文章から地理空間を取り出す
－ 場所を表す言語表現の抽出と地図データベースへの接続 －

大内啓樹*・中谷響・東山翔平***・寺西裕紀****・渡辺太郎*****

Geographical Space in Text
- Extraction of Location Reference Expressions and Connection to Map Databases -

Hiroki Ouchi*, Hibiki Nakatani**, Shohei Higashiyama***, Hiroki Teranishi****, Taro Watanabe*****

Abstract: Text contains vast amounts of location information, including descriptions of human experiences and actions at certain places, impressions received from specific locations, and events occurring in specific locations. Enabling the extraction of such unstructured information and its conversion into well-structured data for use in Geographic Information Systems paves the way for realizing various applications related to geographical spaces. As a first step toward this goal, this study aims to improve the technology of "geoparsing," which serves as a bridge between "the world of texts" and "the world of maps." Specifically, we propose a framework for developing geoparsing systems based on state-of-the-art Natural Language Processing technologies. On the basis of the framework, we have built a specific geoparsing system and investigated the performance of the system.

Keywords: ジオパージング (Geoparsing), 自然言語処理 (Natural Language Processing), 地理情報システム (GIS), 場所参照表現 (Location Reference Expressions), 記号接地 (Symbol Grounding)

１．はじめに

文章で書かれた「場所」を，地理情報システム (GIS) は活用できているだろうか？

文章には，場所に関する豊かな情報が含まれる．ある場所での人間の経験や行動，ある場所から受けた印象，ある場所に存在した事物，ある場所で起こった事象．こうした情報を文章から取り出し，GIS での活用を通じて，新たな価値の創出を目指す．

その第一歩として，「文章の世界」と「地図の世界」の架け橋となる技術「**ジオパージング (Geoparsing)**」の高度化に取り組む．この技術は，ふたつのステップから構成される (Leidner, 2006; Gritta et al., 2020).

(1) 場所を表す言語表現（**場所参照表現**）を文章から抽出する
(2) その表現が表す場所の経緯度を推定する

図 1　文章中の人物の地理的な移動を読み解き，その移動軌跡を地図上に描き出すイメージ

* 正会員　奈良先端科学技術大学院大学・先端科学研究科 (Nara Institute of Science and Technology)
〒630-0101 奈良県生駒市高山町８９１６−５ **hiroki.ouchi@is.naist.jp**
** 非会員　奈良先端科学技術大学院大学・先端科学研究科 (Nara Institute of Science and Technology)
*** 非会員　情報通信研究機構 (National Institute of Information and Communications Technology)
**** 非会員　理化学研究所・革新知能統合研究センター (RIKEN AIP)
***** 非会員　奈良先端科学技術大学院大学・先端科学研究科 (Nara Institute of Science and Technology)

この技術によって，たとえば「近鉄奈良駅」という文字列は，地図上における 1 地点（経緯度），あるいは，領域（複数の経緯度を結んでできる面）を獲得する．つまり，場所を表す文字列に「位置情報を吹き込む技術」と言える．本稿では，最先端の自然言語処理技術を考慮したジオパージングシステムの開発フレームワークを提案し，実際にその実装および性能について報告する．本システムの実装は，我々の研究プロジェクトページ[1] にて広く一般公開する．

1.1. 背景

我々の「地理空間情報と自然言語処理」プロジェクト[2] では，文章中の人物の地理的な移動を読み解き，その移動軌跡を地図上に描き出すシステムの開発を目指している．その処理イメージを図 1 に示す．入力文章には，著者が「近鉄奈良駅」に到着した後，「スターバックス奈良公園店（スタバ）」に寄り，最終的に「ホテニューわかさ」に辿り着いたことが書かれている．開発予定のシステムは，この移動軌跡を描画した地図を出力することを目標とする．

こうしたシステムを実現するための根幹となる技術がジオパージングである．ジオパージングの応用先は多様である．Hu et al. (2022) は，ジオパージングの 7 つの応用領域をまとめている．たとえば，災害対応や感染症サーベイランスがある．前述したように，ジオパージングは，「文章の世界」と「地図の世界」をつなぎ，多様な応用を可能にする点で極めて重要な技術であると言える．

1.2. 技術的課題と解決するためのアプローチ

多様な応用を可能にするには，ジオパージングの高度化が必要である．克服すべき重要な技術的課題として，**場所参照表現の網羅性の低さ**が挙げられる．ジオパージングに関する従来研究では，地名のみを対象としたものが多く，施設名など他の言語表現については対象外とする傾向にある．これに対して，本研究では、**場所に関する言語表現全般**を対象とする汎用的なジオパージングシステムをデザインし，より広範な応用を実現するための足がかりをつくる．

