2J3-NFC2-2

Wikipediaを用いた地名の包含関係情報の抽出

Extraction of geo-spatial relationships among geographical name by using Wikipedia

竹中 均[*1] 吉岡 真治[*2]
Hitoshi TAKENAKA Masaharu YOSHIOKA

[*1]北海道大学大学院情報科学研究科 (現在、ミツミ電機千歳事業所)
Graduate School of Information Science and Technology, Hokkaido University

[*2]北海道大学大学院情報科学研究科
Graduate School of Information Science and Technology, Hokkaido University

In order to construct Information Retrieval (IR) system with geographical name query, it is better to have a mechanism to use geo-spatial relationship for finding out the correspondence between the geographical name in query and documents. In this paper, we propose to use Wikipedia and GeoNames for constructing geo-spatial relationships about Japanese geographical names.

1. はじめに

近年、特定の場所に関連する情報を集めるロケーションベースの情報検索システムの開発などが行われていることなどから、地名に関連する情報を検索するというニーズが高まっている。これらに対応する形で、地理情報に関連する情報検索という課題が提案されている [Gey 10, Purves 10]。

このような検索システムでは、次のような問題を考える必要がある。例えば、日本の観光地を検索したいと考え、「日本　観光地」という検索キーワードを利用した場合を考える。この時、網羅的に文書を獲得するためには、「日本　観光地」を両方含むような文章だけではなく、北海道や東北といった地名が日本の一部であるといったことを考慮して検索を行う必要がある。

これらの検索を行うためには、地名の包含関係に関するデータベースが必要となる。これに対し、英語で利用可能な情報資源としては、GeoNames[*1]や GeoWordNet[Giunchiglia 10] などがあるものの、日本語の地名に関するこのようなデータベースで簡単に利用できるものは少なく、日本語における情報検索システムを構築する際の問題となっている。

本研究では、非常に網羅的に多くの地名が包含関係の情報と共に記述されている GeoNames を基準として、Wikipedia の情報と組み合わせることにより、日本語の地名に関する包含関係の情報を含むデータベースの構築を行う。

2. 地名に関する情報資源

2.1 GeoNames

GeoNames は、Creative Commons attribution ライセンスで開発されている地名情報に関するデータベースである。4/12 現在で、7,665,580 件の世界中の地名の情報が存在している。各地名の情報には、表 1 に示すようなデータ構造を持っている[*2]。

表 1: GeoNames のデータ構造 (概要)

属性名	内容
ID	地名の ID
名前	地名の ASCII 表記
別名	地名の各国語表記
場所	緯度・経度
国名	所属する国名
行政単位	所属する行政単位
タイプ	地名のタイプ

ここで、地名のタイプには、表 2 に示す 9 つの大分類と、さらに、詳細な小分類に分かれている。

表 2: GeoNames の地名タイプ

コード	タイプ	コード	タイプ
A	国、州、県など	S	ビルや農場など
H	川や湖など	T	山や丘など
L	公園など	U	深海
P	町や村など	V	森など
R	道路や鉄道など		

本データベースは、英語のデータベースとしては、非常に有意義なものである一方、日本語の別名がついているデータが 20,034 件 (0.26%) しか存在しない。ただし、大陸の名前 (7 件)、国名 (192 件) などの最も上位のカテゴリーには、全て、日本語の別名が存在する。一方で、州や県などを表すカテゴリーに属する地名 (3822 件) では、218 件 (5.7%) にしか日本語の情報が存在しない。

よって、本データベースを日本語の地名情報データベースとして利用するには、多くの地名、特に、州や県などを表す上位のカテゴリーに属する地名に対する日本語の別名の付加を行うことが望ましい。

連絡先: 吉岡真治, 北海道大学大学院情報科学研究科, 札幌市北区北 14 条西 9 丁目,011-706-7107, yoshioka@ist.hokudai.ac.jp

∗1 http://www.geonames.org/

∗2 詳細については、http://www.geonames.org/を参照のこと

2.2 Wikipedia

Wikipedia[*3]は、Wiki をベースとして作成された百科事典であり、幅広い分野に関する項目を網羅している。この Wikipedia のデータの特性に注目して Wikipedia マイニング [中山 07] による情報・知識の発見の研究が多く行われている。

