学位論文 博士 (工学)

表層情報に基づく
固有表現の属性推定に関する研究

平成 21 年度

慶應義塾大学大学院理工学研究科

佐 野　智 久

目 次

第 1 章　はじめに **1**
1.1 背景 1
1.2 本研究の目的 2
1.3 本研究の構成 2

第 2 章　語句の属性推定の関連研究 **3**
2.1 未知語の認識 3
2.1.1 静的な語句認識 3
2.1.2 動的な語句認識 5
2.2 言語推定 6
2.2.1 文章を対象とした言語推定 6
2.2.2 語句を対象とした言語推定 7
2.3 語句の属性推定処理 7
2.3.1 語句の構文レベルの属性の推定 7
2.3.2 語句の意味レベルの属性の推定 8

第 3 章　表層情報に基づく地名の所属エリア推定 **12**
3.1 固有表現の属性推定 12
3.2 地名の属性推定 12
3.2.1 地名の属性推定の持つ意義 13
3.2.2 地名の所属エリア推定での地名データ 14
3.2.3 地名の所属エリア推定処理 15
3.2.4 ヒューリスティクスに依らない地名の属性推定 16
3.3 本研究で対象とするタスク 17
3.3.1 本研究の所属エリア推定システム 18
3.3.2 本研究で利用する地名コーパス 18
3.3.3 本研究の評価基準 22
3.3.4 本研究の提案手法 26

第 4 章　地名の所属エリア推定における表層情報の利用 (手法*1*) **28**
4.1 地名の表層情報 28
4.1.1 地名コーパス 28
4.1.2 地名の長さ情報 29
4.1.3 地名の n-gram 情報 31
4.2 表層情報を用いた地名の所属エリア推定手法 36
4.2.1 機械学習器を用いた所属エリア候補の推定 36

4.2.2 機械学習に用いる素性 37
4.3 実験結果 38
4.3.1 実験環境 38
4.3.2 実験結果 39
4.4 考察 40
4.4.1 地名の表層情報の有効性 40
4.4.2 文字レベルの n-gram 情報の利用における課題 42
4.4.3 類似した地名を有するエリア間の識別 43
4.5 手法*1*のまとめ 44

第 5 章 表層情報の効果的利用とブロックの提案 (手法*2*，手法*1+2*) 46
5.1 粒度の大きい処理単位としての単語レベルの表層情報 46
5.1.1 地名コーパス 46
5.1.2 単語レベルの表層情報 47
5.2 ブロックの提案および定義 50
5.2.1 新しい処理単位としてのブロックの提案 50
5.2.2 語頭ブロックおよび語尾ブロック 51
5.2.3 ブロックサイズとショートブロック 53
5.3 ブロックの概念を導入した所属エリア推定手法 54
5.3.1 ブロックを用いた所属エリア候補の推定 54
5.3.2 単語レベル TF-IDF に基づく地名のエリア所属確度 55
5.3.3 ブロックレベル TF-IDF に基づく地名のエリア所属確度 56
5.3.4 ブロックレベルと単語レベルでの所属エリア推定結果の比較 . . 57
5.4 ブロックの概念を導入した 2 段階推定手法 58
5.5 実験結果 60
5.5.1 実験環境 60
5.5.2 実験結果 61
5.6 考察 62
5.6.1 ブロックの有効性 62
5.6.2 出力する所属エリア候補数の妥当性 63
5.6.3 ブロックの効果的な利用 64
5.6.4 所属確度推定の妥当性 66
5.6.5 同一言語圏のエリアを含む所属エリア推定 67
5.7 手法*2*および手法*1+2*のまとめ 68

第 6 章 類似エリアからの所属エリア候補の絞込み (手法*3*) 70
6.1 類似エリア問題 70
6.1.1 地名コーパス 71
6.1.2 エリア間の地名の重複 71
6.1.3 n-gram モデルのパープレキシティ 72
6.1.4 類似エリアのグループ 73
6.2 候補絞込みと候補削減からなる 2 段階推定手法 73
6.2.1 生成確率を用いた所属エリア候補の絞込み 74

6.2.2 機械学習器を用いた所属エリア候補の削減 75
6.3 実験結果 76
6.3.1 実験環境 76
6.3.2 実験結果 76
6.4 考察 78
6.4.1 生成確率による所属エリア推定の妥当性 78
6.4.2 生成確率による所属エリア候補の絞込み 78
6.4.3 エリアの類似性に基づくグループの存在の検討 80
6.4.4 グループに着目した所属エリア推定 87
6.4.5 機械学習器の個数 90
6.5 手法*3*のまとめ 90

第 7 章 表層情報に基づく地名の所属エリア推定手法の提案 (手法*4*) 92
7.1 頑健な所属エリア推定手法の提案 92
7.1.1 所属エリア推定手法の統合 93
7.1.2 ブロックレベル情報を用いた所属エリア候補の絞込み 94
7.1.3 類似エリアを対象とした所属エリア候補の削減 96
7.1.4 対象エリア数の増加への対応 96
7.2 実験結果 98
7.2.1 実験環境 98
7.2.2 実験結果 98
7.3 考察 100
7.3.1 ブロックレベル情報を用いた所属エリア候補の絞込みの有効性 . 100
7.3.2 類似エリアを対象とした所属エリア候補の削減の妥当性 102
7.3.3 対象エリア数の増加の影響 103
7.4 手法*4*のまとめ 107

第 8 章 おわりに 109
8.1 結論 109
8.2 今後の展望 110

参考文献 112

図 目 次

2.1 文書中に混在する外国語句に期待される属性のレベル 8
2.2 エリア推定に用いられる手法のアプローチ 10

3.1 地名の属性推定 13
3.2 地名の所属エリア推定タスクの一般的な処理ステップ 15
3.3 所属エリア推定のステップ (従来手法) 16
3.4 所属エリア推定のステップ (提案手法) 17
3.5 本研究での所属エリア推定システムの概要 18
3.6 ウェブユーザ数の言語別推定 21
3.7 推定対象エリアと入力地名の推定結果の関係 22
3.8 二値分類学習器を用いた所属エリア推定実験の評価の例 25
3.9 本論文で提案する 4 手法の関係 26
3.10 手法*1*の概要 27
3.11 手法*2*の概要 27
3.12 手法*1+2*の概要 27
3.13 手法*3*の概要 27
3.14 手法*4*の概要 27

4.1 地名を構成する単語の長さの分布 29
4.2 地名を構成する単語の数と文字の数 30
4.3 unigram 出現傾向 32
4.4 unigram 出現傾向 (母音) 33
4.5 手法*1*の処理の流れ 36
4.6 機械学習に用いる素性 (長さ情報) 37
4.7 機械学習に用いる素性 (n-gram 情報) 38
4.8 n-gram の有する情報 40
4.9 手法*1*と unigram による推定との比較 41
4.10 中国の地名の所属エリア推定結果 43
4.11 台湾の地名の所属エリア推定結果 44

5.1 地名コーパスに出現する構成単語の例 47
5.2 出現頻度毎の単語の割合 49
5.3 手法*2*の処理の流れ 55
5.4 ブロックレベルでの処理結果と単語レベルでの処理結果の比較 58
5.5 手法*1+2*の処理の流れ 59
5.6 手法*1+2*の各フェーズで用いる表層情報 59

5.7 手法*1+2*での所属エリアの絞込み 59
5.8 手法*1+2*とその関連手法 62
5.9 各手法の実験結果の比較 63
5.10 出力エリア数の変化に伴う F 値の変化 64
5.11 ブロックのサイズおよび種類と F 値の変化 65
5.12 ブロックの使用方法の種類 65

6.1 手法*3*の処理の流れ 74
6.2 手法*3*の関連 3 手法の適合率および再現率の比較 77
6.3 生成確率による所属エリア推定の正解カバー率 79
6.4 エリア絞込み候補の順位毎のカバー率 81
6.5 グループに着目した絞込み結果の分類 86

7.1 手法*4*の処理の流れ 93
7.2 階層的クラスタリング 94
7.3 手法*4*とその関連手法での所属エリア推定結果 99
7.4 手法*4*でのエリア毎の実験結果 100
7.5 機械学習器の数と組合せカバー率 103
7.6 実験対象エリアを増やした場合の実験結果 104
7.7 実験対象エリア数の変化による実験結果の変化 105
7.8 実験対象エリア数を増やした場合の手法*4*のエリア毎の実験結果 106

表 目 次

3.1 実験対象エリア 21

4.1 本章で用いる地名コーパス 29
4.2 地名を構成する単語の数と文字の数 31
4.3 頻出 unigram 32
4.4 頻出 bigram 34
4.5 頻出 trigram 35
4.6 所属エリア推定の実験に用いる素性 39
4.7 地名の所属エリア推定の実験結果 39
4.8 unigram のみを用いた地名の所属エリア推定の実験結果 41
4.9 地名の所属エリア推定の成功例と失敗例 42

5.1 本章で用いる地名コーパス 46
5.2 実験対象エリア群で DF 値の高い単語 48
5.3 地名コーパス中での頻出単語 49
5.4 5 文字の語頭ブロックの例 (出現頻度上位 7 個) 51
5.5 5 文字の語尾ブロックの例 (出現頻度上位 7 個) 52
5.6 所属エリア推定の実験に用いる素性 60
5.7 手法*1+2*の実験結果 (ACR+ACEb) 61
5.8 手法*1+2*の実験結果 (確度値順位別集計) 67
5.9 類似エリアを加えた手法*1+2*の実験結果 68

6.1 本章で用いる地名コーパス 71
6.2 エリア間の地名の重なり数 72
6.3 地名コーパスに対するパープレキシティ 72
6.4 各手法を用いた所属エリア推定の結果 76
6.5 生成確率による所属エリア推定の結果 78
6.6 3 エリア組の出現傾向から推測されるグループ 81
6.7 3 エリア組の出現傾向 (グループ A) 82
6.8 3 エリア組の出現傾向 (グループ B) 84
6.9 3 エリア組の出現傾向 (グループ C) 85
6.10 機械学習器による所属エリア推定の結果 (グループ A) 88
6.11 機械学習器による所属エリア推定の結果 (グループ B) 89
6.12 機械学習器による所属エリア推定の結果 (グループ C) 89

7.1 手法*4*とその関連手法での所属エリア推定結果 98
7.2 ACSb モジュールによる絞込み結果 101

第1章　はじめに

本章では，本研究の背景および目的について述べ，本論文の構成を示す．

1.1　背景

現在では，ウェブ上の文書をはじめとして，多種多様な文書に簡単にアクセスすることができる．ニュースやブログの記事にはさまざまな出来事が記述されており，文書中に外国語句が含まれる例も多い．

自然言語処理は，その基本処理として，形態素解析および構文解析を行うことで文書中の要素の特定を行う．これらの処理は，現在，辞書ベースの手法が主流である．辞書ベースの手法では，辞書に記載されていない語 (辞書未登録語) は未知語として処理される．未知語は一般に情報を持たないため，文書中に意味的構文的な空白が含まれることとなり，これが形態素解析や構文解析の誤処理を引き起こしたり，また後に続く機械翻訳等の処理でノイズとなったりする等の問題が起こる．外国語句は一般に，当該言語で外来語としてある程度一般的になったものや専門用語等として辞書に記載されるものを除き，辞書に記載されていない．そのため，文書中に混在する外国語句は辞書未登録語となる．特に地名等の固有名詞は辞書未登録語であることが多く，文書の自動処理における未知語処理の問題の主因の一つとなっている．未知語の問題の解決は自然言語処理の分野で重要な課題のひとつであり，これを解決するために，語句認識や未知語の属性推定の研究が盛んに行われてきた．

語句認識とは，ひとかたまりの語あるいは句を成す文字列を推定する処理であり，これにより未知語となっていた文字列を語句と認識できるようにする．これは，辞書の登録エントリの自動拡充を目的として行うこともあれば，形態素解析や構文解析等の処理の中で動的に行う場合もある．

未知語の属性推定とは，語句に対してその言語や構文情報，意味情報等の属性を推定し付与する処理である．語句は，品詞等の構文情報がなければ，正しい構文解析を行うことができない．また，機械翻訳等の高次の処理を正しく行うためには，語句の意味的な情報が欠かせない．そのため，未知語に対しても，語句としての同定だけではなく，属性の付与を行う必要がある．

未知語の問題は，単に語句認識の問題ではなく，属性推定まで含めて解決する必要がある．本研究では，未知語のひとつとして文書中に混在する外国語句を対象とし，特に問題となりやすい固有名詞に着目して，その属性の自動付与を行う．

1.2　本研究の目的

文書中に混在した外国語句を未知語として認識することは可能だが，これに対して属性推定が正しく行われない場合，この語句は文書中で意味的構文的な空白となり，機械翻訳等の高次の処理で十分に活用することができない．外国語句は，文書の記述言語の辞書に情報がないことが予想され，この情報を適切に取得するには，この語句の属性を正しく推定する必要がある．これが可能になれば，その語句が記載されている辞書を見つけ出すことができ，その語句についての情報を取得することができるようになる．これによって，ノイズの原因となる未知語でしかなかった外国語句が，情報を持つ一般の語句と同様に利用できることになり，文書の意味解析に有用である．

しかし，辞書未登録語は，その属性の推定の手がかりとして，その語句の現れる文脈と，その表層的な情報しか得られない．文脈中に属性が明示されるとは限らず，これを前提として属性推定を行うのは望ましくない．そこで本研究では，表層的な情報のみから外国語句の属性を推定する手法を提案する．表層的な情報のみを利用することで，容易に複数の記述言語に対応することが可能であり，頑健性，汎用性の面で有効である．

本研究では，特に文章中に混在しやすい外国語句として地名に着目し，地名に情報を与えるための適切な属性として，所属する国や地域を自動的に推定する手法を提案する．

1.3　本研究の構成

本論文は，以下の構成をとっている．

第 2 章では，固有表現の自動認識についての関連研究を紹介し，本研究の位置付けを示す．

第 3 章では，本研究のタスクである地名の属性推定について述べる．その上で，第 4 章から第 7 章までで本研究で提案する手法についてそれぞれ概要を説明する．

第 4 章では，本研究の処理対象である各エリアの地名について，その表層情報の持つ情報を調査するとともに，これを利用することで地名の所属エリア推定が可能であることを示す (手法*1*)．

第 5 章では，地名の所属エリア推定処理に適した新しい処理単位を提案し，この処理単位の有効性を実験結果を基に検証する (手法*2*，手法*1+2*)．

第 6 章では，地名の所属エリア候補が複数存在し，この絞込みが困難な状況において，大域的な所属エリア推定処理と局所的な所属エリア推定処理の組合せにより，この課題の解決を試みる (手法*3*)．

第 7 章では，第 4 章，第 5 章，第 6 章の実験結果および考察を考慮し，本研究で対象としている地名の属性の推定を適切に処理するための手法 (手法*4*) を確立する．

第 8 章では，本研究の内容をまとめるとともに，今後の展望について述べる．

第2章　語句の属性推定の関連研究

本章では，語句の属性推定処理について，先行研究を紹介し，本研究との関連を説明する．本研究では，外国語句等，辞書に登録されていない語句への自動属性付与を目的としているため，以下の 3 点についての関連研究を示す．

- 未知語の認識
- 言語推定
- 語句の属性推定

まず，2.1 節で未知語の認識の手法について説明する．次に，自然言語文章を対象とした属性推定のひとつとしての言語推定について 2.2 節で説明する．その上で，本研究と直接的に関わる関連研究として語句の属性推定について 2.3 節で説明する．

2.1　未知語の認識

未知語の認識は，辞書の登録エントリの自動拡充を目的として行う場合と，形態素解析や構文解析の処理の中で動的に行う場合とに大別できる．語句認識では，意味のある文字列の抽出だけでなく，他の辞書登録語と同等に処理対象とできるようにするために，抽出された文字列への構文的意味的な属性の推定を行う必要がある．ここでは，意味のある文字列の抽出について説明し，抽出された文字列の属性推定については 2.3.1 節および 2.3.2 節で述べる．

2.1.1　静的な語句認識

辞書エントリの拡充等の目的で語句認識を行う場合には，大量のコーパスの中から意味のある語句として辞書登録可能な文字列を抽出する処理を行う．

一般的な語句の認識

意味を成すひとかたまりの文字列の抽出は，形態素解析と統計処理とを組み合わせることが多い．辞書登録語の組合せだけでは妥当な形態素解析が行えない場合，辞書未登録語の存在を疑い，その出現箇所を推定することが可能となる．不自然な形態素の並びに対して，辞書未登録ではあるがひとかたまりとなる可能性のある文字列を推定することで，語句の認識が可能となる．ひとかたまりとなる可能性の推定には，大

量のコーパスから得られる統計情報を基にする場合が多い．抽出対象は，主に名詞と名詞句とに二分されており，特に名詞の推定では，専門用語等複合語の推定を行うものと，固有名詞等の未知語の認識を行うものに分けられる．

Argamon らは語句認識の手法として，語句の組成についての形態論的ルールを利用して語句の認識を行う手法 [3] を提案している．また，サブパターンの概念を利用して名詞句をパターンとして認識する記憶ベースの手法 [2] も提案されている．

また，形態素解析処理や構文解析を利用せず，統計情報のみから語句認識を行う試みも行われている．日本語を対象とした手法としては，宇津呂らの手法 [70, 71]，池原らの共起情報を利用して語句の自動抽出を行う手法 [22]，永田の単語出現頻度の期待値を利用して語句の自動抽出を行う手法 [41]，中渡瀬らの手法 [42] 等が挙げられる．Palmer は切り分け規則の学習を行うことで語句切り分けを行う手法を提案し，英語のほか，中国語およびタイ語について実験を行っている [54]．また Lua らは，中国語を対象として，文字単位の隣接 bigram を基に，文の構造を再帰的に組み立てる手法を提案している [34]．

統計情報のみに特化した語句認識手法として， 延澤らは文字単位の共起情報のみを基に文字列がひとかたまりとなる確率を用いる手法を提案している [43, 44, 45, 51]．これは文を文字の並びと捉えてその分断箇所を統計情報を基に推定する手法であり，辞書等の情報を使わないため，一般的な語句に限らずドメイン固有の語句の抽出も可能となった．

専門用語や時事語句等の認識

言語の創造性を考えた場合，新しく出現する語句を自動的に認識する必要性が考えられる．これまでに，新しく出現した語句の自動抽出について，d-bigram と呼ぶ統計情報を用いて自動的に学習する手法を提案し，修士論文等にまとめている [47, 60]．ここで d-bigram とは，隣接 bigram だけでなくギャップ (距離) を持つ bigram を組み込んだ統計モデルであり， bigram に対して要素間のギャップに応じた重み付けを行うことで n-gram のような長い要素列についてもデータを利用可能とするものである [64, 65, 69]．この手法は，訓練コーパスの追加による d-bigram 統計情報の変化から新しく出現した語句の存在を推定するもので，逐次的な辞書の拡充を支援するものである．この手法を用いることで，時系列的に新しく出現する語句 (新語) だけでなく，異なるドメインで頻出する語句 (専門用語) 等の認識も可能となった．

専門用語等ドメイン固有の語句の抽出では，TF-IDF モデル等語句の出現頻度を利用する手法と n-gram による文字の共起頻度を利用する手法等がある．Frantzi らは英語を対象として C-value と呼ぶ評価値を利用して複合語を認識する手法を提案し，さらにこれにコーパスサイズや重み等を利用してランク付けを行っている [15]．日本語を対象とした研究では，湯本らによる専門用語抽出手法 [75]，延澤らの口語語句抽出手法 [46] 等がある．

また，Google はウェブ上の大量なデータから新語や専門用語，固有表現等を機械的，網羅的に収集し，豊富な語彙と統計的言語モデルを基にした日本語入力システムを開

発した[1]．その他，未知語の問題の解消に繋がる研究として，インターネット上で不特定多数が辞書を共有し新語や専門用語の登録を行えるようにすることで辞書の逐次的な拡充を図る手法 [52] 等が提案されている．

固有表現の認識

地名等を含む固有表現の認識は，文書中に埋め込まれている固有表現を抽出するタスクである．このタスクは未知語処理の問題を考える上で重要なタスクの一つである．固有表現の抽出では，固有名詞の前後に出現しやすい語をトリガーワードとして固有表現の認識を行う研究が一般的である．トリガーワードを利用した研究としては，英語を対象とした Wakao らの手法 [72]，中国語を対象とした Chen らの手法 [6, 7] 等が挙げられる．日本語を対象とした研究も，久光らの人名認識手法 [21] 等，数多く提案されている．また，Cucerzan らは文字の並びの情報を利用して，少ない訓練データから固有表現を推定する手法を提案し，ルーマニア語や英語等複数の言語に対して有効であることを示した [10]．Riloff らは，英語を対象として，文法情報を利用したドメイン固有の語句の辞書作成システムを提案している [59]．

他の言語処理タスクと同様に，固有表現認識についても，言語の特徴を利用する手法と言語に独立に処理を行う手法が提案されている．福本らは，固有表現認識に関して，言語による特徴の違いを考察している [16]．また，Fung は中国語と日本語に焦点を当てた手法を提案している [17]．それに対して， Curran らは，maximum entropy tagger を提案し，言語に独立した特徴の抽出を行い，これにより英語以外の言語に対しても効果的に固有表現の認識が可能になったと報告している [12]．

2.1.2 動的な語句認識

辞書未登録語の問題は，完全な辞書の作成が不可能なことから，完全に解決することは不可能である．そのため，形態素解析や構文解析等の文書中の語句の同定処理の中で動的に辞書未登録語の推定を行う必要がある．

延澤らは，辞書未登録語に起因する形態素解析誤りの解決のため，形態素解析処理と並行して語句の自動認識処理を行い，動的に認識された語句を形態素解析処理で利用することで形態素解析処理の精度を向上する手法を提案している [48]．この手法では，形態素解析処理中に平行して行う動的な語句認識処理と，形態素解析処理用の辞書の拡充のための前処理としての静的な語句認識処理とを併用することで，専門用語や新語等の語句を含めた頑健な形態素解析を実現している．

[1] Google 日本語入力: `http://www.google.com/intl/ja/ime/`，Google Japan Blog 2009 年 12 月 3 日「思いどおりの日本語入力 — Google 日本語入力」: `http://googlejapan.blogspot.com/2009/12/google_03.html`.

2.2 言語推定

ウェブ上の文書で用いられる言語の多様化にともなって，自動的な言語識別の重要性が増してきている．その一つの背景には，ウェブの急速な成長にともなう英語以外の文書の増加がある [19]．Internet World Stats による近年のウェブユーザ数の言語別集計の結果[2] によると，2009 年現在では，依然として英語を利用するユーザが最も多く，それに続いて，中国語，スペイン語，日本語，フランス語を利用するユーザが多い．2000 年からの言語別ユーザ数の増加の割合は，アラビア語，ポルトガル語，中国語，フランス語，スペイン語が大きな伸びをみせている．この調査結果は，ウェブ上で用いられる言語の多様性が増していることを示しているといえる．また，ウェブ上の文書の半数以上は英語以外の言語で書かれたものであると同時に，1 個の文書の中で複数の言語が使われていることもある．多種多様な情報源からの情報検索や質問応答，機械翻訳等，ウェブ上の膨大なデータを対象とした自動処理の実現においては，文書の言語の自動推定だけでなく，文書中に出現する固有表現等の外国語句の的確な解析も，自然言語処理の応用分野における精度向上に大きな影響を与える要因となり得る．

2.2.1 文章を対象とした言語推定

言語推定の研究は，音声認識したものを対象とする言語推定の研究 [4, 38] や，文書イメージを対象にした言語推定の研究だけでなく，自然言語で記述された文書テキストを対象としたものもある．Dunning は，ドイツ語の ü やフランス語の ê 等のアクセント記号を用いずに，5,000 バイトのトレーニングコーパスと 500 バイトのテスト用テキストを用いた言語識別の実験で 97% の精度を実現している [14]．Dunning は，英語とスペイン語のコーパスを用いた実験から，20 バイトのテスト用テキストでも 92% の精度での言語識別を実現し，短い文書や数単語で構成される句であっても言語識別が可能であることを示した．

Martins らは，ウェブページに特化した言語識別手法を提案している [37]．Martins らの手法は，n-gram 情報 (n は 1 から 5 まで) のプロファイル間の距離と，ウェブページ固有のヒューリスティクスを用いるものである．12 の言語で 500 のウェブページを用いた実験では，すべての言語で 84% 以上の正解を出したが，スウェーデン語とデンマーク語等の北欧の言語の類似性が若干の精度の低下をもたらしたことを今後の課題として挙げている．

Lins らは，文書中に含まれる複数の単語に対して各言語の辞書中での出現の有無を調べる手法で言語識別を行った [33]．Lins らは言語内で比較的種類が少ないとされている副詞，冠詞，接続詞，感嘆詞，数詞，前置詞，代名詞のみを辞書引きの対象とすることで，高速かつ汎用性の高い手法を提案している．Lins らはこの手法を用いて，ポルトガル語，スペイン語，フランス語，英語の文書 (約 1,000 単語で構成される 600 の文書) を対象とした評価実験を行い，ウェブの文書でも 80% 以上，通常の文書では 90% 以上の精度を達成している．

[2] Internet World Stats: `http://www.internetworldstats.com/stats7.htm`.

Sibun らは，文書イメージから抽出された文字の形状の統計的な分布を利用して言語識別を行った [66]．彼らは，アルファベットの文字の形状を，ベースライン，ボトム，トップ，X ハイトの情報を使って分類し，文書中の文字を Linear Discriminate Analysis を用いて分類した．2,000 から 3,000 文字を含む 23 の言語の文書イメージを用いて文書を構成する言語を識別する実験を行った結果，90% 以上の精度を達成している．

2.2.2 語句を対象とした言語推定

言語推定タスクは一般に文書や文章等を処理単位としており，語句単位での言語推定処理の研究は多くない．これは，1 個の文書の構成言語は 1 個であるとの仮定に則るもので，文書や文章全体に対して 1 個の言語を推定することを自然と捉える考え方のためである．しかし，実際には外国語句が混在する文書は数多い．複数の言語の文が混在するような文書は一般的ではないため，文単位の言語推定を文書単位や語句単位と独立して研究対象とすることの必要性は低いが，文中に混在した外国語句を対象として，その認識および言語推定を行うタスクへのニーズは十分あるものと考えられる．ところが，現実にはこのようなタスクは十分研究されているとはいえず，文書中に混在した外国語句は，通常の未知語処理の枠組みで処理されている．本研究で対象とする地名の所属エリア推定は，言語推定タスクのひとつと位置づけることが可能であるが，地名は一般に 2 単語程度の短い単語列であり利用できる情報が極端に少ないことが，通常の文章を対象とした言語推定と大きく異なる点である．

2.3 語句の属性推定処理

語句の属性推定とは，与えられた語句の有する属性を自動的に推定する処理である．例えば，その語句を語彙に含む言語を推定する言語推定も属性推定の一種と考えることができる． また，その語句の持つ概念，文法的特徴等も，語句の属性である．

一般に，文章中の語句の属性の取得には辞書を用いることが多い．そのため，辞書未登録語となりやすい外国語句は未知語となり，その意味はもちろん，品詞のレベルでも正しい属性情報を持たないまま「未知語」として処理されることも多く，これらの語句の存在が原因となって構文解析，機械翻訳等の処理の精度を低下させる恐れがある．外国語句に適切な属性情報を与えることができれば，処理精度の向上が期待できる．そのために必要な属性情報は，図 2.1 に示す 2 段階に分けて考えることができる．この 2 段階それぞれについて，これまでにさまざまな手法が提案されている．

2.3.1 語句の構文レベルの属性の推定

構文レベルの属性としては，主に品詞や係り受けの推定を行う．例えば， He is Hatoyama. という英文が与えられ未知語 Hatoyama が認識された場合， Hatoyama の品詞が名詞であることがわかれば，英語の SVC 文型に当てはめて he = Hatoyama との解析を行い **彼は** Hatoyama **である。** 等の翻訳を得ることが可能となる．この段階で，

構文レベルの属性:	この語句を含む文または文章の構文解析処理を成功させるために必要なレベルの属性．語句の品詞を正しく推定することができれば，構文解析処理は正しい結果を出力することが可能である．
意味レベルの属性:	この語句を含む文または文章の意味解析処理を成功させるために必要なレベルの属性．語句の意味を正しく推定することができれば，機械翻訳等の処理の精度向上に貢献できる．

図 2.1: 文書中に混在する外国語句に期待される属性のレベル

Hatoyama が名詞の中でも固有名詞であること，さらにその中でも人名に属することがわかれば，適切な処理が期待できる．

語句の品詞推定

現在の形態素解析処理は構文解析処理と同時に行われることが一般的であり，形態素解析処理の過程での未知語認識処理は，構文解析処理の結果を基に行うことが多い．そのため，未知語の認識は品詞の推定に基づいて行われることが多く，この場合には認識した未知語の品詞を改めて推定する処理は一般に行われない．品詞等の構文情報の推定が必要となる語句は，形態素解析処理と独立して認識，抽出される未知語が多く，この場合にはこの語句の出現する文脈等と照らして品詞の推定を行うことが一般的である．未知語はその性質上ほとんどの場合名詞と判断して差し支えない．

固有表現の種別の識別

未知語の主な要素の一つに挙げられる固有表現では，人名，地名，組織名等，その種別の識別が研究対象となっている．統計的な学習をベースとした手法 [5, 35, 39, 40, 68, 74, 76] や，ヒューリスティクスを用いた手法 [11, 67]，これらの手法を組み合わせた手法 [55] 等も提案されている．例えば，Zhou らは，固有表現認識の問題を対象として mutual information independence assumption を用いた HMM をベースとした chunk tagger を提案している [76]．彼らは，英語の固有表現の認識の実験で，96.6% の F 値を達成した．

2.3.2 語句の意味レベルの属性の推定

一般的な語句に対する意味レベルの属性の推定では，その語句の有する意味概念の認識が研究対象となる．これには，既存のシソーラスやオントロジから文脈に合った

意味概念を選択するタスクのほか，シソーラス等を用いずその語句の意味概念の推定を行うタスク等が含まれる． 固有表現の意味レベルの属性としては，人名に対しては姓と名の区別等，組織名に対しては組織の種類等が主な推定対象となっている．また，これらを利用した固有名詞の名寄せ処理等の研究がある．地名を対象とする場合には，その所属エリアの推定[3]を行うタスクが主に研究されている．

人名の所属エリアの推定

地名以外の固有表現として人名に着目し，統計情報を用いた所属エリアの推定を行う研究がある．Nobesawa らは，言語識別の手法を人名に対して適用することで，人名用の言語識別のためのシンプルなシステムを提案し，人名を属するエリアで分類することが可能であることを示した [49]．この手法は，人名文字列の長さや，人名の文字単位の n-gram の情報を活用したものであり，9 種類の言語圏の 12 のエリアに対して 90% 以上の精度を実現することに成功している．また，Nobesawa らは，英語の人名に対して SVM の機械学習器を用いたエリア推定の手法も提案している [50]．