2．本研究の独自性と意義

「実世界の場所を指し示すあらゆる言語表現を抽出し，該当する地図上の位置と結びつけるシステムの開発」を目指す．「あらゆる言語表現」に挑むことが本研究の独自性であり，既存研究における技術的な壁でもあった．本章ではその点を掘り下げる．

2.1. 施設名の網羅性

場所参照表現の中でも「地名」のみに対象を絞った既存研究が非常に多く，「施設名」まで含めたジオパージング研究は例外と言ってよいほど少数である．最大の理由は，**地図データベース（地図 DB）における施設名の網羅性の低さ**にある．

場所参照表現を地図上の位置に紐付ける処理（ジオコーディング）のために地図 DB を活用することが多いが，地図 DB に施設名が十分に収録されていなければ位置情報を取得できない．たとえば，「名前=近鉄奈良，緯度=34.6841971，経度=135.8284104」のように，地名/施設名とその位置情報のペア（エントリ）を収録した地図 DB があれば，DB 内を検索し，適切なエントリの位置情報を取得して，地図上に可視化することができる．代表的な地図 DB のひとつに GeoNames[3] があり，ジオパージング研究で最も頻繁に使用されてきた (Lieberman et al., 2010; Kamalloo and Rafiei, 2018; Wallgrun et al., 2018; Gritta et al., 2020)．しかしながら，施設名の網羅性に課題があった．つまり，仮に文章から施設名を抽出できても，位置情報と結び付けるための地図 DB として十分なものが存在しなかったため，はじめから施設名を抽出の対象外とする研究が多かった．

しかし近年，光明が見えつつある．オープンな地図 DB である OpenStreetMap[4] (OSM) が加速度的に充実してきた．施設名の網羅性が飛躍的に向上し，ジオパージングを目的とした使用に耐えうる基盤が整ってきた．本研究では，地図 DB として OSM を採用する．OSM の豊かな情報を活用したジオパージングシステムを構築し，その可能性と技術的課題を明らかにし，後続する研究への知見を提供する．

---

[1] https://sites.google.com/view/geography-and-language/resources
[2] https://sites.google.com/view/geography-and-language
[3] https://www.geonames.org/
[4] https://www.openstreetmap.org

2.2. 固有名以外の場所参照表現の網羅性

場所参照表現は，地名や施設名などの固有名（名前）に限らない．たとえば，「最寄駅」「空港」「バス停」などの一般名詞句，「ここ」「そこ」などの指示語も場所を表す．例として，以下の文章を考える．

「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。」

観光情報抽出の目的で，観光地「松島」までの経路（あるいは交通アクセス）を文章から自動抽出したい場合，一般名詞句「空港」を対象外としたならば，出発地である「空港」を抽出できない．一方で，仮に「空港」を抽出できても，どの空港かわからないなら，地図上にプロットすることもできないのだから抽出する意味がない，という反論もあり得る．しかしながら，周りの文脈を考えると，この「空港」が「仙台空港」を指すことは，一定の知識を持った人間ならわかる．つまり，文章中に一般名詞句として出現していても，実世界の特定の場所を指すことが文脈上容易にわかる場合も少なくない．「最寄駅」や「バス停」なども同様である．

もうひとつの例として，図 1 の入力文章を考える．5-6 行目の「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。」の「店」は 4 行目の「スタバ」と同じ場所を指し，「ホテル」は 2 行目の「ホテルニューわかさ」と同じ場所を指す．つまり，異なる表記（一般名詞句や指示語）で書かれていても，固有名と同じ場所を指している場合も少なくない．この現象は，まとまった分量の文章に顕著である．図 1 のように，最初は固有名の地名や施設名が書かれ，それ以降は一般名詞句や指示語で同じ場所を指す書き方が多く見られる．したがって，**一般名詞句や指示語をはじめから対象外として貴重な情報を取りこぼすのではなく**，それらも対象に含めた上で，どの程度正確に抽出することが可能か，知見を蓄積することは，学術的にも実応用的にも価値があると考える．このように対象範囲を広げることによって，応用の範囲も広がることが期待できる．

3．提案するフレームワーク

上記の課題に対応するための，汎用的なジオパージング開発フレームワークを提案する．本フレームワークは，次の 3 つのモジュールから構成される．(1) 場所参照表現抽出モジュール，(2) 共参照解析モジュール，(3) ジオコーディングモジュール．本章では，各モジュールについて詳述する．