地名に関する情報についても、広くデータが存在し、2.1 節で述べた GeoNames のデータを作成する際にも、Wikipedia の情報の一部が利用されている。

ここでは、本研究に関連する Wikipedia に関する特徴を説明する。

1. 項目情報の利用
   Wikipedia は百科辞典であるので、その中のテキストとして、地名に関する情報が存在する。特に、国名のリストや州や県などを表す上位のカテゴリーに属する地名に対しては、対応する項目 (ISO3166-1 や ISO3166-2 などの国名・地域名コードに関する項目) が存在し、そのテキストを解析することにより、多くの情報を得ることができる。

2. 言語間リンク・リダイレクトリンクの利用
   Wikipedia のページには、言語間リンクが存在し、日本語で得られた地域名が、他の言語でどのように表記されるかを調べることが可能である。また、表記にぶれがある場合には、リダイレクトリンクにより、代表表記を見つけることが可能である。

3. カテゴリ階層関係の利用
   多くの地理に関する情報は、「日本に関する地理」といったカテゴリのサブである「北海道に関する地理」といった形で、階層関係を持ったかたちで整理されている。この階層関係を使うことにより、階層関係の情報を獲得することが可能である。しかし、「愛知県の地理」→「尾張」→「織田氏」といったような、地名と関係しないカテゴリ階層に移動する場合があるので、注意をする必要がある。

3. GeoNames と Wikipedia を用いた日本語地名データベースの構築

2.1 節で紹介した GeoNames のデータ構造を用いることにより、地名から、関連する州・県あるいは国名などを獲得することができる。この情報は、本研究で利用を想定しているような地理情報に関する情報検索システムで、非常に有用な情報である。

しかし、現在存在する 750 万件を越えるデータについて、その全てを取り扱うことは、現実的ではないため、まず、本研究では、上位レベルのカテゴリである表 3 のカテゴリに注目して、日本語表記の付加を目指す。表 3 からも分かるように、GeoNames では、このような上位レベルの地名に対しても、十分な日本語表記の付加が行われていない。

表 3: 本研究で対象とする地名データ

コード	タイプ	Total	日本語
ADM1	行政単位 (州, 県,...)	3822	218
ADM2	行政単位 (市, 町,...)	26100	1086
PPL	街	2715183	2872

*3 http://ja.wikipedia.org/

3.1 Wikipedia の項目情報の利用

まず、最初に、最上位レベルの ADM1 については、できる限り信頼性の高い情報を利用することが、下位レベルの情報の信頼性の向上につながると考え、これらの情報について記載していると考えられる「ISO3166-2:JP」などの国名に対応する地域コードが記述されているページを利用した[*4]。

「ISO3166-2:国名コード」のページには、国名コードに対応する地域に対して地域コードが割り振られている。多くの場合、この地域名は、GeoNames における ADM1 に対応することが分かったため、以下の手順により、これらの項目情報から ADM1 に対応する日本語地名の付加を行った。

1. 日本語地名候補の抽出
   多くの場合、地名コード、日本語地名、現地表記 (あるいは英語表記) の組合わせが、表、箇条書きのスタイルで記述されていることが確認された。また、日本語地名の多くには、その地名を表すページへのリンクが存在することが確認された。

2. 対訳候補の抽出
   日本語地名候補の抽出の際に、現地表記 (あるいは英語表記) が関係付けられていた場合には、その表記を候補とする。また、地名を表すページへのリンクが存在し、かつ、英語の Wikipedia の言語間リンクが存在する場合には、そのリンク先のページの名前を表記の候補とする。

3. GeoNames とのマッチング
   上記の日本語地名と対訳の候補ペアは、各国ごとに作成されるので、対応する国に属する ADM1 レベルの地名に対する名前・別名と対訳候補を比較し、同じものがある場合には、日本語地名を別名として登録する。

この結果、1105 件の ADM1 の地名に対して、新しく日本語名称を付加することができた。また、このような地名のレベルでの言語間リンクの多くは、ほぼページの内容が 1 対 1 で対応することから、詳細なチェックは今後の課題となるが、ほぼ、間違いのない対訳関係が得られていると考えている。付加できなかったものの多くは、日本語版の「ISO3166-2:国名コード」のページが作成されていない、項目に記載されている情報から対訳候補の抽出が不可能だったものであった。

