Vol.2011-NL-204 No.14
2011/11/22

固有名詞の所属国推定における表層情報の利用

延澤志保[†1,†2] 佐野智久[†2] 菊地弘晶[†2]
松原正樹[†2] 岡本紘幸[†2] 斎藤博昭[†2]

本稿では，固有名詞の表層情報のみを用いてその属性推定を行うことを目的に，地名および人名の表層情報の特徴の差異について検証，考察を行う．ここで対象とする表層情報は，文字単位，単語単位の長さ情報と，n-gram などの頻度情報である．地名および人名，さらにこれらを混合した 3 種類のコーパスを 11ヶ国について用意し，これらに対して，表層情報のみに基づくシンプルな所属国推定実験を行った結果，11ヶ国から平均 1.6 国程度まで出力を絞り込むことに成功し，平均 90%の再現率を得た．

The Use of Surface Information for Area Identification of Proper Nouns

Shiho Hoshi Nobesawa,[†1,†2] Tomohisa Sano,[†2]
Hiroaki Kikuchi,[†2] Masaki Matsubara,[†2]
Hiroyuki Okamoto[†2] and Hiroaki Saito[†2]

Area identification for proper nouns is an attribute estimation task in the field of unknown word processing. Our aim is to validate the effectiveness of surface information such as length data and n-gram data for area identification task of proper nouns, both toponyms and person names. Empirical results showed that a simple area identification method based on surface information succeeded in reducing area candidate from 11 to 1.6 on average, with 90% recall rate. Mixture corpora of toponyms and person names also obtained fair results.

†1 東京都市大学
Tokyo City University
†2 慶應義塾大学
Keio University

1. はじめに

固有名詞の認識およびその属性の推定の一タスクに，固有名詞の所属国の推定処理がある．所属国推定処理は特に地名について研究が盛んだが，十分な文脈や知識情報を前提とすることが一般的で，むしろ人名における名寄せ処理に内容が近いことが多い．人間は，見知らぬ固有名詞であっても，字面の印象である程度その所属国を推定することが可能である．この字面の印象による推定を自動処理として実現することを目標に，固有名詞の持つ表層情報について検証し，これを用いた所属国推定を行う．

これまで，所属国推定は地名または人名のみを対象とするものが多かった．同じ固有名詞である地名と人名が混在した状態で所属国推定ができるようになれば，形態素解析時における属性推定処理の簡素化に繋がるだけでなく，互いを文脈情報として積極的に利用することが期待できる．また地名と人名を同時に処理できれば，コーパスの拡充が容易になるという利点があり，より小さなエリアへ対象を拡大できる可能性が生まれる．そこで本稿では，地名と人名の表層情報と，これらに対してシンプルな所属国推定処理を行った結果とを挙げ，表層情報の面から地名と人名の比較を行う．

2. 固有名詞の所属国推定

2.1 地名の所属国推定

地名の所属国推定では，その地名の出現する文脈に地名の所属国を特定できるような情報が存在することを仮定している研究が多い．例えば，`Sparta` という地名の所属国を推定する場合に，文脈中の `Greece` などの国名を探したり，あるいは `Wisconsin` など他の地名を知識情報として辞書引きを行うことで所属国の曖昧性の解消を図ったりするタスクがこれに当たる．地名の曖昧性の解消を図るタスクでは，Smith らが文書中の地名の地図上での重心を推定する方法を提案している[12] ほか，Li らがパターンマッチングと類似度グラフを利用して十分大きな地名辞書で」探索を行う手法で 93.8%の精度を実現している[5),6)]．文脈を利用する手法では，Garbin らの機械学習器を用いる手法[3)] や出現地名同士の物理的な距離を利用する Zong らの手法[13)]，人口情報などの知識情報を利用する Amitay らの手法[1)] や 2 段階絞込みを行う Ladra らの手法[4)] など，さまざまな手法が提案されている．

これに対して，文脈に頼らず，表層情報のみを利用して所属国の推定を行うアプローチがある．Sano らは，表層情報が所属国推定処理に有効とした[10),11)] うえで，表層情報が類似する国同士の間で判別を行う手法についても検討している[9)]．

