一般社団法人 電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report

文書の時空間 3 次元地図へのマッピング

平山 拓実† 江木 千沙都‡ 難波 英嗣† 竹澤 寿幸†

†広島市立大学大学院情報科学研究科 〒731-3194 広島県広島市安佐南区大塚東 3-4-1
‡広島市立大学情報科学部 〒731-3194 広島県広島市安佐南区大塚東 3-4-1
E-mail: †‡{hirayama, egi, nanba, takezawa}@ls.info.hiroshima-cu.ac.jp

あらまし 本研究では, 地名や施設名などの地名表現を含んだ文書を地図上にマッピングするシステムを構築する．ここで，地名や施設名は年月とともに変わる可能性があるため，古い地名表現を含んだ文書を地図上にマッピングできない場合がある．そこで本研究では，新旧地名・組織名の対をテキストデータベースから抽出し，古い地名表現を含んだ文書のマッピングを実現する．

キーワード マッピング，情報可視化，ジオコーディング

Spatio-temporal Three-Dimensional Mapping of Documents

Takumi HIRAYAMA† Chisato EGI‡ Hidetsugu NANBA† and Toshiyuki TAKEZAWA†

† Graduate School of Information Sciences, Hiroshima City Universiry

3-4-1 Ozuka-higashi, Asaminami-ku, Hiroshima 731-3194 Japan

‡ School of Information Sciences, Hiroshima City Universiry

3-4-1 Ozuka-higashi, Asaminami-ku, Hiroshima 731-3194 Japan

E-mail: †‡{hirayama, egi, nanba, takezawa}@ls.info.hiroshima-cu.ac.jp

Abstract In this study, we have constructed a system for mapping documents that contain geographic information, such as a place name or facility name on the map. However, since the place names and facility names may change over the years, there are cases where it is not possible to map the document that contains the old geographic information on the map. In this study, to extract the pair of old and new place name and organization name from the text database, to realize the mapping of the document that contains the old geographical information.

Keywords Mapping, Information Visualization, Geocoding

1. はじめに

本研究では，地名や施設名などの地名表現を含んだ文書を地図上にマッピングするシステムを構築する．このシステムにより，大量の文書を読むことなく，任意の場所に関する情報を把握することができる．例えば，地元に関する事件や事故，飲食店，書籍といった情報を容易に把握できる．

地図上にマッピングするシステムの一般的な流れとして，対象文書から地名表現を抽出，抽出した地名表現に緯度経度情報の付与（ジオコーディング），付与した緯度経度情報を基に地図上にマッピングを行う．

本研究では，様々な情報を把握できるシステムの構築を目的とするため，対象文書に新聞記事，旅行ブログ，書籍データといった多種の文書を用いる．地名表現の抽出には係り受け解析器 CaboCha の固有表現抽出機能を用いて，地名表現である地名や施設名などを抽出する．また，本研究におけるジオコーディングでは，従来手法で課題とされてきた古い地名表現や曖昧性問題を改善する手法を提案する．曖昧性問題とは，文書中の地名表現が省略されることにより，場所の特定が困難になることである．例として地名表現が「日光」の場合，該当する地名に栃木県，福井県，愛知県の日光が存在する．そのため，地名表現だけで，場所を特定することは困難である．このような曖昧性問題に対して本研究では，まず，文書から抽出した曖昧でない地名表現にジオコーディングを行う．その際，地名表現が属する都道府県の出現頻度を求める．この出現頻度を用いることで，曖昧な地名表現が属する都道府県の推定でき，場所の特定が可能な手法を提案する．古い地名表現においては，現在の地名表現に置き換えることで改善できる．そのため，本研究は大量の

Web ページなどから機械学習手法 Conditional Random Field (CRF)を用いることで，新旧地名対の抽出を行うことで置き換える手法を提案する．

本論文の構成は以下の通りである．2 章では本システムの動作例について述べ，3 章では本研究と関連する研究を述べ，4 章では文書のマッピングについて述べ，5 章では提案手法の評価実験について述べ，6 章で本論文をまとめる．

2. システム動作例

本研究では，書籍や旅行ブログ，新聞記事といった文書を同一地図上にマッピングするシステムを構築する．これにより，視覚的にある場所を視点に起きた出来事を容易に理解できると考えられる．本システムの動作例を図 1 に示す．図 1 では，読売新聞記事に出現した地名表現の地理的位置にピンが立てられている．また,吹き出しには新聞記事の見出しを表示している．

