BERT を用いた場所の説明文に対する位置特定容易性の推定

坂根 和光 [1]　三林 亮太 [2]　川原 敬史 [3]
山本 岳洋 [4]　澤田 祥一 [5]　高階 勇人 [6]
大島 裕明 [7]

本研究では，「計算科学センタービルの入り口の自転車置き場の横」といった，自然言語で書かれたある地点についての説明文に対して，人はどのような特徴があれば地図上から緯度経度の特定が容易となるのかの分析を行う．本研究で用いるデータは，ある地点についての説明文と，それに対して第三者が説明文が示す実際の緯度経度を推定したものである．実際の位置と，推定位置の違いを基に，説明文ごとの特徴を分析し，場所を説明するときに必要な特徴を明らかにする．また，分析の際に本研究では手動で作成したデータを自動で生成可能か，そして，位置特定が容易に可能か BERT を用いた推定タスクに取り組んだ．

1　はじめに

近年，ある地点がどこにあるかを探すこと，および，目的地の共有は，Google マップや Yahoo!地図といった地図サービス，全地球測位システム（GPS）や WiFi などによる位置情報を推定可能なスマートフォンの普及により，容易になってきている．その一方で，「計算科学センタービルの入り口の自転車置き場の横」のように，テキストあるいは口頭で表現される自然言語により自らの場所や目的地を表したいことは多い．たとえば，屋外で交通事故に合い救急車を呼ぶときや，ホームページに道案内を掲載するときなど，日常，非日常を問わず自然言語で場所を人に伝える状況は存在する．

総務省消防庁が刊行した「救急救助の現況」の「事故発生場所別の搬送人員構成比」の道路に注目すると，令和 2 年では 593,991 名の救急救助が発生しており *1，口頭での説明から容易な場所特定を可能とすることは重要であると考えられる．

しかし，口頭で場所を説明するとき，特に，交通事故が発生したときのような非常事態では，場所の説明をする話し手側は困惑しており，上手に場所を説明できるかは聞き手の能力に依存する場合が多い．

そこで，本研究では，自然言語で書かれた場所の説明文から，人が位置特定が容易な場所の説明文にはどのような特徴を有しているのかを明らかにする．

この研究課題に取り組むため，本研究では，以前著者らが収集した 415 件のデータ [21] から場所の説明文として成立していない 5 件のデータを除外した 410 件のデータに新たに収集した 489 件のデータを加え，899 件のデータを対象に分析を行った．本稿では追加した 489 件のデータ収集と，899 件のデータの分析結果について記す．

データの収集にはクラウドソーシングを活用した．具体的には，まず，兵庫県神戸市中央区三宮を対象に，その範囲にある著者らが指定した歩道の 144 カ所の地点に関する説明文を収集した．

加えて，福岡県福岡市博多区の JR 博多駅から半径 1.5km 以内かつ，博多区に該当する地点をランダムに抽出した歩道の 351 カ所の地点を対象とした．その後，収集した説明文を与えてそこから緯度経度を推定するタスクを再度クラウドソーシングによって実施した．これにより，収集した説明文に対して，実際の緯度経度と近い説明文とそうでない説明文を分析することができる．本稿では，分析の対象として，収集した 495 カ所地点に関する説明文中から，著者らが指定した地点の説明文として成立していない 6 カ所を取り除いた 489 カ所地点と以前著者らが収集した 410 カ所地点を対象に，その説明文から推定された緯度経度の基礎的分析を行い，説明文から緯度経度を正確に特定するために必要な説明文の特徴や推定に際して考慮すべき特徴の分析を行った．また，今後，大量のデータを使用した分析に向けて，自動でデータが作成できるように，さらに，位置特定が容易な場所の説明文か推定するタスクを本研究では手作業で作成したデータを基に，BERT を用いて 3 つの推定タスクを行った．

2　関連研究

テキスト情報と地理情報に関連する研究として，様々な分野の研究で，これまで様々な問題が取り組まれてきた．道案内を行う場合は，どのような書き方があるのか，どのような書き方が人にとって分かりやすいか，近年では，様々な地理情報を与えて自動的に緯度経度を推定するジオコーディングの研究も盛んである．

Taylor ら [12] は，場所の説明文は書かれている視点や表現などで主観と俯瞰，この 2 つの混合型の 3 種類のいずれかに分類できるとしている．また，Taylor ら [13] は，目立つ単一のランドマークがあり，かつ，そこへの経路が単調などの，説明文を作成してもらう地点の周辺環境しだいで，書きやすい説明文の視点が変化する可能性を示した．藤井ら [20] は，自然言語で書かれた場所の説明文の分析を行う際に，単語ではなく，句構造に着目し分析を行った．小林 [18] は，機械によるジオコーディングと比較して，人が地図から目視により地名や，施設名を見つけることは，地図の縮尺や，同名の地名が複数あることが要因で，困難であると言及している．若林 [17] は，最初にランドマーク名が記された地図を見せ，次に，ランドマーク名と道路名が記された地図のように