Vol.2011-NL-204 No.14
2011/11/22

2.2 人名の所属国推定

Nobesawa らは，表層的な統計情報のみを利用して人名の言語推定を行う手法を提案し，最大 92.93%，平均 76.67% の精度で言語推定に成功したと結論付けている[7]．Nobesawa らはさらに英語を主に用いる 5 エリアのみを対象とした人名のエリア推定実験を行い，類似エリア間であっても，表層情報のみによって 60% 以上の精度でエリアを識別することは可能と結論付けている[8]．また Bhargava らは線形 SVM によるシンプルな方法で， 13 言語を対象とした実験で 79.9% の正解率を得たと報告している[2]．

3. 固有名詞の表層情報

3.1 コーパス

本稿では，固有名詞として，地名および人名を考える．コーパスとして，国ごとに地名リスト，人名リストの 2 種類のリストを準備した．各リストは地名または人名を 1 行 1 個の形式で羅列したもので，他の情報は含んでいない．リストはすべてラテン文字で記述されており，表記の揺れに対応するため，大文字小文字の区別はしない．またアルファベット 26 文字と空白以外の記号は含まないものとし，ハイフン，カンマ，ピリオドは空白に置き換え，そのほかの記号は削除する．

本稿で検証対象とする国はアジアおよびヨーロッパの 11 カ国（オーストラリア，中国，ドイツ，フィンランド，フランス，ギリシャ，日本，韓国，ロシア，スウェーデン，タイ）である[⋆1]．それぞれの国について，地名および人名それぞれ 4,000 個を無作為に抽出し，実験のためのコーパスとして用いる．

3.2 長さ情報

図 1 に固有名詞コーパスの長さデータを示す． 図 1 の x 軸は固有名詞 1 個あたりの平均語数，y 軸は固有名詞に含まれる語 1 個あたりの平均文字数を示し，それぞれ，各国の地名と人名とを分けて集計している．

本稿で用いる人名コーパスは頭文字のみの語を含む人名を除いてあるため，人名を構成する語の平均数はほぼ 2 前後に集中しており，例外はロシア（平均 2.76 語）と韓国（平均 2.81 語）のみである．各語の文字列長については，欧米系が 6 前後に集中したのに対してそれ以外の地域では平均 8 以上と長くなるか平均 4 程度と短くなるかのどちらかが多くなっ

⋆1 各国を示す略記は以下のとおり．au:オーストラリア，cn:中国，de:ドイツ，fi:フィンランド，fr:フランス，gr:ギリシャ，jp:日本，kr:韓国，ru:ロシア，se:スウェーデン，th:タイ．また，各国を示す略記の前に付いた g は地名を，p は人名を，gp は地名と人名の混合を示す（混合については 4.1 節を参照のこと）．

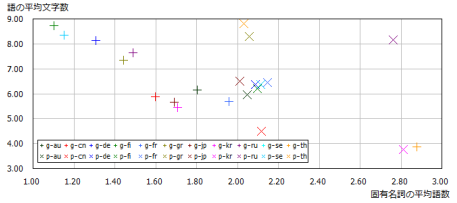

図 1 固有名詞の長さデータ
Fig. 1 Length data of toponyms and person names.

図 2 地名を構成する語の文字列長ごとの出現頻度（左図）および人名を構成する語の文字列長ごとの出現頻度（右図）
Fig. 2 Frequency of toponyms by word (left) and frequency of person names by word (right).

ており，所属国の識別の一助となることが期待される．

それに対して地名コーパスは，平均語数も平均文字列長も国ごとに差異が見られる．アジア系の地名は文字列長が比較的短く，欧米系の地名は文字列長が長く平均語数が少ないため，地名についてはアジア系と欧米系で分かれる傾向がある（図 1）．

図 2 に地名および人名の長さデータを示す． 図 2 のそれぞれのグラフについて，国ごとに語の長さおよび語の出現数の分布には差異が見られる．

地名の場合，タイは 3 文字前後の短い語が数多く出現し（地名に含まれる語の平均文字列長は 2.88）コーパス全体では 11,510 個の語が出現する（1 地名あたりの平均語数は 3.86）．それに対して，ドイツやフィンランドなどは 5 文字から 10 文字程度の比較的長い語が多く，出現する語の総数は 5000 程度にとどまる．