地名の所属エリアの曖昧性の解消

地名は固有表現であり特定の場所を指し示す語句だが，同一の地名が複数のエリアに存在することは珍しくない．例えば，Portsmouth はイギリスとアメリカにあり，Sparta はギリシャとアメリカにある．この場合，例えば Sparta という地名で有名な場所はギリシャのペロポネソス地域にある Sparta であり，アメリカのウィスコンシン州モンロー郡の首都 Sparta はギリシャの Sparta ほど有名ではないが，このどちらもなんらかの文脈で文書中に出現する可能性があり，知名度から文書中に出現する Sparta がギリシャの地名の可能性が高いと推定することはできても，知名度だけでギリシャの地名であると断定することはできない．そこで，このような曖昧性を持つ地名に対してその所属エリアを特定するためにエリア推定を行う必要がある．

地名の所属エリアを特定する研究は，地名辞書の検索と，地名が出現する文書に対する自然言語処理のヒューリスティクスを用いたものがほとんどである．Leidner は，既存の研究で用いられている手法を図 2.2 の 17 種類に分類している [28, 29, 30]．Leidner の分類で示されるとおり，現時点でのエリア推定処理の方針は基本的にヒューリスティクスをベースにしている．

Smith らは地名の曖昧性の解消に焦点を当て，文書中の地名の出現頻度の重心を利用した手法を提案している [67]．これは，地名の出現頻度によって重みが付けられた地図上での重心を計算し，ある閾値よりも離れているものを枝刈りした上で重心を再度計算しなおすことで候補の曖昧性を解消しようとするものである．大きな地名辞書を使うことによって，再現率を高く保てるようにした上で，F 値が 0.81 から 0.96 と

[3]「所属エリア推定」では，地名が指し示す地理的場所が所属する国を推定対象とする場合が多い．ただし，社会的な理由から「国」との記述の代わりに「エリア」を用いることがあるため，本研究では，今後「国」との記述を用いず,「エリア」と記載することとする．

0 Assign unambiguous referent
1 'Contained-in' qualifier following
2 Superordinate mention
3 Largest population
4 One referent per discourse
5 Geometric minimality (minimal bounding polygon/distances)
6 Singleton capitals
7 Ignore small places
8 Focus on geographic area
9 Distance to unambiguous textual neighbours
10 Discard off-threshold
11 Frequency weighting
12 Prefer higher-level referents
13 Feature type disambiguator
14 Textual-Spatial Correlation
15 Default Referent
16 Preference order

図 2.2: エリア推定に用いられる手法のアプローチ

いう結果を出している．しかし，この手法はその重心の付近を示すだけであり，重心のみを使用しただけでは頑健性に欠ける場合があると結論付けている．

地名の曖昧性を解消するための手法として，Li らは，地名表現のパターンマッチングと重み付き地名の類似度グラフの探索，サーチエンジンを用いた地名辞書の補間の 3 個のアプローチを組み合わせる方法を提案している [31]．地名表現のパターンマッチングでは，city of という 2 語の並びの後に地名語が出現するパターン (例: city of Vancouver) や，複数の地名がカンマで接続されて出現するパターン (例: Boston, Massachusetts, USA) といった地名の周辺の表現のパターンを利用している．Li らは大きな地名辞書と地名表現のパターンマッチングを用いることで，93.8% の精度を実現した．また，Li らは，パターンマッチングや共起地名の特徴等を利用することで，地名にタグを付与する統計的な手法を提案している [32]．この手法の評価実験で，ニュース記事や旅行ガイドに対して 96% の精度を達成したと報告している．

Pouliquen らは，ヨーロッパのエリアに限定したマルチリンガルテキストを対象として，場所の判別の精度の向上を目指す手法を提案している [56, 57]．この手法では，And, To, Be 等の一般語として頻出する単語と同形異義語であるような地名を geo-stop list として抽出し，このような地名を候補から排除することで，精度の向上を図っている．また，それ以外にも場所の重要度，人名との区別，地名同士の物理的な距離の情報等を用いて曖昧性の解消を行っている．geo-stop list に登録されている地名を候補か

ら排除することで再現率は低下するが，F 値で 0.77 という結果を示している．

Clough は，複数の地名辞書を用いた場所の判別手法を提案している [9]．複数の地名辞書を優先順位を付けて検索し，stop-list を使って，各候補に対してスコアを与えている．このスコアは，地名表現の周辺の出現パターン，オントロジにおける階層の深さ，地名辞書の優先順位，ユーザのプリファレンスによって計算される．Clough はイギリス，フランス，ドイツ，スイスのみを対象とした実験で 89% の精度を達成した．

また，Rauch らは，辞書への登録状況，言語的な文脈情報等の特徴を用いて，信頼度ベースのエリア推定システムを提案している [58]．

Garbin と Mani は，エリア推定に統計的な機械学習器を用いた手法を提案している [18]．地名のクラスや短縮形の情報，文字の大文字の利用，3 語以内に共起する語の情報等のさまざまな特徴を利用している．人手でタグ付けされたニュースコーパスにおいて，地名の曖昧性解消で 78.5% の精度を達成した．

Zong らは，ウェブページに対してそのページが言及しているエリアを判別する実験を行っている [77]．Zong らはアメリカに関する文書のみを対象とし，地名の周辺の出現パターンと地名同士の物理的な距離を利用することで，地名が 32 個以上 199 個以下含まれるウェブページのみを対象に 760 の地名について実験を行い 88.9% の精度を達成した．

Amitay らは，Web-a-Where と呼ばれるシステムを使い，ウェブページを対象として，地名の曖昧性の解消とウェブページ全体がフォーカスしているエリアの特定を行っている [1]．Amitay らのシステムは，パターンマッチングや，人口情報，one referent per discourse 等のアプローチを組み合わせることで confidence score を計算している．実験では，7,000 以上の地名が含まれる 600 のページを使い，それぞれの地名に対して 82% の精度，ページに対して 91% の精度を達成した．

Ladra らは，まず所属エリア候補の絞込みを行ってから所属エリアの推定を行うという 2 段階の処理手法を提案している [27]．Ladra らの手法でも，処理に際しては地名に関する知識情報を推定の基にしている．

Hauptmann と Olligschlaeger は，音声データを対象とした場所の判別を行う手法を提案している [20, 53]．基本的には地名辞書に含まれる地名のみを対象としているが，同じ地名であっても複数の異なる場所を示す曖昧性がある場合には，同一の会話内に現れる他の地名の情報を活用することによって，その場所の相違を判断している．Hauptmann らの手法では，200 のニュース記事に出現した 357 の地名のうち 269 の地名を正しく判別することができ，75% の精度を達成した．Hauptmann らは，正しく判別することができなかったものは，地名辞書に載っていなかったもの，曖昧性によるエラー，音声認識誤り等が原因であると報告している．

ここに挙げたさまざまな関連研究の所属エリア推定のアプローチはほとんど，地名辞書，文脈情報，その他検索対象とするエリアに関する情報を組み合わせたものを利用するものである．地名辞書や文脈情報等，知識情報を活用したヒューリスティクスを必要とする手法の品質は，利用する知識情報の質と量にその精度を大きく依存するものであるといえる．

第3章　表層情報に基づく地名の所属エリア推定

本研究では，これまで未知語として扱われてきた固有表現に着目し，これに意味的な属性を自動的に付与する手法を提案する．本章では，本研究で対象とするタスクについて述べる．

3.1　固有表現の属性推定

文書中に混在する外国語句を処理対象とする場合，他の未知語と同様に，これらの語句は文字の並び等の表層的な情報以上の情報を有しておらず，これらは一般的な未知語処理の処理対象となる．しかし，外国語句の場合には，その語句を語彙として有する言語が推定できればその言語の辞書等の知識情報から処理に必要な情報を得ることができる可能性がある．その意味で，外国語句に対しては属性推定を行うことによって，通常の未知語処理で得られる以上の情報が得られる可能性がある．

文書中に混在する外国語句は，固有名詞や専門用語が多い．未知語処理を目的とした固有名詞の属性推定では，前述の理由から，対象語句の所属エリア推定が有効である．固有名詞は主に地名，人名，組織名に分けられるが，これらの固有名詞が指し示すものに対して所属エリアを推定することで，その語句に関する情報を知識情報から得ることが可能となる．

また，文書中に混在する外国語句には一般的な語句も考えられる．一般的な語句や専門用語の場合にも，その語句を語彙として有する言語を特定することができれば，辞書引き等を用いてその語句の情報を得ることが可能である．言語推定は主に文章を対象として行われるため，語句レベルの言語推定はほとんど研究されていない．語句レベルの言語推定では，表層情報のみが利用できる場合には，その文字列の短さが問題となる．しかし，語句の文字列の短さが抱える問題は本研究で処理対象とする固有名詞の持つ性質と同じであり，本研究の成果が一般的な語句の言語推定にも活用できるものと考える．

3.2　地名の属性推定

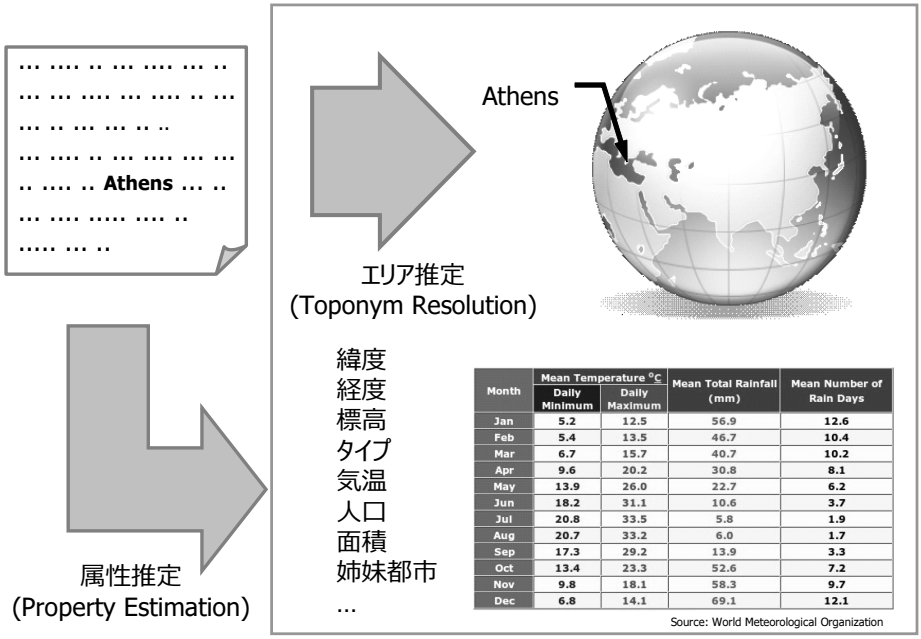
地名，人名，組織名等のどの固有名詞を対象とした場合でも，所属エリア推定の基本的な処理は，ほぼ同じような手法が適用可能であると考えられる．本研究では，その中でも特に混在が自然でありかつその属性の取得が文書の意味解析に重要な意味を

持つ地名に焦点を当てる．ニュース記事やブログ等，ウェブ上の文書では，外国の小さな都市や村が登場する可能性が増え続けているにもかかわらず，現時点では，その地名が所属するエリアを簡単に特定する方法があるとはいえない．

3.2.1 地名の属性推定の持つ意義

固有表現の属性の推定は，単に未知語の問題の解決というレベルのタスクではなく，機械翻訳や情報検索等の高次の自然言語処理タスクの内容に大きく関わるタスクといえる．例えば，前述の He is Hatoyama. の例では， Hatoyama が日本の人名であることがわかれば， Hatoyama に対してその妥当な発音を推定する等の処理が考えられる．さらに， Hatoyama が日本の人名と特定できた段階で，情報検索等の技術を用いて，これが日本の現在の首相の名前であること等，その人物の背景や特徴を引き出す処理への移行が可能である．人物が特定できた時点で， Hatoyama は単なる 8 文字の未知語から， 1 人の政治家を表す意味を持った語となり，機械翻訳処理等の高次の処理の段階でさらに適切な情報が利用可能となることが期待される．

同様に，地名についても，未知語を地名と認識でき，さらにその地名の所属エリアが特定できれば，その地名について既存の知識ベースから情報を引き出すことが可能となる (図 3.1)．ここで I went to Athens. という英文を日本語に翻訳する例を考える．



Month	Mean Temperature °C		Mean Total Rainfall (mm)	Mean Number of Rain Days
	Daily Minimum	Daily Maximum		
Jan	5.2	12.5	56.9	12.6
Feb	5.4	13.5	46.7	10.4
Mar	6.7	15.7	40.7	10.2
Apr	9.6	20.2	30.8	8.1
May	13.9	26.0	22.7	6.2
Jun	18.2	31.1	10.6	3.7
Jul	20.8	33.5	5.8	1.9
Aug	20.7	33.2	6.0	1.7
Sep	17.3	29.2	13.9	3.3
Oct	13.4	23.3	52.6	7.2
Nov	9.8	18.1	58.3	9.7
Dec	6.8	14.1	69.1	12.1

図 3.1: 地名の属性推定

現在の自然言語処理の目標は， Athens を地名として認識し Athens **へ行った。**と正しく機械翻訳することである．ところが，英語の go to の後には地名がくるとは限らず，例えば， All my work went to waste. (**私のやったことはすべて無駄になった。**) のように地名でない名詞がくる場合や， I went to see him. (**私は彼に会いに行った。**) のように

動詞がくる場合等が考えられ， Athens が地名と決まらない段階では， Athens **に行った**。だけでなく Athens **になった**。あるいは Athens **しに行った**。等と訳す可能性を排除できない．そのため，これを正しく翻訳するには，まず，Athens を地名と認識する必要がある．さらに，文章の言語を推定し，例えばこの例であれば英語の文章であることを認識した上で Athens を翻字し**アセンス**のように表記することも考えられる．

しかし，機械翻訳の目的を考えれば， Athens について，ギリシャの地名であることを推定したり過去の翻訳例を参照したりする等して，その日本語での表記が**アテネ**であることを推定し，**アテネへ行った**。と表現することが望ましい．所属エリアが判明している地名に対しては，翻字等を用いて日本語表記に変換するシステムが Google によって既にシステム化されており， 60 ヶ国で 80% 以上のカバー率を達成している[1]．

さらに，地名 Athens の持つさまざまな特徴や文脈等を考慮して，**ギリシャの首都アテネへ行った**。と説明を補うことが可能となれば，機械翻訳は単なる字面の翻訳を超え，伝えたい意味を伝えることのできる高度な処理となり，人間による翻訳に大きく近づくことになる．

このように，外国も含めた地名のエリア属性の推定が可能となることで，その地名の属性として，緯度経度や他の都市との位置関係等その地理上の位置情報や，人口等のその社会的特徴，また他の情報 (関連するニュース等) を引き出すことが可能となる．地名に限らず固有表現はさまざまな情報とリンクしており，これらを積極的に活用することで，固有表現を単に字面で処理するだけでなく，その意味も補うという高レベルの処理が可能となる．前述の例の I went to Athens. のように，地名が未知語として処理されている段階では，この地名は意味的な空白として処理されたり (Athens **へ行った**。のように地名部分はそのまま原語句をはめ込む形で処理される等)，場合によってはノイズとなったりする (Athens **しに行った**。のように地名と判別できず翻訳に誤りが生じる等)．しかし，地名の属性情報を積極的に活用することが可能となれば，単にその地名についての情報が増えるというだけではなく，その地名から得られるさまざまな属性情報から，処理のレベルを飛躍的に向上させる可能性がある．これにより，本研究は，単純な言葉の置き換えレベルの機械翻訳処理ではない，場面を想定した高度な人工知能的レベルの機械翻訳処理の実現の一助となるものと考えている．

3.2.2 地名の所属エリア推定での地名データ

地名については，地図作成や郵便業務等を目的として，ほとんどのエリアでも詳細な辞書が存在し，これを利用することでその地名に付随する国や場所等の属性を得ることが可能である．しかし，文書中に出現する地名はその文書の記述言語を母語とするエリアの地名であるとは限らず，ニュース文書等にあっては理論上全世界のどの地名も現れ得る．そのため，地名の特定にはすべてのエリアの詳細な地名辞書を確認する必要があることとなり，これは効率の面からも辞書の記述方式や記述粒度の不統一の面からも，現実的とはいえない．それにもかかわらず，非常に小さな都市や村の地

[1] Google Japan Blog 2009 年 8 月 28 日「世界の地名を日本語で」: `http://googlejapan.blogspot.com/2009/08/blog-post_28.html`.

名を見つけ出すために全世界の地名辞書のすべてを検索しなければならないとしたら，それは賢明ではない．文書の記述言語に依存せずに，あらゆる地名の所属エリアを自動的に特定する手法を確立することができれば，この検索作業の対象をより少ないエリアの地名辞書に絞り込むことが可能となる．

地名は，人名や組織名等の他の固有名詞と比べてその要素に変動が少なく，詳細な辞書の作成が可能という特徴がある．エリア毎に地名辞書は数多く作成されているため，これらを統合すれば全世界的に地名を網羅することは可能である．しかし，その記述形式や記述内容はさまざまであり，機械処理システム用の辞書としての利用を考えた場合には，これらすべてを統一的に利用する必要があり，やはり現実的ではない．

さらに，橋や運河等の建造物の存在や，複数の名称を持つ地名等を考慮すると，地名は必ずしも固定されたものではなく，新しい語が生まれたり，古い語が廃れたりといった一般的な語句と同様の特徴も備えている．そのため，前述の地名語句の数や粒度，地名辞書記述形式等の問題が解決できたとしても，地名辞書にすべての地名が掲載されていると期待することは難しく，辞書未登録語への対応が不可欠である．

3.2.3 地名の所属エリア推定処理

地名の所属エリア推定の最終的な目標は，その地名が地球上のどの場所を示しているのかを判断することである．文章中の地名の所属エリア推定タスクは，一般に，図 3.2 の 3 段階の処理に分けられる．(1) の地名文字列の認識は固有表現の自動抽出タスクに

(1) 地名文字列の認識
(2) 地名文字列の所属エリアの推定
(3) 地名文字列と場所との対応付け

図 3.2: 地名の所属エリア推定タスクの一般的な処理ステップ

相当する．(2) および (3) は，地名文字列と地球上の場所との対応付けを行う処理であり，(2) でエリアの絞込みを行った後 (3) で具体的な場所へのリンク付けを行う．例を挙げれば，ある文書に対して (1) で文書中の Sparta という語を地名と認識し，(2) で地名 Sparta の所属エリアがギリシャかアメリカである可能性が高いと推定し，(3) でその文書中での Sparta がアメリカのウィスコンシン州モンロー郡の地名を指していることを示すとの手順である．

第 2 章で述べたとおり，(1) の地名文字列の認識については固有表現の認識処理の研究が盛んに行われている．地名と場所との対応付けも既に研究されているが，(2) に当たる地名の所属エリアの推定の研究はほとんどなく，(2) の部分は文脈から明らかとの前提で (3) の処理を行う研究がほとんどである．(3) の地名文字列と場所との対応付けの研究では，あらかじめ対象ドメインや対象言語を制限することで，(2) の所属エリア推定のステップを省略することが多い．2.3.2 節で示したように，(3) としては，複

数のエリアに出現する可能性のある曖昧な地名を対象として，地名の辞書引きを行い所属エリア候補を決め，文脈情報と照らし合わせて地名を特定する手法が主に研究されている．しかし，(2) の段階である全世界的なエリアを対象とした所属エリア推定では，文脈中に所属エリアに関する情報が含まれていることを前提とすることは頑健性の面で問題がある．頑健性の高い処理を実現するためには，空間的な情報や人口の情報等の地名に関する属性のような知識情報や文脈中で共起するエリア名や他の地名等の特定の情報の有無に依存しない所属エリア推定のアプローチを考える必要がある．

本研究は，従来研究のあまり進んでいなかった (2) の所属エリア推定を目的とし，その実現手法を提案するものである．本研究では， (3) の処理の前処理として，地名に対してその所属するエリアを十分に絞り込む手法を提案する．すべての地名についてエリアを一意に絞り込むことは必ずしも妥当とはいえないため，所属エリア候補が複数出力されることはむしろ自然であり，このようにして得られた複数の所属エリア候補からの絞込みは， (3) に当たる従来研究で十分に対応可能である．

3.2.4 ヒューリスティクスに依らない地名の属性推定

2.3.2 節に挙げた関連研究のほとんどは，文書中に出現する地名を対象としており，文脈の情報を用いて地名の所属エリアの判別を行っている (図 3.3)．これらは，その

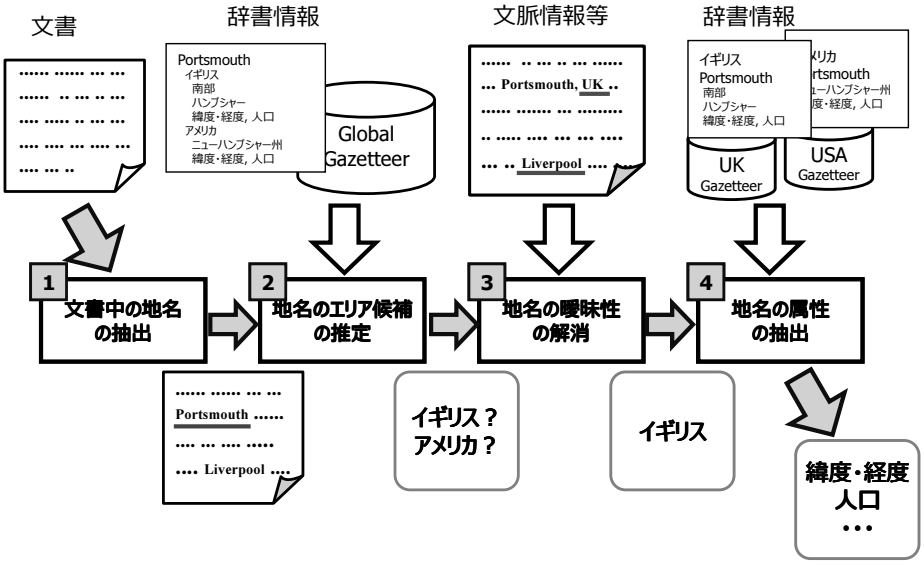


図 3.3: 所属エリア推定のステップ (従来手法)

地名がどんな文脈で出現し，同時に出現するその他の地名とどんな関連があるのかといった情報を積極的に利用する方法である．これらの手法の大きな特徴として，地名の認識および場所の判別に地名辞書を利用していることが挙げられる．地名を表記するときによく用いられるパターンや会話における局所性等の自然言語処理でよく用いられるヒューリスティクスだけではなく，都市の人口数，実際の 2 地点間の距離等の地理的な情報を活用しているものもある．このような辞書ベースのアプローチは，特

定のドメインを対象とした処理の場合には高い精度で場所の判別を行うことが可能である．このように一般的な自然言語処理のヒューリスティクスが適用可能な情報源を用い，かつ，そのドメインに出現し得る地名が変化するスピードが遅く，辞書や地理的なデータの整備を十分に行える場合には，これらの手法は十分に適用可能である．

しかし，全世界のすべての地名を網羅した地名辞書を整備することは現実的でない上に，情報源の多様化のスピードがますます加速している現在では，より頑健性の高い柔軟な手法が必要と考えられる．Rauch らは，知らない地名であっても人間はその所属エリアをある程度推測可能であるという事実を背景として，表層的な統計情報をベースとしたシステムが有効であると主張している [58]．本研究は，Rauch らと同じ主張を共有し，具体的な実現手法を提案するものである．図 3.3 に示すように，従来の所属エリア推定では，推定の前提として直接的にせよ暗黙的にせよ既に所属エリア候補が絞り込まれており，候補となるエリアについて十分な情報が得られる状況を仮定していた．それに対して，本研究で提案する手法は，全エリアを対象とした所属エリア候補の絞込みの実現を目指すものである．そのため，図 3.4 に示すように，所属エリア推定のフェーズでは地名の持つ表層的な情報のみを利用する．これにより，理論的には全世界レベルでの地名の所属エリア候補の絞込みを可能とする．

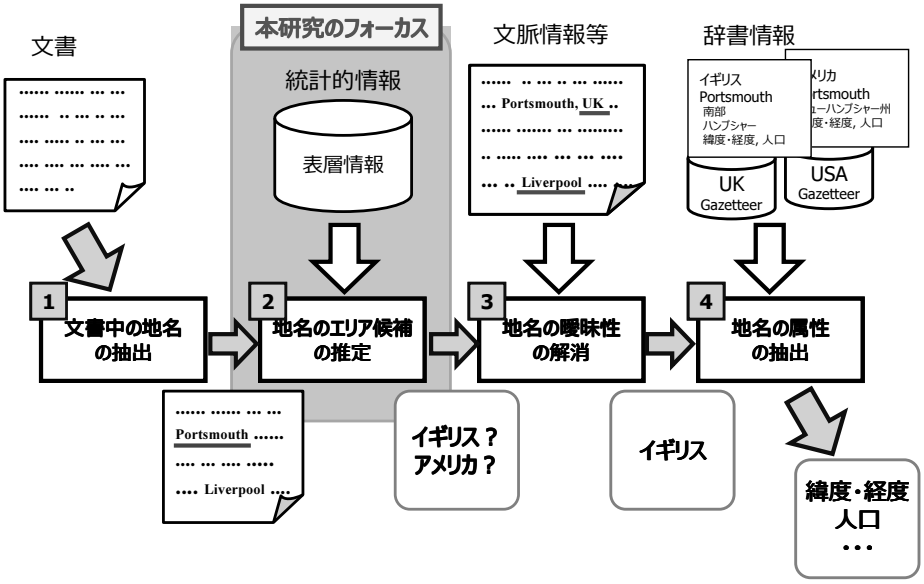


図 3.4: 所属エリア推定のステップ (提案手法)

3.3 本研究で対象とするタスク

人間は，知らない地名であっても，文脈の情報無しにその地名の所属エリアを推定できる場合がある．本研究は，文書中に混在している地名に対して，このような人間の直感的な推定を計算機上で再現するアプローチにより，文脈情報が十分でない場合でも地名語句そのものの持つ表層情報のみからその地名の所属エリアの推定を可能とするものである．

3.3.1　本研究の所属エリア推定システム

人間による直感的な地名の所属エリアの推定が地名を構成する文字の並びを基にしているとすれば，地名に関する統計的な特徴を用いて所属エリアの推定をすることが可能である．本研究では，人間の直感的な推定の根拠に相当するものとして，さまざまな知識ベースから抽出した各エリアに属する地名の例を利用する．具体的には，地名の所属エリア推定処理に用いる情報として，既存の知識ベースから各エリアに所属する地名を複数取得し，エリア毎に地名の持つ表層的特徴を抽出する．これを用いて，入力地名の表層的特徴と各エリアに含まれる地名から抽出した表層的特徴とを比較し，入力された地名の所属エリアの推定を行う．

本研究において推定処理に用いる情報は，既存のさまざまな地名関連の知識ベースから自動的に取得することを想定しており，知識ベースの記述形式等に依らず簡単に取得可能なデータのみとする．図 3.5 に本研究の所属エリア推定システムの概要を示す．本研究では，与えられた地名に対して 1 個あるいは複数の所属エリア候補を出力

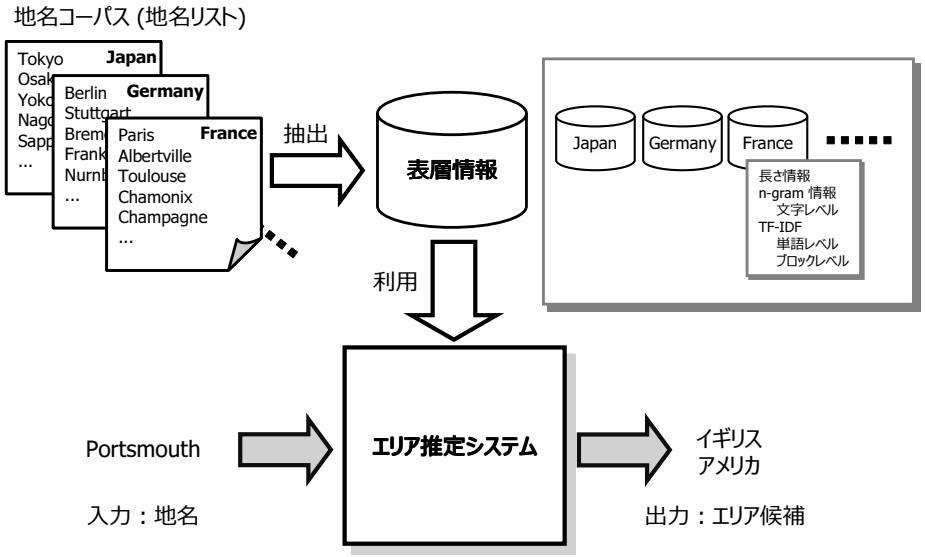


図 3.5: 本研究での所属エリア推定システムの概要

する．これは，同じ地名が複数のエリアに属する可能性に対応するためであり，誤りの可能性の高いエリアは候補に入れず，所属エリアの曖昧性に起因する場合にのみ複数の出力を許すことを目的としている．具体的には，十分な絞込みとして，可能性のあるエリアを 3 個以下に抑えることを目標とする．

3.3.2　本研究で利用する地名コーパス

表層的な情報のみを用いて所属エリアを推定する場合に必要な言語リソースは地名コーパスのみである．本研究の手法は，地名コーパスを辞書引きの対象とするのではなく，地名コーパスからエリア毎の特徴を抽出して処理を行うことで，地名コーパス

に含まれていない地名に対しても所属エリアの推定を行うことを可能にするものである．これは，未知語処理の頑健性，柔軟性の向上に貢献するものである．

本研究のアプローチは地名コーパス以外の情報を一切必要としないことから，処理に必要なデータは既存の地名関連の知識ベースから簡単に作成，拡充が可能である．このことから，本研究のアプローチは，知識ベースの内容に依存する既存の手法と異なり，リソース準備にかかる人的労力や作業期間を大幅に軽減することができる．本研究では，処理対象とするエリアについて，十分な個数の地名を有する地名コーパスを用意する (図 3.5)．これらの地名コーパスは，単純な地名の羅列であり，地名リストと呼ぶレベルの内容である．これらの地名コーパスには，位置情報等，地名に関連するデータは一切記載されていない．

本研究の対象字種

本研究では，対象をラテン文字で表記された地名とする．外国語の文書中に埋め込まれた地名の所属エリア推定を想定しているため，字種情報が所属エリアの推定に影響しないよう，エリアに依らずラテン文字による記述とする．本研究ではラテン文字を処理対象として実験を行うが，本研究のアプローチは表層的な情報のみを用いているため字種に依存せず適用可能であり，その点で高い汎用性を持っている．

ウムラウト等のアクセント記号の添付の有無や大文字小文字の使い分け等は文書の記述言語や記述形式に左右される可能性があるため，アクセント記号等は利用せず，大文字と小文字の区別も行わない．また，記述形式の違いによる統計情報の揺れを防ぐため，地名コーパス内で使用する記号は原則としてすべて空白文字に置き換える[2]．ここで空白文字に置き換えるのはハイフンやカンマ，ピリオド等，語と語の間に置かれる記号で，置換により複数個空白文字が続いて現れる場合には空白文字 1 個とする．したがって，処理の上では用いる記号の種類は計 27 種類となる．例えば，Land Baden-Württemberg という地名は LAND␣BADEN␣WURTTEMBERG として扱い，3 単語 22 文字から成るものとする (空白も 1 文字と数える)．ただし，T'aipei 等，1 語中にアポストロフィ等の記号が挿入されている場合には，これを削除し TAIPEI として扱う．

本研究では，このような地名に対して統計的な言語モデルを適用することで，特徴を定量的に扱う．なお，この記号の制限は，元データに対して機械的に置換処理を行うことで実現可能である．この方法によって地名関連のさまざまな既存の知識ベースから処理に必要なデータを自動的に収集することが可能であり，これによって本研究のアプローチの汎用性，頑健性が損なわれるわけではなく，むしろさまざまな知識ベースの持つデータを効率的に利用する上で有効であるともいえる．

本研究で用いたデータ

本研究の評価実験のための地名コーパスの作成に当たっては，アメリカ国家地球空間情報局[3]が提供する地理空間情報のデータベース GEOnet Names Server (GNS) のデー

[2]空白記号は，説明のためにこれを明示する際には記号 ␣ で記述する．

[3]National Geospatial-Intelligence Agency: `http://www.nga.mil`.