[1] 非会員　兵庫県立大学 大学院応用情報科学研究科
wsakane0531@gmail.com
[2] 学生会員　兵庫県立大学 大学院情報科学研究科
af22h007@guh.u-hyogo.ac.jp
[3] 非会員　兵庫県立大学 大学院応用情報科学研究科
gtiga.goo@gmail.com
[4] 正会員　兵庫県立大学 大学院情報科学研究科
t.yamamoto@sis.u-hyogo.ac.jp
[5] 非会員　三井住友海上火災保険株式会社
shoichi.sawada@ms-ins.com
[6] 非会員　三井住友海上火災保険株式会社
yuto-takakai@ms-ins.com
[7] 正会員　兵庫県立大学 大学院情報科学研究科
ohshima@ai.u-hyogo.ac.jp
*1 https://www.fdma.go.jp/publication/

段階的に地図に情報を足していき，広域の道案内図から所在地の特定に必要な情報を明らかにした．大佛 [19] は，紙媒体とインタネット広告サイトの渋谷駅を出発点とした案内図で，掲載されるランドマークは，視認が容易な建築面積が大きい建物より，曲がり角の建物などのように，目的地と経路の位置関係を考慮して抜粋されていることを明らかにしている．Winter ら [16] は，オーストラリアのビクトリア州の場所の説明文を，自作のゲームを活用し，半年間で約 2000 件の場所の説明文を収集した．説明文作成の際には，ユーザに GPS により現在地を確認してもらう，ビクトリア州の公報に記載されている地名を含まない場所の説明文を除外するなどの処理により場所の説明文の質を一定以上のものにしている．Richter ら [10] は，上記の Winter ら [16] が収集した場所の説明文のクラスタリングを行い，地名を用いた場所の説明，移動の説明，ルート案内の 3 つのタイプの場所の説明がよく使われていることを明らかにした．また，Richter ら [9] は，場所の説明文を国，都市，地区，道路，建物，部屋，家具という 7 つの粒度レベルで分類した結果，86.3% が異なる粒度レベルが含まれており，その大部分が，粒度レベルを順番に並べた階層的な記述となっていることを明らかにした．Lovelace ら [6] は，文章で行う良い道案内とはどのようなものか調査し，複雑なルートを説明する場合は，人による評価は，長い道案内の方が高評価となることを，明らかにした．Noordzij ら [7] は，人は，ルート・ディスクリプション点とサーベイ・ディスクリプションでは，サーベイ・ディスクリプションの方が，詳細な地図をイメージができることを明らかにした．Winter ら [15] の研究では，外出している人に，「今どこにいるか教えてください」という質問で現在地を尋ねると，75% の人が「スワンストン通りとバーク通りの角にいる」のような，住所やランドマーク，空間的な関係を使って自分の位置を説明し，現在地までどのようにして来たのか経緯を詳細にを説明する人は 2% だった．Jordan ら [5] は，重要度の高いランドマークを機械で識別するために，位置情報ゲームを用いて，プレイヤーの移動データを収集し，移動軌跡の重み付けを行い，ページランクアルゴリズムを利用した．Chi ら [1] は，ツイート内の，都市名，国名，ハッシュタグ，メンションなど様々なテキストの特徴を用いて学習した多項式 NaiveBayes 分類器で，ツイートした都市を予測するアルゴリズムを提案した．Thorndyke ら [14] は，人が，地図と，実際に歩いた経験という異なる 2 つの情報源から得られる空間知識と，目的地への方角と距離の推定手順の違いを明らかにして，人が推定を行う場合の難しさを明らかにした．Fan ら [2] は，ツイートを利用して異なる場所で発生した災害被害状況の推移を自動でマッピングするために，ハリケーンの「ハービー」に関するツイートから，固有表現抽出，ファインチューニングされた BERT による分類，グラフベースクラスティング手法による信頼できる情報を識別，これらの処理を行う機械学習パイプラインを提案した．Hoang ら [3] は，ツイートに位置情報が含まれているかどうかを予測するモデルを提案した．地名辞典に登場する単語と固有名詞の直前の前置詞が重要な要素で，ランダムフォレストとナイーブバイアスを利用した分類器が適していることを明らかにした．Indira ら [4] は，Twitter のユーザの自宅の位置，ツイートの位置，ツイートの内容，これら 3 つを考慮して，ツイートの内容からユーザの位置を予測するのに最適な機械学習モデルは，決定木が最適だと結論を出した．Singh ら [11] は，Twitter から災害被害地点を機械が自動抽出するために，助けを求めるツイートをされた時刻から，対象ユーザの過去のツイート内容を参照し，過去の位置情報を基に，自宅付近か職場付近かなどの位置推定を行うモデルを提案した．Gonzalez ら [8] は，ツイート内容からの詳細な位置推定に行うために，そのツイート内容と最も類似したジオロケーションの投稿を集めることで，ツイート地点の位置推定を行う加重多数決アルゴリズムを活用した．

3 データセット構築

本節では，分析に向けたデータセット収集について説明する．本研究では，場所の説明文と，その説明文から人が位置推定した地点の緯度経度を収集するためにクラウドソーシングサービスの 1 つであるランサーズ [*2] を用いた．このクラウドソーシングの目的は，ある地点を説明した説明文および得られた説明文からどの程度その緯度経度を推定可能かを分析することである．