人名の場合，それぞれの人名の構成語の平均数はロシアと韓国の例外を除いてほぼ 2 前後であり，コーパスに含まれる語の総数については上記 2ヶ国以外はほぼ横並びである．しかし，語の長さについては，地名ほど明確でないものの，国ごとの差異が見られる．

Vol.2011-NL-204 No.14
2011/11/22

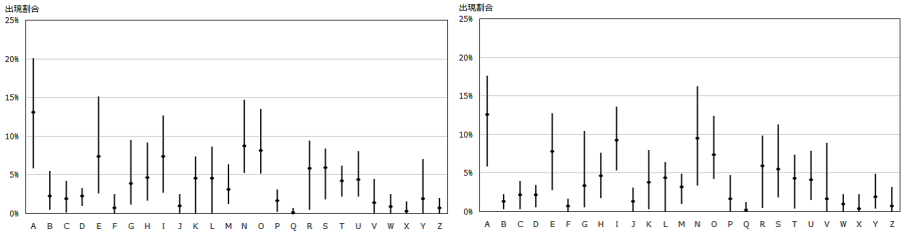

図 3 地名に含まれる文字の出現傾向（左図）および人名に含まれる文字の出現傾向（右図）
Fig. 3 Alphabet ratio in toponyms (left) and alphabet ratio in person names (right).

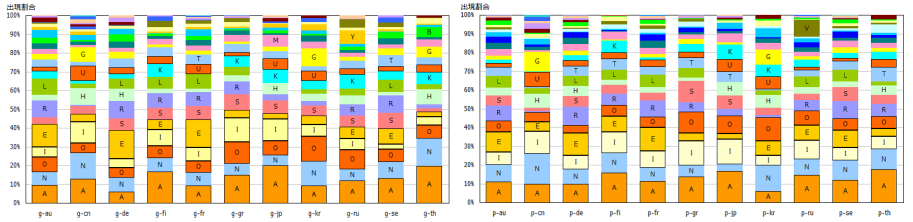

図 4 地名に含まれる文字の国ごとの出現割合（左図）および人名に含まれる文字の国ごとの出現割合（右図）
Fig. 4 Alphabet ratio in toponyms per area (left) and alphabet ratio in person names per area (right).

3.3 *N*-gram 情報

3.3.1 unigram 情報

図 3 に地名および人名に含まれるアルファベット各文字の出現傾向を示す． 図 3 の両グラフともに，アルファベット各文字について各国での出現割合を調べ，その最高値と最低値，平均値を示している．例えばアルファベット `A` は，地名での全文字中の出現割合は 20.0%（日本）が最大，5.9%（ドイツ）が最低であるのに対して人名では最大 17.5%（タイ），最低 5.9%（韓国）であり，いずれにせよ，どの国でも 5%以上の高い割合で出現することを示している．

地名と人名ではそれぞれの各アルファベットの出現の傾向はほぼ同様だが，その最大値，最低値の値には違いが見られる．これは，国ごとにアルファベットの出現傾向を見ると，単に値が上下するだけでなく，同じ国でも地名と人名でアルファベットの出現傾向が異なることがわかってくる．

図 4 に地名および人名に含まれるアルファベット各文字の国ごとの出現割合を示す． 図 4 の両グラフともに，各国について全文字中の各アルファベットの割合を，各アルファベットの地名全体，人名全体の平均出現割合の高いものから順に積み上げたグラフである．例えば，地名でも人名でも一番出現割合の高い文字は `A`，次は `N` であった．図 4 のグラフ中，アルファベットが記載されている項目は，そのコーパスの中での出現割合が 5%を超えていることを示す[⋆1]．基本的には，文字単位で見ると，同じ国のコーパスでは地名も人名も同じような出現傾向を示す．しかし，例えばロシアの場合，地名では `Y` の出現割合の高さ（7.0%）が特徴的であるのに対して人名では `Y` の出現割合は低く（2.2%），逆に人名では `V` の出現割合が高い（8.9%）のに対して地名では `V` はそこまで高くない（4.4%）．後者は，ロシアの人名に `Ivanov` `Vladimir` `Ivanovich` といったような `V` を含む語が多いことに起因するもので，ロシアの人名の重要な特徴と考えられる．