次に、国名の異表記について調べたところ、新聞などで良く用いられる漢字表記の国名の一部が別名として登録されていないことが判明した。これは、新聞記事などの解析を行う上で、大きな問題となるので、この情報についても、Wikipedia の項目情報から追加することとした。Wikipedia には、「国名の漢字表記一覧」という項目があり、この中に、国名の日本語による漢字表記とそのバリエーションが記述されている。しかし、その多くは、実際の文章にはほとんど現れないようなものがほとんどであり、いたずらにこのようなデータを追加することは、無駄な解析結果を産み出すだけと考えた。

これらのデータの内、IPA の地名辞書のバージョン 2.7.0 [*5]に国名として登録されているもののみを、異表記の候補として抽出した。追加した件数は 28 件で、主に、「朝」・「韓」・「中」・「日」・「米」などの一文字表記と、「米国」「豪州」などの中国の表記とは異なるような表記である。

*4 日本語版 Wikipedia の情報としては、2011 年 4 月 20 日付の情報を利用した

*5 http://sourceforge.jp/projects/ipadic/

3.2 Wikipedia の階層関係の利用

前節で述べた項目情報を利用することは、適切なページを見つけることができた場合には有用であるが、一つ一つ、役に立ちそうなページを見つけて処理をしていくのは、非常に手間がかかる。

そこで、Wikipedia のカテゴリの階層関係を利用した地名の包含関係の抽出を行う。Wikipedia のデータを分析したところ、多くの地名の情報が、「〜の地理」というカテゴリの下に分類されていることが確認された。また、「アメリカ合衆国の地理」→「「アメリカ合衆国の各州の地理」→「アラスカ州の地理」といった形の包含関係を示すデータとしても利用可能であることが確認された。一方、「アメリカ合衆国の地理」→「アメリカ合衆国の地震」といった、地名とは関係ないものも同時に含むことが確認された。

よって、単純にサブカテゴリを使う方法では、関係のない情報を多く取得してしまう可能性がある。これに対し、本研究では、GeoNames のデータに存在する所属する国名と行政組織の情報とを利用し、本研究で対象とする ADM1(行政単位 (州, 県,...))、ADM2(行政単位 (市, 町,...))、PPL(街) に対応する対訳データを持つもののみ抽出することとした。この結果、地震の名前などは、包含関係から排除することが可能となる。

以下に、具体的な手続きを述べる。

1. カテゴリーに属する項目の抽出
   「〜の地理」というカテゴリーから各々のカテゴリーに属する項目と、サブカテゴリーを抽出する。サブカテゴリーについては、同様に、カテゴリーに属する項目とサブカテゴリーの抽出を再帰的に行う。あまり、多段にカテゴリーを辿ると、関係のないサブカテゴリーが増えることから、この繰り返しを最大 4 段まで行う。

2. 対応する地名候補の GeoNames からの抽出
   「〜の地理」の「〜」の情報から、抽出した項目が属すべき国名や、ADM1 レベルの行政単位に対応する地名を GeoNames から抽出する。

3. 言語間リンクを用いた対応関係の推定
   抽出したカテゴリーの属する項目の各々について、英語の Wikipedia との間に言語間リンクが存在する場合には、その項目の名前と GeoNames から抽出した地名の名前、別名などと比較する。英語の Wikipedia では、曖昧性がある場合に、”San_Carlos,_Paraguay” の様に、”,”区切りで曖昧性解消に役立つ情報を付加する場合が多い。これらを踏まえて、”,”区切りのデータがある場合には、”,” の前までの名称と対応し比較を行った。この際に、対応が取れる場合には、GeoNames の地名に対する別名として登録することとした。

この結果、84,633 件の GeoNames について Wikipage の項目に対するリンクデータを作成することができた。しかし、行政単位による絞りこみが不十分だった場合に、一つの GeoNames に複数の Wikipedia の項目が対応したり、複数の Wikipedia の項目に一つの GeoNames が対応するといった問題があった。