まず，本研究で実施したクラウドソーシングの概要について説明し，その後実際のタスクの詳細を述べる．

3.1 神戸市中央区三宮データセット構築

さまざまな地点に関する説明文を収集するため，また，「計算科学センタービル」のようなランドマークだけではなく，「計算科学センタービルの入り口と逆側の近く，医療センター方向に向かう交差点を渡る手前の地点」のような，より細かな粒度の地点を説明した説明文を収集する必要がある．本研究では，このような細かい粒度の地点に関する説明文を多数収集するため，我々が指定した場所の説明文をクラウドソーシングのワーカに作成してもらうことを行った．

3.1.1 三宮近辺で対象とした範囲

兵庫県神戸市中央区三宮近辺を対象に，その範囲にある歩道 144 カ所の地点に関する説明文を収集する．図 1 の枠は説明文収集の対象とした三宮近辺の範囲を示したものである．ここから 100m 間隔のメッシュで区切り，区切られた中から説明文を収集する 1 地点を著者らが選択した．

3.1.2 タスク実施手順

具体的なタスクの実施手順は以下の通りである．まず，ワーカに以下の概要文を見せた．

> 本タスクでは，こちらで指定した地点について，その地点を示す地図と，その地点付近のグーグルストリートビューを見ていただき，その地点を特定するような説明文を作成していただきます．指定される地点は 3 カ所あり，兵庫県神戸市の三宮近辺です．各地点において説明文の作成を行っていただきます．本タスクは，兵庫県神戸市の三宮近辺についてご存じない方でもご参加いただけます．

その後，以下の注意事項を見せ，兵庫県神戸市中央区三宮近辺の我々が指定した地点に対しての，説明文を作成してもらった．

[*2] https://lancers.jp

表 1　クラウドーソーシングで三宮近辺を対象に収集した説明文と人手で推定した緯度経度の例

説明文	説明文が示す緯度経度		説明文から人手で推定された緯度経度	
	緯度	経度	緯度	経度
パールストリートのベトナムの国旗が飾られている歩道の地点.	34.696477	135.188070	34.696482	135.188134
行吉学園三宮キャンパスの前にある神戸ハラルフードの入り口の地点	34.696226	135.188956	34.696219	135.188953
New Mode yo から ENERGY DRINK BAR C.R.E.A.M の方へ横断歩道を渡り終わったすぐの地点.	34.696129	135.190909	34.696170	135.190932

図 1　説明文収集対象となる三宮の範囲．地図画像は OpenStreetMap より作成．©OpenStreetMap contributors, Base map and data from OpenStreetMap and OpenStreetMap Foundation/CC BY-SA 2.0

図 2　説明文収集対象となる博多の範囲．地図画像は OpenStreetMap より作成．©OpenStreetMap contributors, Base map and data from OpenStreetMap and OpenStreetMap Foundation/CC BY-SA 2.0

指定した地点は Google マップのマイマップ機能によりワーカに提示した．

- 地図内で「回答地点」と名付けた「青いピン」で指定された地点についてその地点を特定するような説明文を作成してください．
- 他の人（救急車や友人など）にピンポイントにその点に来てもらうために，電話越しに口頭で説明するという状況を想定してご回答下さい．
- 町名や丁目まではすでに特定されているという前提で説明して下さい．
- 作成していただく説明文では，緯度経度や，住所を使わないようにして下さい．
- 町名と丁目が分かっていて，その説明文を見れば，誰にでも，誤差なく，ピンポイントにその地点を特定できるような説明をしてください．
- 作成していただく説明文では，他の指定された地点や，その説明文を参照することはせず，それぞれ独立に回答して下さい．たとえば，「1 か所目から北に 200m ほど先の地点」のような回答はしないようにして下さい．
- 作成していただく説明文では，自動車や人のような短時間で位置が変化すると想定されるものを参照しないで下さい．たとえば，「赤い車のすぐ横」のような回答はしないようにして下さい．
- 地図は，地図上にアイコンを合わせて，マウスを上下左右にドラッグすることで，地図を動かすことができます．また，マウスのホイールボタンや，画面左下の＋，－ボタンで地図の拡大，縮小ができます．
- グーグルストリートビューでは，画面上でマウスを上下左右にドラッグするか，画面右下にあるコンパスの左右にある矢印をクリックすることで，周囲を見回すことができます．また，マウスのホイールボタンや，画面右下の＋，－ボタンで画像の拡大，縮小ができます．

3.1.3　タスクの実施と収集したデータ

今述べたタスクを 2021 年 4 月 29 日に実施した．1 回のタスクあたり 3 カ所の地点について説明文を作成してもらった．対象とした場所は，兵庫県神戸市の三宮近辺である．クラウドソーシングを実施し，144 カ所の説明文を収集した．表 1 の「説明文」と「緯度経度」の列に実際に得られた説明文と緯度経度の例を示す．