図 1. システム動作例

3. 関連研究

本研究では，新聞記事や旅行ブログ，書籍といった様々な文書を地図上にマッピングしている．本研究と関連する研究に郡ら[1]や鎌田ら[2]の研究が挙げられる．郡らは複数の旅行ブログから代表的な行動経路とその行動のテーマを抽出し，地図上にマッピングをしている．鎌田らは Twitter などのつぶやきから経路を抽出し，地図上に経路と投稿された写真を表示するアプリケーションの構築をしている．このように文書をマッピングする研究は多く，マッピングの対象とされる文書は多様である．本研究ではこれらの研究と異なり，複数の種類の文書を同一の地図上にマッピングする．これにより，様々な情報を取得できるシステムの構築が望める．

しかし，文書を地図上にマッピングするには，文書から抽出した地名表現にジオコーディングが必要である．しかし，ジオコーディングでは抽出した地名表現が曖昧な場合，複数の緯度経度情報と対応付く可能性がある．このような曖昧性問題を解消する手法には，河野ら[3]や安田ら[4]，金木ら[5]，平野ら[6]の研究が挙げられる．

河野らは Twitter におけるユーザの一連のツイートから地名を抽出し，ツイートの位置情報を推定する研究を行っている．ツイートから抽出した地名表現が曖昧な場合，前後のツイートから抽出した地名表現，曖昧な地名表現と対応付く地名候補，これら全てを対象にクラスタリングを行う．次に，各クラスタに候補地と曖昧でない地名のスコアを付与する．そして，最も総和の大きいクラスタに含まれる候補地の緯度経度情報を付与する手法を提案している．安田らはキーワードと場所を入力クエリとして，地理的制約を考慮した情報検索手法を提案している．河野らと同様に抽出した曖昧でない地名表現と曖昧な地名表現の候補，全てを対象にクラスタリングを行う．これにより文書の地理的範囲を推定し，関連性の低い地名表現を除去する手法を提案している．本研究はこれらの地理的距離によるクラスタリングを用いた曖昧性問題解消では，複数の都道府県の地名表現が抽出される文書の場合，効果が薄いと考えた．そのため本研究は，抽出される地名表現の都道府県の出現頻度を用いたジオコーディング手法を提案する．

金木らは新聞記事の地域特定を行う研究をしている．ジオコーディングには，地名表現と緯度経度情報の対が登録された地名辞書を参照する一般的な手法を用いている．曖昧性問題解消には次のような手法を提案している．まず，文書から地名表現を抽出し，各地名表現の出現頻度などに応じたスコアを付与する．次に，地名間の地域距離を基にしたスコアを加算する．地域距離算出方法として，図 2 に地名表現「西新宿」と「四谷」を例に示す．図 2 右式の分母は各地名表現の階層の積，分子は共通上位階層の 2 乗としている．この算出方法により，地理的に近い地名は値が大きくしている．上記による手法で高いスコアを得た地名表現が文書を表す地名とする手法を提案している．

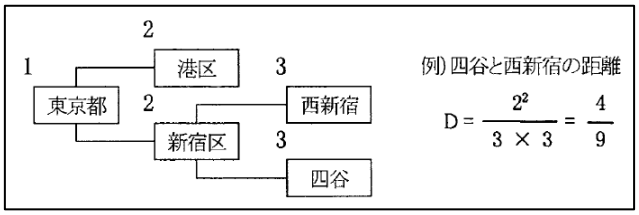

図 2. 地域階層と地域距離の例

平野らは曖昧性問題解消のため地名の有名度と地理的距離を組み合わせた手法を提案している．店舗の

多い地域が有名な場所と考え，有名度には地域に存在する店舗数としている．本研究では，古い文書でも対応できるジオコーディングを目的としているため，現在の有名度を用いる手法では効果が薄いと考えられる．

これらのジオコーディングを用いる多くの研究は，曖昧性問題解消に対する手法を提案している．しかし，地名の旧地名といった時間を考慮する研究は少ない．そのため国分ら[7]のように人手で新旧地名の対応付けを行うことが多い．しかし，人手による対応付けは時間とコストを要する．本研究では，機械学習手法 CRF を用いることで Web ページから自動的に新旧地名の対応付けを行う．

4. 地名表現のジオコーディング

4.1. 新旧地名の対応付け

4.1.1. 基本方針

新地名と旧地名の情報は，テキスト中で[新地名](旧[旧地名])のように記述される．例えば，以下の例では，ポンペイ島はかつてポナペと呼ばれていたことがわかる．

毎年恒例ミクロネシアツアー。今年は太平洋の孤島、ポンペイ島（旧ポナペ）です。

そこで，実際に「(旧)という表現を含む Web や新聞記事中の文を調べたところ，複数の町村が合併してひとつの市が形成される場合には，ひとつの新地名に複数の旧地名が対応する場合があること，また，例えば「小沢一郎前衆院議員(旧自由党党首)」のように,「旧」の前後に地名でない名詞句が出現する場合があることがわかった．そこで，本研究では，新地名と旧地名の対を抽出する課題を，大量のテキスト集合に対し，新地名と旧地名を示す個所に，それぞれ NEW と OLD というタグを付与する，いわゆる系列ラベリング問題としてとらえ，CRF に基づく手法により NEW と OLD タグを付与するシステムを構築することにした．