具体的には、ダグラス郡は、複数のアメリカの州 (ワシントン・カンザスなど) に存在するが、アメリカの地理のサブカテゴリーから抽出を行ったために、ワシントン州、カンザス州のダグラス郡に対応する GeoNames がワシントン州、カンザス州のダグラス郡の Wikipedia の項目群に対応してしまうといった問題があった。

ここで、Wikipedia と GeoNames のデータの冗長性について考える。Wikipedia の項目には、基本的に冗長性がないと考えられるため、一つの GeoNames に複数の Wikipedia の項目が対応することは不適切であると考える。

一方で、GeoNames のデータの中には、例えば札幌市が ADM2(行政単位) であるのと同時に、街 (PPL) であるといった場合や、同じ名前の街のデータが複数存在すると行った場合がある。しかし、行政単位が違う街を同一として扱うことは問題であるので、今回のデータについては、ADM1 のレベルで異なる行政単位に属する事が確認されたデータを持つものについては不適切なリンクデータと判定することとした。

このような重複を考慮して、次の基準で、不適切な関係を削除することとした。

1. GeoNames を基準とした整理
   (a) 同一の GeoNames に対し、複数の Wikipedia の項目が対応する場合に、絞りこみに利用した行政単位を比較する。国名だけで絞りこんだものと、行政単位 ADM1 のコードも利用して絞りこんだものの両方が存在する場合には、国名だけで絞りこんだものを削除する。
   (b) 上記の処理を行った後、同一の GeoNames に対し、複数の Wikipedia の項目が対応する場合には、不適切な対応関係を含む場合があるので、対象の GeoNames に対する全てのリンクデータを削除する。

2. Wikipedia を基準とした整理
   (a) 同一の Wikipedia に対し、複数の GeoNames の項目が対応する場合に、絞りこみに利用した行政単位を比較する。行政単位 ADM1 のコードを利用して絞りこんだものが存在する場合には、そのコードを利用して対応する行政単位 ADM1 のコードと国名を決定する。
   (b) 絞りこみに利用した行政単位 ADM1 のコードや国名が複数存在する場合には、不適切なデータとして、全てのリンクデータを削除する。
   (c) 絞りこみに利用した国名と行政単位 ADM1 のコードが一つの場合は、そのコードと矛盾するリンクデータを削除する。
   (d) 絞りこみに利用した行政単位 ADM1 のコードが存在しない場合には、全ての対応する GeoNames の国名、ADM1 のコードを調べ、矛盾が存在する場合には、全てのリンクデータを削除する。

この判定により、37,415 件の GeoNames と 20,691 件の Wikipedia の項目の間のリンクデータが抽出された。件数が非常に多いため、適当なランダムサンプリングによる検証しかできていないが、ほぼ、適切な Wikipedia のページが獲得されていた。これは、単純に項目名だけで検索するのではなく、国名や行政単位を基準としたカテゴリー情報を使っているためと考えられる。

3.3 英語版 Wikipedia の利用

さらに、3.1 節,3.2 節で述べた方法を、英語版 Wikipedia についても適用を行った[*6]。

*6 英語版 Wikipedia の情報としては、2011 年 4 月 5 日付の情報を利用した

まず、最初に項目情報を利用した結果として、2,989 件の ADM1 について、英語版の Wikipedia との対応関係をつけることができた。また、その中の 1,717 件については、言語間リンクにより対応する日本語表記を得ることができた。

また、階層関係の情報については、「〜の地理」というカテゴリーではなく、「Geography of 〜」というカテゴリーを用いる以外は、全く同じアルゴリズムを用いて対応関係の発見を行った。その結果、499,457 件の GeoNames と 329,364 件の Wikipedia の項目の間のリンクデータを抽出することができた。

英語版の Wikipedia のデータについては、DBPedia 上に Wikipedia と GeoNames とのリンクデータベース 86,546 件[*7]が公開されているため、今回抽出したデータと比較することにより本手法の妥当性の検証が可能となる。

具体的には、本手法で抽出したデータの内、リンクデータベースに含まれる GeoNames に関するものを利用して、再現率、精度を計算したところ、再現率が 71.8%(62,155 件)、精度が 99.5%(62,155/62,474) となった。これらのことから、本提案手法では、非常に精度の良い対応関係が抽出できていることが確認できた。