3.2　福岡市博多区のデータセット構築

本節では 3.1 節で述べたのと同様の方法で，三宮近辺以外でも場所の説明文の収集を行ったことについて述べる．具体的な説明文を収集する地点については，福岡県の自治体が公開している平成 30 年交通事故オープンデータ [*3] にある事故発生地点の緯度経度データ参照し，福岡県福岡市博多区の JR 博多駅を中心とした半径 1.5km 以内かつ，博多区に該当する地点をランダムに抽出した 351 地点を対象とした．また，本研究は，路上での場所の説明を対象としているので，抽出の際は，緯度経度が屋内の駐車場を指しているものと，水上を指している地点についてはあらかじめ除外した．

3.2.1　JR 博多駅近辺で対象とした範囲

福岡県の中心である，JR 博多駅を対象に，その範囲にある歩道 351 カ所の地点に関する説明文を収集する．図 2 の枠は説明文収集の対象とした JR 博多駅を中心とした半径 1.5km の範囲を示したものである．

[*3] https://ckan.open-governmentdata.org/dataset/401000_koutsuujiko2018

3.2.2 タスクの実施と収集したデータ

今述べたタスクを 2021 年 7 月 9 日に実施した．1 回のタスクあたり 3 カ所の地点について説明文を作成してもらった．対象とした場所は，JR 博多駅を中心とした半径 1.5km 圏内の博多区である．クラウドソーシングを実施し，収集した 351 カ所地点に関する説明文中から，著者らが指定した地点の説明文として成立していない 6 カ所を取り除いた 345 カ所の説明文を収集した．

3.3 人手による説明文からの緯度経度の推定

本研究の目的の 1 つは，場所に関する説明文中にどのようなフレーズや単語が出現していれば，その説明文の地点を表す緯度経度（以降，真の緯度経度と呼ぶ）を特定できるのかの知見を得ることである．そこで，前節で収集した説明文を第三者のワーカに提示し，そのワーカに説明文が指し示していると思われる緯度経度を収集するタスクを実施した．

3.3.1 タスク実施手順

具体的なタスクの手順は以下の通りである．まず，ワーカに以下の概要文を提示した．

> 本タスクは，場所について述べた説明文を見て，その説明文が指し示す地点を特定して回答していただくものです．ある程度の住所の情報は与えられています．その住所の中で，説明文に該当する地点を，Google マップを開いて探していただき，緯度経度を取得していただいて，回答していただきます．緯度経度を回答していた後，説明されている地点をどの程度自信を持って特定できたかについて，4 つの選択肢から 1 つを選択することで回答していただきます．

次に，3.1 節で収集した説明文を 1 つ提示し，用意した Google マップのリンクを開き，説明文が指し示していると思われる緯度経度を回答してもらった．このとき，ワーカに緯度経度を推定する地点の範囲が分かるように，Google マップの初期状態として，説明文の真の緯度経度に対して，リバースジオコーディングを行い，得られた住所に関する字（あざ）の範囲を，字が存在しない場合は，大字の範囲を，また，博多区のどちらも存在しない地点は，博多区全域ハイライト表示した．リバースジオコーディングには Yahoo!リバースジオコーダ API [*4]を利用した．たとえば，緯度 34.689246, 経度 135.501673 であれば，得られる町丁名は「大阪府大阪市中央区伏見町３丁目」となり，字までの情報が得られ，緯度 34.924161，経度 135.710621 であれば，得られる町名は「京都府長岡京市神足」となり，大字までの情報を得る．また，リバースジオコーディングで得られた範囲に真の緯度経度が含まれていない場合は，著者らが住所を直接修正し，字や大字，あるいはそれに近い範囲の領域がハイライトされるハイライト表示の範囲を修正した．

緯度経度を回答してもらった後は，どの程度自信をもって特定できたか以下の 4 つの選択肢からワーカの自己評価で選んで回答してもらった．

- とても自信あり
- やや自信あり
- やや自信なし
- ほとんど自信なし

図 3　自信度と誤差の大きさの関係

これは，真の緯度経度とワーカが推定した地点の緯度経度のずれが小さい場所の説明文は，位置の特定を容易にできるのかを分析するためである．

3.3.2 タスク実施と収集したデータ

今述べたタスクを 2021 年 6 月 15 日と 2021 年 7 月 16 日にランサーズを用いて実施した．具体的には，3.1 節，および，3.2 節で収集した単語のみで回答した 6 件を除外した，合計 489 件の説明文に対して，説明文からの緯度経度の推定のタスクを実施した．後の分析で，説明文から推定される緯度経度のワーカごとのばらつきを分析するために，1 地点につき 3 名のワーカが緯度経度の回答を行い，どの程度自信をもって特定できたのかワーカに自己評価で行ってもらった．また，1 つのタスクにつき 1 名のワーカは 4 件の説明文に対する緯度経度を回答してもらった．また，1 名のワーカがタスクに参加できるのは 4 回までとした．

本研究では，このようにして得られた，ある地点の説明文と，そして説明文から推定された緯度経度のデータを分析することで，位置特定が容易な場所の説明文の特徴を明らかにする．

4 分析結果

本稿では，以前著者らが収集したデータ [21]，3 節で収集した説明文およびその説明文から人が推定した緯度経度のデータ，合計 899 件を分析することで，人手による緯度経度の推定に有用な情報として，どのような言語的特徴が説明文に表れるのかの分析を行った．

