適切なエントリを検索できれば良い
- 主流のアプローチのひとつは「文字列マッチ」による検索
 - 典型的には**入力クエリ**とDB側の各**エントリ**との文字列の**類似度**を計算



「ベクトル検索」を採用

■ ベクトル検索の概要

- 典型的には**入力クエリ**とDB側の各**エントリ**の**ベクトル間類似度**を計算
- ここでいう「**ベクトル**」は「**数値のリスト**」と考えてよい
 - 例) 3次元のベクトル表現: [1.2, -0.4, 3.9]

■ (事前学習済み) ベクトルを使うことの**利点**

- 「**意味的類似性**」を捉えられる (と期待できる)
- 例) 「仙台」は, 「宮城」と「宮崎」ならどちらのほうが類似度が高いか?
 - 「宮城」だと期待される
 - 理由: 自然言語処理におけるベクトル表現の作り方が, 登場する文脈が似ている語句同士の類似度が高くなるように学習する場合が多いから

解析誤りの傾向

■ 同名エントリ問題

- たとえば「興福寺」 もたくさんある（以下，OSMのIDを列挙）

- node/2787766967 (広島県)
- node/3382683779 (静岡県)
- node/3828906063 (長崎県)
- node/6364194307 (大阪府)
- node/8663676920 (宮城県)
- way/1134439456 (奈良県)

...

同名エントリから適切なものを検索できていない
現在，位置情報や文脈情報を考慮するように改良中

訪問状態予測

訪問行動を予測する意義

- 著者が訪れた場所とそうでない場所を区別したい場合に有用

奈良公園にはたくさんの鹿がいました。



「奈良公園」を著者が
訪れたと解釈できる

奈良公園は人気の観光スポットです。



一般的な事実を表しており
著者の訪問を意味しない

場所に対する訪問状態（訪れた、訪れていない、訪れる予定、など）
を判定するタスクに取り組む

機械学習モデルの訪問行動予測の性能調査

■ リサーチクエスチョン

訪問行動に関する文脈上の解釈・判断を、
機械学習モデルはどの程度正確に行えるだろうか？

■ 機械学習モデルとして最近の言語モデルを想定

- 最近の言語モデル: 文章理解タスクで高い正確性
- 代表的な2種類の言語モデルを調査

A) Masked Language Model (MLM)

例) BERT, RoBERT, LUKE

B) Causal Language Model (CLM)

例) GPT, Llama, Mistral

訪問状態ラベルをデザイン

- | | |
|------------------|--|
| 1. Visit | 無事に <u>赤岳山頂</u> に着きました！ |
| 2. PlanToVisit | <u>穂高神社</u> に向かいます。 |
| 3. See | <u>硫黄岳</u> が近くに見えて来ました。 |
| 4. Visit-Past | 数年前に訪れた <u>フランス・ルルドの泉</u> を思い出しました。 |
| 5. Visit-Future | 今度は、 <u>松江</u> と <u>米子</u> に来てみたいものです。 |
| 6. UnkOrNotVisit | <u>糸魚川駅</u> 行きの車両は1両です。 |

※ Unk = Unknown

訪問状態ラベルの種類別事例数

- [ATD-MCL \(Arukikata Travelogue Dataset with Geographic Entity Mention, Coreference, and Link Annotation\)](#) に対して、新たに訪問状態ラベルを付与した

- ATD-MCLとは？

- 「[地球の歩き方旅行記データセット \(Arukikata Travelogue Dataset\)](#)」に場所参照表現等の情報を付与したデータセット

	1.Visit	2.Plan	3.See	4.Past	5.Future	6.UN
学習データ	2,577	358	212	10	6	619
開発データ	332	48	46	1	4	74
評価データ	748	121	59	10	4	160

場所参照表現に訪問状態ラベルを付与

入力: 穂高神社に向かいます。

機械学習モデル

出力: PlanToVisit

MLMとCLMによる解法

穂高神社に向かいます。

LUKE (MLM)

多クラス分類

- | | |
|------------------|-----|
| 1. Visit | 9% |
| 2. PlanToVisit | 64% |
| 3. See | 6% |
| 4. Visit-Past | 1% |
| 5. Visit-Future | 18% |
| 6. UnkOrNotVisit | 2% |

確率最大の
ラベルを選択

PlanToVisit

穂高神社に向かいます。

Llama (CLM)

言語生成

Plan

To

Visit

1単語(サブワード)ずつ生成

PlanToVisit

訪問状態予測の性能評価

- どちらのモデルも同等の性能

モデル	正解率
LUKE (MLM)	78.9%
Llama (CLM)	78.1%

- パラメタ数の観点ではLUKEの方が効率的
 - LUKEは5.6億パラメタ（ノートパソコンでも遅いけど動くレベル）
 - Llamaは70億パラメタ（大規模計算機じゃないときびしいレベル）
- 解き方の違いが関係してると考えられる
 - 一般に多クラス分類より言語生成の方が難しい