タを使用している．ただし，GNS にはアメリカの地名が含まれていないため，アメリカの地名については，アメリカ地名委員会[4]が提供する Geographic Names Information System (GNIS) のデータを使用している．GNS は全体としては 550 万件程度の地名データ，GNIS は 200 万件程度の地名データを有している．

どちらのデータも英語をベースとして作成されているため，英語圏以外のエリアの地名であっても Elbe River のように地名中に英語語句が含まれていることがあり， mountain や river のような地名に関連する英語語句がどのエリアのデータでも散見される．このような英語語句の利用は徹底されたものではなく，例えば**富士山**は Fuji Mountain と記載されているのに対して**浅間山**は Asama yama と記載される等，現地の語句のラテン文字表記と英語を用いた表記とが混在している．本研究では，地名コーパス作成時の制約を極力排除し本手法の汎用性，頑健性を保つため，このような利用データ固有の特徴もそのまま吸収できるような手法の提案を目指す．そのため，英語語句の混入についても，排除等の処理はせず，そのまま利用する．

元データからの地名コーパスの作成

処理対象とする各エリアについて，元のデータ中での地名の個数は異なる．地名の個数の違いによる条件の不均衡を避けるため，本研究では，各エリアから同数の地名をランダムに抽出し地名コーパスを作成する．この際，同じ地名コーパス中に同一地名文字列が複数回記述されることはない．ただし，同一の地名が複数のエリアで出現し得るという実情を反映するため，エリアの地名コーパス間での同一地名の有無については考慮しない．また地名コーパスを作成する際，地名の粒度 (州名と町名等行政区域レベルの区別や，町か山かの区別，町の大きさ等) や地名の重要度 (首都等の有名な地名か，ごく小さな地名か) 等は考慮しない．

実験対象エリア

本研究では，全世界的な所属エリア推定を目標としているため，効率的に調査，実験を行うことができるように実験対象エリアを絞る．ウェブ上で多く使用される言語と，その比較対象としてそれと類似性の高い言語を対象とし，地名の所属エリア推定を試みる．Internet World Stats による 2009 年 9 月 30 日現在のウェブユーザ数の言語別推定の結果[5]を図 3.6 に示す．この結果から，本研究では，全体として，表 3.1 に示す 20 エリアを実験対象とする．地名の所属エリア推定に関する複数の課題に対して順に解決を図るため，各実験ではそれぞれの目的に応じて実験エリアを選ぶ形をとる．その際，実験毎の結果の比較を可能にするため，条件に関わらず 8 エリアは固定とする．表 3.1 のエリア名の左にある丸印は，固定 8 エリアを示す．ウェブ上の使用言語で最も多いものは英語であり，その意味で英語圏のエリアは固定エリアとすべきである．しかし本研究では，以下の理由から，英語圏のエリアを固定エリアに含めない．第一に，本研究で実験に用いた地名コーパス群は，英語ベースの知識データを基に作成

[4]United States Board on Geographic Names: `http://geonames.usgs.gov/`.
[5]Internet World Stats: `http://www.internetworldstats.com/stats7.htm`.

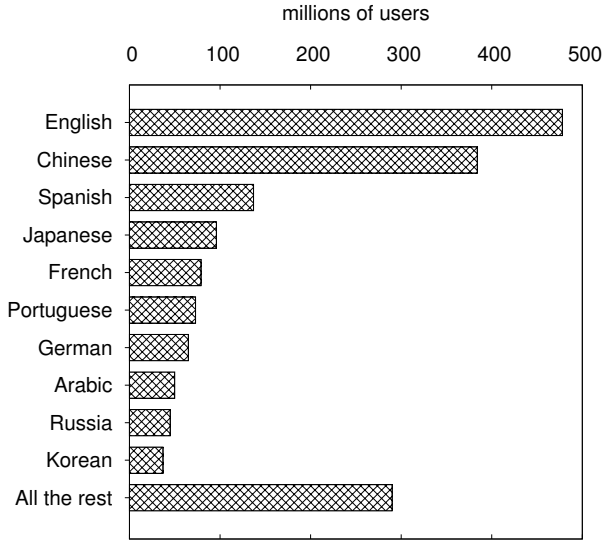


図 3.6: ウェブユーザ数の言語別推定

しているため，英語の影響を受けている．そのため，英語圏のエリアを実験対象に含めることで，コーパスの影響とエリアの影響とを区別することが難しくなる．第二に，英語圏のエリアとして代表的なアメリカは移民国家であり，主な使用言語は英語だが地名にはネイティブアメリカンの言語も含めさまざまな言語が反映されているとの特徴がある．このことから，固定エリアとしての利用よりも，比較対象エリアとして利用することが適切と判断する．なお，実験結果等を示す図や表では，表 3.1 に示す略記でエリアを表すことがある．

表 3.1: 実験対象エリア

	エリア	略記	主な使用言語		エリア	略記	主な使用言語
○	中国	CH	中国語	○	フィンランド	FI	フィンランド語
	台湾	TW	中国語	○	フランス	FR	フランス語
	韓国	KR	韓国語		イタリア	IT	イタリア語
○	タイ	TH	タイ語	○	スペイン	ES	スペイン語
	マレーシア	MY	マレー語		チリ	CL	スペイン語
○	日本	JP	日本語		ポルトガル	PT	ポルトガル語
	インド	IN	ヒンディー語	○	ドイツ	DE	ドイツ語
	エジプト	EG	アラビア語		オランダ	NL	オランダ語
○	ギリシャ	GR	ギリシャ語		イギリス	GB	英語
	ロシア	RS	ロシア語		アメリカ	US	英語

3.3.3 本研究の評価基準

実験コーパスの評価形式

本研究における所属エリア推定の実験では，5 分割クロスバリデーションを採用している．例えば，各エリアの地名コーパスに含まれる地名の数が 10,000 個の場合，各エリアの地名コーパスはそれぞれ 2,000 個ずつの地名リストに分割して利用することになる．その上で，ある 2,000 個 の地名をテスト用コーパス，その残りをトレーニングコーパスとして実験を行う手順を，テスト用コーパスとする 2,000 個の地名を変えながら 5 回繰り返す．この方法で各エリアの地名コーパスに含まれる 10,000 個のすべての地名に対して所属エリア推定を実施し，その結果を集計することで，オープンテストの実験を行う．

実験の出力結果の評価形式

地名の所属エリア推定の実験の出力結果は，正解エリアで正しく正解 (positive) と判定することができたかと，不正解エリアで正しく不正解 (negative) と判定することができたかの双方から評価する必要がある．そこで本研究では，各手法における実験の出力結果の評価に際して，出力結果に含まれる所属エリア候補それぞれについて上記の 2 種類の立場での評価を行い，それらを組み合わせて処理精度の評価を行う．例えば，Tokyo という日本コーパスに含まれる地名に対する評価は，出力結果に日本が含まれていれば正解を正しく推定したものと考えられ，また出力結果にフランス等の他のエリアが含まれていなければ不正解を正しく推定したものと考えられる．

tp: 正解エリアを対象とした判定で正しく positive と判断された地名の数 (true positive)
fn: 正解エリアを対象とした判定で誤って negative と判断された地名の数 (false negative)
fp: 不正解エリアを対象とした判定で誤って positive と判断された地名の数 (false positive)
tn: 不正解エリアを対象とした判定で正しく negative と判断された地名の数 (true negative)

図 3.7: 推定対象エリアと入力地名の推定結果の関係

ここで，推定対象エリアと入力地名との関係から，実験結果を図 3.7 の 4 種類に分類する．図 3.7 の 4 種類の状況は，それぞれ式 (3.1) から式 (3.4) で表される．ここで A はエリア全体の集合，a は各実験の対象エリア，t は入力地名，C はテスト用コーパス (地名の集合)，C_a はエリア a のコーパスを示し，$\mathcal{IR}_a(t)$ は地名 t に対してエリア a を対象エリアとして推定した結果の値で，所属エリア候補としての可能性があると

判断されれば p，そうでなければ n とする．この分類にしたがえば，tp および tn の割合を向上させることが，実験の目的となる．

$$tp = \left|\left\{t \in C_a : \mathcal{TR}_a(t) = \mathrm{p}\right\}\right| \tag{3.1}$$

$$fn = \left|\left\{t \in C_a : \mathcal{TR}_a(t) = \mathrm{n}\right\}\right| \tag{3.2}$$

$$fp = \sum_{a' \in A-a} \left|\left\{t \in C_a : \mathcal{TR}_{a'}(t) = \mathrm{p}\right\}\right| \tag{3.3}$$

$$tn = \sum_{a' \in A-a} \left|\left\{t \in C_a : \mathcal{TR}_{a'}(t) = \mathrm{n}\right\}\right| \tag{3.4}$$

本研究での実験結果の評価には，適合率 P，再現率 R，正解率 A，F 値 F を用いる．適合率は，出力要素中の正解要素の割合であり (式 (3.5))，正解と誤って出力される要素の数が増えるほど適合率は低下する．そのため，適合率を上げるためには，出力要素を十分に絞込み妥当な要素のみ出力するように調整する必要がある．

$$P = \frac{tp}{tp + fp} \tag{3.5}$$

再現率は，全正解要素中いくつの正解要素を出力できたかを示す割合であり (式 (3.6))，出力の絞込みが強すぎると取りこぼす正解要素が増えるため再現率が低下する．そのため，再現率を上げるためには，正解要素を除去してしまうことのないよう，出力判定に十分な余裕を設ける必要がある．

$$R = \frac{tp}{tp + fn} \tag{3.6}$$

F 値は，トレードオフの関係となる適合率と再現率の双方を考慮した精度を表す値で [13, 36]，式 (3.7) で定義される．

$$F = \frac{2PR}{P + R} \tag{3.7}$$

正解率とは，正解要素を正解要素と出力する割合と，不正解要素を不正解要素と出力する割合を合わせたものであり，全要素に対して正しく出力するかしないかの判定を行ったか否かを示す (式 (3.8))．

$$A = \frac{tp + tn}{tp + fp + fn + tn} \tag{3.8}$$

本研究では，対象エリア毎に評価を行うと同時に，全体をまとめた状態での評価も行う．式 (3.5) から式 (3.8) までの定義を本研究の実験内容に当てはめた場合の意味付

けを式 (3.9) から式 (3.14) に示す．各対象エリアについての適合率は式 (3.9)，再現率は式 (3.10)，正解率は式 (3.11) との意味を持つ．

$$P = \frac{\text{対象エリアの地名のうち正しく positive と判断された地名の数}}{\text{対象エリアの地名に対する所属エリア推定の結果 positive と判断された数}} \tag{3.9}$$

$$R = \frac{\text{対象エリアの地名のうち正しく positive と判断された地名の数}}{\text{対象エリアのすべての地名の数}} \tag{3.10}$$

$$A = \frac{\begin{array}{c}\text{対象エリアの地名のうち正しく positive と判断された地名の数}\\ +\ \text{対象エリアの地名のうち正しく negative と判断された地名の数}\end{array}}{\text{対象エリアのすべての地名の数} \times \text{全エリア数}} \tag{3.11}$$

また，実験対象エリア全体に対しての適合率は式 (3.12)，再現率は式 (3.13)，正解率は式 (3.14) に示したような意味を持つ．実験結果の中で全体の値は，各エリアの平均値ではなく，ここで示した式 (3.12)，式 (3.13)，式 (3.14) の値である．

$$P = \frac{\text{全エリアの地名のうち正しく positive と判断された地名の総数}}{\text{全エリアの地名に対する所属エリア推定の結果 positive と判断された総数}} \tag{3.12}$$

$$R = \frac{\text{各エリアの地名のうち正しく positive と判断された地名の総数}}{\text{全エリアのすべての地名の総数}} \tag{3.13}$$

$$A = \frac{\begin{array}{c}\text{全エリアの地名のうち正しく positive と判断された地名の総数}\\ +\ \text{全エリアの地名のうち正しく negative と判断された地名の総数}\end{array}}{\text{全エリアのすべての地名の総数} \times \text{全エリア数}} \tag{3.14}$$

地名の所属エリア推定では， 1 個の地名が複数のエリアに所属する可能性があるため，所属エリア候補を 1 個に限定すると正しい所属エリアを候補から除去してしまうリスクがある．したがって，地名の所属エリア推定では，まず第一に再現率を十分に確保し，正解エリアが除去されるリスクを抑える必要がある．その上で，誤った所属エリア候補の出力を抑えるよう精度の向上を目指し，結果的に適合率を向上させることが目標となる．

機械学習器としての二値分類学習器の利用

本研究では，この評価方法に合う機械学習器として，二値分類学習器を用いる．二値分類学習器を用いた所属エリアの推定処理では，対象エリアそれぞれに対して 1 個の二値分類学習器を作成し，与えられた入力がその対象エリアの地名である可能性について判定を行う．したがって，学習の段階では，対象エリアの地名を正事例，それ以外のエリアの地名を負事例として学習を行う．例えば，日本を対象エリアとする二値分類学習器では， Tokyo 等の日本コーパスに属する地名は正事例， Paris 等日本以外のコーパスに属する地名は負事例として学習する．またフランスを対象エリアとする二値分類学習器では，フランスコーパスに属する Paris は正事例とし，フランス以外

のコーパスに属する Tokyo 等は負事例となる．同じ地名が対象エリアと対象以外のエリアに同時に出現していた場合には，この地名は正事例および負事例の双方として学習されることになる．

エリア毎に作成された二値分類学習器は，与えられた地名に対して，そのエリアに属する地名か否かの判定を行い，positive (そのエリアに属する可能性があるとの判定) または negative (そのエリアに属する可能性があるとはいえないとの判定) の判定を出力する．そのため，1 個の地名に対して複数のエリアで positive の判定が出た場合，それらのエリアがすべてその地名の所属エリア候補として出力されることになる．

ここで，図 3.8 に実験結果の例を挙げる．図 3.8 において，p は positive を，n は

テスト用コーパス		エリア			
		日本	ドイツ	···	ギリシャ
日本	Tokyo	p	n	···	n
	Osaka	p	n	···	n
	Nagoya	n	n	···	p
	⋮	⋮	⋮	⋮	⋮
ドイツ	Berlin	n	p	···	n
	Hamburg	n	p	···	p
	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮
ギリシャ	⋮	⋮	⋮	⋮	⋮
	Athens	p	n	···	p

図 3.8: 二値分類学習器を用いた所属エリア推定実験の評価の例

negative を示す．例えば，図 3.8 の 1 行目は，日本コーパスに属する地名である Tokyo に対して各エリアの二値分類学習器を用いて推定実験を行った結果として，日本の二値分類学習器では p (positive)，ドイツの二値分類学習器では n (negative)，というように各エリアの二値分類学習器毎がそのエリアに属する可能性の有無を推定し，結果的に所属エリア候補となり得るエリアが日本であると出力される状況を示している．図 3.8 中，白抜きの部分は正解エリアを，グレーの網掛け部分は不正解エリアを示しており，白抜きの部分 (正解エリア) で p が出力されるか，グレー部分 (不正解エリア) で n が出力された場合に，エリアの推定に成功したことを示している．Tokyo の例では，正解エリアは日本であり，日本でのみ p との推定が出ているため，所属エリアを正しく推定したことになると同時に，不正解エリアが誤って所属エリア候補として出力されることも防いだとの結果となる．

3.3.4 本研究の提案手法

本研究では，地名の所属エリア推定に関連する複数の問題を順に解決しながら精度を向上させる形で説明を行う．それぞれの問題およびその解法を明確にするため，それぞれについて手法を提案し，その結果の比較を行う (図 3.9).

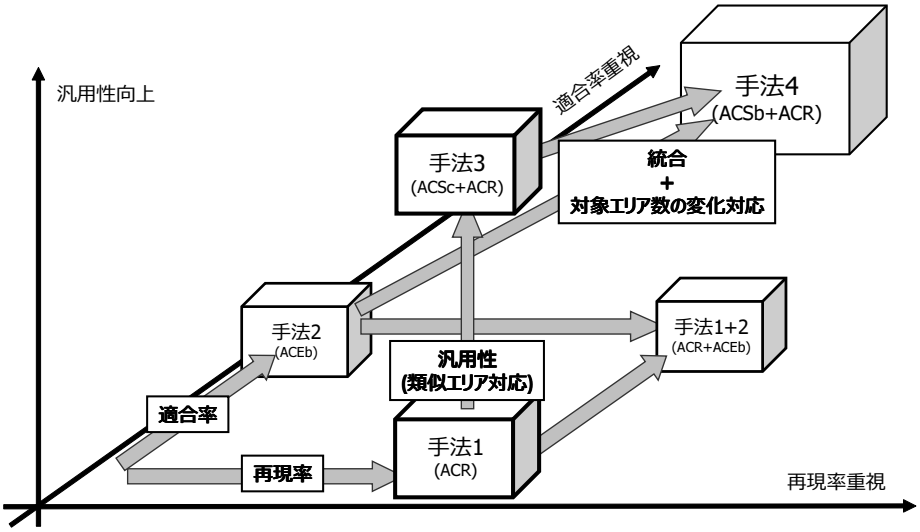


図 3.9: 本論文で提案する 4 手法の関係

第一に，地名の持つ表層情報の特徴を調査し，表層情報のみに基づく地名の所属エリア推定の可能性を確認する (第 4 章)．地名の所属エリア推定では，極論を言えば，すべてのエリアを候補とすれば再現率は 100% となる．所属エリア推定の基本的な方針は，正解エリアを推定し抽出する方向と，不正解エリアを除去することで候補を絞り込む方法の 2 種類ある．第 4 章では地名辞書等の知識情報を必要としない頑健な地名の所属エリア推定の基本的な手法となるよう，再現率を極力低下させずに地名の所属エリア推定を行う手法として，不正解と推定されるエリアの除去による所属エリア候補の絞込み手法を検討する (手法*1* [63]: 図 3.10).

第二に，地名の所属エリア推定の精度の向上を目指し，地名の処理に適した新しい処理単位の提案を行う (第 5 章)．文字列が比較的短いこと，固有表現のためデータスパースネスの問題が避けられないこと等の地名特有の問題のため，従来の処理単位では地名から十分な表層的統計情報を得ることが難しい．そこで，地名の処理に適した新しい処理単位としてブロックの概念を提案し，これを導入した手法を用いてその有効性を検証する (手法*2*: 図 3.11)．さらに，提案する新しい処理単位と既存の処理単位を適切に組み合わせることで，再現率を高く確保したまま適合率を向上できることを示す (手法*1+2* [62]: 図 3.12).

第三に，地名特有の問題として，言語的，地理的，歴史的背景等から類似した地名を多く有するエリアが複数存在することに着目し，この問題の解決を図る (第 6 章)．地名の所属エリア候補を 1 個に絞ることは必ずしも可能ではなく，また過剰な絞込みを行うことで正解エリアを候補から外してしまうリスクがある．しかし，例えばスペイ

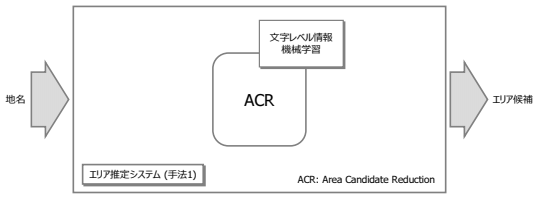


図 3.10: 手法*1*の概要

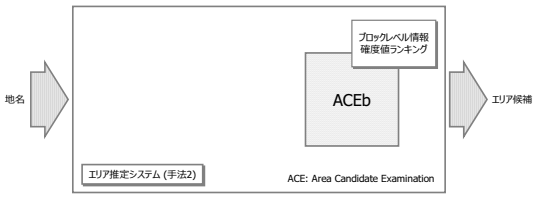


図 3.11: 手法*2*の概要

ンとチリのように同一言語圏であったり，歴史的になんらかの繋がりがある等の理由で相互に同一あるいは類似した地名を有するエリアのグループが存在することは明らかであり，こういったグループのエリアをすべて所属エリア候補として出力することは望ましくない．そこで，可能な限り適切に所属エリア候補の絞込みを行うため，このような複数の候補を取り得る状況の調査を行い，類似したエリア間の識別を図る．類似したエリア間の識別を実現するために，大域的な所属エリア推定と局所的な所属エリア推定の 2 段階の処理を行い，関連性の高いエリアが含まれるような場合においても高い精度で所属エリア推定が可能であることを示す (手法*3* [61]: 図 3.13).

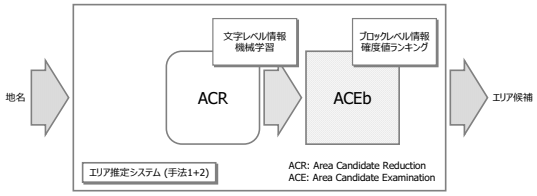


図 3.12: 手法*1+2*の概要

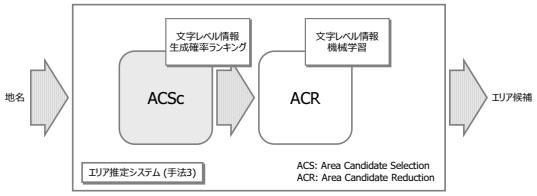


図 3.13: 手法*3*の概要

最後に，これらの成果を統合し，頑健性と汎用性を備えた高精度の所属エリア推定手法を確立する (第 7 章)．ここでは，類似エリアに対応するための 2 段階推定処理と，再現率および適合率向上のための複数の処理単位を組み合わせた推定との双方の特徴を考慮し，これらをお互いを損なわない形で効果的に組み合わせ，精度と汎用性の両面で十分な結果を得られることを示す (手法*4*: 図 3.14)．この手法が本研究の最終的な

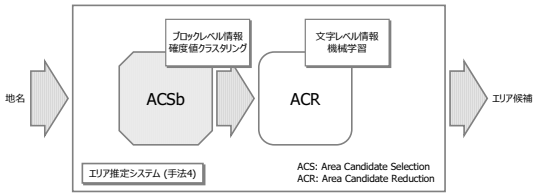


図 3.14: 手法*4*の概要

提案手法であり，これについて，実験結果を基に，精度，複数候補の問題等について検証し，本手法の有効性，頑健性，汎用性について考察する．

第4章　地名の所属エリア推定における表層情報の利用 (手法*1*)

本章では，頑健かつ汎用性の高い地名の所属エリア推定を実現するため，地名の表層情報として文字レベルの情報に着目し，その効果的な抽出方法および活用方法について調査した結果を報告する．その上で，取得した文字レベルの表層情報を利用した地名の所属エリア推定の実験結果を報告し，高い再現率で所属エリア推定が可能であることを示す.

4.1　地名の表層情報

地名は，一般的に数単語で構成され，数文字から十数文字から成る．Dunning は文書を対象とした実験結果からたかだか 20 文字あれば言語推定が可能と述べている [14] が，Dunning の実験の対象は一般の語句の並びであり，言語識別の指標となる言語特有の頻出語や頻出文字列の存在や，複数の語句の並びを含むことから得られる構文的な情報等を考慮できる点で地名文字列とは性質が異なる．同じ文字数であっても，地名は固有表現であるため個々の語句の出現頻度が低いこと，構文的な情報を含んでいないこと等から，地名文字列の持つ情報はほぼ音韻的な表層的情報のみに限られることが予測できる．このことから，地名文字列からの所属エリア推定は，一般語句の文字列を対象とした言語推定に比べて難易度の高いものと考えられる.

本節では，地名文字列の持つ表層的情報を調査し，所属エリア識別の可能性を検討する．文字レベルの表層情報を考えた場合，この地名コーパスから簡単に取得可能な情報は，地名を構成する文字の個数に関連する長さ情報と，連続する共起文字の頻度を示す n-gram 情報の 2 種類である．本節では，4.1.2 節で地名の長さ情報について，4.1.3 節で地名の n-gram 情報について説明する.

4.1.1　地名コーパス

地名の表層情報の調査の対象エリアは中国，台湾，タイ，日本，ギリシャ，フィンランド，ドイツ，フランス，スペインの 9 エリアとする．ここでは，比較対象として中国の類似エリアとして台湾を加えた．調査に用いる地名コーパスのサイズを表 4.1 に示す．地名の平均長 (文字数) は，最小がフィンランドのコーパスで 9.73，最大がスペインのコーパスで 13.36 であり，全体の平均は 11.42 文字である．本章では，ここから得られる表層情報から，地名の所属エリア推定を行うために必要な情報を推定する.

表 4.1: 本章で用いる地名コーパス

エリア	地名数	単語数	文字数
中国	10,000	15,943	98,609
台湾	10,000	22,757	123,056
タイ	10,000	29,021	131,159
日本	10,000	16,892	102,193
ギリシャ	10,000	14,631	109,532
フィンランド	10,000	11,053	97,317
フランス	10,000	19,776	121,947
スペイン	10,000	22,372	133,575
ドイツ	10,000	13,269	109,987

4.1.2 地名の長さ情報

地名を構成する文字の数に関連する情報は，その地名が所属するエリアあるいはその言語によってその特徴に差異が見られる．地名の所属エリア推定を行うためには，この特徴の差異を手がかりとすることが考えられる．図 4.1 に，各エリアにおける地名の構成単語の長さの分布を示す．図 4.1 の x 軸は地名の構成単語 1 個あたりの文字数

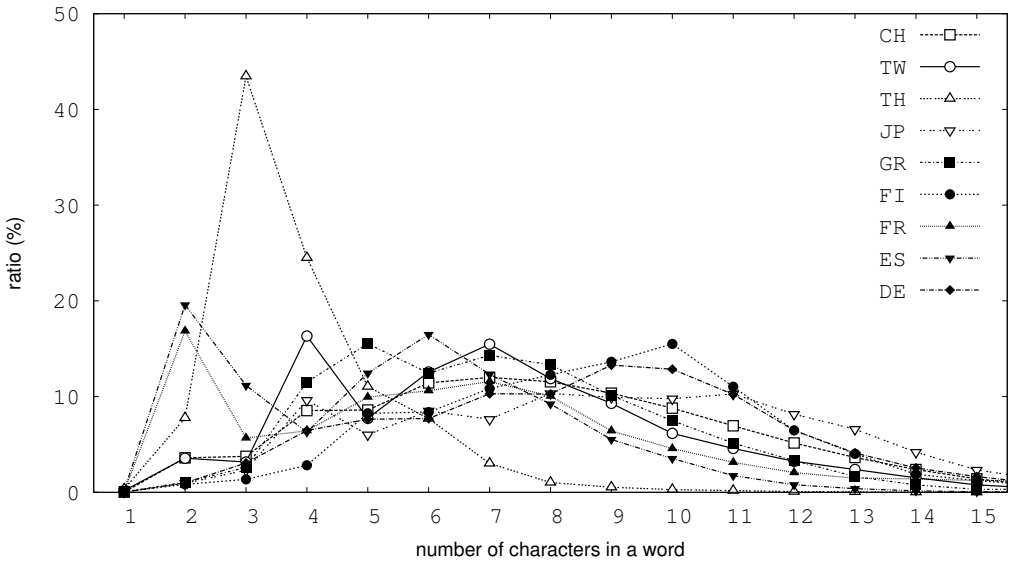


図 4.1: 地名を構成する単語の長さの分布

(number of characters in a word)，y 軸は各文字数の構成単語のエリア毎の割合 (ratio) を示している．構成単語とは，地名の先頭と末尾および空白文字で区切られた文字の並びと考えることができ，単語の長さとは，その文字の並びに含まれる文字の数を表している．

図 4.1 を見ると，地名を構成する語の長さはだいたい 2 文字から 14 文字程度の範囲で丘状に分布していることがわかる．その中で，特に高いピーク値を持つエリアと，全体になだらかに分布するエリアとに分けられることが見て取れる．最も特徴的なのはタイの地名の分布であり，3 文字または 4 文字の単語が多く，この 2 種類だけで全体の 68% を占める．またフランスやスペインの地名では，他のエリアの地名と比較して，単語の長さの分布のピークが左に偏っている．これは，この 2 エリアで 2 文字から 3 文字の単語が頻出するためである．フィンランドやドイツの地名では，他のエリアの地名と比較して，単語の長さの分布のピークが右に偏っている．

全体としては，1 語あたりの文字数のピークが 2 文字から 4 文字程度のところにあるエリア群，5 文字から 7 文字程度のところにあるエリア群，8 文字から 10 文字程度のところにあるエリア群に大別できる．このように，単語を構成する文字数のピークはエリアによって傾向の違いがあり，地名を構成する単語の長さは，所属エリア推定の要素になり得る．

地名を構成する単語数と文字数の関係を図 4.2 に示す．図 4.2 の x 軸は地名あたりの構成単語の平均個数 (number of words per toponym)，y 軸は地名あたりの構成文字の平均個数 (number of characters per toponym) を示している．これらのエリアは，地

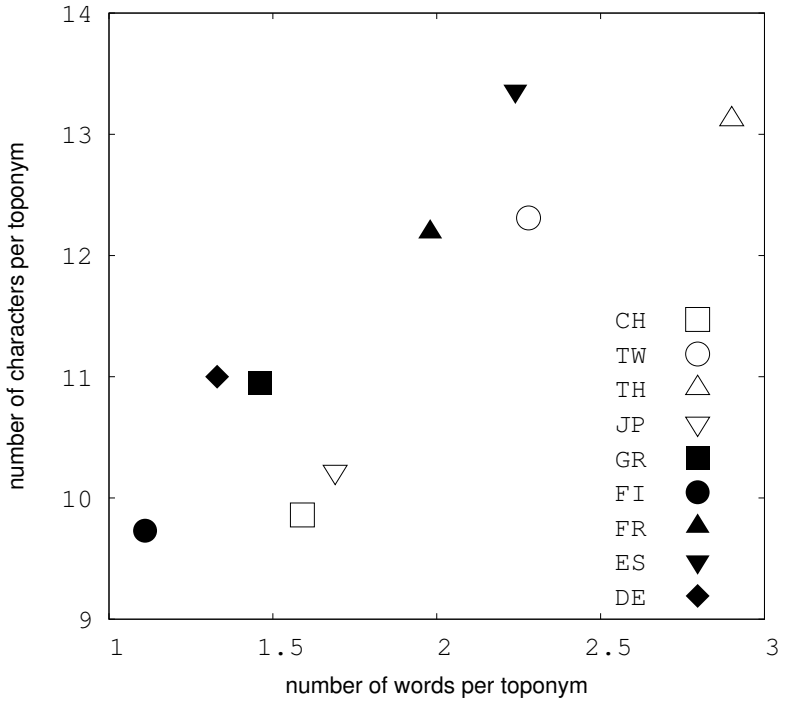


図 4.2: 地名を構成する単語の数と文字の数

名 1 個あたりの平均語数が 2 個未満のエリアと 2 個を超えるエリアとに大別できる．当然，平均語数の比較的多いエリアは地名 1 個あたりの平均文字数も増える傾向にあるが，平均語数 2 個未満のエリアで 10 文字から 11 文字程度，平均語数 2 個以上のエリアで 12 文字から 13 文字程度と，その差はさほど大きくない．表 4.1 で得られた平均 11.42 文字との地名長と比較しても，地名長の長いエリアと地名長の短いエリア