4.1 誤差の計算方法

分析にあたり説明文が示す真の緯度経度と，説明文からワーカが推定した緯度経度との誤差を求めた．誤差は，地球の楕円体モデルの表面上での最短距離である測地線距離を 2 地点間で算出した．具体的には，Python のジオコーディングライブラリ GeoPy [*5]が提供する楕円体上の 2 点間の距離計算法 Karney 手法を用いた．測地系は日本の法律でも採用され，多くの国が基準としている GRS80 を用いた．また，算出する数値の単位はメートルとした．

本研究では，1 つの説明文に対して 3 名のワーカが緯度経度を

[*4] https://developer.yahoo.co.jp/webapi/map/openlocalplatform/v1/reversegeocoder.html

[*5] https://github.com/geopy/geopy

推定している．3 名のワーカそれぞれについて真の緯度経度との距離を測定し，その中央値を算出することで，説明文と説明文から推定される緯度経度の誤差を計算した．この値が小さいほど，人にとって説明文から真の緯度経度を特定することが容易であると考えられる．

4.2 真の緯度経度と説明文と誤差の関係

4.1 節で述べた方法を用いて収集した説明文に対してワーカが推定した地点と真の緯度経度との誤差を求めた．

また，3.3 節でも述べた通り，場所の説明文に対して，どの程度自信をもって位置特定を行えた（以降，自信度と呼ぶ）かを調査しており，自信度と推定位置と真の位置からのずれは図 3 にあるように，自信度が高いほど誤差の中央値が小さくなっており，位置特定が容易な場所の説明文には何らかの特徴があると考えられる．

4.3 場所の説明文の特徴

本研究では，関連研究を基に，場所の説明文の視点が主観で書かれているか，俯瞰で書かれているのか分類して分析を行った．また，説明文を構成する要素として，7 種類のタグを用いて，たとえば，「鵠沼薬局」のような固有名詞の建物には「固有名詞の点か面の地物」，のようなタグ付けを行い，分析を行った．

4.3.1 2 種類の場所の説明文

場所の説明文は Taylor ら [12] の研究によると，ルート型説明文（以降，主観説明文と呼ぶ），とサーヴェイ型説明文（以降，俯瞰説明文と呼ぶ）のどちらか，もしくは，両者の混合のいずれかに分類できる．そこで，当研究を基に，著者らが手動で 3 節で収取した場所の説明文を主観説明文と俯瞰説明文のいずれかに分類した．また，両者の特徴を備えている場所の説明文は主観説明文と分類した．分類した結果は，主観説明文の割合が 39% で，俯瞰説明文の割合が 61% と俯瞰説明文が多い結果となっている．主観説明文とは，たとえば，「サミットの駐輪場を出て，右折で歩道に出る地点」のような，始点を定めて，実際に歩くことを想定して行う場所の説明である．また，俯瞰説明文とは，たとえば，「北側に柏井保育園，南側には墓地のある JR 中央線桂林寺踏切の南側 T 字路地点.」のような，現在地から地図を確認し，目印となるものを列挙する場所の説明文である．

まず，人は地図上から位置特定を行う場合，場所の説明文はどちらの視点で書かれた方が容易となるのか分析を行う．

場所の説明文には， 4.1 節で述べたように，1 つの説明文に対して 3 名のワーカが緯度経度を推定している．3 名のワーカそれぞれについて真の緯度経度との距離を測定し，3 名の中央値を，推定の誤差として，それぞれの場所の説明文にデータとして付与した．

容易性について本研究では，真の緯度経度と説明文から 3 名のワーカが推定した緯度経度との誤差の中央値が 25m 未満に収まった場所の説明文を，人による位置特定が容易な説明文とし，25m 以上を位置特定が困難な説明文とした．分類した結果は表 2 に示す．

まず，3 名のワーカが推定した緯度経度との誤差の中央値が 25m 未満に収まった場所の説明文の主観説明文と俯瞰説明文の割合は，主観説明文が 37% で，俯瞰説明文が 63% であった．それに対して，誤差の中央値が 25m 以上の場所の説明文の主観説明文と俯瞰説明文の割合は，主観説明文が 47% で，俯瞰説明文が 53% と誤差の中央値が 25m 未満と比較すると主観説明文の割合が増加する結果となった．そして，有意性を検証するためにカイ二乗検定を行ったところ，P 値が 0.01 となり，統計的に有意であると言える結果となった．

このことから，場所の説明文は俯瞰で書く方が位置特定が容易となる可能性が高いと言える．

表 2　主観説明文と俯瞰説明文を分類した結果のクロス集計表

	主観説明文	俯瞰説明文	合計
位置特定容易	252	427	679
位置特定困難	103	117	220
合計	355	544	899

表 3　場所の説明文に付与したタグの種類と具体例

タグ名	具体例
固有名詞の点か面の地物	鵠沼薬局，浜田町
一般名詞の点か面の地物	駐車場，コンビニ
固有名詞の線の地物	有馬街道，元町商店街
一般名詞の線の地物	歩道，高架
方向または位置表現	西，右手側
移動表現	渡って，右折して
数的表現	100m，1 つ目