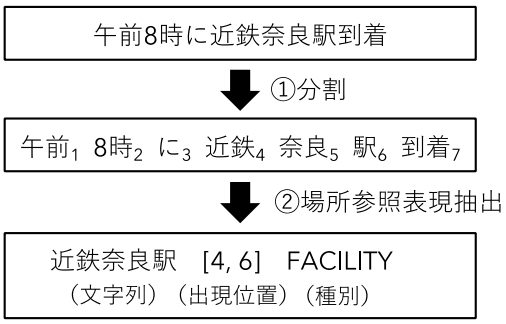

図 2 場所参照表現抽出の例

3.1. 場所参照表現抽出モジュール

本モジュールでは，文章を入力とし，場所を指し示す言語表現（場所参照表現）を抽出する．

3.1.1. 問題設定

図 2 に具体例を示す．入力は文章の文字列である．目標出力として，入力文章中の各場所参照表現の「文字列」「出現位置」「種別」を認識する．まず，入力文章をある単位（単語や文字などの単位）に分割する．次に，場所参照表現として，その「文字列」と文章中での「出現位置」，表現の「種別」を出力する．図 2 の例では，入力文章が単語単位に分割され，「近鉄奈良駅」という文字列が抽出されている．出現位置 [4, 6] は，分割された文章の 4 番目の単語「近鉄」から 6 番目の単語「駅」までが該当文字列であることを意味する．種別「FACILITY」は「施設名」であることを表す．種別は，問題の設計者が任意の種別を設定できる．たとえば，「施設名」の他にも，「LOCATION」（地名；たとえば都道府県・市区町村名）「LINE」（線状の地物；たとえば道路や河川）「TRANSPORTATION」（交通手段；たとえばバスや電車）を設定し，予測対象とすることも可能である．

3.1.2. 主流のアプローチ

これまでの自然言語処理研究を鑑みると，(a)辞書ベースのアプローチと(b)機械学習ベースのアプローチが主流である．辞書ベースのアプローチでは，辞書に含まれている場所参照表現の文字列と，入力

文章中の文字列を照合するのが基本である. 例えば, 辞書に「近鉄奈良駅」という文字列があれば, 同一の文字列が入力文章中に含まれるか探索し, 一致すればその出現位置と種別を出力する. 機械学習ベースのアプローチでは, 文章を機械学習モデルに入力し, モデルが場所参照表現に該当すると判断した文字列の出現位置と種別を列挙する形で出力する.

提案フレームワークでは, 機械学習ベースのアプローチを推奨する. このアプローチの利点は, 辞書に含まれていないような文字列でも場所参照表現として適切に抽出されうる点である. 機械学習モデルは, 学習に使用するデータセット（学習データ）に登場する場所参照表現を丸暗記するのではなく, **学習事例に共通する特徴や傾向を学ぶ**. 言い換えると, 学習データに登場しないような場所参照表現も適切に抽出できるような「汎用性」を獲得する可能性を秘める. この点を重視し, 4 章で説明する我々の実装でも機械学習ベースのアプローチを採用した.

3.2. 共参照解析モジュール s

文章には, 同じ場所を指す異なる複数の言語表現が現れる. 図 3 の例では,「スタバ」と「店」は同じ場所を指す. 同様に,「ホテルニューわかさ」と「ホテル」は同じ場所を指す. このように同じ物や概念を指す異なる言語表現をグルーピングする処理を**共参照解析**という.

3.2.1. 共参照解析を行う利点

この処理を挟むことの利点はなんだろうか？例えば, 一般名詞句「ホテル」を, その文字列情報だけから適切な地図 DB エントリと紐づけるのは難しい. しかしながら,「ホテル」が「ホテルニューわかさ」と同じグループに属する（同じ場所を指す）ことがわかっており, かつ,「ホテルニューわかさ」が適切なエントリと紐づいているなら,「ホテル」も同じエントリと紐づくはずである（同一の場所は同一の地図 DB エントリに対応するため）. 言い換えると, 固有名「ホテルニューわかさ」を介して,「ホテル」も**適切なエントリに間接的に紐付けられる**.

さらに, 共参照解析の副産物として,「ホテルニューわかさ」に関するより詳しい情報を取り出すことも可能になる. 例として,「ホテルニューわかさ」に

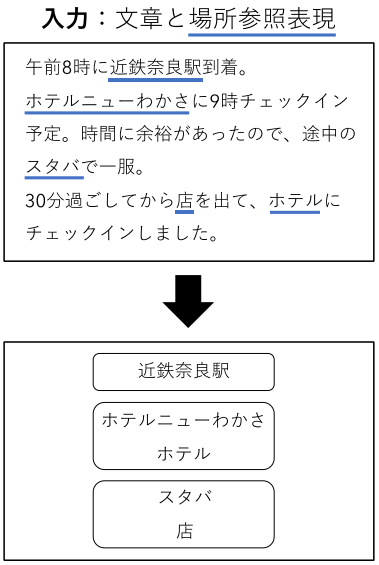

図 3　共参照解析の例

チェックインした事実やタイミングを特定したい場合を考える.「ホテルニューわかさ」が登場する一文を見ただけではわからないが,「ホテル」が登場する文を根拠として特定することが可能になる. 仮に「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認識することは困難になるだろう. したがって, 同じ場所を指す異なる複数の言語表現がグループ化されることによって, **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**. このように, 共参照グループ自体を有効に利用することによって, より豊かな情報抽出が可能になる. 本研究の中心課題はジオパージングであるため, この観点を深掘りしないが, 将来的には, 場所に関する豊かな情報を抽出するために共参照グループを利活用する方向性も模索したい.