⋆1 例えば，図 4 の地名コーパス（左図）の日本のグラフに `M` との文字があるのは，`M` が全文字中 5%以上出現した（6.3%）ことを示す．図 4 の人名コーパス（右図）では日本のグラフにも `M` ラベルは記載されていないが，これは同じ日本でも人名コーパス中では `M` は 5%未満だった（4.9%）ためである．

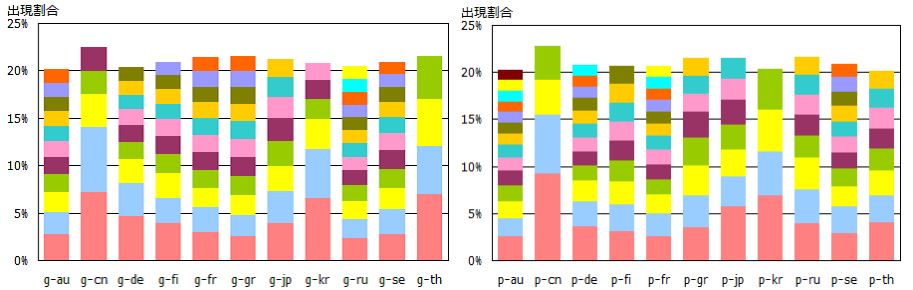

図 5 地名に含まれる文字 bigram の出現傾向（左図）および人名に含まれる文字 bigram の出現傾向（右図）
Fig. 5 Bigram ratio in toponyms (left) and bigram ratio in person names (right).

3.3.2 bigram 情報

図 5 に地名および人名に含まれる文字 bigram の出現傾向を示す． 図 5 の両グラフともに，各文字 bigram について各国での出現割合を調べ，出現割合の高いものから順に，累計出現割合が 20%を超えるまで積み上げている．地名，人名ともに，bigram の出現の分布は各国で差がある．中国，韓国，タイなどアジア系の国では少数の頻出 bigram が高い割合で出現するのに対して，欧米系の国では特に 3 位以下に似たような頻度の bigram が多く並ぶ．

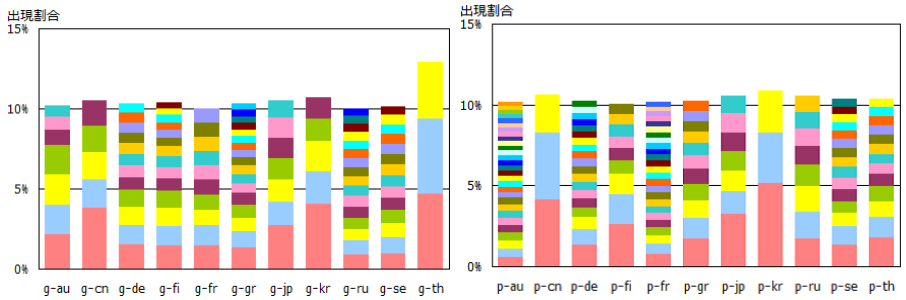

図 6 地名に含まれる文字 trigram の出現傾向（左図）および人名に含まれる文字 trigram の出現傾向（右図）
Fig. 6 Trigram ratio in toponyms (left) and trigram ratio in person names (right).

3.3.3 trigram 情報

図 6 に地名および人名に含まれる文字 trigram の出現傾向を示す． 図 6 の両グラフともに，各文字 trigram について各国での出現割合を調べ，出現割合の高いものから順に，累計出現割合が 10%を超えるまで積み上げている．trigram では，bigram よりもより明確に，アジア系と欧米系の差が見られる．アジア系では特定の trigram が特に高い頻度を示して上位数種類で全 trigram の 10%をカバーするのに対して，欧米系では 10%カバーするのに必要な trigram の種類が明らかに多い．例えば人名コーパスでは，オーストラリアは 10%をカバーするには上位 29 種類の trigram を必要としたが，中国および韓国では上位 3 種類の trigram で 10%をカバーしている．この傾向は人名で特に顕著である．

人名コーパスで上位 10%をカバーするのに最も多くの trigram を要したのはオーストラリアであり，このことは，オーストラリアの人名を構成する文字列が多岐に渡ることを示している．オーストラリア地名コーパスを構成する 4,000 の地名の中には `creek` という文字列が 726 回出現しており，これがこのコーパスのサイズに対して特に多かったため，出現割合が 2%前後と高い trigram が出現し，上位 4 位を占めた．試みに上位 5 位以降のみの出現割合を足し合わせて 10%を超える種類数を調べると 20 となり，上位 1 位からの場合の種類数 7 個に比べ，人名の場合の結果に近づく．