の差は十分大きいとはいえない．またこの値は，Dunning が言語推定に必要とした 20 文字 [14] の約半分であり，また図 4.1 から地名の構成文字数がたかだか 15 文字程度であることを考えると，地名の所属エリア推定の難しさがうかがえる．

ここで，図 4.1 および図 4.2 のデータを数値で表したものを表 4.2 に示す．図 4.1 や図 4.2 のグラフから，エリア間の長さのデータの差は小さいことがわかるが，表 4.2 のデータから，語数や文字数の傾向がエリア毎に異なることが読み取れる．フィンランドやドイツの地名は，単語数が少なく文字数が多いことから，単語あたりの文字数が多い．フィンランドでは 1 個の単語が平均 8.8 文字，ドイツでは 1 個の単語が平均 8.3 文字で構成されている．それに対して，タイの地名は単語数が多く単語あたりの文字数が少ないため，単語の平均長は 4.5 文字と，フィンランドやドイツの半分程度である．これらのデータから，地名の持つ語や文字の長さの情報が所属エリア推定に利用できる可能性が示唆される．

表 4.2: 地名を構成する単語の数と文字の数

エリア	地名あたりの平均単語数	地名あたりの平均文字数	単語あたりの平均文字数
中国	1.59	9.86	6.19
台湾	2.28	12.31	5.41
タイ	2.90	13.12	4.52
日本	1.69	10.22	6.05
ギリシャ	1.46	10.95	7.49
フィンランド	1.11	9.73	8.80
フランス	1.98	12.19	6.17
スペイン	2.24	13.36	5.97
ドイツ	1.33	11.00	8.29

4.1.3 地名の n-gram 情報

統計情報を用いた自然言語処理においては，効果的な特徴量として連続する文字や単語の共起頻度を表す n-gram 情報が一般的に用いられる．本研究では，n-gram 情報を利用することで，文字レベルでの共起頻度を特徴として捉え，エリア間での差異を見出すことで地名が所属するエリアを識別する手がかりとする．本節では，n-gram 情報として unigram，bigram，trigram を各エリアの地名コーパスから取得し，その特徴を述べる．

unigram

表 4.3 に頻出する unigram を示す．‘10%’ のカラムは，地名コーパス中で 10% 以上の頻度で出現した unigram の数である．全体の傾向として，どのエリアでも母音の出

表 4.3: 頻出 unigram

エリア	10%	上位 5 個				
中国	3	N	A	I	H	U
台湾	3	I	N	A	H	U
タイ	2	A	N	H	O	K
日本	2	A	I	O	K	S
ギリシャ	2	A	I	O	S	R
フィンランド	1	A	I	R	K	N
フランス	1	E	A	R	L	N
スペイン	2	A	E	R	O	L
ドイツ	1	E	R	N	H	S

現頻度が高いことがわかる．各エリアで 10% 以上の頻度で出現した unigram を見ると，中国，タイ，台湾で N が挙がっているほかは，すべて母音であった．表 4.3 からは，中国と台湾が unigram の出現状況に強い類似性があることがわかる．また，日本とギリシャの間にも類似性が見られる．

図 4.3 に各エリアでの unigram の出現割合 (ratio) を示す．図 4.3 では，空白を含む 27 種の文字をアルファベット順に並べている (空白は Z の後，表の一番下に示す)．5 個の母音は細いストライプの背景で示している．また特に頻度の高かった子音は粗いストライプで示している．ここで，母音に着目してその特徴を検討する．子音よりも

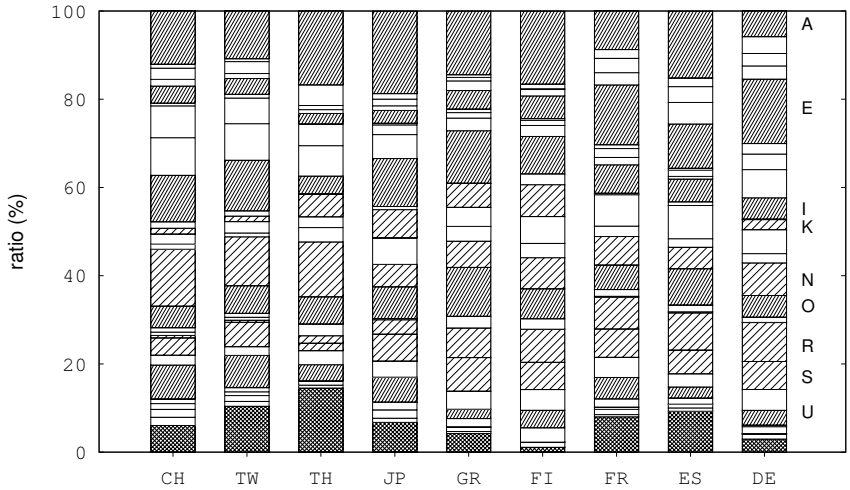


図 4.3: unigram 出現傾向

母音が出現する頻度が高いことは自然であり，実際どのエリアでも母音の出現頻度は他の文字と比べて高い．しかし，それぞれのコーパスにおいて母音の使われ方の傾向には差がある．表 4.3 を見ると，A はどのエリアでも 3 位以内に入る頻出 unigram だが，ドイツでは A の出現割合は 5.83% と他のエリアに比べて低く，5 位に入っていない．この値は他の母音と同じ程度の出現割合 (I: 4.84%，O: 4.92%，U: 3.37%) であり，他のエリアで A が高い出現割合を示す中で特徴的である．フランスはドイツと同様に E の出現割合 (13.59%) が A の出現割合 (8.78%) を上回っている．また，9 エリア中 5 エリアで I が 3 位以内に入っているがこれらの地域では E は 5 位以内に入っておらず，逆に E が 2 位以内に入る 3 エリアでは I は 5 位以内には入らない．

図 4.4 に母音 unigram の出現割合 (ratio) を示す．図 4.4 から，どのエリアでも母

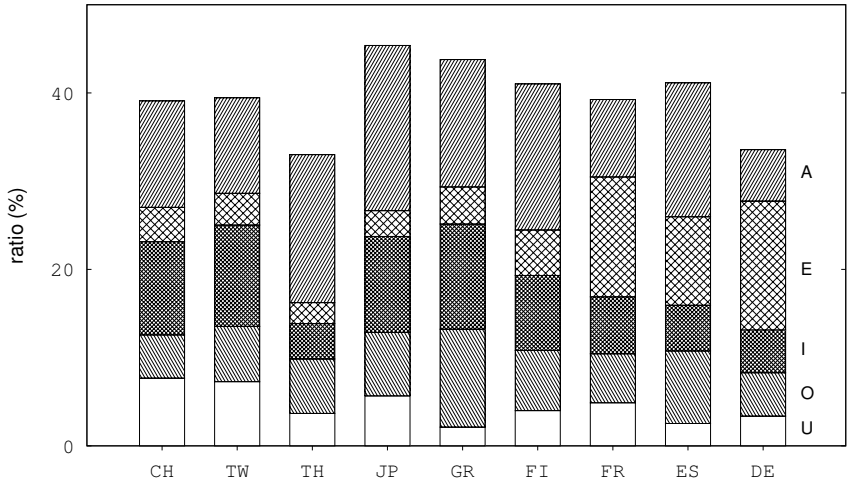


図 4.4: unigram 出現傾向 (母音)

音 unigram が全体の 40% 程度の出現割合を占めることがわかる．その内訳はエリアによって差が見られ，A の割合が特に多いエリア，E の割合が高いエリア等の特徴が現れている．また，タイのように母音が比較的少ないエリアも見られる．E が頻出したエリアはドイツ，フランス，スペインとどれもヨーロッパ地域のエリアであり，言語は異なるが言語的，地理的，歴史的に関連が深いことを考えると，エリア同士の地名の類似度にも大小があることが仮定できる．

bigram

表 4.4 に頻出する bigram を示す．‘3%’ のカラムは，地名コーパス中で 3% 以上の頻度で出現した bigram の数である．また，表中の記号 ␣ は空白文字を表している．bigram では，unigram と比べて組合せ数が大きいため，頻出 bigram の上位 5 個にはエリア毎の相違が現れる．また，高い頻度で現れる bigram の存在の有無がエリア毎に異なる．タイ，中国，台湾では 3% 以上の頻度を持つ bigram が 3 個を超え，bigram の分布に偏りがあることがわかる．それに対して，フランスやギリシャでは 3% の頻度で出現する bigram はなく，中国等に比べると bigram の分布の偏りは小さい．中国

表 4.4: 頻出 bigram

エリア	3%	上位 5 個				
中国	3	AN	NG	IA	SH	CH
台湾	4	AN	NG	IA	AO	SH
タイ	5	AN	NG	N␣	BA	G␣
日本	2	HI	MA	AK	SH	KA
ギリシャ	0	ON	OS	IS	RA	OU
フィンランド	1	AR	LA	KA	AN	EN
フランス	0	E␣	LE	ER	ES	RE
スペイン	2	DE	␣D	E␣	LA	A␣
ドイツ	2	ER	EN	CH	BE	TE

のコーパスでの頻出 bigram は AN と NG であり，それぞれ 7.18% と 7.00% の高い割合で出現している．この特徴は，台湾のコーパスでも同様である．この頻度の割合は他のエリアと比較しても大きな値である．

スペインのコーパスでは，頻出 bigram に A，E，D，L と空白が多く含まれている．これは，スペインのコーパスに Montes de Leon や Jerez de la Frontera のように DE, LA, DEL 等の冠詞および前置詞が多く含まれていることに起因する．同様に，フランスの地名では，E␣ (出現頻度 2.88%) に続いて LE (2.58%) の出現頻度が高いが，これは Le Havre，Le Mans 等地名中での冠詞 LE の頻出のほか，îles (**島**) や ville (**町**) 等の地名関連語句に LE が含まれることによるものである．DE や LE 等の冠詞および前置詞の出現は，図 4.1 でスペインとフランスの地名が短い単語の部分でピークを示していることにも関連している．

ギリシャと日本の間にあった unigram における類似性は，bigram では見られなくなっている．ギリシャでは，ON (2.28%)， OS (2.28%) につづいて IS (2.07%)， RA (2.07%) が頻出している．日本では HI の頻度が高いが，これは HI が**ひ**のラテン文字表記として使われるだけでなく**し** (SHI) と**ち** (CHI) のラテン文字表記にも含まれるためと考えられる．

trigram

表 4.5 に頻出する trigram を示す．‘1%’ のカラムは，地名コーパス中で 1% 以上の頻度で出現した trigram の数である．表 4.5 から，頻出する trigram の情報はエリアによってその傾向が大きく異なることがわかる．これは，unigram 情報や bigram 情報と比較して，その情報がもたらす情報量が大きいことを表す．

trigram では bigram に比べても組合せ数は膨大となり，それらの出現頻度はエリア毎に大きく異なる．ドイツでは SCH というドイツ語特有の文字の並びが頻出 trigram として抽出されている．また，ラテン文字表記にした場合原則として子音 (C) 1 文字

表 4.5: 頻出 trigram

エリア	1%	上位 5 個				
中国	12	ANG	HAN	IAN	NG␣	ING
台湾	11	IAO	NG␣	HAN	ANG	SHA
タイ	8	AN␣	BAN	NG␣	ONG	ANG
日本	9	SHI	AMA	AWA	AKI	CHI
ギリシャ	1	OS␣	NIS	ION	ISI	MOS
フィンランド	4	JAR	ARV	RVI	KAR	SKA
フランス	4	␣DE	DE␣	LLE	LE␣	LA␣
スペイン	5	␣DE	DE␣	LA␣	ARR	EL␣
ドイツ	4	BER	ERG	SCH	ER␣	ACH

と母音 (V) 1 文字とで 1 個の音に相当させる日本では，1 音分に相当する CV の組合せ (SHI 等[1]) と，前の音の母音が頭に着いた形の VCV の組合せ (AMA 等) が頻出 trigram として現れている．それに対して，似た傾向を示すエリア同士は，有する地名の類似度が高いことが予想される．例えば中国と台湾では HAN，ANG，NG␣ が共通して上位 5 個までに出現している．これらは SHANG␣ や CHANG␣ 等の一部と考えられ，これらの文字列が中国や台湾で共通に出現していることが推測できる．同様にフランスとスペインで ␣DE␣ が共通して頻出することが予測されるが，フランスでは LLE␣ といった文字列の頻出が予測できるのに対してスペインではこれは見られない等，エリア間の類似の度合いや理由もさまざまであることが推測できる．

trigram を用いることである程度確度の高い情報が期待できる半面，それぞれの trigram が出現する頻度はさらに低下している．1% 以上の高い出現頻度を持つ trigram を多く有する中国と台湾では，上位 5 個までの頻出 trigram だけでそれぞれ出現 trigram の 10.76% と 10.58% をカバーしているが，出現頻度の高い trigram の少ないギリシャでは，出現頻度が 5 個までの trigram をすべて合わせても全体の 4.59% しかカバーしていない．中国のコーパスで 1% 以上の頻度で出現する trigram は 12 個あり，この 12 個の trigram で全体の 19.06% をカバーしている．同じように，台湾のコーパスで 1% 以上の頻度で出現する trigram は 11 個あり，これらで全体の 18.37% をカバーしている．一方で，ギリシャのコーパスでは，1% 以上の頻度で出現する trigram は 1 個しかなく，全体の 1.18% をカバーしているにすぎない．このように，エリア間の trigram 頻度の差も，表層情報を用いた統計的な手法における所属エリア推定の精度の差となり得る．

[1]SH 等はラテン文字表記上 2 文字で記述されるが，音としては ʂ 等の子音 1 文字に相当する．

4.2 表層情報を用いた地名の所属エリア推定手法

本章では，地名の表層情報のみを用いた所属エリア推定の基本的な手法として，高い再現率の実現を目指した手法を提案する．確実に再現率を向上させるためには，所属する可能性が多少でもあるエリアをすべて出力すればよい．しかし，地名の所属可能性について類似した特徴を有するエリアが複数存在することは珍しくないため，可能性がまったくないエリアと判断されるエリアはむしろ少ないことが予想され，可能性のあるエリアをすべて残す方法では十分な絞込みができず適合率を検討する意味がない．そのため，ここでは，所属エリア推定というタスクのゴールを，可能な限り所属エリア候補を減らすことと定義する．

ここでは，地名の表層情報として地名コーパスから抽出した特徴情報を機械学習することで，各エリアの地名の特徴をモデル化することを考える．所属する可能性の高い所属エリア候補を絞り込むアプローチとして，本章では，二値分類学習器を用いて地名の表層情報の特徴を機械的に学習する手法を提案する．

4.2.1 機械学習器を用いた所属エリア候補の推定

本章で提案する手法*1*のシステムの概要を図 4.5 に示す．手法*1*は所属エリア候補の

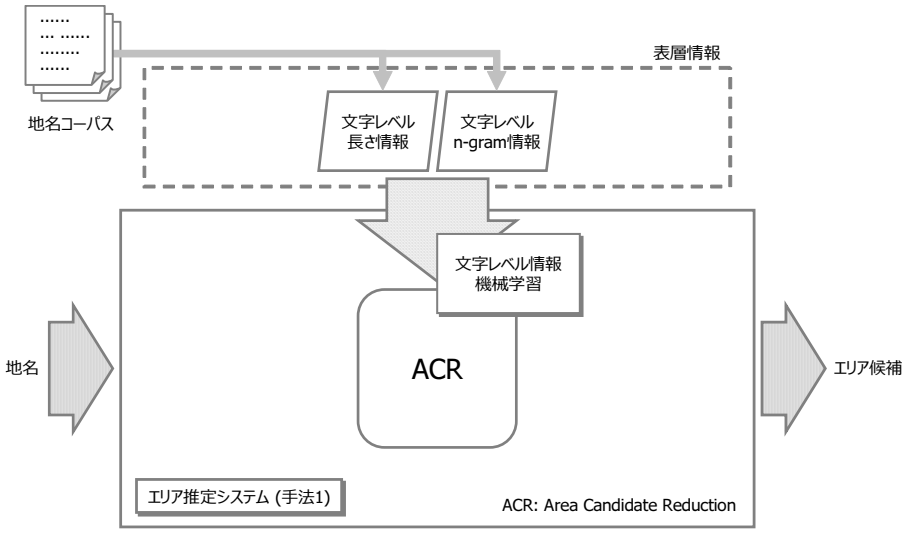


図 4.5: 手法*1*の処理の流れ

削減を目的とする area candidate reduction (ACR) モジュールのみから成る．手法*1*の ACR モジュールは機械学習器であり，エリア毎にその特徴を学習する．その上で，入力地名に対してそれぞれのエリア専用の機械学習器が独立に所属の可能性の推定を行う．ACR モジュールの機械学習器には，サポートベクタマシン [23, 26] TinySVM[2]を利用した．各入力地名に対して，各エリアに対応した二値分類学習器を用いて実験対象エリアの所属の可能性の判定処理が行われる．出力は入力地名に対して所属の可能

[2]TinySVM: `http://chasen.org/~taku/software/TinySVM/`.

性があるとの判定を行ったエリアであり，出力される所属エリア候補数は 0 個 (すべてのエリアの二値分類学習器が negative と判定した場合) から最大で実験対象エリアの数 (すべてのエリアの二値分類学習器が positive と判定した場合) までとなる．

4.2.2 機械学習に用いる素性

機械学習に用いる素性は，4.1.2 節と 4.1.3 節の調査結果に基づき，地名の表層情報として，その長さ情報と n-gram 情報を用いる．

地名の有する長さ情報

4.1.2 節で述べた地名の持つ長さ情報の特徴を示す素性を，図 4.6 のように 4 種類定義する．この 4 種類を素性とすることで，単語数 (feature_{L1}) や文字数 (feature_{L2}) の情報だけでなく，単語そのものの出現傾向の特徴や単語の分布の情報を利用することができる．例えば，フランスの地名では LA という単語が頻繁に出現するが，そのうち約半数が地名の 1 番目の単語として登場する．このように，単語によっては，その

feature_{L1}: 地名を構成する単語数
feature_{L2}: 地名を構成する文字数
feature_{L3}: n 番目の単語に含まれる文字数 ($1 \leq n \leq 16$)
feature_{L4}: 地名に含まれる m 文字の単語の数 ($1 \leq m \leq 32$)

図 4.6: 機械学習に用いる素性 (長さ情報)

単語が出現する地名内での位置に特徴がある．これを素性に表したものが feature_{L3} である．feature_{L3} は 1 個の素性ではなく，n 個の独立した素性として扱う．表 4.2 より，地名の構成単語数は 3 個から 4 個程度であるため，ここでは特に長い地名も含めて確実にカバーするため，n の範囲を $1 \leq n \leq 16$ とする．

また，図 4.1 に示したように，タイの地名は他のエリアの地名と比較して 3 文字および 4 文字で構成される単語の割合が圧倒的に高い等，地名に含まれる単語の文字数もエリアの特徴となり得る (feature_{L4})．図 4.1 より，地名の構成文字数はたかだか 20 文字程度と考えられ，単語の構成文字数はそれよりもさらに短いことが予想できる．m の範囲については，特に長い地名も含めて確実にカバーするため，$1 \leq m \leq 32$ とする．feature_{L4} も feature_{L3} と同様に m 個の独立した素性として扱う．

地名の有する n-gram 情報

4.1.3 節で示した n-gram 情報の特徴から，n-gram 情報に関する素性を図 4.7 のように 3 種類定義する．ここで用いる 3 種類の n-gram 情報はすべて文字レベルである．

feature_{N1}: 文字レベルの unigram 情報
feature_{N2}: 文字レベルの bigram 情報
feature_{N3}: 文字レベルの trigram 情報

図 4.7: 機械学習に用いる素性 (n-gram 情報)

feature_{N1} から feature_{N3} の各素性は，それぞれ n-gram 毎に計算する．したがって，feature_{N1} は unigram の異なり数の素性の集合であり，feature_{N2}，feature_{N3} はそれぞれ unigram の異なり数の 2 乗，3 乗の個数の素性の集合となる．各素性の値 v はエリア a のコーパスの n-gram の出現頻度をベースに定義する (式 (4.1))．

$$v_a(t,s) = P_a(s) \times N(t,s) \tag{4.1}$$

s を長さ n $(1 \leq n \leq 3)$ の文字列とした場合，$P_a(s)$ はエリア a において文字列 s が出現する頻度，$N(t,s)$ は地名 t の中で文字列 s が出現した回数を示す．素性の値は n-gram 情報と対象とする地名に依存しており，ある文字列のそのエリアでの出現しやすさと，その地名の中での出現しやすさを示している．

4.3 実験結果

4.3.1 実験環境

地名コーパス

本節では，表層情報の調査に用いた地名コーパス (表 4.1) を用いて所属エリア推定実験を行う．4.1.1 節で述べたとおり，本実験の対象エリアは中国，台湾，タイ，日本，ギリシャ，フィンランド，ドイツ，フランス，スペインの 9 エリアである．

ACR モジュールの設定

本実験で機械学習に用いる素性を表 4.6 に示す．素性の個数は計 30,833 個である．本実験では，空白のほかアンパサンド (&)，アポストロフィ (')，カンマ (,)，ピリオド (.) もそれぞれ別の記号として数えている．そのため，unigram の異なり数はアルファベット 26 種と記号 5 種の計 31 種となる．これらの記号の出現頻度は低く，また基データの記述形式によってその使われ方もまちまちのため，本章の実験以外では，アルファベット以外に使用する記号は空白文字のみとした (3.3.2 節)．

表 4.6: 所属エリア推定の実験に用いる素性

素性		素性の数
feature_{L1}	地名に含まれる文字の数	1
feature_{L2}	地名に含まれる単語の数	1
feature_{L3}	n 番目の単語に含まれる文字数	16
feature_{L4}	地名に含まれる m 文字の単語の数	32
feature_{N1}	文字レベルの unigram 情報	31
feature_{N2}	文字レベルの bigram 情報	961
feature_{N3}	文字レベルの trigram 情報	29,791

4.3.2 実験結果

表 4.7 に所属エリア推定実験の結果を示す．表 4.7 にあるとおり，すべてのエリアについて F 値はほぼ 0.70 を獲得し，タイに関しては 0.93 を達成した．また，すべて

表 4.7: 地名の所属エリア推定の実験結果

エリア	F 値	適合率 (%)	再現率 (%)	正解率 (%)
中国	0.75	62.00	95.60	93.00
台湾	0.72	58.93	92.75	92.01
タイ	0.93	87.66	98.00	98.24
日本	0.76	64.11	94.75	93.52
ギリシャ	0.69	55.09	93.40	90.81
フィンランド	0.70	55.80	92.35	91.02
フランス	0.74	59.63	97.05	92.37
スペイン	0.73	60.72	92.15	92.51
ドイツ	0.81	68.66	98.60	94.84
全体	0.75	62.64	94.96	93.15

のコーパスにおいて 90% 以上の正解率を得ることに成功した．全体での実験結果は，F 値 0.75 である．この結果は，n-gram 情報と長さ情報とのごくシンプルな表層情報のみを用いた所属エリア推定が十分に可能なことを示している．表 4.7 では，すべてのエリアで，適合率よりも再現率が高いとの結果となっている．手法*1*は再現率重視の手法であり適合率は考慮していないため，この手法で F 値 0.75 を得られたことは，この段階としては十分な結果といえる．

手法*1*で重視した再現率は，全体で 94.96%，すべてのエリアで 92% 以上，最も結果のよかったドイツでは 98.60% との結果を得ることができた．このことから，長さ情報および文字レベルの n-gram 情報というシンプルな表層情報のみで，正解エリアを排除

せず positive と判定するために十分な情報を得ることが可能であると考えられる．それに対して，適合率は全体で 62.64%，低いエリアでは 55% 程度に留まった．これは，所属エリア候補として誤ったエリアが多く出力されていることを示している．手法*1*では，所属エリア候補を積極的に削減することはなく，可能性があると判断したエリアはすべて出力するため，適合率が低くなる傾向があることは否めない．手法*1*のシンプルな手法で F 値 0.75 を得られた理由は，高い再現率が貢献したものであり，適合率については改良の余地が大いにある．

4.4 考察

4.4.1 地名の表層情報の有効性

本章では，地名の所属エリア推定に利用する地名の表層情報として，長さ情報と n-gram 情報に着目した．4.3.2 節の結果から，これらの表層情報を用いることで，地名の所属エリア推定は十分可能であるといえる．

n-gram 情報は，図 4.8 に示す 2 種類の情報を有している．前者は，各エリアで地名に使われる文字の種類の情報を示し，例えば L や Q 等の文字が入っていれば日本の地名の可能性は低いといった抑制要因として働いたり，母音としては出現割合の低い U が使われている地名であれば中国や台湾の地名の可能性が比較的高いとの推進要因として働いたりすることが期待できる．それに対して後者は，文字列の並びの情報を持ち，各エリアで頻出する文字列の推測が期待できる．文字の出現頻度と文字の共起頻

(1) 各 unigram の出現頻度 (unigram)
(2) 各 unigram の共起情報 ($n > 1$ の n-gram)

図 4.8: n-gram の有する情報

度が持つ情報には意味合いに違いがある．固有表現である地名は文字数が少ない上に，頻出単語の出現頻度が比較的少ないため，文字の共起頻度の情報が所属エリア推定に十分な有効性を発揮しない可能性がある．それに対して，固有表現であっても文字の出現頻度はそのエリアの特徴を反映しているため，地名の所属エリア推定では，文字の出現頻度がベースとなるものと考えられる．そこで，unigram 情報のみを用いて所属エリア推定を行った場合の実験結果を表 4.8 に示し，長さ情報と n-gram 情報の有効性を検証する．表 4.8 に示すとおり，unigram 情報のみに基づいて所属エリア推定を行った場合の F 値は全体で 0.69 であり，表 4.7 に示した手法*1*の F 値 0.75 に比べても大きな差がないことがわかる．これは，n-gram 情報，特に文字の出現割合が所属エリア推定処理において重要な役割を担っていることを示している．

F 値の算出に用いる適合率および再現率を比較すると，unigram のみで所属エリア推定を行った場合，手法*1*と比較して適合率，再現率ともに 6% 程度劣る (表 4.7，表 4.8)．

表 4.8: unigram のみを用いた地名の所属エリア推定の実験結果

エリア	F 値	適合率 (%)	再現率 (%)	正解率 (%)
中国	0.67	53.36	90.50	90.16
台湾	0.65	51.11	88.45	89.32
タイ	0.87	78.90	97.40	96.82
日本	0.68	55.12	88.00	90.71
ギリシャ	0.64	51.42	83.20	89.40
フィンランド	0.67	52.62	90.50	89.89
フランス	0.61	48.11	83.55	88.16
スペイン	0.69	56.79	87.45	91.21
ドイツ	0.77	64.04	96.95	93.61
全体	0.69	56.03	89.56	91.03

これをエリア毎に比較した結果を図 4.9 に示す．x 軸は適合率 (precision)，y 軸は再現率 (recall) を表している．図 4.9 で，丸印のマーカは手法*1*の所属エリア推定の実験結果，四角のマーカは unigram のみによる所属エリア推定の実験結果を表す．大きな丸印および四角はそれぞれの手法での全体の値を示す．図 4.9 では，unigram のみの実験結果がほとんど再現率 80% から 90%，適合率 50% から 60% のあたりに固まっているのに対して，手法*1*の実験結果はすべてのエリアについて再現率，適合率ともに unigram のみの結果に比べて向上している．手法*1*での再現率，適合率の向上の度合いは，エリアによって異なる．これは，4.1.2 節および 4.1.3 節で述べたように，エリア

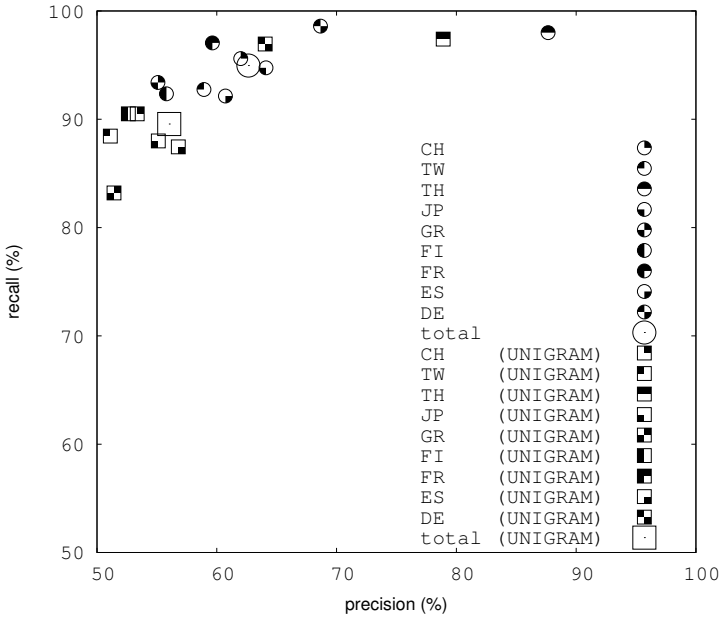


図 4.9: 手法*1*と unigram による推定との比較

毎に表層情報の特徴の出方の差異に起因するものである．長さ情報に特徴が見られたエリア，trigram で特徴が見られたエリア等，エリア毎の特徴が複合的に効果を示すことによって，手法*1*での実験結果においてエリア間で差異が現れたものである．

図 4.9 では unigram のみでの所属エリア推定実験でタイとドイツが他のエリアに対して突出した値を示しているが，タイもドイツも unigram 情報，特に母音の出現傾向が特徴的であったことがよい結果の要因と考えられる．これらから，地名の所属エリア推定では，unigram 情報が基本的な推定情報として用いられ，$n > 1$ の n-gram の持つ文字の共起情報や地名文字列の長さ情報が再現率，適合率双方の向上に有効に働くものであるということができる．また，エリア毎に表層的な特徴の出方には差異があることがわかり，タイやドイツの例のように表層的特徴を的確に捉える手法が提案できれば，所属エリア推定の精度の向上が期待できることがわかった．

4.4.2 文字レベルの n-gram 情報の利用における課題

表 4.9 に，本実験での所属エリア推定の成功例と失敗例を挙げる．ここでの成功とは正解エリアが正しく所属エリア候補として出力されたもの，失敗とは正解エリアが排除されたものを指す．表 4.9 を見る限り，所属エリア推定に成功した地名と失敗した

表 4.9: 地名の所属エリア推定の成功例と失敗例

エリア	成功例		失敗例	
中国	Qingdao	Hangzhou	Amoy	Maocao
台湾	Tainan	Kao Hsiung	Keelun	Jhongjheng
タイ	Bangkok	Nakhon Pathom	Ayuthaya	Chanthaburi
日本	Nagoya	Kagoshima	Shuzenji	Horobetsu
ギリシャ	Santorini	Thessaloniki	Plemenaki	Manatadhes
フィンランド	Helsinki	Tempere	Issalo	Heinonen
フランス	Toulouse	Bourgogne	Grandris	Narthoux
スペイン	Galicia	Valencia	Migueláñez	Cortijo Juan Gomez
ドイツ	Sachsen	Rotenberg	Maximiliansau	Hanau am Main