3.2.2. 主流のアプローチ

近年は機械学習ベースのアプローチが主流である. その中でも Lee et al. (2017) の手法がシンプルかつ強力であるため, 多くの研究者が採用している. 詳しい手法の説明は当該論文に譲り, ここでは手法のエッセンスを直感的に説明する. 例として, 図 3 中の「店」と同じ場所を指す表現を予測する場合を考える. 基本は, 候補の場所参照表現の中から適切な

ものを選択する．まず,「店」より前の文脈に登場した場所参照表現を候補とする. つまり,「近鉄奈良駅」「ホテルニューわかさ」「スタバ」が候補となる．これらに追加で，同じ場所を指す表現がないことを表す特殊な記号「None」も候補となる．次に，これらの中から適切なものを選ぶ．もし適切なものが複数ある場合は，正解のうちどれを選んでも良い．選ぶ際は,「店」の文脈情報を手がかり（特徴量）として，各候補に対して尤もらしさを表すスコアを算出し，最大スコアのものを選ぶ．ここでは，直前に登場している「スタバ」を選べば正解である．選ばれたもの同士を共参照グループとして出力する．同様に，「ホテル」と同じ場所を指す表現を予測する場合は，「近鉄奈良駅」「ホテルニューわかさ」「スタバ」「店」「None」の 5 つ中から選ぶ.「ホテルニューわかさ」を選べば正解である．

3.3. ジオコーディングモジュール

本モジュールでは，文章と共参照グループを入力とし，共参照グループの位置情報を推定する．

3.3.1. 主流のアプローチ

主流のアプローチはふたつある．(a) 位置情報（経緯度）を直接出力するアプローチと，(b) 地図 DB のエントリを介して位置情報を出力するアプローチである．前者のアプローチでは，機械学習モデルに経緯度を直接出力させる方法をとる．後者のアプローチでは，検索モデルに検索クエリを入力して地図 DB 内を検索し, 適切なエントリを突き止め, そのエントリに登録されている経緯度を出力する．

我々の提案フレームワークでは，後者のアプローチを推奨する．その利点として，**システムの保守のしやすさ**がある．なんらかの理由で登録されているエントリの位置情報を変更したい場合でも，その情報を変更するだけで対処でき，検索モデル自体に変更を加えなくて済む. 一方で前者のアプローチだと，機械学習モデルを一から作り直す必要がある．地図 DB の保守としてエントリの修正・追加を含めた更新作業が必要なため，後者のアプローチの方がこの点に柔軟に対応できる．以降，地図 DB に基づくアプローチを前提に種々の説明をする．

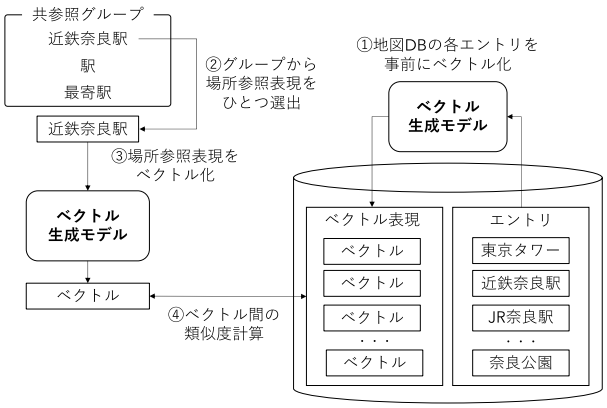

図 4　ジオコーディングの例

3.3.2. 処理の流れ

本モジュールでは, 共参照グループを入力として，地図 DB の適切なエントリに紐づける．例を図 4 に示す．最初のステップでは，事前に地図 DB 内の各エントリをベクトル化する．自然言語処理におけるベクトル化とは，単語や文などの文字列を，事前に決められた次元数（通常，数百〜数千程度）の実数値ベクトルに変換する処理を指す．同様に次のステップでは, 入力の共参照グループをベクトル化する．ここで，各共参照グループは複数の場所参照表現から構成されることも多く，それらからどのように 1 本のベクトルを作るかを考えなければならない．単純には，共参照グループから 1 つの場所参照表現を選び，選んだ表現をベクトル化する方法が考えられる. あるいは, グループに属するN個の場所参照表現をすべてベクトル化してN本のベクトル表現を得て, そのN本のベクトル表現を平均して 1 本のベクトル表現を作るという方法も考えられる．最後のステップでは，地図 DB のエントリのベクトルと共参照グループのベクトルの類似度を計算し，その類似度に基づいてエントリを順位付けする. 最終出力として, 上位ベストk個のエントリを出力してもよいし, 1 位のエントリだけを出力してもよい．想定する応用先を鑑みて任意に決める．