4.3.2 場所の説明文の要素

場所の説明文に付与したタグは表 3 の通り 7 種類である．

これは，4.1 節で算出した誤差の中央値が小さい場所の説明文，および，誤差の中央値が大きい説明文にはどのような要素が見られるのか調査するためである．

差が顕著に現れたタグは，3 名のワーカによる推定の誤差の中央値が 25m 以上の場所の説明文では，固有名詞の点か面の地物のタグと，方向または位置表現のタグの出現割合が，誤差の中央値が 25m 未満に収まっている場所の説明文と比べると減少し，一方で，移動表現のタグの出現割合が増加している．分類した結果は表 4 に示す．これらのことから，人が場所の説明文からの位置特定を行う際に，容易となる要因として考えられることは，2 点あり，まず 1 点は，位置特定をするときに基準となる点である，「計算科学センタービル」のようなピンポイントで地点特定が可能な情報の出現頻度が高く，もう 1 点は，「西」のような地図上だと瞬時に方向が判断できる表現を用いており，一方で，位置特定が困難な説明文では方向を表す際に，「左折してから信号まで直進してそこから右折…」のような，思考が必要で，経路移動が長い複雑な説明となっていることが傾向として見られた．この 2 点が要因で位置特定容易となると考えられる．分析の有意性を検証するために，誤差の中央値 25m を境に，固有名詞の点か面の地物，方向または位置表現，移動表現の出現割合でカイ二乗検定を行ったところ，P 値が 4.7 × 1/10 の 5 乗となり，統計的に有意であると言える結果となった．

表 4　位置特定容易な説明文と困難な説明文で差が顕著に表れたタグのクロス集計表

	固有名詞の点か面の地物	方向または位置表現	移動表現	合計
位置特定容易	1,214	1,567	499	3,280
位置特定困難	305	432	198	935
合計	1,519	1,999	697	4,215

このことから，固有名詞の点か面の地物が書かれていて，そして，そこから説明文が示している地点に向かうの方向または位置表現が説明文に情報として出現していることが要因だと明らかにした．

本研究では，手作業でデータを作成し，分析を行った．今後の展望として，場所の説明文を入力として，人による位置特定が容易か困難かを自動で判断して，出力を行うモデルを作成を考えている．これにより，人にとって位置特定が容易な説明文か評価を受けることができ，位置特定が容易な説明文作成の手助けとなる．また，タグの付与，および，主観俯瞰判定を自動で行えた場合，大量のデータを利用しての分析が可能となる．

5　BERT を用いた推定タスク

5.1　タスクの概要

本節では，分析結果に関する 3 つの推定タスクについて説明する．これまで，タグ，主観俯瞰，容易性の 3 つの分析において，いくつか有効な分析結果が得られた．しかし，これらの分析を行うにあたって使用したデータは，すべて手作業によって作成したものであり，大量のデータに対して分析を行う際には，自動でのデータ作成が必要となる．そこで，本節では，分析した 3 つのタスクに対して，自動でデータが作成できるように 3 つの推定タスクを行う．推定タスクは以下の通りである．

- タグ推定タスク
- 主観俯瞰推定タスク
- 位置特定容易性推定タスク

まず，タグ推定タスクでは，4.3.2 節で解説した，7 つのタグの推定を行う．タグ推定タスクの入力は，場所説明文であり，出力はタグが付与された場所説明文である．タグ推定タスクはいわゆる系列ラベリングタスクとして考えることが出来る．そこで，本研究ではタグ推定タスクを自動化するための手法として，BERT をファインチューニングする手法を提案する．

次に，主観俯瞰推定タスクに取り組む．主観俯瞰推定タスクとは，4.3.1 節で解説した，説明文の主観と俯瞰を推定するタスクである．主観俯瞰推定タスクの入力は，場所説明文であり，出力は主観か俯瞰かを示す 2 値である．主観俯瞰推定タスクはいわゆる文書分類タスクとして考えることが出来る．そこで，本研究では主観俯瞰推定タスクを自動化するための手法として，BERT をファインチューニングする手法を提案する．

最後に，位置特定容易性推定タスクに取り組む．3.3 節では，人手による説明文からの緯度経度の推定について述べた．ここで，人が正しく位置特定出来た説明文は，位置特定が「容易」な場所説明文と考えられる．一方で，人が正しく特定出来なかった説明文は，位置特定が「容易でない」な場所説明文と考えられる．位置特定容易性推定タスクは場所説明文が入力された時に，位置特定が「容易」か「容易でない」かを 2 値分類するタスクである．このタスクは，前述の主観俯瞰推定タスクと同様に文書分類タスクとして考えることが出来る．そこで，本研究では位置特定容易性推定タスクを自動化するための手法として，BERT をファインチューニングする手法を提案する．

各タスクでのファインチューニングには，4 節で分析の対象とした，899 件のデータを使用する．本研究では，899 件を 8:1:1 に分割し，それらを訓練データ，検証データ，テストデータとして扱う．訓練データと検証データを用いて，ファインチューニングを実施する．その後，テストデータに対して推論を行うことで，モデルの性能を評価する．以下では，上記で述べた手法の詳細について記述する．