4.1.2. 教師データの作成

本研究では，NTCIR-5Web 検索タスクや The ClueWeb09 のデータセットである Web ページと読売新聞記事(1993-2012)から「(旧)を含む文を抽出し，これらを，新旧地名の対を抽出する対象とした．

教師データには，人手で新地名に NEW，旧地名に OLD タグを付与したデータを用いた．4.1 節の例の場合，以下のようにタグを付与する．

毎年恒例ミクロネシアツアー。今年は太平洋の孤島、<NEW>ポンペイ島</NEW>（旧<OLD>ポナペ</OLD>）です。

4.1.3. 新旧地名対の抽出

CRF の素性は，ターゲットとなる単語から前後 k 個の単語の形態素，品詞，固有表現の 3 つとした．固有表現抽出には，日本語係り受け解析器 CaboCha を利用し，今回は，地名(LOCATION)または組織名(ORGANIZATION)の 2 種類の固有表現を用いた．また，本研究では予備実験の結果から k=4 と定めた．図 3 では，説明のため k=2 の場合を例として示している．これにより，CRF は教師データと素性に基づき，与えられた文に対して NEW・OLD タグを付与する．このタグ付け結果を用いて新旧地名対抽出を行い，4.2 節のジオコーディングに用いる．

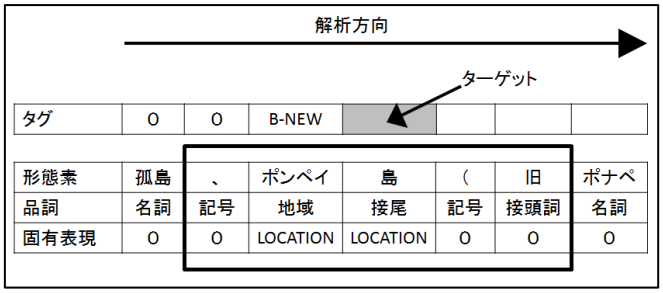

図 3. CRF を用いた新旧地名対の抽出

4.2. 曖昧性問題を考慮したジオコーディング

本研究のジオコーディングでは，都道府県・市区町村・町丁名や駅名，山の名称などと対応する緯度経度情報の対が登録された地名辞書，4.1 節で抽出した新旧地名対を用いて，時空間を考慮したジオコーディングを行う．本研究の地名辞書には，国土地理院から収集した都道府県・市区町村の計 1,957 件，言語資源協会から収集した町丁や施設名の計 117,061 件，Wikipedia から収集した駅名や山などの計 44,930 件が登録されている．また，ジオコーディングは抽出された地名表現と地名辞書内の各地名を比較し，部分一致した地名の緯度経度情報を付与する．

しかし文書から抽出される地名表現には，古い地名表現や曖昧な地名表現の場合がある．古い地名表現の場合，地名辞書に登録されていないためジオコーディングが不可能である．そのため 4.2 節で抽出した新旧地名対を用いて，地名辞書に登録されている現在の地名表現に置き換えることで改善できる．図 4 を例に説明する．まず，新旧地名対を参照し，地名表現「与野市」や「大宮赤十字病院」と全ての古い地名表現を比較する．一致すれば抽出された地名表現を新しい地名表現「さいたま市」や「さいたま赤十字病院」に置き換える．

次に曖昧な地名表現の場合，1 章で述べたように該当する候補が複数あるため,場所の特定が困難となる．

そこで本研究は，曖昧でない地名表現から文書の地理的範囲を推定する手法を提案する．曖昧性問題解消における提案手法を図 5 を例に説明する．図中①では，部分一致を用いたジオコーディングを曖昧でない地名表現を対象に行う．図中②では，ジオコーディングされた地名表現における都道府県名を基に文書中の出現頻度を求める．図中③では，②で求めた出現頻度を参照し，出現頻度の多い都道府県名に属する候補が正しい緯度経度情報であると推定し，曖昧な地名表現に付与を行う．本研究では，出現頻度の多い都道府県名ほど文書の代表的な地理的範囲であると考え，出現頻度の多い都道府県名を優先して推定を行う．