技術的ポイントに挙げられるのは，語句のベクトル表現を用いる点である．この利点はふたつある．ひとつは，曖昧な文字列マッチを実現できることである．つまり，入力の場所参照表現と完全一致しな

い地図 DB のエントリ名であっても，その文字列同士が類似していれば高い検索スコアを与えることができる．ただ，これは曖昧な文字列一致技術でも実現可能であるため，語句のベクトル表現に特有の利点とは言えない．より重要なのはもうひとつの点である．すなわち，**意味的に類似している語句同士に高い検索スコア**を与えることができる．たとえば，「仙台」という語句は，「宮崎」よりも「宮城」のほうが高い検索スコアとなることが期待される．この理由として，現在の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（あるいは共起する）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である．「宮城」のほうが，登場する文脈が「仙台」と似ているため（あるいは「仙台」と共起するため），「宮崎」よりも高い類似度スコアがつけられる場合が多い．今後は，位置情報や地理的属性情報などをベクトル表現に反映させるような学習を行うことによってさらなる飛躍が期待できるため，将来に渡るポテンシャルが非常に高いアプローチであると考える．

4．開発したシステム

提案フレームワークに基づいてシステムを実装した．本章では，そのシステムについて詳述する．

4.1．場所参照表現抽出モジュールの実装

4.1.1．モデルの構築

入力文章が与えられたとき，場所参照表現の出現位置とその種別ラベルを正確に予測できるモデルを構築したい．そのためには，場所参照表現の出現位置と種別ラベルの情報が付与されたデータセットを用いて「教師あり学習」を行う必要がある．つまり，出現位置と種別ラベルを正確に予測できるようになるまで，モデルに繰り返し教え込む必要がある．

そのために，「地球の歩き方旅行記データセット[5]」(Arukikata. Co. Ltd., 2022; Ouchi et al., 2023) を使用した．同データセットは日本語テキストデータであり，4,500 の国内旅行記と 9,500 の海外旅行記から構成され，全体で 3,100 万単語を超える規模である．このうち，200 の国内旅行記に，場所参照表現の出現位置と種別ラベルの情報を付与したデータを作成し，学習/開発/評価データに分割した．これらの情報を正確に予測できるようにモデルの学習を行った．具体的なモデルとして，multilingual LUKE[6] (mLUKE) (Ri et al., 2022) を採用した．

表 1　提案システムで扱う中心的な種別ラベル

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

4.1.2．種別ラベルの設計

各場所参照表現は，その種別を表すラベルの情報を持つ．表 1 に我々のシステムで扱う中心的な種別ラベルを示す．アンダーバー「_」の前半と後半で異なる情報を表し，全体でひとつのラベルを構成する．

前半部分は 4 種類に分かれる．(a) LOC (Location; 地名)，(b) FAC (Facility; 施設名)，(c) LINE (Line; 路線名)，(d) TRANS (Transportation; 乗り物名)．後半部分は次の 2 種類に分かれる．(i) NAME (Name; 固有名)，(ii) NOM (Nominal; 名詞語句)．

さらに，以下の 4 つの種別ラベルも設計した．

- LOC_ORG (Location_Organization)：語句単体だと土地を指しうるが，当該文脈上は組織を指す表現．例えば「広島」は，その語句単体だと土地（行政区）を指す可能性もあるが，「広島はセ・リーグを制覇した」という文脈ではプロ野球のチーム（組織）を指す．純粋に土地を指す表現と区別をするために本ラベルを定義した．
- FAC_ORG (Facility_Organization)：語句単体だと施設を指す可能性もあるが，当該文脈上は組織を指す表現．例えば「奈良先端大」は，その語

[5] https://www.nii.ac.jp/dsc/idr/arukikata/

[6] https://huggingface.co/studio-ousia/mluke-large-lite

句単体だと施設を指す可能性もあるが，「奈良先端大は卒業式を延期した」という文脈では大学組織を指す．純粋に施設を指す表現と区別をするために本ラベルを定義した．

- LOC_OR_FAC：土地と施設の両方を指しうる名詞語句．例えば「観光地」という名詞語句は，土地である可能性と施設である可能性の両方を持つ．このように，土地を指すか，施設を指すか，決められない場合に付与するラベルとして，本ラベルを定義した．
- DEICTIC：指示語．例えば「そこ」「ここ」など．

上記のようなラベル体系を導入することによって，出力結果の中から所望の種類の場所参照表現だけを選択して，応用先に沿った利活用が可能になる．たとえば，施設を指す固有名のみを利用したい場合は，FAC_NAME のラベルが付与された出力結果のみを利用すればよい．土地を指す固有名と名詞語句を利用したい場合は，LOC_NAME と LOC_NOM が付与された出力結果を利用すればよい．