5.2　タグ推定タスク

5.2.1　手法

本研究では，場所説明文から表 3 に示すような場所特徴タグをアノテーションする手法として，汎用言語モデル BERT をファインチューニングする手法を提案する．

テキストを BERT に入力する際には，事前学習と同じ条件でテキストをトークンレベルに分割する必要がある．以後，この分割のことをトークン化と呼ぶ．従って，BERT への入力はトークン化されたトークン列である．

表 3 に示す通り，場所特徴タグは全部で 7 種類のカテゴリによって構成される．しかし，実際の場所説明文では，どのタグにも当てはまらない要素も存在する．そこで本研究では，ファインチューニング時に限り，「その他」タグを追加のタグとして用いる．従って，タグは全部で 8 種類のカテゴリと見なす．

タグ推定タスクを解くファインチューニングのネットワーク図を図 4 に示す．図 4 における FC 層とは全結合層の略語である．ファインチューニングの際には BERT 最終層の上に追加の全結合層を設置する．BERT から出力される特殊トークン以外のトークンに対応した 768 次元のベクトルが全結合層への入力となる．全結合層では入力されたベクトルを 8 次元のベクトルに変換することで，各入力トークンに付与されるタグの推論を行う．誤差逆伝播に用いる損失値は，損失関数クロスエントロピーロスによって算出する．

テキストのトークン化の際にはサブトークンとして分割されるトークンがある．たとえば，「ダイワロイヤル」をトークン化すると「ダイ」，「##ワ」，「##ロイ」，「##ヤル」に分割される．ここでトークンの頭に「##」が付いているのがサブトークンである．本研究で取り組むファインチューニングでは，サブトークンに対する予測結果は損失値計算の対象から除外した．

5.2.2　結果

ファインチューニングモデルのテストデータに対する正解率は 0.918 であった．この結果から，場所説明文からのタグ推定は高い正解率で推定できることがわかる．今後は，タグが付与されていない場所説明文に対しても，タグを付与し，さらなる傾向の分析を行う予定である．

図 4　場所特徴タグ推定タスクにおけるファインチューニング

図 5　主観俯瞰推定タスクにおけるファインチューニング

図 6　場所推定容易性推定タスクにおけるファインチューニング

5.3　主観俯瞰推定タスク

5.3.1　手法

主観俯瞰推定タスクでは，BERT を用いて，場所説明文の主観と俯瞰を分類する．入力は場所説明文であり，出力は主観か俯瞰かの 2 値分類を行う．

ファインチューニングにおけるネットワーク図を図 5 に示す．ファンチューニングの際には，BERT の最終層の上に追加の FC 層を設置する．主観俯瞰推定タスクを解くためのファインチューニングでは，[CLS] トークンに対応した 768 次元のベクトルが FC 層への入力となる．768 次元のベクトルを FC 層入力する際には，Dropout を行う．今回の実験では，Dropout 率を 0.1 に設定した．全結合層で 768 次元のベクトルを 2 次元に変換することで主観か俯瞰かの 2 値分類問題を解く．

図 7　主観俯瞰推定タスクの混同行列

図 8　容易性推定タスクの混同行列

5.3.2　結果

主観俯瞰推定のテストデータ 90 件に対する，正解率は 0.877 であった．この結果から，場所説明文を用いた主観俯瞰推定は，高い正解率で推定できることがわかる．推定結果の混同行列を図に 7 に示す．テストデータの 90 件のラベルは，主観が 35 件，俯瞰が 55 件であり，やや俯瞰が多い傾向にあったが，混同行列から，主観 31 件，俯瞰 50 件の正しい推定がおこなわれていることがわかる．

5.4　容易性推定タスク

5.4.1　手法

容易性推定タスクでは，BERT を用いて，場所説明文の位置特定の容易性を推定する．入力は場所説明文であり，出力は，位置特定が「容易」か「容易でない」かの 2 値である．場所説明文に対して，位置特定が「容易」か「容易でない」かを判断する基準としては，4.1 節で計算した「人が推定した緯度経度」と「真の緯度経度」の誤差により決定する．本研究では，誤差が 25m 未満の場所説明文は「容易」，25m 以上では「容易でない」とした．ファインチューニングにおけるネットワーク図は，図 6 に示す．

5.4.2　結果

容易性推定タスクのテストデータ 90 件に対する推定結果の混同行列を図 8 に示す．正解率は 0.711 であったが，混同行列を確認すると，作成したモデルは全てのテストデータに対して「容易」と判断してしまっており，全く分類が行えていないことが分かる．これは，この容易性推定タスクが困難であるということを示していると考えられる．また，訓練データにおけるラベルの分布の偏りがあることも，推定がうまく行えていない原因として考えられる．

6 まとめと今後の課題

本研究では，自然言語で書かれたある地点についての説明文から，その地点がどこであるかの緯度経度を特定する際に，容易な位置特定に必要な特徴を明らかにする問題の取り組みにあたって必要となるデータの収集と，収集されたデータの分析を行った．その結果，人による位置特定が容易になる要素として，場所の説明文中に推定の際に基準となる点として，「計算科学センタービル」のようなピンポイントでの地点特定が可能な情報が出現し，かつ，地図を見ているときに「西」のような目的地に対する方向が瞬時に判断できる情報が出現する．同時に，場所の説明文の視点が俯瞰で書かれていると，位置特定が容易となる可能性が高いことを明らかにした．今後も引き続き説明文を収集するとともに，得られた説明文を基に人が推定する緯度経度についても収集する．その後，収集された説明文と緯度経度において，位置を推定することに資する情報がどのように出現するかの分析をさらに進める．