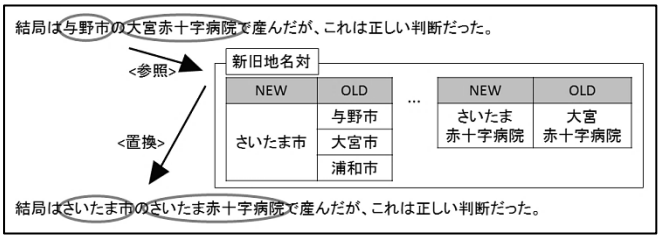

図 4. 古い地名表現の置き換え

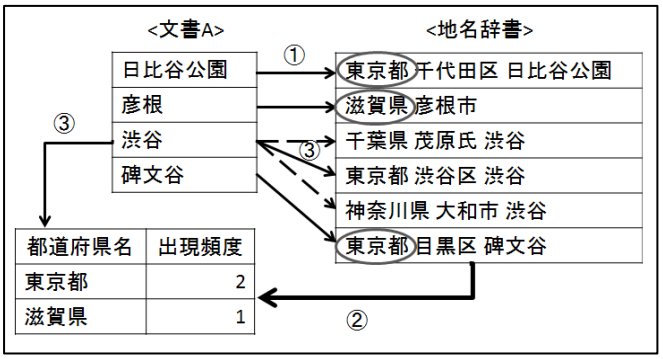

図 5. 曖昧な地名表現のジオコーディング

5. 評価実験

5.1. 新旧地名対抽出実験

本節では 4.1 節新旧地名対の対応付け手法の評価実験を行う．5.1.1 節では実験方法について述べ，5.1.2 節では実験結果と考察について述べる．

5.1.1. 実験方法

本実験では，4.1.2 節で作成した教師データを用いて 5 分割交差検定を行う．作成した教師データは，NTCIR-5 において使用された Web ページから「(旧」を含む文 6,914 文に対して人手で NEW タグ(3,958 件)，OLD タグ(4,002 件)が付与されている．そして，4.2.2 節の手法を用いて NEW と OLD タグを自動で付与し，結果を本実験で評価する．評価尺度には，精度と再現率，F 値を用いる．

5.1.2. 結果と考察

5 分割交差検定を行った結果を表 1 に示す．表 1 を見ると，NEW，OLD ともに全ての値が 8 割を超える結果が得られた．また，精度を見ると約 9 割あり，高い数値といえる．

表 1. 新旧地名抽出の実験結果

	精度	再現率	F 値
NEW	0.8887	0.8045	0.8445
OLD	0.8804	0.8231	0.8508

次に，新旧地名自動タグ付与システムを NTCIR-5，ClueWeb09 で使用された Web ページや読売新聞における「(旧)を含むすべての文に適用し，計 43,333 対の新旧地名対を抽出した．NTCIR-5 や ClueWeb09 では「HP – コンパック」といったコンピュータ企業の名称の対が読売新聞より比較的多く得られた．これは，2002 年に買収されたコンピュータ企業の旧名称が Web ページで広く使われていたためと考えられる．また，「HP – コンパック」の対は，NTCIR-5 と比べ ClueWeb09 は全抽出中での件数比率が下がっていた．これは NTCIR-5 の Web ページ収集時期が 2005 年，ClueWeb09 が 2009 年であったためと考えられる．読売新聞からは「ロシア – ソ連」，「新生銀行 – 日本長期信用銀行」といった会社名や国名が他と比べ多く抽出された．

これらのことから，新旧地名対の対応付け手法は，対象とするデータを様々な情報媒体や収集時期にすることで効率的かつ網羅的な対の収集が可能であると考えられる．また，精度の高さや自動抽出可能な観点から古い地名表現のジオコーディングに適切であると考えられる．

5.2. ジオコーディング実験

本節では 4.2 節における提案手法の評価実験を述べる．5.2.1 節では実験方法について述べ，5.1.2 節では実験結果と考察について述べる．

5.2.1. 実験方法

本実験では読売新聞記事 5 年間(1993,1998,2003,2008,2012)の内，各 10 件の記事を用いて 4.3 節における提案手法の評価実験を行う．正解データには，4.1 節で抽出した地名表現に対して地名辞書を用いて人手で緯度経度情報を付与したデータを用いる．ただし，緯度経度情報が付与できない地名表現は本実験のデータからは除く．本実験に用いる手法を以下に示す．