地名の間に大きな差異は見受けられない．例えば日本の地名を見ると，Horobetsu (**幌別**) が所属エリア推定に失敗し，日本の地名ではないと判定されている．Horobetsu はアイヌ語に由来する北海道の地名であり，これが失敗の原因と考えられるが，実際にはアイヌ語由来の北海道の地名も地名コーパスに多く記載されており，HORO (**ポロ**: **大**) や BETSU (**ペッ**: **川**) 等の文字列も珍しくない．同様に，他のエリアの失敗例を見ても，そのエリアの地名として不自然ではないものが多く混じっており，これらは十分に推定することが可能であると考えられる．

これは，文字列の並びの情報が n-gram で評価されたことに起因する．前述の Horobetsu の例では，HORO および BETSU がそれぞれに意味を持っているが，n-gram で

は語の中の意味的な区切り目は考慮せず機械的に n の値に応じて切り分けて数えている．本実験では n の値の範囲を $1 \leq n \leq 3$ としているため，4 文字の HORO，5 文字の BETSU はそれぞれ 1 文字から 3 文字の n-gram に切り分けられた状態で評価され，その結果，アイヌ語由来で他の地名に含まれる文字列に比べて頻度が低い HORO 等の文字列の頻度が十分に考慮されず，Horobetsu の推定の失敗に繋がったものと考えられる．したがって，このような意味的なかたまりを考慮するためには，n-gram の n の値を十分に大きくする必要がある．これは，単語レベル等，より大きな粒度の n-gram 情報の有効性を示唆するものである．

手法*1*では n-gram 情報として文字レベルの情報のみを用いている．実験結果から，文字レベルという粒度の細かい n-gram 情報は，エリア特有の表層情報の抽出に有効であり，再現率の確保に大きく貢献しているものといえる．それに対して，前述の北海道の地名のようにエリアの中でも特殊な性質を持つ地名に対しては文字レベルは粒度が細か過ぎ，ある程度の出現頻度が見込める文字列についても他の頻出文字列との関係からその情報を掬いきれない状況があることが認められる．

4.4.3 類似した地名を有するエリア間の識別

本実験では，中国と類似した地名を含む可能性のあるエリアとして，同じ中国語圏である台湾コーパスを実験対象に加えた．この 2 エリア間の識別について論ずるため，中国と台湾の地名について，所属エリア推定の結果を図 4.10 および図 4.11 に示す．y 軸は各エリアの地名に対する推定結果の種類の割合 (ratio) を表している． 中国の地名

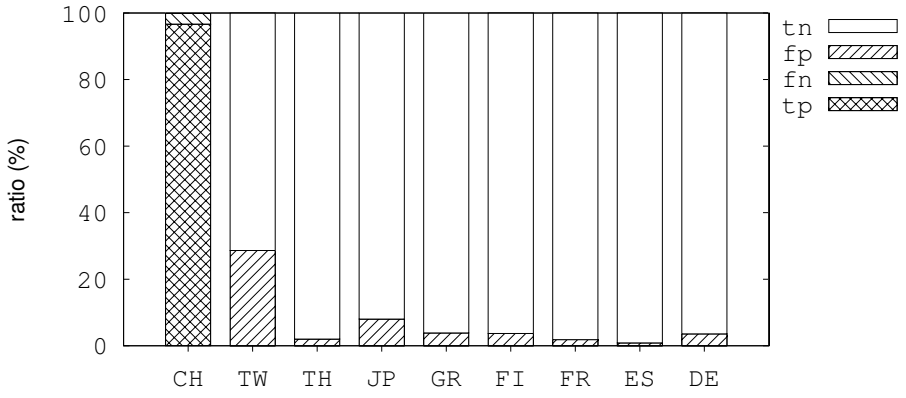


図 4.10: 中国の地名の所属エリア推定結果

に対しては，中国エリアを正しく positive と出力した割合 (tp) が 96.62%，台湾エリアを誤って positive と出力した割合 (fp) が 28.62% であった (図 4.10)．それに対して，台湾エリアを正しく positive とした割合 (tp) が 93.38%，中国エリアを誤って positive とした割合 (fp) が 41.15% であった (図 4.11)．このように，中国エリアも台湾エリアもそれぞれ正解エリアを正しく positive と判定する割合は 90% と十分高いが，同時に，他のエリアに比べて明らかに高い割合でお互いを誤って positive と判定している．

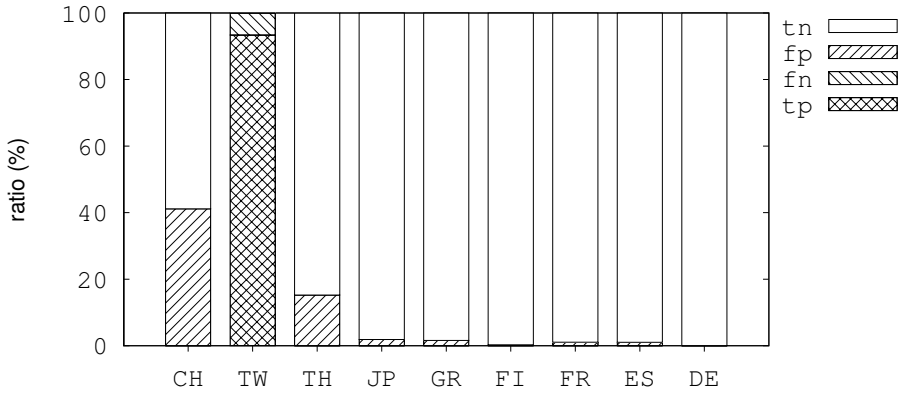


図 4.11: 台湾の地名の所属エリア推定結果

このことから，地名の持つ表層情報だけでは，似た地名を含むエリア間の識別を十分に行うことは難しいことがわかる．

この結果は逆に，同じ言語圏のエリアであっても，エリア間の識別が行える可能性があることを示唆しているとも読める．4.1.3 節で示した n-gram データからは，中国と台湾の n-gram の出現傾向に多少の差異が認められる．この差異は他のエリアとの間の差異に比べて明らかに小さいが，これが両者の識別を可能にしている．中国と台湾は中国語圏であり，ラテン文字表記はもともとの原言語の表記ではない．そのため，中国語からラテン文字表記への変換を行う際の変換ルールの差が，こういった差異に繋がっている可能性もある．4.1.2 節の長さ情報のデータから，台湾の地名は中国の地名に対して平均単語数や平均文字数が多いことがわかる (表 4.2) が，例えば表 4.9 の中国エリアの成功例として挙げられている Qingdao (**青島**) や Hangzhou (**杭州**) と台湾エリアの成功例として挙げられている Kao Hsiung (**高雄**) を比べると，台湾エリアでは漢字 1 文字をラテン文字表記では 1 語として記述する例が多く，このことが長さ情報の差異として現れている可能性がある．このようにラテン文字表記への変換が必要な地名については，その変換ルールは必ずしも徹底されていないのが一般的で，台湾エリアでも**台南**は Tainan と 1 語で表現されている場合もあり，また今回基データとした知識ベースとは別の地図データでは**高雄**が Kaohsiung と表現されている．しかし，このような変換ルールについては，エリア内の表記の揺れ以上にエリア間の変換ルールの差異が大きい可能性は十分にあり，これが類似エリア間の自動識別に利用できる可能性がある．これらのことから，類似エリア間の自動識別を実現するには，表層情報の粒度に着目し，各エリアの表層的な特徴をさらに確実に抽出する手法を提案し，適合率の向上を図る必要があると考えられる．

4.5 手法*1*のまとめ

本研究では，単純な地名コーパスから自動的に抽出可能な表層的な特徴のみに基づいて地名の所属エリア推定を行うシンプルな手法を提案することを目標としている．本章は，地名の持つ表層情報の特徴の調査結果を示し，これを所属エリア推定に有効に

活用する方法に関して実験および考察するものである．手法*1*は地名の表層情報として n-gram 情報と長さ情報とを用い，9 エリアを対象とした所属エリア推定実験ですべてのエリアに対して 90% 以上の正解率を達成した．特に，類似エリアを含む実験にも関わらずすべてのエリアで 90% 以上の正解率を得ることができたことは，本章で着目した再現率に関して，長さ情報および n-gram 情報等の表層情報の有効性を示すものと考える．このことから，地名の表層情報のみによる所属エリア推定は可能といえる．

手法*1*では，n-gram 情報として文字レベルの情報のみを利用している．低い適合率を改善するためには，より確度の高い情報を利用することが必要となる．文字レベルの情報は，連続する文字の並びを局所的に観測することによって地名の特徴を抽出するものである．手法*1*では n-gram として unigram，bigram，trigram のみを用いており，たかだか 3 文字の並びに着目しているにすぎない．実験結果から，表層情報の粒度の細かさが高い再現率に貢献していると同時に，低い適合率を生み出している可能性が認められた．そのため，適合率の改善を目指し，文字レベルだけでなくさらに大きな粒度の処理単位の導入が課題である．また，類似エリア間の識別についても，さらに詳しく調べる必要がある．

第5章　表層情報の効果的利用とブロックの提案 (手法*2*，手法*1+2*)

手法*1* (第 4 章) では，地名の表層情報が所属エリア推定を行うのに十分な情報を有していることを確認したが，適合率の向上には処理単位の検討が必要なことも示した．そこで本章では，所属エリア推定に適した処理単位の提案を行う．

まず，文字よりも粒度の大きい処理単位として単語に着目し，その特徴および所属エリア推定での活用を検討する．その結果を基に，文字と単語の間の粒度を持つ処理単位としてブロックの概念を提案する．所属エリア推定に適した単位としてのブロックを定義するとともに，これを用いた所属エリア推定の実験結果を報告する (手法*2*)．さらに，第 4 章で用いた手法*1*に手法*2*を組み合わせることで再現率を低下させることなく適合率を向上させることができることを示す (手法*1+2*)．

5.1　粒度の大きい処理単位としての単語レベルの表層情報

5.1.1　地名コーパス

本章で用いる地名コーパスを表 5.1 に示す．表 5.1 の 11 エリア中，台湾およびチリ

表 5.1: 本章で用いる地名コーパス

エリア	地名数	単語数	文字数
中国	10,000	15,975	98,773
台湾	10,000	22,518	122,630
タイ	10,000	28,974	131,128
日本	10,000	16,960	102,311
ギリシャ	10,000	14,612	109,818
フィンランド	10,000	11,050	97,447
フランス	10,000	19,820	121,945
スペイン	10,000	22,797	135,947
チリ	10,000	22,450	139,561
ドイツ	10,000	13,206	109,615
アメリカ	10,000	25,622	165,509

は類似エリアの混在への影響を確認するために用いるもので，本章の基本的な実験対象エリアは台湾，チリを除いた 9 エリアとする．

5.1.2 単語レベルの表層情報

第 4 章では，地名の所属エリア推定処理の処理単位として，文字のみを利用した．本節では，文字よりも粒度の大きな処理単位としての単語の持つ特徴を検討する．

地名の TF-IDF 情報

地名コーパスに出現する単語の特徴を，TF-IDF モデルの概念にしたがって検討する．TF-IDF モデルとは，文書の特徴を示す重要語の推定に使われるモデルで，ある単語の各文書内での出現頻度 (term frequency: TF) と，この単語が出現する文書の数 (document frequency: DF) とから，一般的な頻出語と各文書を特徴付ける頻出語とを区別するものである．基本的に，DF が低くある文書でのみ TF が高い単語はその文書の特徴を現す重要語と考えられるため，TF-IDF モデルでは TF と IDF (inversed document frequency) とを基に各単語の評価値の算出を行う．ここではエリアの特徴を示す単語の抽出を考え，TF としては単語の各エリア内での出現頻度，DF としては単語が出現するエリアの数を用いる．

例として，Mt. Fuji San という地名の所属エリア推定を行うと仮定する．Fuji という単語が日本の地名コーパスにのみ出現し，他の地名コーパスに出現することがなければ，この地名は日本の地名である可能性が強いと考えることができる．逆に，すべての地名コーパスに含まれる単語は，エリアの特徴を示す可能性は低くなる．図 5.1 の

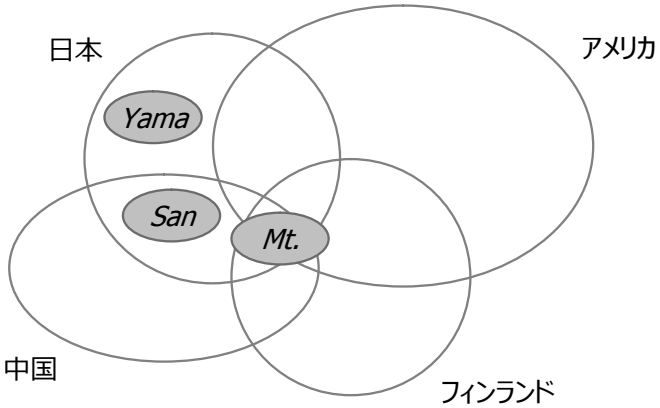


図 5.1: 地名コーパスに出現する構成単語の例

例のようにすべての地名コーパスに Mt. が含まれる場合，Mt. という単語を含むことは，Mt. Fuji San がアメリカの地名であることの証拠とはならない．これは Mt. の存在は Mt. Fuji San がフィンランドの地名であることの証拠ともフィンランドの地名でないことの証拠ともならないことと同じである．このように，DF の高い単語は地名の

所属エリア推定では影響を小さく抑えるべきである．それに対して，図 5.1 のように Yama が日本のコーパスにしか現れないような場合，推定対象となる地名が Mt. Fuji Yama のように Yama を含むことは，所属エリアを日本と推定する根拠となり得る．また図 5.1 の San のように，一部のエリアでは出現するがすべてのエリアで出現するわけではない単語は，エリア毎に TF，DF それぞれを考慮して，適切に処理する必要がある．

表 5.2 に地名コーパス全体で出現する単語の例を示す．'DF' カラムの数字は，単語が出現したエリアの数を表す．例えば，'DF' の 9 は 9 エリアすべてで出現した単語を意味し，CANAL がすべての地名コーパスで出現したことを示す．表 5.2 が示すよう

表 5.2: 実験対象エリア群で DF 値の高い単語

DF	出現文書数の多い単語
9	CANAL
8	RIVER POINT
7	I BAY ISLAND LAKE LA DE SAN
6	A E O STATION PASS SOUTH CAPE CHANNEL NORTH TO WEST EN

に，高い DF 値を持つ単語のほとんどが英単語である．これは，英語の地名データから地名コーパスを作成したためであり (3.3.2 節)，Biwako Canal のように地名の中に英単語が含まれる例が見られることから，基データの性質を反映した自然な結果といえる．DF 値は各エリアで出現したか否かを示すのみで，これらの単語の出現頻度の分布はエリア毎の地名コーパスで違いがある．表 5.2 中，DE や LA は 7 エリアで出現しているが，これらが 3% を超える頻度で出現しているのはフランスとスペインの地名コーパスのみであった．これらの単語は，高い DF 値を持つが，各エリアでの TF 値からフランスやスペインのエリア特有の単語として捉えることができる．

表 5.3 に各エリアでの頻出単語の上位 5 個を示す．なお，類似したエリア間の比較のため，表 5.3 には台湾とチリを含める．表 5.3 では，頻出単語のエリア間の重複はほとんどなく，頻出単語はほぼエリア固有の単語と考えられる．このように，地名に含まれる単語の TF と DF の情報は，所属エリア推定に貢献する可能性がある．またその中で，DE と LA が同じように頻出するフランスとスペイン等，同じ単語が頻出するエリアには強い相関関係が示されることが予測される．

地名の構成単語の出現頻度

地名は，固有表現としての性質から，地名の特定が可能になるよう，**山**や**町**等に相当する地名特有の語以外の構成単語の出現頻度は低いことが予想される．図 5.2 に，5.1.1 節の 9 エリアの地名を構成する単語のうち頻度の少ないものが占める割合 (ratio) を示す．図 5.2 の凡例の `1` はそれぞれの地名コーパスの中に 1 回のみ出現した単語の割合を，`4+` は各地名コーパス中の出現頻度が 4 回以上の単語の割合を表す．`TOTAL` は，

表 5.3: 地名コーパス中での頻出単語

エリア	頻出単語				
中国	SHAN	HSIEN	HO	CHEN	XIAN
台湾	SHAN	CIAO	QIAO	LI	HSI
タイ	BAN	KHLONG	HUAI	KHAO	NONG
日本	YAMA	MURA	SAKI	GAWA	SHIMA
ギリシャ	AKRA	ORMOS	NISIS	CAPE	REMA
フィンランド	ISO	STORA	SODRA	STOR	NORRA
フランス	DE	LA	LE	SAINT	LES
スペイン	DE	LA	DEL	ARROYO	RIO
チリ	CERRO	DE	QUEBRADA	ESTERO	PUNTA
ドイツ	BERG	BACH	BAHNHOF	SEE	WALD
アメリカ	CREEK	CHURCH	CEMETERY	SCHOOL	LAKE

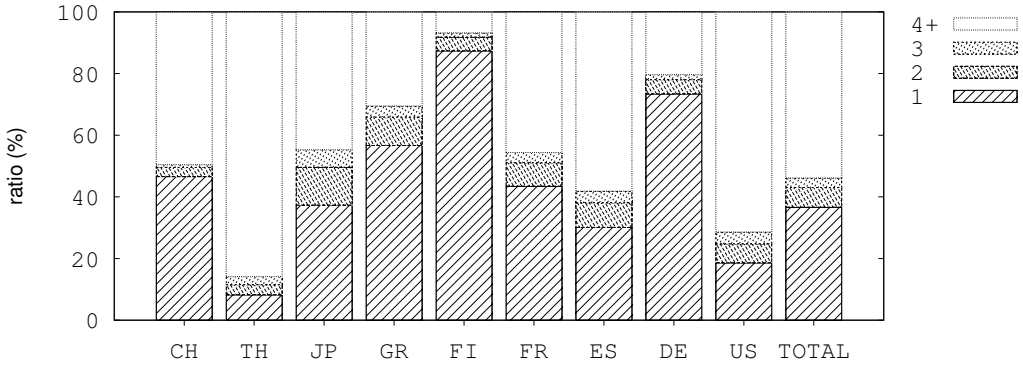


図 5.2: 出現頻度毎の単語の割合

9 個の地名コーパス全体の中での出現頻度を示している．すなわち，TOTAL の 1 は，9 個の地名コーパス全体の中で 1 回のみ出現した単語の割合を示す．

図 5.2 を見ると，地名コーパスに出現する単語のうち，30% 以上は出現頻度が 1 である．4.1.2 節で述べたとおり，多くの地名は 3 単語以下で構成されているため，1 個の未知語が致命的な結果をもたらす可能性があり，単語レベルの情報を用いた処理はリスクが高い．このようなデータスパースネスの問題が，単語レベルの表層情報の利用の課題である．

地名の構成単語の出現頻度には，エリア毎に大きな偏りが見られる．例えばフィンランドでは 80% 以上の単語が出現頻度 1 なのに対して，タイでは出現頻度 4 以上の単語が 80% を超えている．エリア間の差異が大きいため，全体の値は平均程度の意味しか持たず，所属エリア推定での頻出単語の有効性はエリアによって大きく異なるものと推測できる．

5.2 ブロックの提案および定義

5.1 節で述べたとおり，単語レベルの表層情報は第 4 章で利用した文字レベルの表層情報とは異なる情報を含むものと考えられ，適合率の向上に効果を示す可能性がある．しかし，5.1.2 節に示したとおり単語レベルの表層情報はデータスパースネスの問題を抱える．そこで本章では，単語に代わる地名の処理単位として，ブロックと呼ぶ新たな単位を導入する．

5.2.1 新しい処理単位としてのブロックの提案

ここで提案するブロックとは，n-gram の一種である．第 4 章では，文字レベルの表層情報として，その出現頻度に関わらずすべての n-gram を対象とした．またそのため，n のサイズは $1 \leq n \leq 3$ に限定していた．それに対して，本章では n のサイズの拡大を許し，頻出 n-gram を固定の文字の並び (ブロック) と定義し，これを用いて地名の構成単語に含まれる特徴的な文字列の認識を行う．

例えば，Yamagata 等，日本の地名コーパスでは語頭に YAMA を含むものが数多く存在する．また同様に，Heidelberg 等，ドイツの地名コーパスでは語尾に BERG を含むものが多数存在する．Yamagata や Heidelberg 自身は頻出単語ではないが，これらに含まれる部分文字列である YAMA や BERG は，そのエリアで頻出する文字列であり，そのエリア固有の表層情報を有すると考えられる．このように，地名の構成単語は，その文字列中にそのエリア固有の文字列を有している可能性がある．ブロックの概念は，このような隠れた固有文字列の抽出を可能にするものである．したがって，ブロックは単語と比較してより高い頻度で出現することが期待され，その点が統計情報をベースとするアプローチの重要なポイントとなる．

本章では，地名の特徴を適切に捉えるブロックとして，特に，語頭ブロックと語尾ブロックの 2 種類を考える．語頭ブロックとは前述の Yamagata の YAMA のように地名の構成単語の語頭に現れるブロック，語尾ブロックとは Heidelberg の BERG のように地名の構成単語の語尾に現れるブロックを指す．ブロックは，例えば**東**や**新**等の説明的な接頭辞や，**町**や**山**のような地名固有の語が接尾辞的に付着したもの等，意味的，言語学的に定義する接頭辞や接尾辞に対応する可能性があるが，それに限るものではない．地名の構成単語に含まれる意味的なかたまりを明示的に認識するために辞書等の知識情報を利用することは本研究の趣旨ではないため，この意味的なかたまりをブロックを用いて再現する．本研究では，ブロックをあくまでも統計的に処理することとし，地名中の単語の語頭や語尾に現れる文字レベルの頻出 n-gram と定義する．これにより，機械的な認識処理が可能となる．

ブロックはさまざまな長さの文字列となり得る．短い n-gram ほど出現頻度が高くなりやすいため，ブロックの長さ (ブロックサイズ) の異なるブロック同士を単純に出現頻度で比較することは適切ではない．そこで，ブロックは，n-gram の n 毎に出現頻度を考慮して抽出することとする．単語 w の n 文字サイズのブロック $B^{*n}(w, i)$ を

式 (5.1) で定義する．

$$B^{*n}(w,i) \quad = \quad c_i \cdots c_{i+n-1} \quad \text{for } i \geq 1, i+n-1 \leq l \tag{5.1}$$

ここで i はブロックの開始位置，c_k は単語 w 中の k 番目の文字，l は単語 w ($w = c_1 \cdots c_l$) の単語長 (単語に含まれる文字の数: $1 \leq k \leq l$) とする．また，式 (5.1) を用いて，単語 w の語頭ブロック $B_P^{*n}(w)$ および語尾ブロック $B_S^{*n}(w)$ はそれぞれ式 (5.2) と式 (5.3) で定義する．

$$B_P^{*n}(w) \quad = \quad B^{*n}(w,1) \qquad \text{for } n \leq l \tag{5.2}$$

$$B_S^{*n}(w) \quad = \quad B^{*n}(w,l-n+1) \qquad \text{for } n \leq l \tag{5.3}$$

5.2.2 語頭ブロックおよび語尾ブロック

表 5.4 および表 5.5 に，5.1.1 節の地名コーパス中に出現する 5 文字の語頭ブロックと語尾ブロックの例を示す．ブロックは DF を考慮せず各エリア内で独立に出現頻度を

表 5.4: 5 文字の語頭ブロックの例 (出現頻度上位 7 個)

エリア	語頭ブロック						
中国	HSIEN	SHANG	HUANG	CHANG	CHUAN	CHIAN	SHUIK
台湾	GUOXI	HSIAO	GUOSI	CHUNG	CHIAO	SHANG	CHUAN
タイ	KHLON	AMPHO	MUANG	CHANG	THUNG	RIVER	LUANG
日本	SHIMA	MACHI	MISAK	NISHI	SHIMO	HIGAS	SHIRA
ギリシャ	ORMOS	NISIS	VRACH	POTAM	NISOS	NISID	AYIOS
フィンランド	STORA	FJARD	SODRA	NORRA	LILLA	SVART	VASTE
フランス	SAINT	RIVIE	RUISS	GRAND	FORET	VILLE	CHATE
スペイン	ARROY	CASER	VILLA	PUNTA	SIERR	BARRA	CORTI
チリ	CERRO	QUEBR	ESTER	PUNTA	CALET	BAHIA	ISLOT
ドイツ	GROSS	FORST	KLEIN	UNTER	BAHNH	NIEDE	STEIN
アメリカ	CREEK	CHURC	CEMET	SCHOO	SPRIN	BRANC	MOUNT

調べて抽出するが，表 5.4，表 5.5 ともに各エリアのブロックには重複が少なく，ブロックがエリアの特徴を有することが期待できる．例えば日本のコーパスからは，SHIMA (**島**) や MACHI (**町**) が語頭，語尾双方のブロックとして抽出されている．また，NISHI (**西**) や HIGASHI (**東**) の一部のように地名の前後に付く文字列もブロックとして抽出された．HIGASHI は語頭，語尾ともに 5 文字までの部分文字列として抽出されているが，これはブロックが文字列の持つ意味を考慮せずに定義されているためである．提案する定義ではそれだけで意味を持つ文字列か否かの推定は行わず頻度のみでブロックを認識するため，それだけでは意味を成さない部分文字列であっても，頻出する文

表 5.5: 5 文字の語尾ブロックの例 (出現頻度上位 7 個)

エリア	語尾ブロック						
中国	HUANG	HSIEN	CHANG	XIANG	HIANG	HUIKU	CHENG
台湾	HUANG	HSIAO	OXIAO	OSIAO	CHUNG	CHANG	SIANG
タイ	HLONG	MUANG	MPHOE	THUNG	LUANG	RIVER	CHANG
日本	SHIMA	MACHI	ISAKI	BETSU	GASHI	SHIMO	INAMI
ギリシャ	ORMOS	NISIS	ISIDA	NISOS	AIIKA	TAMOS	ORION
フィンランド	JARVI	NIEMI	VAARA	SAARI	LAHTI	SELKA	OLMEN
フランス	SAINT	VILLE	VIERE	SSEAU	COURT	FORET	IERES
スペイン	RROYO	SERIO	PUNTA	IERRA	RANCO	SANTA	CASAS
チリ	CERRO	BRADA	STERO	PUNTA	ALETA	BAHIA	UERTO
ドイツ	AUSEN	SDORF	INGEN	NDORF	FORST	NBERG	HNHOF
アメリカ	CREEK	HURCH	ETERY	CHOOL	RANCH	PRING	ENTER

字列はブロックとして抽出される．ドイツのコーパスからは，GROSS (**大**) と KLEIN (**小**)，UNTER (**下**) 等の接頭辞的な文字列が語頭ブロックとして抽出されているほか，STEIN (**石**) 等が抽出された．ドイツの語尾ブロックとして SDORF や NDORF のように似た文字列が入っているが，DORF には**村**という意味があり，これに S や N が付着した状態で切り出されたものと考えられる．また，語頭ブロック，語尾ブロックから，Bahnhof (**駅**) という単語がドイツのコーパスに多く出現したことがうかがえる．フランスでも**町**に相当する VILLE がブロックとして抽出されているが，語尾だけでなく語頭ブロックとしても頻出しているのは，Ville が 1 単語として出現する頻度が高いためではなく，Thionville や Lunéville のように VILLE が語尾に付く地名と，Villeneuve-d'Ascq や Villeneuve-sur-Lot のように VILLE が語頭に付く地名とが多く出現したためである[1]．ドイツの BAHNHOF のように 1 単語の形で出現する頻出文字列は単語レベルの統計情報で反映できるが，フランスの VILLE のように 1 単語ではほとんど出現しない頻出文字列は単語レベルの統計情報では認識できない．ブロックは，このような頻出文字列を所属エリア推定に効果的に活用することを目的としている．

表 5.4 および表 5.5 には，比較対照として，中国と同じ言語圏の台湾と，スペインの同じ言語圏のチリとを含めた．表 5.4 と表 5.5 の双方で，出現頻度上位 7 個までに同一ブロックが出現したのはこの 2 組のほか，中国，台湾，タイの 3 エリアで同じ語尾ブロック が 1 個 (CHANG) 出現したのみであった．中国と台湾の間の共通ブロックは語頭の SHANG と語尾の HUANG および CHANG の 3 個である．これら 3 個のブロックはすべて ANG を含んでおり，4.1.3 節に現れた頻出 n-gram の傾向に一致する．

[1]フランスのコーパスに含まれる 10,000 地名について VILLE の出現頻度を調べると，Sainte-Eugénie-de-Villeneuve のように VILLE が語頭に現れた地名が 92 (単語として現れた頻度も合わせると 118)，Tourville-sur-Sienne のように VILLE が語尾に現れた地名が 236 (単語として現れた頻度も合わせると 262) なのに対して，La Ville aux Dames のように VILLE が独立した 1 語として現れた地名は 26 に過ぎなかった．このことから，VILLE は地名構成要素としては独立した 1 語として扱われるよりも単語の部分文字列として含まれることが多いと考えられる．

語頭の SHANG は，例えば Shanghai (**上海**) の**上**に相当する．

スペインとチリの間の共通ブロックは，語頭と語尾の双方で出現した PUNTA (**岬**) のみであった．PUNTA はスペインの Punta del Cerro de la Cruz やチリの Punta de Díaz のように Punta で 1 単語として使われる 5 文字の文字列のため[2]，5 文字のブロックとしては語頭ブロックおよび語尾ブロックの双方として数えられている．スペインおよびチリで頻出した Punta (**岬**) のように，地名特有の単語や文字列が同じ言語圏のエリアで共通して出現することは自然であり，このような単語や文字列の出現は言語圏の特徴を示すものである．さらに，同じ言語圏でもブロックの出現状況に偏りが見られたことは，n-gram 情報が類似した特徴を示すようなエリア間の識別において，このブロックが有効であることを示唆するものである．

5.2.3 ブロックサイズとショートブロック

ブロックはブロックサイズ (ブロックの文字列の長さ) 毎に抽出を行うが，ブロックサイズによっては，そのサイズより短い単語が出現する可能性がある．この場合，この単語の扱いとしては，(1) ブロックサイズに満たないため，そのブロックサイズではブロック候補から除外する，(2) ブロックサイズに満たないため，その単語全体を 1 個のブロックとして扱う，との 2 とおりが考えられる．ここで，(2) のアプローチをとる場合，この単語をショートブロックと定義する．

ショートブロックを含んだブロックの定義を式 (5.4) に示す．

$$B^n(w,i) = \begin{cases} c_i \cdots c_{i+n-1} & \text{for } i \geq 1, i+n-1 \leq l \\ w & \text{for } n > l \end{cases} \tag{5.4}$$

ここで，i は単語 w におけるブロックの開始位置，c_k は単語 w における k 番目の文字，l は単語 w $(c_1 \cdots c_l)$ の単語長 (単語に含まれる文字の数: $1 \leq k \leq l$) を表している．また，式 (5.4) のショートブロックを含むブロックの定義を用いて，単語 w の語頭ブロック $B_P^n(w)$ および語尾ブロック $B_S^n(w)$ を，式 (5.5) と式 (5.6) で定義する．

$$B_P^n(w) = B^n(w,1) \tag{5.5}$$

$$B_S^n(w) = B^n(w,l-n+1) \tag{5.6}$$

例えば，ブロックサイズが 5 の場合，Yama という単語は単語長が 4 のため 5 文字のブロックの切り出しができない．この場合，YAMA を 5 文字ブロックのショートブロックとして抽出し，語頭ブロックおよび語尾ブロックの双方として扱う．この場合のショートブロックとは，あくまでもブロックサイズに満たない単語に対して適用さ