4.1.3. 評価方法

入力文章から所望の言語表現を抽出する問題では，適合率（Precision）と再現率（Recall），それらの調和平均である F 値によって性能を評価することが一般的である．場所参照表現抽出も同様の評価方法を採用する．適合率は，モデルが抽出した言語表現の中で，実際に正解の言語表現だった割合である．再現率は，正解の言語表現の中で，モデルが抽出できた言語表現の割合である．より正確に説明すると，モデルは各場所参照表現の出現位置と種別ラベルを予測するが，両方とも正解である場合のみを正解とした．言い換えると，どちらか片方のみが正解の場合は正解と見なさないこととした．たとえば，出現位置は正解であるが種別ラベルが誤っていた場合は不正解と判定した．

4.1.4. 比較モデル

提案モデルの性能が相対的に高いか，低いかを示すため，既存ツールのモデルと比較する．1 つめの比較モデルは，既存の日本語自然言語処理オープンソースライブラリ GiNZA[7] “ja_ginza” (version 5.1.2) (松田他, 2019) の固有表現抽出モデルである．2 つめの比較モデルは，既存ツール KWJA[8] “base” (version 2.1.1) (Ueda et al., 2023) の固有表現抽出モデルである．これらのモデルは，地名や施設名など場所に関連する固有表現だけでなく，人名などの固有表現も抽出することができる一般的なモデルである．

表 2　場所参照表現抽出の性能比較

モデル	種別ラベル	P	R	F
GiNZA	Overall	.574	.277	.374
	*_NAME	.574	.548	.560
KWJA	Overall	.279	.352	.311
	*_NAME	.279	.695	.398
提案	Overall	.813	.817	.815
	*_NAME	.828	.813	.821
	*_NOM	.832	.826	.829
	LOC_OR_FAC	.731	.711	.721
	DEICTIC	.616	.896	.730

4.1.5. 性能比較結果

表 2 に評価データにおける各モデルの性能を示す．我々のモデルは「地球の歩き方旅行記データセット」の一部にあたる学習データを学習に使用しており，他モデルに比べて有利な状況であるものの，評価データは，全モデルにとって未知の文章である．

まず，各モデルの全体的（Overall）な F 値について議論する．提案モデルは.815 を記録し，GiNZA は.374，KWJA は.311 を記録している．つまり，提案モデルの性能は，2 つの比較モデルを大きく上回っていると解釈できる．この結果は自然であり，実験前からある程度予見できたことだ．なぜなら，提案モデルは場所参照表現の抽出に特化したモデルであり，評価データと同じ傾向を持つ旅行記データで学習しているためである．逆に言えば，比較モデルは人名などを含めた幅広い固有名を抽出するモデルであり，かつ，旅行記データで学習をしていない点が，我々のモデルと比べて不利な条件だと言える．

[7] https://github.com/megagonlabs/ginza
[8] https://github.com/ku-nlp/kwja

また，比較モデルは，名詞語句を抽出の対象外としている．そこで次に，比較モデルも抽出対象としている固有名（*_NAME）のみに対する実験結果（F 値）について議論する．提案モデルは.821 を記録し，GiNZA は.560，KWJA は.398 を記録している．この実験結果においても，提案モデルが比較モデルを大きく上回っている．この理由としても，上述した理由と同じく，我々のモデルが場所参照表現に特化したモデルであることが挙げられる．

4.2. 共参照解析モジュールの実装

4.2.1. モデルの構築

入力文章中の場所参照表現が抽出済みであることを前提とし，同じ場所を指す場所参照表現をグルーピングするためのモデルを構築したい．そのためには，場所参照表現抽出と同様，共参照グループの情報が付与されたデータセットを用いて「教師あり学習」を行う必要がある．そのためのデータセットとして，「地球の歩き方旅行記データセット」のうち，200 の国内旅行記に共参照グループの情報が付与されたデータを用いた．これらの共参照グループを正確に予測できるようにモデルの学習を行った．具体的なモデルとして，multilingual LUKE (mLUKE) (Ri et al.,2022) を採用した．共参照グループの予測方法として，Lee et al. (2017) の手法を採用した．

4.2.2. 比較モデル

比較モデルとして，ふたつのルールベースモデルと，ひとつの既存ツールのモデルを取り上げる．

- Rule-1：場所参照表現をそれぞれ単独のグループとする．つまり，N 個の場所参照表現があった場合，N グループが出来上がる（各グループの構成要員は 1 つの場所参照表現のみ）．
- Rule-2：同一文字列の場所参照表現をグループ化する．
- KWJA：既存ツール KWJA に含まれる共参照解析モデル．

4.2.3. 評価方法

共参照解析で一般的な評価指標である「B^3」(Bagga and Baldwin, 1998) で評価する．図 5 は，正解の共参照グループとモデルが予測した共参照グループを示している．「4: 京都駅」の適合率と再現率を