4. 表層情報を利用した所属国推定

4.1 固有名詞の所属国推定実験

ここでは，地名コーパス，人名コーパス，混合コーパスの 3 種類のコーパスについて所属国推定実験を行い，所属国推定における地名と人名の表層情報の効果について考察する．実験に用いるコーパスは 3 節で挙げたものと同じである．

（地名: g）各国について，地名のみ 4,000 個を要素とするコーパス．
（人名: p）各国について，人名のみ 4,000 個を要素とするコーパス．
（混合: gp）各国について，地名 2,000 個，人名 2,000 個の計 4,000 個を要素とするコーパス，ただし地名および人名はそれぞれ（g）および（p）の要素からランダムに抜粋している．

4.2 二値分類器による所属国推定

本稿では，Sano らの提案したシンプルな推定手法[11] を用いて固有名詞コーパスに対して所属国推定を行った．Sano らの手法では，それぞれのコーパスの文字単位および単語単位の長さ情報，文字単位の n-gram 情報を素性とする二値分類器を利用している．これはそれぞれのコーパスについて，ある 1 国のコーパスに含まれる要素を正事例，他の国のコーパスに含まれる要素をすべて負事例として機械学習を行う二値学習で，それぞれの入力文字列に対して実験対象とする国すべての二値分類器による判定を行い，それぞれの二値分類器が独立に結果を出力する．11ヶ国についてそれぞれ 4,000 個の固有名詞をコーパスとして実験対象とする本稿の実験の場合では，事前に，そのうち 1ヶ国の固有名詞 4,000 個を正事例，残り 10ヶ国の固有名詞計 40,000 個を負事例とする二値分類学習器を，正事例とする国を変えて計 11 個用意する．その上で，すべての固有名詞に対してこの 11 個の二値分類学習器を利用して各国に属するか否かを推定する．そのため，各固有名詞に対して，最低 0 個（すべての二値分類器が negative と判定した場合）から最大 11 個（すべての二値分類器が positive と判定した場合）の所属国候補を出力として得る．

この手法では，対象とする国の数が増えると正事例に対して負事例の割合が大きくなりバランスが悪くなる問題があるものの，Sano らは十分な再現率を実現できるとしている．所属国推定処理では，言語的，知識的，歴史的な背景などから同じ固有名詞が複数の国に含まれる可能性があるため，出力は必ずしも 1 個に絞れるとは限らず，適合率については 100%を実現することはできない．したがって，所属国推定処理では，再現率を十分保証しながらどこまで出力数を絞り込めるかが問題となる．

4.3 固有名詞の所属国推定結果

図 7 に，3 種類のコーパスでの実験結果の F 値を示す． 地名，人名，混合の 3 種類とも，F 値の平均は 0.70 程度であり，地名および人名については最大値 0.92 程度，最小値 0.57 程度，混合では最小値は地名や人名と同程度だったが最大値が 0.81 と低かった（**表 1**）．

Vol.2011-NL-204 No.14
2011/11/22

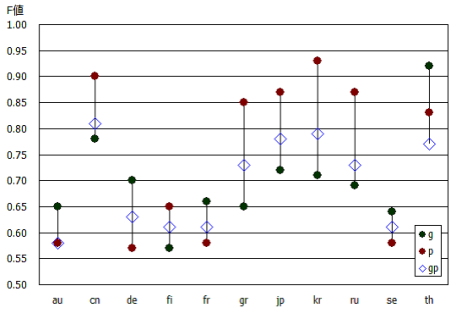

図 7 3 種類のコーパスの F 値の比較
Fig. 7 Comparison of F-measure values for toponym, person-name and mixed corpora.

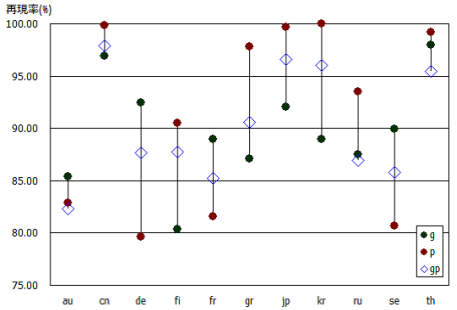

図 8 3 種類のコーパスの再現率の比較
Fig. 8 Comparison of recall values for toponym, person-name and mixed corpora.