[2]スペインのコーパスの 10,000 の地名のうち PUNTA が 1 単語として現れた地名は 273 個なのに対して，単語の語頭の部分文字列として現れた地名は 12 個で，単語の語尾の部分文字列として現れた地名は無かった．海岸線の長いチリでは，コーパスの 10,000 の地名のうち PUNTA が 1 単語として現れた地名は 626 個なのに対して，語頭の部分文字列として現れた地名は 8 個，語尾の部分文字列として現れた地名は 2 個だった．このことから，PUNTA は地名の構成要素としては基本的に 1 単語として扱われる文字列と考えられる．

れる考え方である．したがって，例えば Yamagata から YAMA を切り出す等，ブロックサイズを満たす長さの単語からブロックサイズより短い文字列を切り出すことはしない．このように，ショートブロックは，短いブロックサイズのブロックとは異なる概念である．

ブロックサイズを 1 個に固定し，ショートブロックを考慮せずにブロックの抽出を行う場合，短い単語の情報が排除される．例えばブロックサイズが 5 の場合，Yamagata からは YAMAG という語頭ブロックと AGATA という語尾ブロックを抽出することができるが，Mie は単語長が 3 のためサイズ 5 のブロックの抽出ができず，地名の構成単語 Mie の存在がブロックに反映できない．

またブロックサイズを複数設定した場合でも，同様の問題が生じる．推定処理に際して複数のブロックサイズを利用する場合，同じ単語から長さの異なる複数のブロックが抽出されるが，これを同時に推定処理に用いることは同じ単語の影響を複数のブロックによって与えることになる．例えば Yamagata に対してブロックサイズが 3 から 5 のブロックを推定に利用する場合，語頭として YAM，YAMA，YAMAG の 3 種類，語尾として ATA，GATA，AGATA の 3 種類のブロックがそれぞれ推定に影響を与えることになる．この場合，語頭の YAM 部分，語尾の ATA 部分はそれぞれ 3 重に影響を与えることになる．これに対して，同じ条件でも Mie のような短い単語は，ブロックサイズ 4 以上のブロックは抽出できないため，語頭，語尾としてそれぞれ MIE が 1 回ずつ抽出されるだけとなり，地名の構成単語 Yamagata と地名の構成単語 Mie それぞれの出現頻度が同じ場合でも，Yamagata の方が Mie よりも大きな影響を与えることになり不均衡が生じる．

ショートブロックは，この問題の解決のために導入する概念である．ショートブロックを導入することで，ブロックサイズに満たない短い単語についてもブロックの情報を反映させることが可能となる．例えば Mie に対しては，ブロックサイズが 3 から 5 の場合，サイズ 3 のブロックとして語頭語尾それぞれで MIE が抽出されるほか，サイズ 4，サイズ 5 のブロックとしても MIE がショートブロックとの扱いで語頭語尾それぞれについて抽出されるため，十分な単語長を持つ単語と同等の影響力を持つことができるようになる．

5.3 ブロックの概念を導入した所属エリア推定手法

本節では，ブロックの情報を利用して所属エリア推定を行う手法について説明する．ここでは，まず TF-IDF モデルにしたがって単語レベルの統計情報を基に地名の所属エリア推定を行う手法を提案し，これをブロックに拡張する形でブロックレベルの情報を用いた所属エリア推定手法を提案する．

5.3.1 ブロックを用いた所属エリア候補の推定

本節で提案する手法*2*のシステムの概要を図 5.3 に示す．手法*2*は所属エリア候補の判定を目的とする area candidate examination (ACE) モジュールのみから成る．各エ

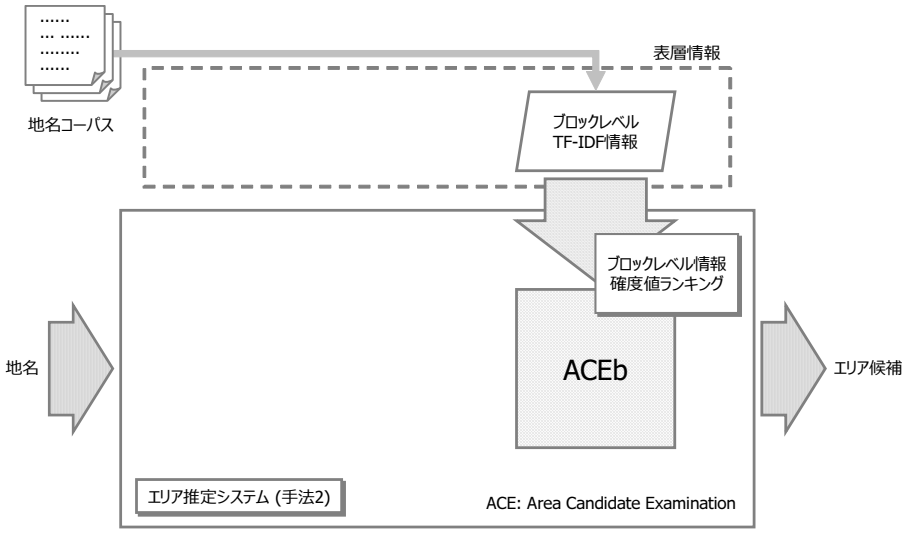


図 5.3: 手法2の処理の流れ

リアにおける頻出ブロックの情報を基にしている ACEb モジュールは，そのエリアに所属する可能性の高さを表す所属確度の指標として TF-IDF モデルによる特徴量を採用している．複数のエリアで共通して出現するブロックは，そのブロックがエリア固有の特徴を表すものではないと判断され，その重みが低くなる．それに対して，特定のエリアでのみ頻出するブロックは，エリア固有の特徴を表すものであると判断される．ブロックレベルの表層情報 (5.2 節) は，文字レベルの表層情報よりも粒度が大きく，より確度が高い情報を保持している．これにより，エリア固有のブロックの情報を活用して，より適合率の高い地名の所属エリア推定を実現することを目指す．

ACEb モジュールは，入力となる地名に対してその構成ブロックから得られる情報を基にエリアへの所属確度を算出する．各エリアに対して算出された所属確度を比較することで，高い所属確度を持つエリアを所属エリア候補として出力する．出力は相対的な評価によって決定され，所属確度の上位のエリアが 1 個以上選択される．

また，同様に確度が高い情報を保持する単語レベルの表層情報 (5.1 節) を用いた場合との比較も行う．各エリアにおける頻出単語の情報を基にしている ACEw モジュールは，ACEb モジュールと同様の方法で，エリアに所属する可能性の高さを表す所属確度を算出する．ACEw モジュールでは，入力となる地名に対してその構成単語から得られる情報を基にエリアの所属確度を算出し，相対的に高い所属確度を持つエリアが 1 個以上選択される．

5.3.2 単語レベル TF-IDF に基づく地名のエリア所属確度

地名 t がエリア a に所属する可能性の高さを表す所属確度の指標として，単語レベルの表層情報を用いた確度値 $E_{\mathrm{w}}(a,t)$ を定義する (式 (5.7))．

$$E_{\mathrm{w}}(a,t) = \frac{1}{n}\sum_{k=1}^{n} e_{\mathrm{w}}(a,w_k) \tag{5.7}$$

ここで w_k は地名 t の k 番目の構成単語，n は地名 t に含まれる単語の数，$e_{\mathrm{w}}(a, w_k)$ はエリア a に対する単語 w_k の特徴量を表す．単語の特徴量は，単語 w がエリア a の表層的特徴をどれだけ有しているかを示す指標であり，TF-IDF をベースに式 (5.8) で定義する．

$$e_{\mathrm{w}}(a, w) = TF_{\mathrm{w}}(a, w) \log \frac{|A|}{DF(w)} \tag{5.8}$$

ここで $TF_{\mathrm{w}}(a, w)$ はエリア a における単語 w の出現頻度，$DF(w)$ は単語 w を含む地名コーパスの数，$|A|$ は地名コーパスの総数を示す．確度値 $E_{\mathrm{w}}(a, t)$ はエリアに対する単語の特徴量 $e_{\mathrm{w}}(a, w)$ の平均を意味する．

5.3.3 ブロックレベル TF-IDF に基づく地名のエリア所属確度

確度値を適切に算出するためには，その地名の構成要素が適切な推定を行うために十分な情報を有している必要がある．しかし，5.1.2 節で述べたとおり，地名を構成する単語は十分な出現頻度を示すことは期待できない．つまり，5.3.2 節で定義した単語レベルの表層情報を用いた確度値 $E_{\mathrm{w}}(a, t)$ (式 (5.7)) では適切な所属エリア推定ができない可能性がある．

そこで，単語レベルの確度値 $E_{\mathrm{w}}(a, t)$ と同様に，ブロックの特徴量を用いて，地名 t がエリア a に所属する可能性の高さを表すブロックレベルの確度値 $E^*_{\mathrm{b}}(a, t)$ を定義する (式 (5.9))．

$$E^*_{\mathrm{b}}(a, t) = \frac{1}{n} \sum_{k=1}^{n} \sum_{m=m_{\min}}^{m_{\max}} \frac{e^*_{\mathrm{b}}(a, B^{*m}_P(w_k)) + e^*_{\mathrm{b}}(a, B^{*m}_S(w_k))}{2} \tag{5.9}$$

ここで n は地名 t の長さ (単語数)，$B^{*m}_P(w_k)$ は地名 t の k 番目の構成単語 w_k のブロックサイズ m の語頭ブロック ($m_{\min} \leq m \leq m_{\max}$)，$B^{*m}_S(w_k)$ は同じ条件でのブロックサイズ m の語尾ブロックを示す．ブロックレベルの特徴量 $e^*_{\mathrm{b}}(a, B^{*m}(w, i))$ は，サイズ m のブロック $B^{*m}(w, i)$ がエリア a の表層的特徴をどれだけ有しているかを示す指標であり，TF-IDF をベースに式 (5.10) で定義する．

$$e^*_{\mathrm{b}}(a, B^{*m}(w, i)) = TF_{\mathrm{b}}(a, B^{*m}(w, i)) \log \frac{|A|}{DF(B^{*m}(w, i))} \tag{5.10}$$

$TF_{\mathrm{b}}(a, B^{*m}(w, i))$ はエリア a におけるブロック $B^{*m}(w, i)$ の出現頻度，$DF(B^{*m}(w, i))$ はブロック $B^{*m}(w, i)$ を含む地名コーパスの数，$|A|$ は地名コーパスの総数を示す．

また，5.2.3 節 で定義したショートブロックの概念を取り入れたブロックレベルの確度値 $E_{\mathrm{b}}(a, t)$ を以下のように定義する (式 (5.11))．

$$E_{\mathrm{b}}(a, t) = \frac{1}{n} \sum_{k=1}^{n} \sum_{m=m_{\min}}^{m_{\max}} \frac{e_{\mathrm{b}}(a, B^{m}_P(w_k)) + e_{\mathrm{b}}(a, B^{m}_S(w_k))}{2} \tag{5.11}$$

ここで n は地名 t の長さ (単語数)，$B^{m}_P(w_k)$ は地名 t の k 番目の構成単語 w_k のブロックサイズ m の語頭ブロック ($m_{\min} \leq m \leq m_{\max}$)，$B^{m}_S(w_k)$ は同じ条件でのブロッ

クサイズ m の語尾ブロックを示す．ショートブロックの概念を取り入れたブロックレベルの特徴量 $e_{\mathrm{b}}(a, B^m(w,i))$ は，サイズ m のブロック $B^m(w,i)$ がエリア a の表層的特徴をどれだけ有しているかを示す指標であり，TF-IDF をベースに式 (5.12) で定義する．

$$e_{\mathrm{b}}(a, B^m(w,i)) \;=\; TF_{\mathrm{b}}(a, B^m(w,i)) \log \frac{|A|}{DF(B^m(w,i))} \tag{5.12}$$

ここで $TF_{\mathrm{b}}(a, B^m(w,i))$ はエリア a でのブロック $B^m(w,i)$ の出現頻度，$DF(B^m(w,i))$ はブロック $B^m(w,i)$ を含む地名コーパスの数，$|A|$ は地名コーパスの総数を示す．

5.3.4 ブロックレベルと単語レベルでの所属エリア推定結果の比較

ACE モジュールを用いた所属エリア推定には，単語レベルの表層情報を用いた ACEw モジュールと，ブロックレベルの表層情報を用いた ACEb モジュールの 2 種類ある (5.3.1 節)．図 5.4 に，ACEw モジュールを用いた手法と ACEb モジュールを用いた手法 (手法*2*) のそれぞれの所属エリア推定の実験の結果を示す．x 軸は適合率 (precision)，y 軸は再現率 (recall) を表している．ACEw モジュールは単語レベルの表層情報を用いた所属エリア推定のモジュールであり，式 (5.7) で定義される確度値を用いたものである．それに対して，ACEb モジュールはブロックレベルの表層情報を用いた所属エリア推定のモジュールであり，式 (5.11) で定義される確度値を用いたものである．図 5.4 の丸印のマーカはブロックレベルの結果を，四角のマーカは単語レベルの結果を示している．それぞれの手法についてすべてのエリアをまとめて評価した結果は大きなマーカで示している．小さな丸印および四角のマーカは，各エリアの結果を示す．手法*2*のブロックサイズは 5 のみとし，ショートブロックを含めて推定を行っている．利用対象ブロックには語頭ブロック，語尾ブロックの双方を含める．

ACEw モジュールの場合でも ACEb モジュールの場合でも，適合率は約 90% を達成しており，適合率については十分によい結果が得られた．適合率で見ると，ブロックレベルで処理を行った場合に比べて単語レベルでの結果がよいが，その差は大きくなく，ブロックレベルでも十分単語レベルに近い適合率を得ることができることがわかる．これに対して，再現率の結果は手法によって大きく異なる．単語レベルのアプローチでは，適合率 92.96%，再現率 52.77%，F 値 0.67 だったのに対して，ブロックレベルのアプローチでは，適合率 83.84%，再現率 90.51%，F 値 0.87 であった．各エリアの結果を見ると，ブロックレベルのアプローチでは再現率は最低でも 70% 程度だったのに対して，単語レベルのアプローチでは再現率が 10% に満たないエリアもあり，そのばらつきの大きさが憂慮される．この差は，5.3.3 節で述べたように，単語レベルの情報が持つデータスパースネスの問題と未知語の問題に起因する．図 5.2 で示したとおり，地名の構成単語の出現頻度の傾向はエリアによって大きく異なる．出現頻度が低い構成単語の割合が大きいエリアにおいては，単語レベルの情報を用いた処理では確度値の算出に必要な情報が十分に得ることができず，結果的に低い再現率に繋がっている．このようなエリアでも，構成単語がすべて互いに独立な文字列とは限らない．むしろこのようなエリアでは，地名特有の意味を持つ文字列等のエリア固有の文字列が，

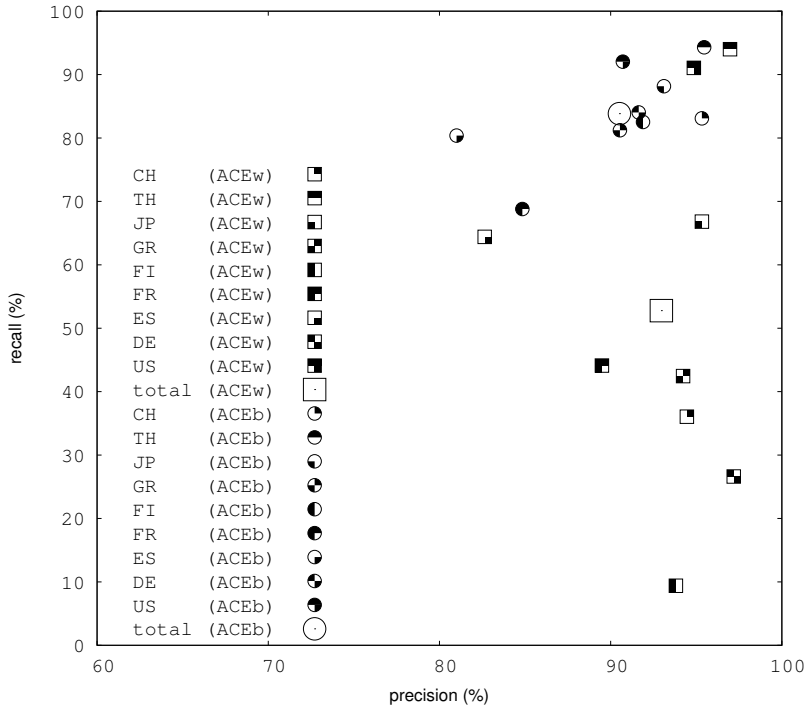


図 5.4: ブロックレベルでの処理結果と単語レベルでの処理結果の比較

単語の一部分として単語中に含まれているような表記が多い．このような場合，このような隠れた部分文字列を抽出，利用することによってブロックレベルの情報を活用し，データスパースネスの問題を解決することが可能である．図 5.4 の結果から，単語レベルの情報ではなく，ブロックレベルの情報を利用することで，再現率を低下させることなく適合率を改善することが可能であるといえる．

5.4 ブロックの概念を導入した 2 段階推定手法

手法*2*の再現率 83.84% は，第 4 章の手法*1*の再現率 94.96% (表 4.7) を下回る．これは，手法*1*が再現率のみに着目した手法であるのに対して，手法*2*が適合率の向上のみを目的としたアプローチであるためである．手法*1*の適合率が 62.64% に留まったのに対して，手法*2*は適合率 90.51% を得ており，手法*1*の高い再現率と手法*2*の高い適合率の双方を同時に実現可能な手法が期待される．そこで本節では，再現率と適合率の双方で十分な結果を得ることを目的として，再現率に着目した手法*1*と適合率に着目した手法*2*を組み合わせる手法 (手法*1+2*) を検討する．

本章で検討する手法*1+2*のシステムの概要を図 5.5 に示す．手法*1+2*は，所属エリア候補の削減を目的とする area candidate reduction (ACR) モジュールによる処理と，所属エリア候補の判定を目的とする area candidate examination (ACEb) モジュールによる処理の 2 フェーズで構成する．ACR モジュールでは手法*1*をベースとした処理を行う．まず，手法*1*で実現した高い再現率を活かし，実験対象エリア群の中から可能性の低い所属エリア候補を排除することで，所属エリア候補の絞込みを行う．次の ACEb

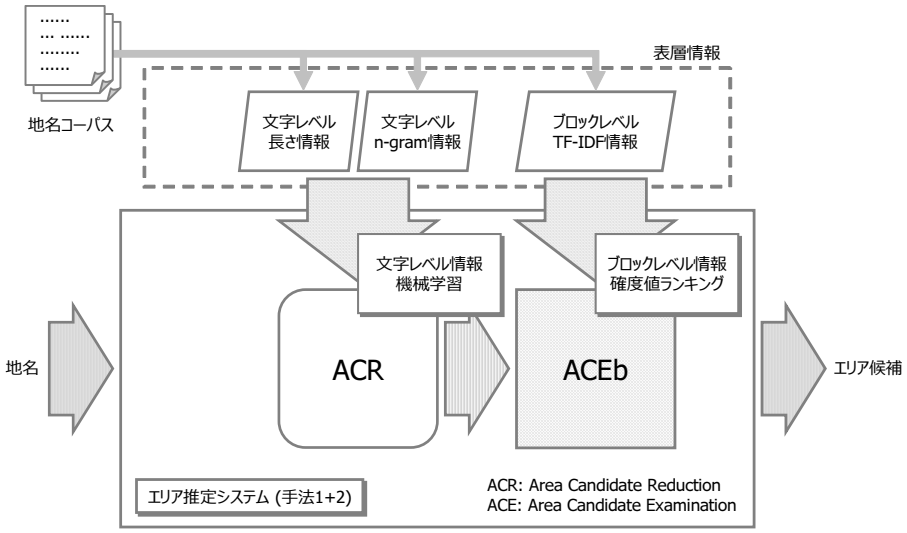


図 5.5: 手法*1+2*の処理の流れ

モジュールでは手法*2*をベースとした処理を行う．各フェーズで用いる表層情報を図 5.6 に示す．手法*2*の高い適合率を活かすため，ACR モジュールから出力された所属エリア候補に対してその確度値を算出し，その確度値の高い順にあらかじめ定めた出力エリア数だけ候補を出力する．この結果をシステム全体の推定結果として出力する．

第 1 フェーズ (ACR モジュール) : 文字情報と長さ情報
第 2 フェーズ (ACEb モジュール): ブロック情報

図 5.6: 手法*1+2*の各フェーズで用いる表層情報

Madrid という地名を入力とした場合を例にとって手法*1+2*の処理の流れを説明する (図 5.7)．この例では，実験対象エリアを 9 エリア，出力エリア数を 2 エリアとして所属エリア推定処理を行い，結果としてスペインとドイツが所属エリア候補として出力されたものとする．まず，ACR モジュールでは，文字レベルの n-gram 情報と長さ情

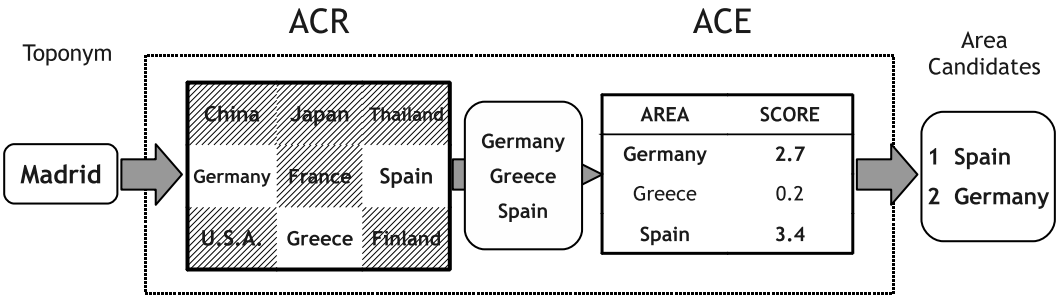


図 5.7: 手法*1+2*での所属エリアの絞込み

報を用いて，Madrid が所属する可能性の低い 6 エリアを排除する．この段階で所属エリア候補として残ったエリアの 3 エリア (ドイツ，ギリシャ，スペイン) が次の ACEb モジュールの処理対象となる．ACEb モジュールでは，絞込み後の 3 個の所属エリア候補に対して，ブロックレベルの出現頻度情報を基に Madrid の各エリアに対する確度値を算出する．この例では，出力エリア数が 2 エリアと設定されているため，最も確度値の高い所属エリア候補としてスペインを，第 2 位の確度値の値を持つ所属エリア候補としてドイツを最終的に出力する．また，それ以外の 7 エリアは最終的な所属エリア候補とならず，出力されない．

5.5 実験結果

5.5.1 実験環境

地名コーパス

本節では，表層情報の調査に用いた地名コーパス (表 5.1) を用いて所属エリア推定実験を行う．5.1.1 節で述べたとおり，本実験の対象エリアは中国，タイ，日本，ギリシャ，フィンランド，ドイツ，フランス，スペイン，アメリカの 9 エリアである．

ACR モジュールの設定 (第 1 フェーズ)

ここでは手法*1+2*の ACR モジュールとして，手法*1*と同じモジュールを用いる．本実験で機械学習に用いる素性を表 5.6 に示す．素性の種類および取り得る値は表 4.6 と同じだが，利用する記号の種類を 3.3.2 節 で示した 27 種類としたため，素性の個数は計 20,489 個である．

表 5.6: 所属エリア推定の実験に用いる素性

素性		素性の数
feature_{L1}	地名に含まれる文字の数	1
feature_{L2}	地名に含まれる単語の数	1
feature_{L3}	n 番目の単語に含まれる文字数	16
feature_{L4}	地名に含まれる m 文字の単語の数	32
feature_{N1}	文字レベルの unigram 情報	27
feature_{N2}	文字レベルの bigram 情報	729
feature_{N3}	文字レベルの trigram 情報	19,683

ACEb モジュールの設定 (第 2 フェーズ)

手法*1+2*の ACEb モジュールとして，手法*2*と同じモジュールを用いる．ここではブロックサイズを 5 のみとし，ショートブロックを含めて推定を行う．利用対象ブロックには語頭ブロック，語尾ブロックの双方を含める．

複数のサイズのブロックを併用した場合，同じ単語から長さの違う複数のブロックを切り出すことができる．ここでは語頭および語尾に限ってブロックを抽出するため，頻出単語の語頭および語尾のブロック情報が数多く切り出されることになり，語頭部分および語尾部分の影響が大きくなり過ぎる可能性がある．本実験では，ブロックサイズに範囲を与えず，ブロックサイズを 1 個に固定することで，複数のブロックサイズの併用による頻出単語の語頭および語尾の過剰な影響を排除した．

第 2 フェーズで出力される所属エリア候補数は 1 とする．すなわち，システムは最終的な所属エリア候補として，確度値の最も大きなエリア 1 個のみを出力する．

5.5.2 実験結果

表 5.7 に手法*1+2*による地名の所属エリア推定の実験結果を示す．全体では，適合率 85.54%，再現率 91.92%，F 値 0.89 の結果を得た．エリア別に見た場合，最も F 値

表 5.7: 手法*1+2*の実験結果 (ACR+ACEb)

エリア	F 値	適合率 (%)	再現率 (%)
中国	0.92	88.61	96.23
タイ	0.98	98.39	96.78
日本	0.90	86.74	94.09
ギリシャ	0.86	82.79	88.66
フィンランド	0.84	79.29	89.73
フランス	0.81	77.16	84.41
スペイン	0.85	82.86	87.86
ドイツ	0.90	83.89	96.68
アメリカ	0.92	92.12	92.85
全体	0.89	85.54	91.92

が低かったのがフランスで 0.81 であり，最も F 値が高かったのはタイで 0.98 であった．最も結果が悪いケースでも 0.81 の F 値を得ることができ，オープンテストの結果としては，十分によい結果といえる．

5.6 考察

5.6.1 ブロックの有効性

手法*1+2*の有効性を示すため，これと関連する手法 (図 5.8) との比較の形で考察する．図 5.9 は，それぞれの手法での所属エリア推定の実験結果を図示したものである．

ACR	: ACR モジュール (手法*1*) のみ
ACEw	: 単語レベルの ACEw モジュールのみ
ACEb	: ブロックレベルの ACEb モジュール (手法*2*) のみ
ACR+ACEw	: ACR モジュール (手法*1*) と 単語レベルの ACEw モジュールの 2 段階処理
ACR+ACEb	: ACR モジュール (手法*1*) と ブロックレベルの ACEb モジュール (手法*2*) の 2 段階処理 (手法*1+2*)

図 5.8: 手法*1+2*とその関連手法

x 軸は適合率 (precision)，y 軸は再現率 (recall) を表している．全体の結果は，大きなマーカで表示する．図 5.9 から，2 段階処理 (ACR+ACEw，ACR+ACEb) は，高い再現率に加え，実験結果のエリア間のばらつきが小さいことがわかる．これは，文字レベルの情報を用いた ACR モジュールの特徴が現れたものである．また，ACEw モジュールと ACEb モジュールは，高い適合率をもたらす能力を有していることが見て取れる．これは，確度の高い情報としてブロックレベルあるいは単語レベルの表層情報を利用していることに起因している．

ブロックレベルの情報を用いた処理 (ACEb，ACR+ACEb) は，単語レベルの情報を用いた処理 (ACEw，ACR+ACEw) よりも全体的によい結果を示している．これは，ブロックレベルの頻度情報が，単語レベルの情報を利用するアプローチの課題であるデータのスパースネスの問題を解消しているといえる．また，ACEw モジュールの実験結果では，再現率が低くなっているが，フィンランドやドイツ等の地名が，未知語となりやすい長い単語を含むことが多いことに起因している．一方で，ACR+ACEw は高い再現率を示している．これは，ACR モジュールによって所属エリア候補の数をある程度削減することに成功している (ACR モジュールでの処理後の所属エリア候補数は，フィンランドの地名では平均 1.6，ドイツの地名では平均 1.5 であった) ことが，その要因である．このことから，ACR モジュールが所属エリア推定において重要な役割を担っており，ACEw モジュールが適合率の改善をサポートしている処理であると考えることができる．このことから，本手法は，どのエリアの地名に対しても効果的に動作し，高い再現率を確保しながら高い適合率を実現したといえる．

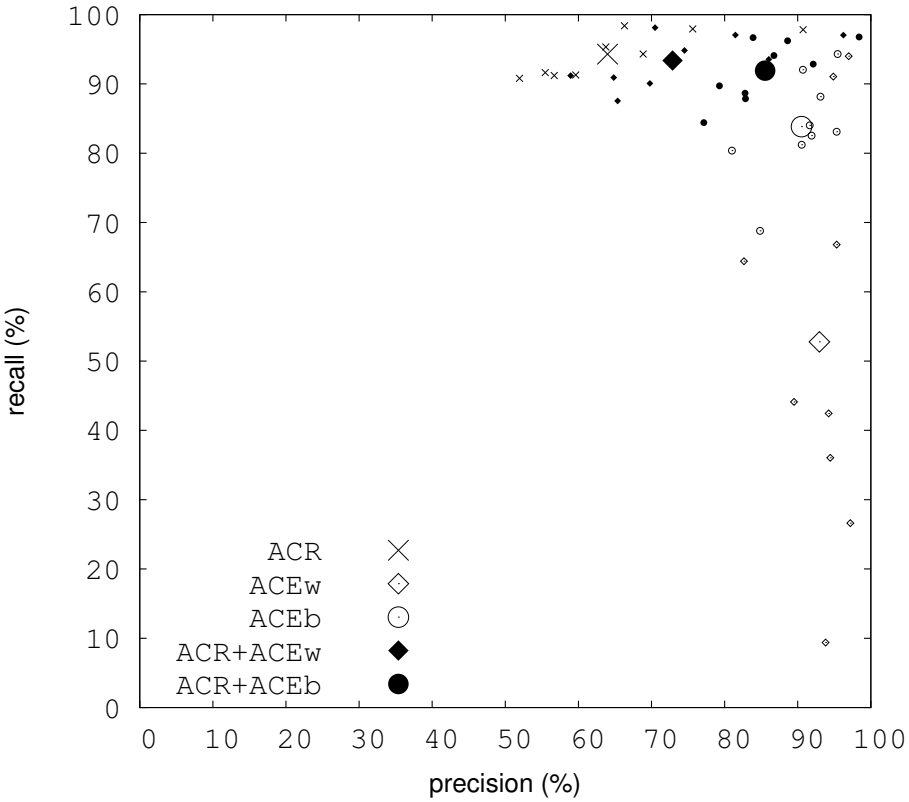


図 5.9: 各手法の実験結果の比較

5.6.2 出力する所属エリア候補数の妥当性

ACEb モジュールは，適合率と再現率を考慮し，確度値に基づいて，複数の所属エリア候補を出力することができるように設計されている．図 5.10 に，出力する所属エリア候補の数を変化させた場合の F 値の変化を示す．黒いマーカの線は ACR+ACEb の所属エリア推定結果を，白いマーカの線は ACEb の結果を示す．x 軸は出力エリア数 (maximum number of outputs: ACEb モジュールが出力する所属エリア候補の最大数) を表し，y 軸は各出力エリア数での F 値 (F-measure) を表している．