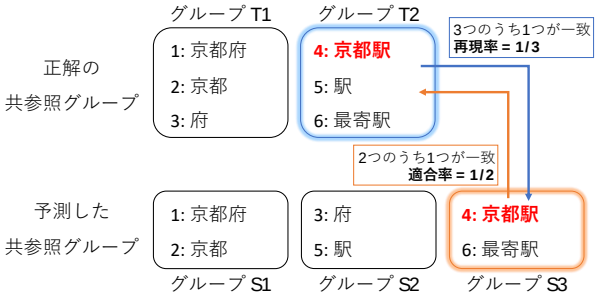

図 5　共参照解析の評価指標 B^3 の算出途中の例

求める場合，正解グループのうち「4: 京都駅」を含むグループ T2 と，予測グループのうち「4: 京都駅」を含むグループ S3 の重なり具合から算出する．まず，適合率を求める．予測グループ S3 は構成員として 2 つの場所参照表現を含み，このうち「4: 京都駅」のみが正解グループ T2 の構成員と重なるため，適合率は 1/2 となる．次に再現率を求める．正解グループ T2 は構成員として 2 つの場所参照表現を含み，その唯一の構成員である「4: 京都駅」が予測グループ S3 の構成員と重なるため，再現率は 1/3 となる．このように，各場所参照表現に対して，適合率と再現率を算出する．最終的に，すべての場所参照表現に対して再現率と適合率を算出し，それらの平均をとって全体の再現率と適合率，さらにはその調和平均である F 値を算出した値が B^3 における再現率，適合率，F 値となる．実験では F 値のみを報告する．

4.2.4. 性能比較結果

表 3 に評価データにおける各モデルの性能を示す．まず，Rule-1 モデルの結果.755 をデータセットの特性に絡めて議論する．本モデルは，各場所参照表現をそれぞれ単独のグループとする．このような単純なルールでも，比較的高い解析結果と言える.755 を記録している．これはデータセットの偏りに由来する．構成員が 1 のグループが評価データに支配的（約 7 割）であるため，単純に各場所参照表現を独立のグループとしただけで評価指標の値が高くなる．

このような状況を踏まえると，より重要なのは構成員が 2 以上（≥2）の共参照グループを適切に予測できるかという点である．この場合も，提案モデルは.733 を記録し，比較モデルを上回る結果となった．この数値は，ある場所参照表現に着目したとき，そ

表 3　共参照解析の性能比較

モデル	サイズ	B^3 (F 値)
Rule-1	≥ 1	.755
	≥ 2	.000
Rule-2	≥ 1	.840
	≥ 2	.613
KWJA	≥ 1	.839
	≥ 2	.661
提案	≥ 1	.875
	≥ 2	.733

の場所参照表現が属する共参照グループの構成員の約 7 割が正解と重なるイメージに近い．つまり，多くの場合は適切にグルーピングされていると言える．

4.3．ジオコーディングモジュールの実装

4.3.1．モデルの構築

入力文章中の場所参照表現が抽出済みで，かつ，共参照グループも予測済みであることを前提とし，各共参照グループに該当する地図 DB のエントリを検索したい. 本研究の実装では, 地図 DB として OSM を用いる．共参照グループをベクトル化するにあたり，共参照グループから場所参照表現を 1 つ選び，その場所参照表現をベクトル化する方法を採用する．場所参照表現の選び方は，グループ内に固有名である場所参照表現（*_NAME）があればそれを選び，なければ最も文字列として長い場所参照表現を選ぶ（長い文字列のほうが情報量が多いと想定できるため）．OSM のエントリと共参照グループの場所参照表現をベクトル化するため，言語モデルの一種である BERT (Devlin et al., 2018) の日本語モデル[9] を用いる．ベクトル間の類似度としてコサイン類似度を使用し，その類似度が高い順に OSM のエントリを順位付けする．なお，本モデルは既存の公開モデルを使用したものであり,「地球の歩き方旅行記データセット」を用いた学習は行っていない．

4.3.2．比較モデル

比較モデルとして，シンプルなルールベースのモデルを実装した．提案モデルと同様，各共参照グループから 1 つ場所参照表現を選び, OSM のエントリと文字列一致検索を行う．もし複数のエントリが完全一致した場合は，それらすべてを出力する．

表 4　ジオコーディングの性能比較

モデル	R@1	R@5	R@10	R@100
ルール	.221	.323	.345	.362
提案	.219	.366	.399	.482

4.3.3．評価方法

検索タスクで一般的な評価指標である Recall@k（R@k）で評価する．これは，モデルにk個のエントリを出力させ，そのk個に正解エントリが含まれていれば正解とする指標である．OSM では，「国道 1 号」のような地物は，部分区間ごとに別々のエントリとして登録されている. このようなエントリでは，旅行記中の場所参照表現に対応づけるには粒度が細かすぎるため，ヒューリスティクスにより，実質的には同一の施設を指すと考えられるエントリ全体を一つのグループにまとめ上げる処理を行った．したがって, 実際にはエントリのグループ単位での予測・評価を行うタスク設定を採用した．