ラベルごとの性能評価

	事例数	LUKE	Llama
Visit	332	93.0%	90.3%
PlanToVisit	48	66.6%	47.9%
See	46	76.0%	82.6%
Visit-Past	1	0.0%	0.0%
Visit-Future	4	0.0%	0.0%
UnkOrNotVisit	74	54.0%	39.1%

- 最も多いラベル Visit については9割以上正解できている
- See についても8割程度正解できている
- PlanToVisit と UnkOrNotVisit については改善の余地あり

ラベルごとの正解事例数

	両モデル 正解	LUKEだけ 正解	Llamaだけ 正解	両モデル 不正解
LUKE Llama	○ ○	○ ×	× ○	× ×
Visit	287	22	13	10
PlanToVisit	18	14	5	11
See	34	1	4	7
Visit-Past	0	0	0	1
Visit-Future	0	0	0	4
UnkOrNotVisit	28	12	1	33

- Visit については両方のモデルで正解できている場合が多い
- UnkOrNotVisit については両方とも不正解の場合も多い

両方のモデルで不正解だった事例の傾向

■ UnkOrNotVisit に着目

ホテルの目の前の気になるお店ですが、今回はパスしました。

予測: Visit

正解: UnkOrNotVisit

今回は、松江と米子は素通りします。

予測: Visit

正解: UnkOrNotVisit

- 「～パスしました」「～素通りします」等の語句の意味を正確に捉えられなかった？

機械学習モデルによる訪問行動の予測

■ リサーチクエスチョン

訪問行動に関する文脈上の解釈・判断を,
機械学習モデルはどの程度正確に行えるだろうか？

■ 機械学習モデルとして最近の言語モデルを想定

- MLMとCLMで同等の性能(正解率80%弱)
- 「訪問 (visit)」は正解率90%超
- 「訪問不明/非訪問 (UnkOrNotVisit)」の認識に改善の余地あり

今後の展望（現在進行中）

訪問順序の予測

入力: 文章

午前8時 **近鉄奈良駅** 到着。
Visit①

ホテルニューわかに
PlanToVisit

9時 チェックイン予定。

時間に余裕があったので、

途中の**スタバ**で一服。
Visit②

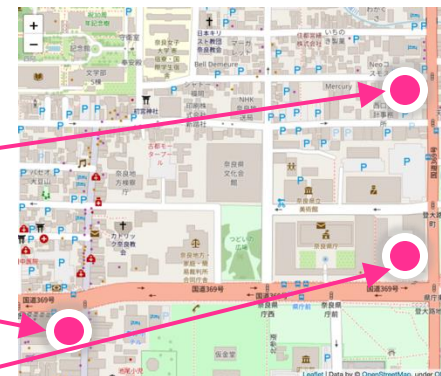
30分過ぎしてから **店**を出て、

ホテルにチェックインして。
Visit③

2023年

本発表

現在進行中

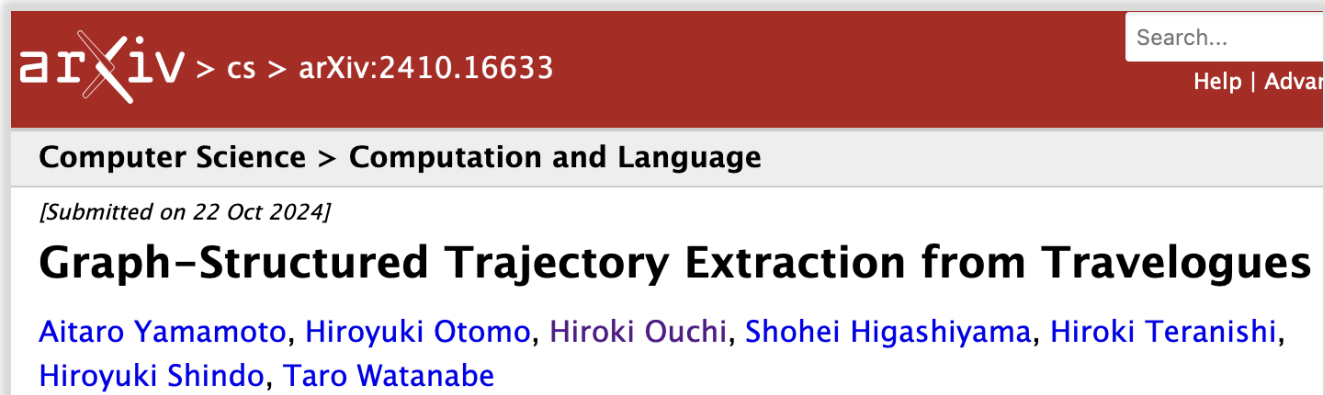


地図データベース
OpenStreetMap

- ① **場所を表す表現**を抽出
- ② 同じ場所を指す表現の認識
- ③ 各場所の表現と 地図上の対応箇所をつなぐ 
- ④ 各場所を訪問したか否かを予測 **Visit/PlanToVisit/See...**
- ⑤ 訪問場所を時系列順に整序 **①, ②, ③...**

今後の展望（現在進行中） 訪問順序の予測

- 基盤技術に関する論文を公開しました



<https://arxiv.org/abs/2410.16633>



- 実システムの開発を頑張ります