どちらの場合でも，出力する所属エリア候補数が 1 の場合に最もよい結果を示し，また，その時の F 値は大きな差はない．出力数を増加させるにつれて，その結果は大きく異なるが，これは，ACEb モジュールの場合には適合率の低下がその原因である．図 5.10 から，出力する所属エリア候補数を増やすことによる再現率の向上よりも，適合率の低下が与える影響が大きいことを意味している．そこで，本システムでは，適合率の観点から所属エリア推定の結果として 1 個の所属エリア候補のみを出力することとした．

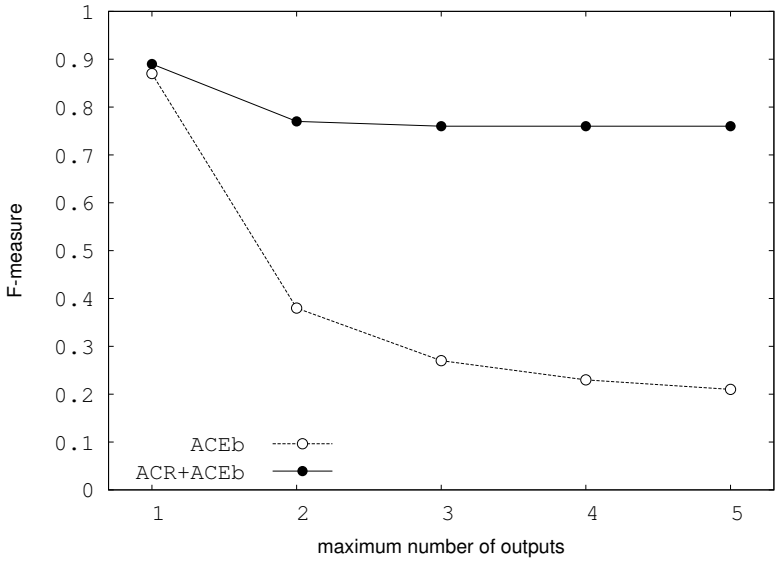


図 5.10: 出力エリア数の変化に伴う F 値の変化

5.6.3 ブロックの効果的な利用

5.3.3 節では，地名の所属エリア推定に有効な処理単位として，ブロックの概念を提案した．ブロックは，サイズを考慮して抽出可能な処理単位であり，ブロックサイズによってその影響に差が出ることが予想される．また，5.3.3 節の提案の中で，特に有効と考えられるブロックとして，語頭ブロックと語尾ブロックを提案した．同時に，短い単語の利用を考慮し，ショートブロックの概念を提案した．本節では，これらの所属エリア推定への影響を考察する．

図 5.11 に，ブロックサイズ，ブロックのタイプ (語頭ブロックと語尾ブロック)，ショートブロックの扱いによる F 値の変化を示す．x 軸はブロックサイズ (block size)，y 軸は ACE モジュールによる処理における F 値 (F-measure) を表す．このグラフの 6 本の線は，それぞれ図 5.12 のブロックの使用方法を示す．図 5.11 から，全体として，ブロックサイズは 4 から 5 の辺りで F 値の極大値を取ることがわかる．特に結果のよいショートブロックを含めるブロック定義ではブロックサイズ 5 で最もよい F 値を得ることから，ブロックサイズは 5 程度とすることが妥当と判断する．

ショートブロックを含めるブロック定義に従う手法では，ショートブロックを用いない場合に比べて，F 値が高いという結果になった．このことから，ショートブロックの利用が所属エリア推定に効果を持つことがわかる．図 5.11 から，ショートブロックを使わない場合にはブロックサイズが大きくなるにつれて F 値が大きく下がるのに対して，ショートブロックを含めたブロック定義を用いる場合にはブロックサイズが大きくなった場合の F 値の低下が十分抑えられていることから，ショートブロックの利用によってブロックサイズに満たない単語の影響を加味することでブロックの概念が有効に活用できていることがわかる．ここまでの検討から，ブロックの導入によるノイズを最大限抑えると同時にその効果を最大限引き出すには，ブロックサイズに幅を持たせず 5 に固定し，ショートブロックを用いる方法が適切であると考えられる．

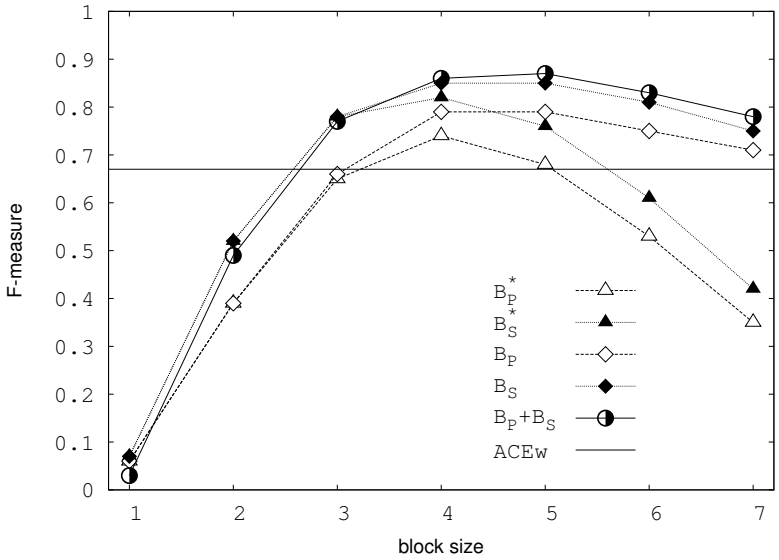


図 5.11: ブロックのサイズおよび種類と F 値の変化

B_P^*	語頭ブロックのみを使用 (ショートブロックを含まない)
B_S^*	語尾ブロックのみを使用 (ショートブロックを含まない)
B_P	語頭ブロックのみを使用 (ショートブロックを含む)
B_S	語尾ブロックのみを使用 (ショートブロックを含む)
$B_P + B_S$	語頭ブロックと語尾ブロックの両方を使用 (ショートブロックを含む)
ACEw	単語レベルの所属エリア推定 ACEw の結果 (F 値 0.67)

図 5.12: ブロックの使用方法の種類

地名の構成単語中に含まれる意味のあるブロックの長さは 5 と限らないため，ブロックサイズ 5，ショートブロック込みの設定でのブロックの影響を考察する．地名の構成単語中，長さ 5 の意味のある接頭辞的文字列および接尾辞的文字列 (Nishifunabashi (**西船橋**) の NISHI 等) は，サイズ 5 のブロックで適切に抽出できるため問題はない．それに対して，5 を超える長さの意味のある文字列 (Minamisenju (**南千住**) の MINAMI 等) は，サイズ 5 のブロックでは途切れてしまう (MINAM)．しかし，n-gram は n の増加に伴い各 n-gram の出現頻度が小さくなるため，サイズ 5 以上の n-gram は少なく，途切れたブロックであっても元の文字列の代用となり得る．例えば 5 文字の語頭ブロック MINAM は日本コーパスの 10,000 個の地名の中で 252 回出現したが，この頻度は 6 文字の語頭ブロック MINAMI の出現頻度 247 とほぼ変わらない．したがって，語頭ブロック MINAM の後に続く文字はほぼ I と断定できることになり，5 文字ブロック MINAM と 6 文字ブロック MINAMI はほぼ同等の情報量を持つ．これは 7 以上の文字列でも同じであり，7 文字の語頭ブロック HIGASHI の出現頻度は 358，6 文字の

語頭ブロック HIGASH の出現頻度は 359，5 文字の語頭ブロック HIGAS の出現頻度は 366 と，7 文字以上の長さを持つ文字列も 5 文字ブロックでその存在を十分代用でき，推定処理に反映することができる．

ブロックサイズが異なれば抽出されるブロックの異なり数が異なるため，同じ文字列を指すブロックでも出現割合はブロックサイズによって変化する．そのため，ブロックサイズが大き過ぎる場合には大きなブロックの出現割合が過剰に反映されてノイズの原因となることが，図 5.11 でのカーブの下降に繋がるものと考えられる．また逆に，意味のある文字列のうちブロックサイズより小さいもの (Kitayamagata (北山形) の KITA 等) は，ブロックサイズに満たないためそのままの形では抽出されず，KITAY のように余分な文字を付けた形でブロックとされる．4 文字ブロック KITA は十分な出現頻度が見込めるため影響力を持つことが期待されるが，5 文字ブロック KITAY は出現頻度が低く影響力は小さい．この場合には KITAY の他に KITAS (KITASENJU (**北千住**) の語頭ブロック) や KITAK (KITAKOGANEI (**北小金井**) の語頭ブロック) 等，冒頭に KITA を含む複数の 5 文字ブロックが推定処理に利用されることになり，KITA の後に出現しやすい文字が含まれるか否かの情報も含めてブロックの持つ情報として影響することになる．このように，ブロックサイズより短い文字列も，ブロックを用いることで十分影響を持ち得る．

単語の長さ自体がブロックサイズよりも短い場合 (Kita (**北**) が 1 語として出現するような場合) には，通常のブロックの抽出ではこの単語の情報は利用することができない．したがって，ブロックサイズが大きくなると，単語長がブロックサイズに満たず推定に反映されない単語の割合が大きくなり，長い単語の影響力が過剰に大きくなる．ショートブロックを利用しない場合に図 5.11 でブロックサイズが大きくなるにつれて曲線が大きく下降するのはこのためである．この問題は，ショートブロックを導入することで解決できることが図 5.11 から明らかである．このように，ブロックサイズを適切に設定できれば，ブロックサイズに幅を持たせる代わりにショートブロックを導入することで，さまざまな長さの地名の構成単語に固有の文字列を同時に評価することが可能である．ブロックサイズに幅を持たせることは，同じ文字列がブロックサイズ毎に抽出されて影響力を持つことになり，過剰な影響力を持つ可能性が出てくる．このことから，地名の所属エリア候補推定では，ブロックサイズを 1 個に固定しショートブロックを導入する手法が適切と判断できる．また，語頭ブロックと語尾ブロックについては，ショートブロックの利用の有無に関わらず，全体に語尾ブロックの方が語頭ブロックよりも高い F 値を引き出すことがわかった．さらに，ほとんどのブロックサイズにおいて，語頭ブロックと語尾ブロックの両方を同時に利用することで，語尾ブロックのみの場合に比べて F 値を向上させる可能性があることも示された．

5.6.4 所属確度推定の妥当性

ACEb モジュールによる所属確度推定の妥当性を検討するため，出力エリア数を上限 (実験対象エリア数) に設定し，すべての地名の順位付けの結果を示す (表 5.8)．表 5.8 の ‘1’ のカラムは，正しい推定結果が最も大きな確度値を得ている割合を示している．

‘unknown’ のカラムは第 2 フェーズの ACEb モジュールでブロックがコーパス中に存在しないために確度値を算出することができなかった地名の割合を示しており，全体で 7.1% が ‘unknown’ に該当した．‘negative’ のカラムは第 1 フェーズの ACR モジュールで正しいエリアが所属エリア候補から除外されてしまった割合を示しており，全体で 5.7% が ‘negative’ に該当した．‘negative’ に該当する場合には，正解エリアが第 2 フェーズの ACEb モジュールの入力に含まれないため，正解エリアに対する ACEb モジュールによる確度値の算出は行われていない．表 5.8 から，全体で 84.8% の地名に

表 5.8: 手法*1+2*の実験結果 (確度値順位別集計)

エリア	1	2	3	4 – 9	unknown	negative
中国	86.2	1.7	0.0	0.0	10.0	2.1
タイ	96.1	1.0	0.0	0.0	0.6	2.2
日本	88.7	1.2	0.0	0.0	5.4	4.7
ギリシャ	79.4	2.4	0.1	0.0	9.3	8.8
フィンランド	80.2	1.8	0.1	0.0	9.6	8.4
フランス	73.8	6.3	0.1	0.0	10.6	9.2
スペイン	80.4	3.3	0.1	0.0	7.5	8.7
ドイツ	86.2	1.6	0.1	0.0	10.5	1.6
アメリカ	92.0	1.4	0.1	0.0	0.9	5.7
全体	84.8	2.3	0.1	0.0	7.1	5.7

ついて正解エリアを最も高い所属確度を持つエリアと推定していることがわかる．1 位に正解エリアが選ばれた割合はエリア毎に異なるが，どのエリアでも 1 位に正解エリアが入る割合は十分に高い．確度値の算出の結果正解エリアが 1 位とならなかった場合でも，ほぼすべての場合で正解エリアは 2 位に入っており，出力エリア数を 2 位までとすれば再現率は十分高くなる．

ACEb モジュールでは，確度値の算出に必要な情報をブロックの出現状況に依存している．そのため，入力地名中に情報を持つブロックが含まれていない場合 (‘unknown’)，正しい確度値の算出ができない．単語レベルでの処理の場合と比べて状況は大幅に改善しているものの，ブロックレベルでも unknown 扱いとなる割合はエリアによっては 10% になり，改良の余地が残る．この点の改良が今後の課題である．

5.6.5 同一言語圏のエリアを含む所属エリア推定

本実験では，エリアとその言語の間に 1 対 1 の対応があることを前提とした (5.1.1 節)．しかし，実際には，スペイン語圏としてスペインやチリ，中国語圏として中国や台湾のように，同じ言語圏に複数のエリアが存在する．これらのエリアには，言語的，地理的あるいは歴史的要因によって，同じ地名もしくは類似した地名が数多く含まれる可能性があり，これらの類似エリアを考慮する必要がある．

5.5.2 節の 9 エリアにスペインと中国の類似エリアとしてチリと台湾を追加した 11 エリアでの実験の結果を表 5.9 に示す．表 5.9 のカッコ内の数字は，9 エリアでの実験 (表 5.7) に対する割合を示す．カッコ内が 100% であれば，同一言語圏のエリアを含

表 5.9: 類似エリアを加えた手法*1+2*の実験結果

エリア	F 値	適合率 (%)	再現率 (%)
中国	0.78 (84.8%)	74.53 (84.1%)	82.50 (85.7%)
台湾	0.83 (-)	82.29 (-)	83.20 (-)
タイ	0.97 (99.0%)	97.36 (99.0%)	96.56 (99.8%)
日本	0.90 (100.0%)	86.99 (100.3%)	93.57 (99.5%)
ギリシャ	0.85 (98.8%)	82.80 (100.0%)	88.30 (99.6%)
フィンランド	0.84 (100.0%)	79.27 (100.0%)	89.34 (99.6%)
フランス	0.80 (98.8%)	76.29 (98.9%)	83.36 (98.8%)
スペイン	0.75 (88.2%)	71.73 (86.6%)	78.12 (88.9%)
チリ	0.80 (-)	81.39 (-)	78.81 (-)
ドイツ	0.90 (100.0%)	83.82 (99.9%)	96.48 (99.8%)
アメリカ	0.92 (100.0%)	91.62 (99.5%)	92.43 (99.6%)
全体	0.85 (95.5%)	82.36 (96.3%)	87.52 (95.2%)

まない 9 エリアでの実験結果と同じ結果を得たことを示す．表 5.9 では，チリと台湾の追加によって，スペインと中国の地名の所属エリア推定の精度が元の 9 エリアでの実験結果の 85% 程度まで低下している．この結果から，手法*1+2*だけでは類似エリアの識別は難しく，類似エリアの識別にはさらに手法の検討の必要があることがわかる．ただし，スペインや中国以外の地名に対する所属エリア推定の結果については，ほとんど精度の低下が見られない．これは手法*1+2*の頑健性を示すものである．

5.7 手法*2*および手法*1+2*のまとめ

手法*1*(4 節) では，地名の持つ表層情報を文字レベルで取得し，これを用いて所属エリア推定を行った．その結果，表層情報は所属エリア推定に有効な特徴を有していること，適合率を向上させるには文字よりも大きな粒度の処理単位の導入が必要であることがわかった．本章では，まず手法*1*の結果および考察に基づき地名の所属エリア推定に適した処理単位を検討し，ブロックの概念を提案した．この新しい処理単位を導入した手法*2*では，高い適合率を得ることに成功した．さらに，手法*1*と手法*2*を組み合わせることで，再現率を保ったまま適合率を向上させることに成功した (手法*1+2*)．

手法*1+2*では，まず文字レベルの特徴の利用により，可能性の低いエリアを排除することで所属エリア候補を絞り込む．次に，この絞り込まれたエリアに対してブロックレベルの表層情報を用いた所属エリア推定処理を適用する．高い再現率で絞り込ま

れたエリアに対して，高い適合率を実現するブロックレベルの処理を適用することで，F 値 0.89 (適合率 85.54%，再現率 91.92%) を達成した．しかし，所属エリア推定の対象エリアとして，同一言語を使用するエリアを 2 個加えた実験では，F 値 0.85，適合率 82.36%，再現率 87.52% まで低下することがわかっている．手法*1+2*は表層情報の特徴が異なるエリアに関しては十分に頑健だが，類似エリアに対してはそれらを識別する能力が低く，この部分への対応が必要である．

第6章　類似エリアからの所属エリア候補の絞込み (手法*3*)

第 4 章，第 5 章の実験結果から，地名の所属エリア推定では，言語的，地理的，歴史的に関係があり，類似した地名を共有するエリア間の識別が問題となることがわかった．地名の所属エリア候補を 1 個に絞ることは必ずしも可能ではなく，また過剰な絞込みを行うことで正解エリアを候補から外してしまうリスクがある．類似した地名を有するエリアのグループが存在することは明白であり，このグループ内のすべてのエリアを所属エリア候補とすることで，高い再現率を確保することは可能である．しかし，適合率を考えた場合には，グループ内のすべてエリアを候補として出力することは望ましいとはいえない．そこで本章では，可能な限り適切に所属エリア候補の絞込みを行うため，このような複数の候補を取り得る状況の調査を行い，類似したエリア間の識別を図る．類似したエリア間の識別を実現するために，大域的な所属エリア推定と局所的な所属エリア推定の 2 段階の処理を行うことで，関連性の高いエリアが含まれるような場合においても高い精度で所属エリア推定が可能であることを示す (手法*3*).

6.1　類似エリア問題

地名の所属エリア推定では，(1) 同一の地名が複数のエリアに存在する場合，(2) 地名文字列に類似した特徴を持つエリアが複数ある場合に，地名の所属エリア候補を一意に絞ることができないことがある．前者は地名の曖昧性，後者はエリアの類似性の問題と言い換えることができる．どちらの場合でも，複数のエリアに対して高い所属可能性を示すような地名が存在するため，これらの地名に対しては該当する複数のエリアを候補として出力するのが妥当であり，複数のエリアを候補として出力することはシステムの精度を落とすものではない．それに対して，明らかに所属エリアである可能性が低いエリアは出力に含まれるべきではない．単に複数の候補を出力するだけでは，高い再現率は得られても適合率は下がり，結果的にシステムの信頼性は落ちる．

地名は，その所属エリアの言語的，地理的，歴史的影響を受ける．そのため，同一の言語圏に所属するエリア同士や，地理的に近いエリア同士，なんらかの歴史的関係を持つエリア同士等では，同一あるいは類似した地名や地名構成単語を共有する可能性がある．これらのエリア同士では，地名の類似性のため，エリアの持つ地名の表層情報の類似度が高くなり，表層情報に基づく所属エリア推定が困難となる場合がある．

地名は固有表現のため，基本的には同一の地名が複数箇所に点在することは望ましくない．しかし，実際には同一あるいは類似した地名が複数箇所に存在することは珍

しくなく，これが複数のエリアに跨ることもあり得る．しかし，同一言語圏，あるいは地理的に近いエリア同士等の間の識別は，所属エリア候補の数をできる限り抑えるためには必要な処理である．本章では，こういった類似エリア間の識別を目的として，手法の検討を行う．

6.1.1 地名コーパス

本章では，中国，台湾，タイ，日本，ギリシャ，フィンランド，フランス，スペイン，ドイツ，アメリカの 10 エリアを実験対象エリアとする．本節で用いる地名コーパスを表 6.1 に示す．この地名コーパスは，第 5 章で用いた地名コーパス (表 5.1) と同じものである (チリを除く)．

表 6.1: 本章で用いる地名コーパス

エリア	地名数	単語数	文字数
中国	10,000	15,975	98,773
台湾	10,000	22,518	122,630
タイ	10,000	28,974	131,128
日本	10,000	16,960	102,311
ギリシャ	10,000	14,612	109,818
フィンランド	10,000	11,050	97,447
フランス	10,000	19,820	121,945
スペイン	10,000	22,797	135,947
ドイツ	10,000	13,206	109,615
アメリカ	10,000	25,622	165,509

6.1.2 エリア間の地名の重複

表 6.1 のコーパス全体に含まれる計 100,000 個の地名は，それぞれのエリアでは重複はないように選ばれているが，複数のエリアで出現する地名は存在し，全体での異なり数は 99,794 である (表 6.2)．地名の重複は，全体としては 0.2% 程度だが，言語的に近いと考えられる，中国と台湾，フランスとスペインの間で他の組合せよりも大きな数値を示す．特に，中国と台湾は 10,000 の地名のうち 97 個と，1% 近くが重複していることになる．

これらの重複した地名は，それ自身の所属エリア推定が困難なだけでなく，これらの表層情報が両方のエリアの特徴として反映されるため，他の地名の所属エリア推定の誤判定を引き起こす可能性がある．

表 6.2: エリア間の地名の重なり数

エリア	CH	TW	TH	JP	GR	FI	FR	ES	DE	US
中国	–	97	0	5	2	4	3	2	1	1
台湾		–	0	1	0	0	0	0	0	1
タイ			–	0	2	0	0	0	0	1
日本				–	8	4	2	3	0	2
ギリシャ					–	3	3	4	1	4
フィンランド						–	2	8	5	3
フランス							–	16	8	6
スペイン								–	0	4
ドイツ									–	1
アメリカ										–

6.1.3 n-gram モデルのパープレキシティ

各エリアの n-gram モデルのそれぞれの地名コーパスに対してのパープレキシティを表 6.3 に示す．パープレキシティは，情報理論に基づいた n-gram モデルの評価尺度であり，地名を構成する文字列のエントロピから計算される [24]．本研究で用いてい

表 6.3: 地名コーパスに対するパープレキシティ

エリア	CH	TW	TH	JP	GR	FI	FR	ES	DE	US
中国	8.95	9.96	25.48	27.13	62.12	84.63	75.10	61.78	89.02	84.21
台湾	14.71	6.83	39.65	42.81	90.83	166.26	76.20	63.26	114.52	63.89
タイ	60.73	55.23	6.44	37.17	60.86	101.99	96.27	67.86	117.10	73.44
日本	80.26	79.49	54.02	7.72	74.81	102.27	98.40	65.81	134.83	113.93
ギリシャ	78.23	79.10	40.35	33.20	9.72	43.25	34.45	25.80	64.10	45.88
フィンランド	66.64	66.10	36.78	32.34	26.24	10.33	44.33	38.50	39.35	48.94
フランス	62.66	68.02	42.29	68.08	41.22	65.53	10.23	15.75	31.78	30.02
スペイン	85.01	94.66	80.65	65.59	54.05	100.54	28.79	8.91	89.28	64.20
ドイツ	80.54	91.41	44.64	64.97	48.08	45.00	33.27	41.17	9.73	31.88
アメリカ	59.15	62.26	30.27	38.06	34.07	46.05	22.10	21.12	28.56	8.89

る地名の n-gram モデルにおいては，パープレキシティは文字の平均分岐数を表している．この評価尺度は，n-gram モデルの性能を評価する一つの指標であり，パープレキシティの値が小さいほどよいモデルであると考えられる．各エリアの n-gram モデルは，対応するエリアのコーパスに対して高い性能を示しており，n-gram モデルが所属エリアを推定するための情報を有していることがわかる．このパープレキシティの値は，対応するエリアに対してだけでなく，中国と台湾やフランスとスペイン等の間

でも比較的小さな値となった．これは，これらのエリアを区別することが困難であることを示し，n-gram 情報のみで所属エリアを判断することは，所属エリア推定の精度に限界があることを示唆している．

6.1.4 類似エリアのグループ

第 4 章で述べた地名の長さ情報の特徴 (4.1.2 節) や地名の n-gram 情報の特徴 (4.1.3 節)，表 6.3 に挙げた n-gram モデルのパープレキシティ等を見ると，類似エリアは 2 エリアのペアとは限らず，長さ情報の特徴が似ている類似エリア群，n-gram 情報が似ている類似エリア群といったように，類似した特徴を持つエリア同士が複数のエリアを含む緩いグループを成している可能性が高いと考えられる．また，同じグループにまとめられるエリア間でも，エリアの間の類似性の高さや類似する特徴の傾向は差があるものと予測できる．例えば，4.1.3 節では中国と台湾，タイの 3 エリアの間で n-gram の類似性が見られることが示されているが，4.1.2 節ではこの 3 エリアの間の長さ情報の特徴はむしろ距離があることが示されており，長さ情報の類似性の面から台湾と最も近いと考えられるエリアはフランスである．表 6.3 を見ると，パープレキシティの値も，例えばタイから見た中国や台湾のパープレキシティと，中国や台湾から見たタイのパープレキシティの値には開きがあり，類似性の強さには方向性があると考えられる．

このことから，同じグループに属するエリア同士であっても，直接エリア同士を比較することができれば，互いを識別できる可能性がある．ここでのグループとは，例えばアジアグループとヨーロッパグループ等，実験対象エリアをあらかじめ固定したグループに分けるものではない．実験対象エリア全体を対象とした所属エリア推定処理では，互いに類似した複数のエリアが所属エリア候補として残る可能性があるが，これらの所属エリア候補群の組合せが入力地名毎に完全に異なることは考えにくい．むしろ，所属エリア候補群として同時に残りやすい組合せにはある程度の傾向があると考えるのが自然であり，これを広い意味でのグループとすれば，実験対象エリア全体を対象とした所属エリア候補の推定は，グループの推定の意味を持つものと考えることができる．本章では，このようなグループの存在を仮定し，実験対象エリア全体を対象とした所属エリア推定実験を通してグループの存在や特徴を検討する．その上で，グループの存在を意識した 2 段階処理手法を提案し，類似エリア問題の解決を図る．

6.2 候補絞込みと候補削減からなる 2 段階推定手法

本章では，類似エリアのグループの存在を仮定し，入力地名に対する所属エリア推定処理を所属グループの判定とグループ内での所属エリアの識別の 2 段階に分けて行うことで類似エリア間の識別の問題の解決を図る手法 (手法*3*) を検討する．

本章で用いる手法*3*のシステムの概要を図 6.1 に示す．手法*3*は，所属エリア候補の絞込みを行う area candidate selection (ACSc) モジュールによる処理と，所属エリア候補の削減を目的とする area candidate reduction (ACR) モジュールによる処理の，

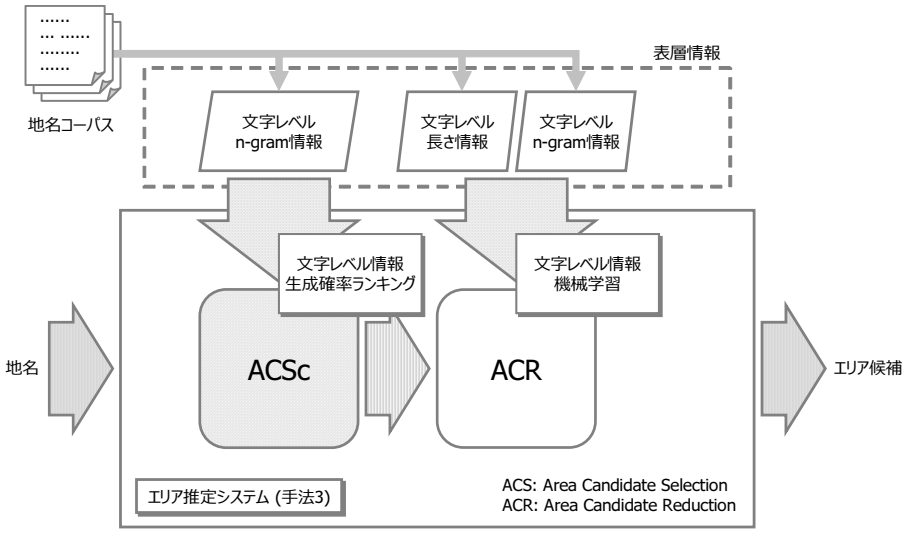


図 6.1: 手法*3*の処理の流れ

2 フェーズで構成する．ACSc モジュールでは，入力地名について，所属エリア候補の絞込みを行い実験対象エリア全体からあらかじめ設定した候補数まで所属エリア候補を絞り込む．本手法では，絞込み後のエリア候補群が意味のあるグループを成すものと仮定し，エリアの組合せの出現傾向を調べることでグループの存在や特徴を調査する．ACSc モジュールは，文字レベルの n-gram 情報をベースとした生成確率を用いて絞込み処理を行う．次の ACR モジュールでは，絞込み後の所属エリア候補に対して所属の可能性を推定し，可能性があると推定されたものを所属エリア推定の結果として出力する．ACR モジュールでは，手法*1*と同様に，文字レベルの n-gram と長さ情報を用いて推定を行う．本システムの出力は，入力地名文字列が所属する可能性のあるエリアであり，推定結果に応じて 0 個 (第 2 フェーズで候補のすべてが排除された場合) から第 1 フェーズでの絞込み個数まで (第 2 フェーズで候補のすべてが排除されなかった場合) となる．

6.2.1 生成確率を用いた所属エリア候補の絞込み

手法*3*の 2 段階処理の最初のフェーズは，所属エリア候補の絞込みを行う ACSc モジュールによる処理である．ここでは，地名の表層情報として文字レベルの n-gram のみを用いて所属エリア候補の絞込みを行う．地名を構成する文字の並びに対して，各エリアについてその文字列が生成される確率を算出する．これにより，この生成確率の値が高くなるような統計情報を持つエリアを所属エリア候補として抽出する．所属エリアの絞込み個数はあらかじめ与えるものとする．n 個の文字で構成される地名 $t_1^n = c_1 \cdots c_n$ のエリア a における生成確率 $P_a(t_1^n)$ は n-gram 情報を用いて次のように定義される (式 (6.1))．

$$P_a(t_1^n) = P_a(c_1)P_a(c_2|c_1)P_a(c_3|c_1c_2)\cdots P_a(c_n|c_{n-2}c_{n-1}) \tag{6.1}$$

ここで，$P_a(c_n|c_{n-2}c_{n-1})$ はエリア a において $c_{n-2}c_{n-1}$ の後に c_n が出現する頻度を表している．文字レベルの n-gram 情報は，Modified Kneser-Ney ディスカウンティング [8, 25] を用いて，低次の n-gram 情報を補間しながら平滑化を行っている．

6.2.2 機械学習器を用いた所属エリア候補の削減

手法*3*の第 2 フェーズである ACR モジュールでの処理では，絞込み処理後の所属エリア候補を対象に，文字レベルの n-gram 情報と長さ情報を用いた機械学習器による所属エリア候補の削減処理を行う．機械学習器はエリア毎に作成し，入力地名が対象エリアに属する可能性の有無を二値で判定する．この結果に応じて，所属エリア候補の絞込みを行う．機械学習器は，手法*1*の ACR モジュールと同じものを用いる．学習に用いる素性は，地名コーパスから得られる表層情報であり，手法*1*と同様に，地名の長さ情報と，文字レベルの n-gram 情報の 2 種類を利用する．手法*1*では，機械学習器のみを用いて所属エリア推定を行うため，実験対象エリアすべてを対象とした機械学習を行った．実験対象エリアの数を $|A|$ とすると，機械学習器の数は $|A|$ であり，それぞれの機械学習器が対象エリア a に属する地名を正事例，対象エリア以外のエリア $A-a$ に属する地名を負事例として機械学習を行う．この場合，機械学習器は，他のエリアすべてに対する対象エリアの特徴を学習することになり，類似エリアとの間の小さな差異が十分に反映されない恐れがあった．