表 1 所属国推定処理の結果（F 値）
Table 1 Results of area identification (F-measure).

コーパス	最大値	最小値	平均値
地名	0.92(th)	0.57(fi)	0.70
人名	0.93(kr)	0.57(de)	0.75
混合	0.81(cn)	0.58(au)	0.70

表 2 所属国推定処理の結果（再現率）
Table 2 Results of area identification (recall).

コーパス	最大値	最小値	平均値
地名	97.95% (th)	80.33%(fi)	89.78%
人名	100.00%(kr)	79.62%(de)	91.39%
混合	97.92% (cn)	82.33%(au)	90.21%

所属国推定結果の F 値の値には大きなばらつきがあり，特に人名については，0.60 以下となる国が 11ヶ国中 4ヶ国あったのに対して，0.85 以上の国は 5ヶ国，うち 2ヶ国は 0.90 を超える高い値を示すなど，国によって結果が大きく異なる．それに対して，地名では，タイが 0.92 と高い値を示したほかはすべて 0.80 に届かず，11ヶ国中 7ヶ国が 0.65–0.75 の範囲に集中した．地名と人名の混合コーパスでは，ほぼすべての国で人名と地名の結果の間の値をとったが，タイについては混合コーパスの結果が他を下回っている．このことから，基本的に，人名については所属国推定の精度は国によって大きな差が出る可能性があること，人名と地名を混在させて所属国推定を行った場合には最低でも人名のみの場合と地名のみの場合の精度の低い方と同程度の精度が期待できることが示された．ただし，タイのように，混在させることで結果が悪くなる場合がある．

F 値に関係する適合率，再現率のうち，所属国推定処理で特に重要となるのは再現率である．**図 8** に，3 種類のコーパスでの実験結果の再現率を示す． 地名，人名，混合の 3 種類とも，F 値の平均は 90%前後，最大値 97%以上，最小値 80%前後となった（**表 2**）． F 値同様，地名については，フィンランドが唯一地名について 80.3%と低い値を示しているほかはすべて 85–95%のあたりに集中しているのに対して，人名は再現率のとる値が 80%前後と 95%以上にほぼ 2 分化されている．再現率でも，タイのみ混合コーパスが地名のみおよび人名のみのコーパスよりも低い値を示したほかは，混合コーパスの値は少なくとも地名のみ人名のみのどちらか低い方と同程度となっている．

適合率は，positive と出力された全候補国中の正解国の数で算出される．所属国推定では必ずしも出力を一意に絞り込むことは可能ではなく，複数の国が推定結果として出力されることはむしろ自然であり，問題はこれをどこまで絞り込めるかにある．本実験の適合率は，例えばオーストラリアの地名コーパスでの実験の結果は 53.12%であったが，この内訳はオーストラリアの地名に対してオーストラリアが正しく推定結果として出力された数 3,417 件およびオーストラリア以外の地名に対してオーストラリアが誤って推定結果として出力された数 3,018 件から算出されている．実験対象は 11ヶ国なので，この結果は，オーストラリアを正解とする地名 4,000 個のうち 3,417 個が正しい出力を得，オーストラリア以外の地名 40,000 個のうち 3,018 個が誤ってオーストラリアを出力結果に含んだことを意味しており，正解を正しく出力した正解率（positive accuracy）は 85.43%，不正解を正し

表 3 所属国推定処理の結果（適合率）
Table 3 Results of area identification (precision).

コーパス	p/p	p/n	平均出力数	適合率
地名	89.78%	6.90%	1.59	56.5%
人名	91.39%	5.92%	1.51	60.7%
混合	90.21%	7.12%	1.61	55.9%

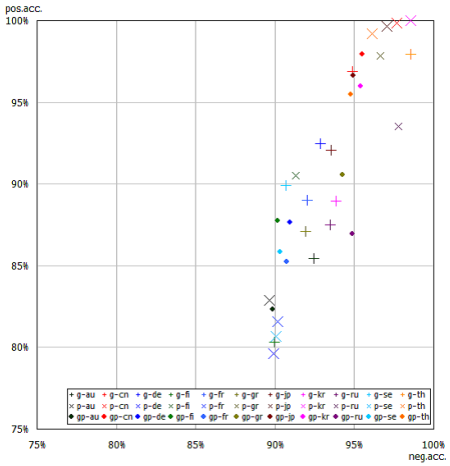

図 9 3 種類のコーパスの正解率の比較
Fig. 9 Comparison of accuracy values for toponym, person-name and mixed corpora.