4.3.4．性能比較結果

表 4 に性能評価実験の結果を示す. この実験でも，場所参照表現抽出や共参照解析と同様,「地球の歩き方旅行記データセット」の一部を評価データとして使用している．全体の傾向として，提案モデルがルールベースモデルよりも良い結果を記録している．特に，R@kのkが大きい場合にその差が顕著である．これは，k個のエントリを順位付けして出力する際に，提案モデルの出力したkの中に正解エントリが含まれる場合が相対的に多いことを意味する．言い換えれば，正解のエントリが上位にくるような順位付けになっていることを示している．この理由は，前述したベクトル表現によるマッチングの効果であると推測できる．つまり，曖昧な文字列一致（ファジーマッチ）と意味的な類似性の考慮がある程度実

[9] https://huggingface.co/cl-tohoku/bert-base-japanese-whole-word-masking

現できており，その結果として上位に正解エントリを順位付けることに成功している可能性が高い．

一方で，R@1 の結果は改善の余地が大きい．提案モデルは，文字列の類似度のみから順位付けをしており，「文脈情報」や「地理的な位置関係」を考慮していない．したがって，今後はそれらの点を考慮し，より洗練されたモデルを構築する必要がある．

５．おわりに

本研究は，文章と地図をつなげる技術「ジオパージング」の高度化に取り組んだ．具体的には，最先端の自然言語処理技術を考慮したジオパージングシステムの開発フレームワークの提案と，その実装の性能評価実験を行った．今後の課題として，特にジオコーディングの実装について，地理的な位置関係を考慮したモデルに改善することが挙げられる．

今後の展望を述べる．ひとつは，1 章で述べたように，本研究のジオパージングシステムに基づいて，文章から人間の「地理的移動」を読み取り，地図上に描画するシステムを開発する予定である．もうひとつは，場所参照表現とそれに関連する種々の情報を文章から抽出し，より高度な空間分析へとつなげていきたい．自然言語処理と地理空間情報が交差する領域で，新たな価値を創出することを目指す．

謝辞 Acknowledgement

本研究は JSPS 科研費 JP22H03648 の助成を受けたものです．

参考文献 References

Jochen L Leidner (2006). An evaluation dataset for the toponym resolution task. Computers, Environment and Urban Systems, 30(4):400–417.

Milan Gritta, Mohammad Taher Pilehvar, and Nigel Collier (2020). A pragmatic guide to geoparsing evaluation: Toponyms, named entity recognition and pragmatics. Language Resources and Evaluation, 54:683–712.

Xuke Hu, Zhiyong Zhou, Hao Li, Yingjie Hu, Fuqiang Gu, Jens Kersten, Hongchao Fan, Friederike Klan (2022) Location reference recognition from texts: A survey and comparison. arxiv:2207.01683.

Ehsan Kamalloo and Davood Rafiei (2018). A coherent unsupervised model for toponym resolution. In Proceedings of WWW '18, page 1287–1296.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer (2017). End-to-end neural coreference resolution. In Proceedings of the 2017 Conference on EMNLP, pages 188–197.

Arukikata. Co., Ltd (2022). Arukikata travelogue dataset. Informatics Research Data Repository,National Institute of Informatics. https://doi.org/10.32130/idr.18.1.

Hiroki Ouchi, Hiroyuki Shindo, Shoko Wakamiya, Yuki Matsuda, Naoya Inoue, Shohei Higashiyama, Satoshi Nakamura, and Taro Watanabe (2023) Arukikata travelogue dataset. arXiv:2305.11444.

Ryokan Ri, Ikuya Yamada, and Yoshimasa Tsuruoka (2022). mLUKE: The power of entity representations in multilingual pretrained language models. In Proceedings of the 60th Annual Meeting of the ACL, pages 7316–7330.

松田寛・大村舞・浅原正幸 (2019). 短単位品詞の用法曖昧性解決と依存関係ラベリングの同時学習. 言語処理学会 第 25 回年次大会 発表論文集.

Nobuhiro Ueda, Kazumasa Omura, Takashi Kodama, Hirokazu Kiyomaru, Yugo Murawaki, Daisuke Kawahara, and Sadao Kurohashi (2023). KWJA: A unified japanese analyzer based on foundation models. In Proceedings of the 61st Annual Meeting of the ACL: System Demonstrations.

Amit Bagga and Breck Baldwin (1998). Algorithms for scoring coreference chains. In The first international conference on language resources and evaluation workshop on linguistics coreference, volume 1, pages 563–566.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the NAACL, pages 4171–4186.