謝辞

本研究の一部は JSPS 科学研究費助成事業 JP21H03775, JP21H03774, JP21H03554, JP18H03244 の助成を受けたものです．ここに記して謝意を表します．

参考文献

[1] Lianhua Chi, Kwan Hui Lim, Nebula Alam, and Christopher J. Butler. Geolocation prediction in twitter using location indicative words and textual features. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, pages 227–234, 2016.

[2] Chao Fan, Fangsheng Wu, and Ali Mostafavi. A hybrid machine learning pipeline for automated mapping of events and locations from social media in disasters. *IEEE Access*, 8:10478–10490, 2020.

[3] Thi Bich Ngoc Hoang, Véronique Moriceau, and Josiane Mothe. Can we predict locations in tweets? a machine learning approach. *International journal of computational linguistics and applications*, 9.

[4] K Indira, E Brumancia, and P Siva kumar. Location prediction on twitter using machine learning techniques. In *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)*, pages 700–703, 2019.

[5] Klaas Ole Jordan, Iaroslav Sheptykin, Barbara Grüter, and Heide-Rose Vatterrott. Identification of structural landmarks in a park using movement data collected in a location-based game. In *Proceedings of The First ACM SIGSPATIAL International Workshop on Computational Models of Place*, pages 1–8, 2013.

[6] Kristin L. Lovelace, Mary Hegarty, and Daniel R. Montello. Elements of good route directions in familiar and unfamiliar environments. In *Spatial Information Theory Cognitive and Computational Foundations of Geographic Information Science*, pages 65–82, 1999.

[7] Matthijs L. Noordzij and Albert Postma. Categorical and metric distance information in mental representations derived from route and survey descriptions. *Psychological Research*, 69(3):221–232, 2005.

[8] Jorge David Gonzalez Paule, Yashar Moshfeghi, Joemon M. Jose, and Piyushimita (Vonu) Thakuriah. On fine-grained geolocalisation of tweets. In *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval*, pages 313–316, 2017.

[9] Daniela Richter, Maria Vasardani, Lesley Stirlng, Kai-Florian Richter, and Stephan Winter. Zooming in–zooming out hierarchies in place descriptions. In *Progress in location-based services*, pages 339–355, 2013.

[10] Daniela Richter, Stephan Winter, Kai-Florian Richter, and Lesley Stirling. How people describe their place: Identifying predominant types of place descriptions. In *Proceedings of the 1st ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information*, pages 30–37, 2012.

[11] Jyoti Prakash Singh, Yogesh K. Dwivedi, Nripendra P. Rana, Abhinav Kumar, and Kawaljeet Kaur Kapoor. Event classification and location prediction from tweets during disasters. *Annals of Operations Research*, 283(1):737–757, 2019.

[12] Holly A Taylor and Barbara Tversky. Spatial mental models derived from survey and route descriptions. *Journal of Memory and Language*, 31(2):261–292, 1992.

[13] Holly A Taylor and Barbara Tversky. Perspective in spatial descriptions. *Journal of memory and language*, 35(3):371–391, 1996.

[14] Perry W. Thorndyke and Barbara Hayes-Roth. Differences in spatial knowledge acquired from maps and navigation. *Cognitive psychology*, 14(4):560–589, 1982.

[15] Stephan Winter, Allison Kealy, Matt Duckham, Abbas Rajabifard, Kai-Florian Richter, Tim Baldwin, Lesley Stirling, Lawrence Cavedon, and Daniela Richter. Starting to talk about place. In *Infrastructure Engineering - Research Publications*, 2011.

[16] Stephan Winter, Kai-Florian Richter, Tim Baldwin, Lawrence Cavedon, Lesley Stirling, Matt Duckham, Allison Kealy, and Abbas Rajabifard. Location-based mobile games for spatial knowledge acquisition. 2011.

[17] 若林芳樹. 道案内図を用いた地理情報の伝達とナヴィゲーションの成立条件. volume 10, pages 171–176, 2001.

[18] 小林亘. 平成 29 年 7 月九州北部豪雨の道路被災場所の特定への道路ジオコーダの適用評価. **災害情報**, 17(1):31–40, 2019.

[19] 大佛俊泰. 歩行経路案内図生成のための地図構成要素抽出モデルについて. **日本建築学会計画系論文集**, 70(593):117–122, 2005.

[20] 藤井晴行 and 青木義次. ことばによる経路案内の統語論的分析. **日本建築学会計画系論文集**, 66(549):199–206, 2001.

[21] 坂根 和光, 山本 岳洋, 澤田 祥一, 大塚 一路, 山本 光穂, and 大島 裕明. 自然言語からの緯度経度推定に向けた説明文収集とパターン分析. In **データ工学と情報マネジメントに関するフォーラム**, 2021.