く棄却した正解率（negative accuracy）は 92.46%と，この結果は十分に高い値と考えられる．地名，人名，混合の 3 種類とも，positive accuracy の平均は 90%前後，各国について誤ってその国を候補とした割合は 6–7%であり，平均すると 1 固有名詞あたり 1.6 個程度の国を所属国候補として出力していることになる（**表 3**，ただし p/p は正解国を出力した事例の割合，p/n は誤った国を出力した事例の全負事例中の割合を示す）． 実験対象国が 11ヶ国であること，固有名詞の所属国推定では候補が複数になり得ることを考えると，1.6 程度の出力は妥当な絞込み結果と考えられる．

図 9 に，3 種類のコーパスでの実験結果の正解率を示す． 図 9 に示したとおり，3 種類のコーパスのすべてについて，どの国でも positive accuracy は 80%程度以上，negative accuracy は 90%程度以上を実現しており，表層情報のみを用いた所属国候補の絞込みは十

表 4 人名コーパス中の頻出文字 bigram 上位 10 個
Table 4 Top 10 frequent bigrams in person-name corpora.

国	au	de	fr	se
1	AN	ER	AN	AN
2	AR	AN	ER	ER
3	ER	CH	E_	AR
4	N_	RI	IE	N_
5	ON	AR	RI	IN
6	IN	IN	AR	ON
7	EN	MA	RE	RI
8	NE	R_	IN	EN
9	EL	EL	N_	ST
10	LE	ST	LA	MA

分機能していることがわかる．ここでの positive accuracy は再現率であり，この値をさらに上げて 100%に近づける必要がある．negative accuracy は，誤った出力をどれだけ抑えられているかを示すため，適合率に直結する．これは，先に述べたとおり，複数の国に所属し得る固有名詞が存在することから 100%とすることは不可能である．negative accuracy については，これをさらに上げることよりも，実験対象国の数が増えた場合にも高い値を保持することがより重要である．

4.4 所属国推定での表層情報の利用

4.4.1 類似する特徴を持つ国の識別

図 7 および図 8 で示したとおり，所属国推定実験の結果は，F 値が最大 0.70 程度の国と，F 値が最小 0.70 程度の国とに大別できる．特に差の大きい人名について見ると，人名の F 値が地名の F 値を下回る 4ヶ国（オーストラリア，ドイツ，フランス，スウェーデン）は再現率の低さがその要因となっていることがわかる．これらの 4ヶ国は，図 6 で示したように，人名中に特徴的な文字 trigram が存在していない．bigram で見ても特徴的といえるものは少なく（図 5），さらにこの 4ヶ国の間で分布が似ている（**表 4**）ため混同され，再現率の低下に繋がったものと考えられる．このような類似国間の混同は Sano らの研究でその改善方法について提案されており[9]，さらに検討する必要がある．

4.4.2 混合コーパスの利用の不適切な事例

地名と人名は同じ固有名詞として扱われるが，同じ国であっても，これらの表層情報が似た傾向を示すとは限らない．図 2 の 2 グラフを比べると，同じ国でも地名と人名で語の長さおよびその出現頻度に差異が見られることがわかる．11ヵ国中，韓国とタイは明らかに語

長の分布が地名と人名で異なっている．このような場合，同じ国であっても，地名と人名を混合して扱うことはノイズの原因となる可能性が高い．

図 4 の 2 グラフを比べると，unigram についても，地名と人名とで差異が出ることがわかる．特に，ロシアでは地名に多い K と Y，人名に多い L と V に頻度の差が見られ，タイでは地名に多い B，G，K と人名に多い R，T が特徴的と考えられる．

図 7 ではタイが唯一混合コーパスでは値を下げている．これは，タイについては長さ，n-gram ともに地名と人名の差異が明確であり，混合することで互いがノイズになった可能性が高いものと考えられる．