- ベースライン手法
  曖昧な地名表現に対して候補からランダムで選択した緯度経度情報を付与する手法

- 提案手法

  4 章で述べたジオコーディング手法

評価尺度には精度を用いる．算出法を式(1)に示す．

$$精度=\frac{正しく付与された地名表現数}{付与された地名表現数} \quad (1)$$

5.2.2. 結果と考察

実験結果の表 2 を見ると，全ての年代において曖昧な地名表現が存在するため，精度が向上する結果が得られた．

表 2. 地名への緯度経度情報付与の実験結果

発行年 (地名表現数)	曖昧な 地名表現数	ベース ライン	提案手法
1993 年(22 件)	6 件	0.727	0.909
1998 年(41 件)	2 件	0.951	0.951
2003 年(31 件)	1 件	0.968	1.000
2008 年(31 件)	3 件	0.900	0.935
2012 年(52 件)	8 件	0.840	0.942

次に，都道府県名の出現頻度を用いたジオコーディングについて，考察する．本実験では対象データを新聞記事としているため，曖昧な地名表現が少ない．そのため，曖昧でない地名表現における都道府県名の出現頻度の取得が容易であった．結果として，提案手法では精度が高くなったと考えられる．

旧地名の新地名への置き換えについて考察を述べる．提案手法において置き換えを行ったが，置き換えられた件数は非常に少ない結果であった．これには，1993 年から 2012 年までの間に名称が変わった地名があまり多くなかったことが考えられる．しかし，置き換えられた地名表現の例に「久留米市-田主丸町」がある．この置き換えにより，ベースライン手法では行われなかったジオコーディングが提案手法では行われる結果が得られた．これらのことから有効性を示すため，対象データをより古い時期に発行されたデータもしくは歴史を言及するようなデータに変更する必要性があると考える．

以上のことから，都道府県名の出現頻度を用いた手法はベースライン手法を上回る精度が得られ，有効性が示せたといえる．また，新旧地名の対応付けによる新地名への置き換え手法は異なるデータを対象とし，再実験の必要があると考えられる．

6. おわりに

本研究では，新聞記事や旅行ブログ，書籍といった様々な文書を同一地図上にマッピングするシステムの構築を行った．また文書から抽出された地名表現に対して，曖昧性問題や古い地名表現を考慮したジオコーディングを提案した．曖昧性問題には，文書における都道府県名の出現頻度を用いる手法を提案した．また古い地名表現には，機械学習手法 CRF を用いることで新旧地名の対を抽出し，古い地名表現を対の新しい地名表現に置き換える手法を提案した．実験の結果より，新旧地名対の抽出では NEW タグや OLD タグの自動付与の F 値はともに 0.800 を超える結果が得られた．ジオコーディングでは，読売新聞の記事を対象に行った結果，全てにおいて精度 0.900 を超える結果が得た．

今後は，古い地名表現の置き換えの有効性が示せるデータを対象に実験を行う．また，新旧地名対の抽出を様々なデータから抽出することを課題とする．

7. 謝辞

本研究では，国立情報学研究所主催の第 5 回 NTCIR ワークショップ Web 検索タスクで提供されている Web 文書を利用させていただいた．ここに記して，謹んで感謝の意を表す．

文　献

[1] 郡宏志, 服部峻, 手塚太郎, 田島敬史, 田中克己, “ブログからのビジターの代表的な行動経路とそのコンテキストの抽出，”情報処理学会研究報告データベース，Vol.2006, No.78, pp.35-42, 2006.

[2] 鎌田早織，坂本寛幸，井垣宏，中村匡秀，“マッシュアップ API を用いた異なるライフログサービスの連携，” 電子情報通信学会技術研究報告 LOIS, Vol.109, No.450, pp.91-96, 2010.

[3] 河野愛樹，中村健二，小柳滋，“マイクロブログから抽出した地物情報と投稿間隔を考慮した位置情報推定，” 情報処理学会全国大会講演論文集, Vol.2011, No.1, pp.785-787, 2011.

[4] 安田宜仁, 戸田浩之, “検索位置のごく周辺を対象とした地理情報検索”, 人工知能学会論文誌, Vol.23, No.5, pp.364-373, 2008.

[5] 金木雄太, 山田剛一, 絹川博之, 中川裕志, “地名辞書を利用した地名特定方式，” 情報科学技術フォーラム一般講演論文集, Vol.3, No.2, pp.181-182, 2004.

[6] 平野徹, 松尾義博, 菊井玄一郎, “地理的距離と有名度を用いた地名の曖昧性解消，” 情報処理学会全国大会講演論文集，Vol.70, No.2, pp.285-286, 2008.

[7] 国分芳宏，岡野弘行，“複数の観点で分類した自然言語処理用シソーラス，” 自然言語処理, Vol.17, No.1, pp.247-263, 2010.