また、対応関係が異なる 359 件について調べてみると、単純な間違い以外に、次のようなものが多く見られた。

- DBPedia 上の情報が古いため、対応するページの名前が変更されていたり、曖昧性解消のページになってしまっている。
- GeoNames のタイプの違うデータ間に対応をつけている (例えば、地域の名前にたいし、同じ名前を持つ街の名前のページを対応づける)

前者については、今回のデータが正解と判断しても良いと考えられるので、実際の精度は、もう少し、向上すると考えて良い。一方、後者については、Wikipedia のページ情報から、GeoNames のタイプを推定することなどを行うことにより、精度の向上が期待される。

本研究で抽出した Wikipedia と GeoNames とのリンクデータと Wikipedia の言語間リンクを用いることにより、38,485 件については対応する日本語表記を得ることができた。

これらの手続きの結果、全体として、新たに、41,580 件の地名について日本語を付加することができた。また、特に重要と考えた下記のタイプについては、表 4 に示す数の地名に日本語を付加することができた。

特に、ADM1 では、初期の段階では 37ヶ国 (実際の国名ではなく、ISO3166 で定義される国名コードに対応する国名) に対してのみにデータが存在するだけだったのに対し、今回のデータでは、195ヶ国の国の地方名に対してデータを付与することができた。

表 4: 本研究の結果日本語の別名を登録したデータ数

コード	タイプ	Total	日本語
ADM1	行政単位 (州, 県,...)	3,822	2,441
ADM2	行政単位 (市, 町,...)	26,100	3,143
PPL	街	2,715,183	20,710

これらの地名については、GeoNames のコード体系を利用することにより、地名の包含関係を表す情報として利用可能である。

∗7 http://wiki.dbpedia.org/Downloads36#linkstogeonames

4. おわりに

本研究では、Wikipedia と GeoNames という多言語対応の情報資源を組み合わせることによって、日本語の地名に関する包含関係の情報を含むデータベースの構築を試みた。本手法で生成する Wikipedia と GeoNames のリンクデータについては、英語版による評価ではあるが、かなり精度良く抽出できることが確認できた。

しかし、現時点では、日本語版 Wikipedia の項目が不足しているために、全体の登録数に比較して、十分な数の日本語訳を与えることはできていない。また、Wikipedia の項目に与えるカテゴリの情報が不十分な場合には、同表記の地名の分類がうまくできない場合があることも確認された。

今後は、GeoNames のタイプや座標の情報と Wikipedia 上に記述されているを比較したリンクデータの妥当性検証法などを検討すると共に、さらなる対応関係の生成方法についても検討していきたいと考えている。

また、本研究の成果として作成したリンクデータならびに、日本語地名データについては、ある一定の妥当性検証が終わった段階で公開を予定している。

謝辞

本研究の一部は、科研費基盤研究 (B) 21300029 により行われた。

参考文献

[Gey 10] Gey, F., Larson, R., Kando, N., Machado-Fisher, J., and Sakai, T.: NTCIR-GeoTime Overview: Evaluating Geographic and Temporal Search, in *Proceedings of the 8th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Quesiton Answering, And Cross-Lingual Information Access*, pp. 147–153 (2010)

[Giunchiglia 10] Giunchiglia, F., Maltese, V., Farazi, F., and Dutta, B.: GeoWordNet: A Resource for Geo-spatial Applications, in Aroyo, L., Antoniou, G., Hyvonen, E., Teije, ten A., Stuckenschmidt, H., Cabral, L., and Tudorache, T. eds., *The Semantic Web: Research and Applications*, Vol. 6088 of *Lecture Notes in Computer Science*, pp. 121–136, Springer Berlin / Heidelberg (2010)

[Purves 10] Purves, R., Clough, P., and Jones, C.: Highlights from GIR 2010: The 6th Workshop on Geographic Information Retrieval (Zurich, Switzerland - February 18–19, 2010), *SIGSPATIAL Special*, Vol. 2, pp. 17–23 (2010)

[中山 07] 中山 浩太郎, 原 隆浩, 西尾 章治郎：人工知能研究の新しいフロンティア：Wikipedia, 人工知能学会誌, Vol. 22, No. 5 (2007)