5. おわりに

本稿では，固有名詞の所属国推定処理について，地名と人名の表層情報について比較するとともに，これらを利用した推定実験を行った結果を報告した．地名と人名との混合コーパスを処理に用いることができればコーパスの拡充が容易になるため，本稿では混合コーパスについても同様の推定実験を行い，地名のみ，人名のみのコーパスを用いた推定実験の結果と比較した．その結果，どのコーパスについても最も結果の悪い国で 80%前後，平均では 90%前後，最大 97%以上の再現率を得，地名，人名それぞれ表層情報のみで所属国推定が可能と考えられること，地名と人名を混合した場合でも表層情報のみを基にした所属国推定は可能であることが示された．実験結果から，地名と人名の表層情報の類似性については国によって差があること，そのため混合することで結果が悪くなるケースが考えられること，地名と人名の間の表層情報の差異が大きくない場合には混合することで精度が上がる可能性があることが示された．

今回の実験は規模も小さく，また類似した固有名詞を持つ国の間の判別について考慮していないため，地名，人名それぞれについて，今後さらに精度を上げることができるものと期待する．また混合コーパスの利用については，表層情報の効果的な利用について，さらに検討する余地があると考えられる．

参考文献

1) Amitay, E., Har'El, N., Sivan, R. and Soffer, A.: Web-a-Where: Geotagging Web Content, *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '04)*, New York, NY, USA, ACM, pp.273–280 (2004).
2) Bhargava, A. and Kondrak, G.: Language Identification of Names with SVMs, *Proceedings of the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, ACL, pp.693–696 (2010).
3) Garbin, E. and Mani, I.: Disambiguating Toponyms in News, *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT '05)*, pp.363–370 (2005).
4) Ladra, S., Luaces, M.R., Pedreira, O. and Seco, D.: A Toponym Resolution Service Following the OGC WPS Standard, *Proceedings of the 8th International Symposium on Web and Wireless Geographical Information Systems (W2GIS '08)*, Berlin, Heidelberg, Springer-Verlag, pp.75–85 (2008).
5) Li, H., Srihari, R.K., Niu, C. and Li, W.: Location Normalization for Information Extraction, *Proceedings of the 19th International Conference on Computational Linguistics (COLING '02)*, pp.1–7 (2002).
6) Li, H., Srihari, R.K., Niu, C. and Li, W.: InfoXtract Location Normalization: A Hybrid Approach to Geographic References in Information Extraction, *Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References*, pp.39–44 (2003).
7) Nobesawa, S. and Tahara, I.: Language Identification for Person Names Based on Statistical Information, *Proceedings of the 19th Pacific Asia Conference on Language, Information and Computation (PACLIC 19)*, pp.289–296 (2005).
8) Nobesawa, S. and Tahara, I.: Area Identification of English Person Names Based on Statistical Information, *Proceedings of the 19th IEEE Canadian Conference on Electrical and Computer Engineering (CCECE '06)*, pp.1688–1691 (2006).
9) 佐野智久，延澤志保，岡本紘幸，鈴木宏哉，松原正樹，斎藤博昭：候補間の表層的差異に着目した地名の所属国推定，言語処理学会自然言語処理， Vol.17, No.1, pp.29–54 (2010).
10) Sano, T., Nobesawa, S.H., Okamoto, H., Susuki, H., Matsubara, M. and Saito, H.: Robust Toponym Resolution Based on Surface Statistics, *IEICE Transactions on Information and Systems*, Vol.E92-D, No.12, pp.2313–2320 (2009).
11) Sano, T., Nobesawa, S.H. and Saito, H.: Automatic Country Identification of Area Names Based on Surface Features, *Proceedings of the 7th International Symposium on Natural Language Processing (SNLP '07)*, pp.13–18 (2007).
12) Smith, D.A. and Crane, G.: Disambiguating Geographic Names in a Historical Digital Library, *Proceedings of the 5th European Conference on Research and Advanced Technology for Digital Libraries (ECDL '01)*, London, UK, Springer-Verlag, pp.127–136 (2001).
13) Zong, W., Wu, D., Sun, A., Lim, E.-P. and Goh, D. H.-L.: On Assigning Place Names to Geography Related Web Pages, *Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL '05)*, pp.354–362 (2005).