これに対して，手法*3*の第 2 フェーズの ACR モジュールでの処理では，第 1 フェーズで絞り込まれた所属エリア候補のみを対象にした機械学習を行う．実験対象エリアの数を $|A|$ とし，第 1 フェーズの ACSc モジュールでの絞込み個数を p とする場合，第 2 フェーズでは，ACR モジュールとしてこの p 個のエリアを対象としてエリア毎に所属の可能性を判定する機械学習器を利用する[1]．対象となる $|A|$ 個のエリアの中から異なる p 個のエリアを選び出し (${}_{|A|}C_p$ とおり)，選ばれた p 個のエリアに対して所属エリアであるかどうかをそれぞれの機械学習器によって判定する．本フェーズでは，選ばれた p 個の組合せ毎に学習対象が異なるため，それぞれの組合せ毎に p 種類の機械学習器を用意する．第 2 フェーズでは，第 1 フェーズの絞込みによって所属エリア候補とされたエリア群のみを対象とした機械学習器による判定を行い，可能性を棄却された候補を所属エリア候補から外した上で所属エリア候補の削減を行う．

[1]例えばある入力地名に対して第 1 フェーズで所属エリア候補を 3 個に絞った結果，中国，台湾，タイの 3 エリアが第 2 フェーズでの推定の対象となったとする．この場合，第 2 フェーズでは，中国，台湾，タイの 3 エリア組を対象としてこの地名が中国に所属する可能性を判定する機械学習器，同じ 3 エリア組を対象として台湾に所属する可能性を判定する機械学習器，同じ 3 エリア組を対象としてタイに所属する可能性を判定する機械学習器の 3 個の機械学習器を利用して，この 3 エリアそれぞれについてその所属の可能性の有無の推定を行うことになる．第 1 フェーズで排除されたエリアは，第 2 フェーズでは考慮せず，機械学習器の学習対象に含めない．

6.3 実験結果

6.3.1 実験環境

地名コーパス

本節では，表 6.1 に示した地名コーパスを用いて所属エリア推定実験を行う．6.1.1 節で述べたとおり，本実験の対象エリアは中国，台湾，タイ，日本，ギリシャ，フィンランド，フランス，スペイン，ドイツ，アメリカの 10 エリアである．

ACSc モジュールの設定 (第 1 フェーズ)

ACSc モジュールの所属エリア候補の絞込み数は 3 個とする．

ACR モジュールの設定 (第 2 フェーズ)

本章の実験では，実験対象エリア数を 10 とする．第 1 フェーズの ACSc モジュールでの絞込み数を 3 個としたため，ACR モジュールでは，全体で 360 種類 (${}_{|A|}C_p \times p = {}_{10}C_3 \times 3$) の機械学習器を用いる．素性の種類および取り得る値は表 5.6 と同じである．

6.3.2 実験結果

表 6.4 に手法*3*による地名の所属エリア推定の実験結果を示す．比較として，手法*3*の第 1 フェーズである 6.2.1 節の生成確率のみを用いた手法 (ACSc) での結果と，手法*3*の第 2 フェーズに相当する機械学習器のみを用いた手法 (ACR) での結果を示す．生成

表 6.4: 各手法を用いた所属エリア推定の結果

手法	F 値	適合率 (%)	再現率 (%)
ACSc のみ	0.49	32.49	97.46
ACR のみ	0.74	61.58	94.05
ACSc+ACR	0.83	74.08	95.03

確率 (ACSc) と機械学習器 (ACR) を組み合わせた手法*3*(ACSc+ACR) の F 値は 0.83 となった (表 6.4)．この結果は，ACSc モジュールのみの結果と ACR モジュールのみの結果のどちらよりも高い値であった．ACR モジュールのみの場合と手法*3*の場合では再現率にはほとんど違いがないが，適合率は ACR モジュールの場合と比べて向上しており，これが F 値の改善に繋がった．

手法*3*では，まず第 1 フェーズとして ACSc モジュールによる絞込みを行い，その結果を第 2 フェーズの ACR モジュールで評価して所属エリア候補を決定するという

流れである．ACSc モジュールのみを用いた手法では F 値は 0.49 と低いが，その主因は適合率の低さにある．ACSc モジュールによる処理は再現率が高く，このことから，手法*3*で ACSc モジュールを処理の第 1 段階の絞込みフェーズとして用いたことは妥当であるといえる．ACR モジュールのみを用いた手法では，ACSc モジュールのみの場合と比べ，再現率は若干劣る．ACR モジュールのみの場合には機械学習器の学習対象は実験対象エリア全体となるが，手法*3*では十分に所属エリア候補を絞り込んだ状態で ACR モジュールを利用しており，理想的には ACR モジュールは類似エリア間での識別のみを担当することとなる．手法*3*の再現率は ACR モジュールのみの場合とほぼ同等であり，絞込みフェーズで再現率が 97% 程度に落ちることを考え合わせると，手法*3*の ACR モジュールによる処理での正解の脱落は十分少ないことがわかる．図 6.2 に所属エリア推定の実験結果を示す．x 軸は適合率 (precision)，y 軸は再現率 (recall) を表している．図中の丸印のマーカは，ACSc モジュールと ACR モジュール

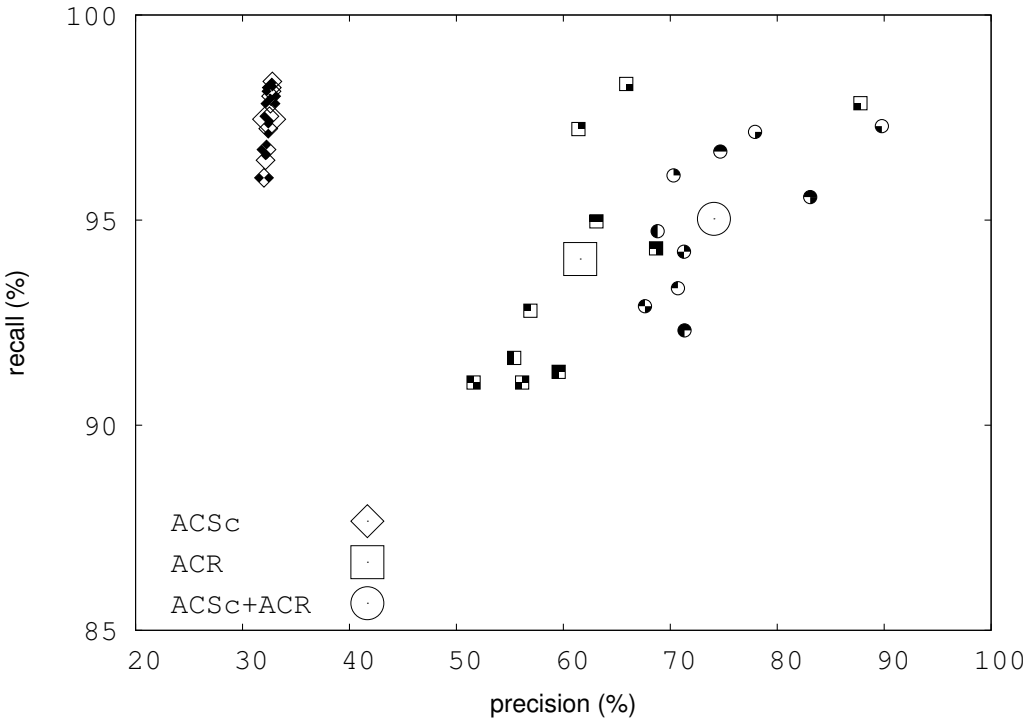


図 6.2: 手法*3*の関連 3 手法の適合率および再現率の比較

を組み合わせた手法*3*の実験結果であり，菱形のマーカは ACSc モジュールのみを用いた手法の実験結果，四角のマーカは ACR モジュールのみを用いた手法の実験結果である．図 6.2 では，エリア毎に比較した場合でも同様の傾向が見られ，それぞれの傾向は各手法の特徴と捉えることができる．

これらの結果から，複数の所属エリア候補をまず可能性の高い所属エリア候補のみに絞り込む第 1 フェーズと，絞込み後の候補のみに対象を絞って機械学習器を用いる第 2 フェーズから成る手法*3*は，第 1 フェーズでの可能性の低い候補の排除と，第 2 フェーズによる類似エリアの間の差異を用いた候補削減とがそれぞれ有効に働き，高い再現率を保ったまま適合率の向上を実現したといえる．

6.4 考察

6.4.1 生成確率による所属エリア推定の妥当性

手法*3*の第 1 フェーズでは，文字レベルの n-gram 情報による生成確率を用いて所属エリア候補の推定を行う．本節では，生成確率による所属エリア候補の推定結果について考察する．

ある地名 t に対して，その生成確率 $P_a(t)$ (式 (6.1)) が最大となるようなエリアをその地名が属するエリアと推測し，これを $\hat{a} = \underset{a}{\operatorname{argmax}}\, P_a(t)$ とする．この定義により n-gram 情報のみを用いて $\hat{a}$ を推定した結果を表 6.5 に示す．生成確率に基づいて所属エリア推定を行った結果，最も高い生成確率を持つエリア ($\hat{a}$) が正解エリアとして出力された割合は全体で 85.96% であった．表 6.5 に示されるように，生成確率による

表 6.5: 生成確率による所属エリア推定の結果

エリア	$\hat{a}$ が正解した割合 (%)
中国	72.70
台湾	73.54
タイ	95.37
日本	93.90
ギリシャ	87.23
フィンランド	88.80
フランス	81.10
スペイン	85.14
ドイツ	90.98
アメリカ	90.86
全体	85.96

推定手法を用いることで，約 86% と高い正解率を得ることに成功した．生成確率を用いた推定手法で $\hat{a}$ の推定に成功した割合が最も高い国はタイで 95.37%，最も低い国は中国で 72.70% であった．中国と台湾の正解率が低い原因は，4.1.3 節で述べた両国の n-gram の類似性によるものと考えられる．

6.4.2 生成確率による所属エリア候補の絞込み

地名の所属エリア推定では，同一の地名が複数のエリアに存在する可能性がある等，所属エリア名を唯一出力する形式は馴染まない．唯一の正解を判定するのではなく，所属エリア候補を絞り込むことを目的とした場合，上位 n 個を候補として残すことで正解を所属エリア候補から不適切に排除されるのを防ぎ，再現率を高めることができる．

図 6.3 に生成確率による推定結果上位 3 個までを候補として残した場合の正解のカバー率 (coverage: 正解エリアが出力された所属エリア候補に含まれた割合) を示す．図 6.3 の `Top 1` は正解エリアが最も高い生成確率値を得た割合，`Top 2` は正解エリアが 2 位の値を得た割合，`Top 3` は正解エリアが 3 位の値を得た割合を示し，`Top 3` までの値が，生成確率値に基づく順位付けで正解エリアが上位 3 個までに入った割合 (上位 3 個での正解カバー率) を示している．また，全体の平均で，上位 2 個までを候補

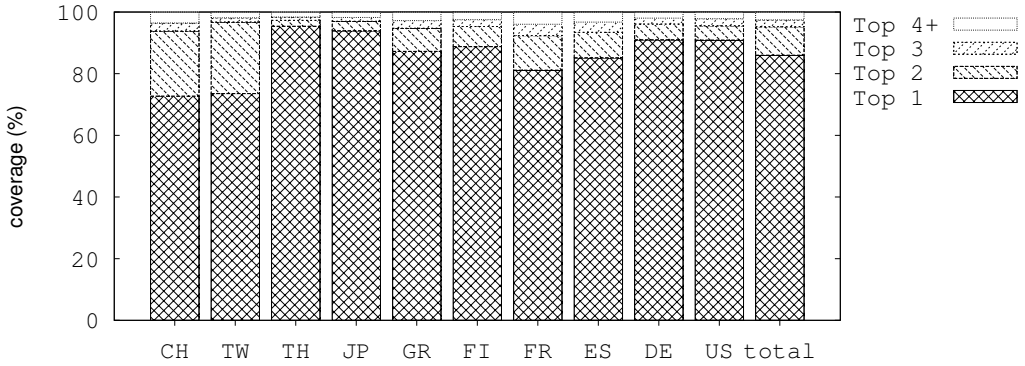


図 6.3: 生成確率による所属エリア推定の正解カバー率

とした場合 95.22% をカバーし，上位 3 個までを候補とした場合には 97.46% をカバーしている．エリア毎に推定結果を見た場合でも，上位 3 個までを所属エリア候補とした場合では最低でも 96.03%，最高で 98.38% の再現率を達成した．これにより，生成確率を利用した手法で所属エリア候補を上位 3 個までに絞り込むことで，98% 程度の再現率を得ることができることがわかる．

出力される所属エリア候補の数を増やすことは，適合率を下げるだけでなく，可能性の低いエリアを候補に残すリスクがある．したがって，所属エリア推定の精度の向上を図るには，各地名に対して適切な所属エリア候補を必要十分な個数だけ出力することが望ましい．生成確率による所属エリア推定では，各エリアに対して推定値 (生成確率の値) が求まる形であり，可能性の有無を二値で判断するものではない．したがって，生成確率値から所属エリア候補とするか否かを判断する必要があるが，これには (1) 閾値等による絶対的評価，(2) 順位付け等，エリア毎の生成確率値の比較による相対的評価，の 2 種類の判断方法がある．さらに，(2) の相対的評価では，(2-1) 入力地名毎に適切な候補数を決定する，(2-2) 入力地名に依らずあらかじめ設定した数の候補を出力する，との 2 種類の判断ができる．生成確率値から正解エリアを判断する妥当な閾値を決めることができれば (1) や (2-1) によって入力地名毎に適切な個数の所属エリア候補を決めることができる可能性があるが，本実験ではこの段階にないため，(2-2) の方法で所属エリア候補を決定する．

生成確率を用いて所属エリア推定を行う場合，図 6.3 で述べたとおり，所属エリア候補を上位 1 個のエリアのみとした場合でも平均 86% 程度の正解率を得ることができる．図 6.3 の結果から，98% 程度の再現率を得るには，所属エリア候補は上位 3 個程度とすることが妥当と判断する．

6.4.3 エリアの類似性に基づくグループの存在の検討

手法*3*では，エリア間の類似性によるグループの存在を仮定し，第 1 フェーズでグループの推定を行い，第 2 フェーズでグループ内の識別を行うとの立場で 2 段階処理を行った (6.1.4 節)．第 1 フェーズでの絞込みが妥当な結果であれば絞込み後の所属エリア候補の組合せに特徴が現れるはずである．手法*3*の第 1 フェーズでは，すべての入力地名について，同じ数 (本実験では 3 個) の所属エリア候補を出力している．出力される所属エリア候補の組合せは入力地名毎に異なるが，すべての組合せが均一に出現することは考えにくく，同時に出力されやすいエリアの組合せが存在するものと考えられる．これを類似エリア群のグループとして定義する．第 1 フェーズでは n-gram 情報のみに基づいて推定処理を行っているため，ここで得られるグループは，エリア間の n-gram 情報の類似性に基づいて現れるものとなる．グループの存在が認められれば，第 2 フェーズで第 1 フェーズと異なる方法で推定を行うことで，第 1 フェーズで利用できなかった差異を基にグループ内のエリア間の識別を行うことができる可能性があり，類似エリア間の識別の可能性が生まれる．手法*3*では，第 2 フェーズとして機械学習器を用いることで，グループ内の差異を機械学習しエリア識別を行う手法を提案している．

所属エリア候補絞込みの組合せの分布

本実験では実験対象エリア数を 10，第 1 フェーズでの絞込み数を 3 とした．そのため，第 1 フェーズでの出力エリアの組合せ数は ${}_{10}C_3 = 120$ となる．1 回の推定 (入力地名 1 個に対する推定) で出力される 3 エリアの組合せを 3 エリア組と呼ぶことにすると，グループの存在は 3 エリア組の出現頻度の偏りとなって表出するものであるといえる．

ここで，3 エリア組を出現頻度順に順位付けすることを考える．絞込み処理を行った結果の 3 エリア組の出現傾向を，3 エリア組の順位毎のカバー率の形で示す (図 6.4)．x 軸は 3 エリア組の順位 (candidate reduction result) を，y 軸はカバー率 (coverage) を表す．例えば x 軸の 5 の点でのカバー率は，出現割合が上位 5 個までの 3 エリア組の出現割合を累積したものである．図 6.4 では，120 とおりの 3 エリア組のうち上位 30 個までの累積出現割合を示している．図 6.4 から，すべてのエリアについて，出現割合上位 30 個までの 3 エリア組だけで 100% 近いカバー率を示すことがわかる．これは，どのエリアについても，理論上 120 とおり存在する 3 エリア組のうち 30 とおり程度しか実際には出力されていないということである．このことから，各エリアについて，出現割合の高い 3 エリア組の存在が裏付けられる．図 6.4 では，エリア毎にカバー率に差はあるものの，上位 10 個程度までの 3 エリア組の出現割合が比較的高く，それ以降徐々に各 3 エリア組の出現割合の差が小さくなっていくような出現傾向が見て取れる．中国，台湾は上位 8 個まで 3 エリア組の累積で 90% 程度のカバー率に達する．それに対して，累積のカーブの緩いエリアは順にタイ，フィンランド，ギリシャ，日本であり，これらの地域が他の特定のエリアとの結びつきが薄いことが考えられる．

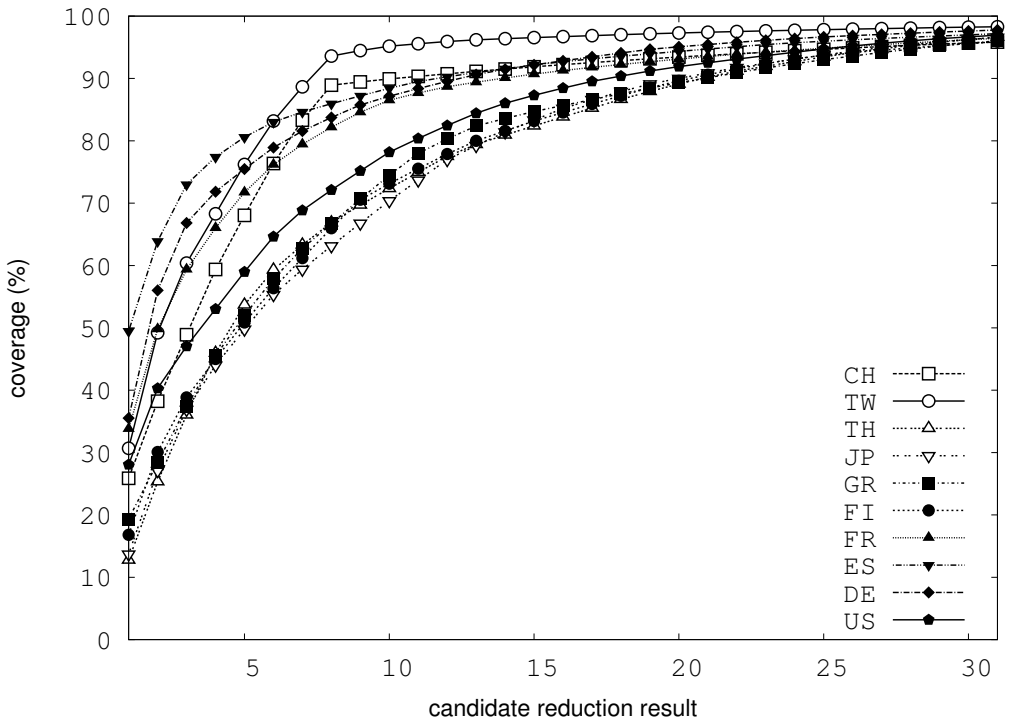


図 6.4: エリア絞込み候補の順位毎のカバー率

所属エリア候補絞込みの組合せからのグループ推定

本実験での出力される 3 エリア組の組合せの数 120 とおりのうち，正解エリアを含む組合せは 36 とおりである．図 6.3 で示したとおり，第 1 フェーズによる絞込みの結果，約 98% の割合で正解が絞込み後の 3 エリア組に含まれていた．このことから，どのエリアについても，正解エリアを含む 36 種類の 3 エリア組の出現頻度が他の 3 エリア組に対して十分に高いことがうかがえる．

ここで，手法*3*の第 1 フェーズ (ACSc) を用いた地名の所属エリア候補の推定結果からグループの存在を検討するため，実験対象エリアのグループを表 6.6 のように推測する．6.1.4 節 で述べたとおり，このグループ分けは各エリアの地理的関係等に基づいて決めたものではなく，3 エリア組の出現傾向から仮定したものである．

表 6.6 のグループ A のエリアについて 3 エリア組の出現状況をまとめたものを表 6.7 に示す．ここでは，120 種類の 3 エリア組のうち，正解エリアを含む 3 エリア組 36 種

表 6.6: 3 エリア組の出現傾向から推測されるグループ

グループ	エリア			
A	中国	台湾	タイ	
B	フランス	スペイン	ドイツ	アメリカ
C	日本	ギリシャ	フィンランド	

表 6.7: 3 エリア組の出現傾向 (グループ A)

中国				台湾				タイ			
順位	CH 以外		%	順位	TW 以外		%	順位	TH 以外		%
1	TW	TH	25.87	1	CH	TH	30.68	1	CH	TW	12.76
2	TW	US	12.37	2	CH	US	18.53	2	CH	US	12.57
3	TW	JP	10.67	3	CH	JP	11.17	3	US	GR	10.72
4	TW	DE	10.46	4	CH	DE	7.92	4	FR	US	9.91
5	TW	GR	8.67	5	CH	GR	7.90	5	DE	US	7.72
6	TW	ES	8.33	6	CH	ES	6.98	6	CH	GR	5.62
7	TW	FR	6.94	7	CH	FI	5.48	7	CH	JP	4.01
8	TW	FI	5.62	8	CH	FR	4.93	8	TW	US	3.56
9	DE	US	0.52	9	FR	US	0.88	9	US	FI	2.84
10	TH	US	0.48	10	ES	US	0.70	10	FI	GR	2.64
11	DE	FR	0.41	11	TH	US	0.40	11	US	JP	2.60
12	TH	FI	0.40	12	US	GR	0.36	12	CH	FI	2.55
13	US	FI	0.39	13	FR	ES	0.27	13	CH	DE	1.92
14	FI	JP	0.38	14	US	FI	0.18	14	TW	JP	1.52
15	US	GR	0.36	15	DE	US	0.17	15	FR	GR	1.45
16	ES	JP	0.35	16	TH	JP	0.16	16	CH	FR	1.45
17	FR	US	0.33	17	FR	GR	0.16	17	TW	FR	1.41
18	TH	JP	0.31	18	FI	JP	0.15	18	GR	JP	1.41
19	FI	GR	0.30	20	US	JP	0.13	19	DE	FI	1.26
20	DE	FI	0.28	23	TH	FI	0.10	20	DE	GR	1.20
21	FR	ES	0.26	24	ES	GR	0.10	21	ES	US	1.09
22	TH	DE	0.26	25	FR	FI	0.09	22	DE	FR	0.97
23	US	JP	0.25	27	TH	ES	0.09	23	FI	JP	0.77
24	GR	JP	0.24	28	GR	JP	0.08	24	ES	JP	0.77
25	FR	GR	0.23	29	FI	GR	0.08	25	FR	JP	0.73
26	TH	GR	0.22	31	DE	FI	0.07	26	FR	ES	0.71
27	TH	FR	0.21	36	TH	DE	0.05	27	TW	GR	0.61
28	DE	JP	0.21	37	TH	GR	0.05	28	CH	ES	0.53
29	ES	GR	0.20	39	FR	JP	0.04	29	TW	DE	0.50
30	FR	FI	0.16	57	ES	FI	0.03	30	DE	JP	0.46
32	DE	GR	0.15	58	TH	FR	0.03	31	FR	FI	0.37
34	ES	US	0.14	60	DE	GR	0.03	32	TW	FI	0.30
38	FR	JP	0.12	71	DE	FR	0.02	33	ES	GR	0.28
40	ES	FI	0.11	85	ES	JP	0.01	34	TW	ES	0.24
48	DE	ES	0.08	88	DE	ES	0.01	35	ES	FI	0.16
73	TH	ES	0.04	93	DE	JP	0.00	60	DE	ES	0.04

類についてのみその出現割合を記載し，同じグループに属するエリアを他と区別するため下線で示している．

表 6.7 の中国の結果 (左端) では，最も出現頻度の高かった 3 エリア組は中国，台湾，タイの 3 エリアの組合せで，その出現頻度の割合は中国の地名を対象とした推定実験全体の 25.87% であったことがわかる．同様に中国を対象とした実験では，2 番目に頻度の高かった 3 エリア組は中国，台湾，アメリカの組合せで出現割合は 12.37%，3 番目に頻度の高かった 3 エリア組は中国，台湾，日本の 3 エリア組で出現割合は 10.67% となっている．中国と台湾では両者の双方を含む 3 エリア組は 8 種類あるが，どちらのエリアでもこの 8 種類の 3 エリア組が上位 8 個までを占めており，これだけで 90% 程度 (中国エリア 88.93%，台湾エリア 93.59%) の出現割合を占めている．このことから，この 2 エリアについては 90% 程度の地名について中国語圏であることがこの段階で判別できることがわかる．同じ東アジア地域である日本が 3 エリア組の中に残る場合を調べると，中国，台湾，日本の 3 エリアの出現割合は，正解が中国の場合で 10.67% (3 位)，台湾の場合で 11.17% (3 位)，日本の場合で 13.71% (1 位) であった．それに対

して，中国，台湾，タイの 3 エリア組は，正解が中国の場合で 25.87% (1 位)，台湾の場合で 30.68% (1 位)，タイの場合で 12.76% (1 位) の割合であり，中国，台湾，日本の組合せの場合よりも高い値を示した．これは，中国語圏の地名が中国，台湾，タイの 3 ヶ国に絞られる割合が高いことを示している．

表 6.7 は各正解エリアが含まれる 3 エリア組のみを記載しているため，表の最下位に記載されている 3 エリア組が，正解エリアを含む 3 エリア組のうち最も低い出現割合を持つものを示すことになる．図 6.3 の結果から正解エリアは出力されやすいことがわかっているが，表 6.7 を見ると，正解エリアを含むにもかかわらず 120 組中 93 位で出現する 3 エリア組があることがわかる (台湾が正解エリアの場合での台湾，日本，ドイツの 3 エリア組)．このことから，この 3 エリアが同時に出現する頻度が相当に低いことがうかがえる．

中国と台湾が同時に出現する頻度が非常に高い反面，タイとこの 2 エリアとが同時に出現する頻度はそこまで高い値は示さない．表 6.7 を見ると，中国と台湾では上位 8 個までの出現割合が高く，上位 9 個目以降は 1% 以下に落ち込んでいる．これは，図 6.4 の曲線からも読み取れる．図 6.4 を見ると，上位 5 個程度までの部分では他のエリアの方が高いカバー率を示しているが，中国，台湾では上位 8 個までがほぼ線形に上昇し，その後急に角度を変えてなだらかな曲線を描く．このことから，中国，台湾については，この 2 エリア以外との組合せは特に出現頻度の高い上位 1 位と 2 位以外はどれもほぼ同じ程度の出現頻度であり，中国，台湾，タイの 3 エリア組の可能性が強いことがわかる．それに対して，タイは 1 位の 3 エリア組の出現割合が 12.76% と実験対象エリア全体の中で最も低い値を示しており，2 位以下の出現割合の減少の幅も比較的小さい．図 6.4 でもタイの曲線は最も下に位置しており曲線のカーブも緩く，タイと他のエリアの類似性は低いものと考えられる．

同様に，表 6.6 のグループ B のエリアについて 3 エリア組の出現状況をまとめたものを表 6.8 に示す．グループ B としてまとめられたのは，フランス，スペイン，ドイツ，アメリカの 4 エリアである．表 6.8 で示されるとおり，ドイツ，フランス，スペインにアメリカを加えた 4 エリアは，互いに候補として残る頻度が高い．特に，1 位の 3 エリア組は他と比べて高い出現割合を示している．

フランスとスペインの 2 エリアでは，フランス，スペイン，アメリカの 3 エリア組が特に高い出現割合を示して 1 位となっている．この 3 エリア組の出現割合は他のどのエリアの 1 位の出現割合よりも高く，この組合せが特に強い類似性を持つことがうかがえる．フランスでは 33.81% とほぼ 1/3，スペインでは 49.49% とほぼ 1/2 の出力がこの 3 エリア組であった．これに対して，ドイツとアメリカの 2 エリアでは，ドイツ，フランス，アメリカの 3 エリア組の出現割合が 30% 前後と高い値を示している．この 4 エリアからなる 3 エリア組は 4 種類存在し，表 6.8 にはそのうち 3 種類が記載されている[2]が，その出現の傾向は必ずしも 4 エリアで共通するものではない．フランスでは該当する 3 エリア組は 1 位，2 位，4 位に出現しており，この 3 個の 3 エリア組の累計出現割合は 56.46% である．スペインでは 1 位，5 位，10 位に出現し，累計は 53.89%，ドイツでは 1 位，6 位，9 位に出現し，累計は 40.97%，アメリカでは

[2]4 種類の 3 エリア組のうち，正解エリアが入らない 3 エリア組は表 6.8 には記載されていない．

表 6.8: 3 エリア組の出現傾向 (グループ B)

フランス				スペイン				ドイツ				アメリカ			
順位	FR 以外		%	順位	ES 以外		%	順位	DE 以外		%	順位	US 以外		%
1	ES	US	33.81	1	FR	US	49.49	1	FR	US	35.52	1	DE	FR	28.07
2	DE	US	15.99	2	FR	GR	14.41	2	US	FI	20.49	2	FR	ES	12.23
3	US	GR	9.57	3	US	GR	9.09	3	FR	FI	10.82	3	FR	GR	6.76
4	DE	ES	6.66	4	FR	FI	4.42	4	US	GR	5.01	4	TW	FR	5.97
5	ES	GR	5.69	5	DE	FR	3.18	5	FR	GR	3.63	5	DE	FI	5.95
6	US	FI	4.46	6	FI	GR	2.43	6	FR	ES	3.44	6	TW	DE	5.65
7	ES	FI	3.25	7	FR	JP	1.62	7	FI	GR	2.66	7	DE	GR	4.21
8	DE	FI	2.75	8	US	FI	1.31	8	CH	US	2.17	8	TH	DE	3.27
9	FI	GR	2.41	9	GR	JP	1.25	9	ES	US	2.01	9	FR	FI	3.07
10	DE	GR	1.92	10	DE	US	1.22	10	CH	FR	1.41	10	CH	DE	3.00
11	TW	US	1.17	11	CH	FR	1.08	11	TW	US	1.21	11	TH	FR	2.19
12	TH	US	0.99	12	TW	FR	0.78	12	TH	US	1.17	12	CH	FR	2.09
13	CH	US	0.72	13	FI	JP	0.65	13	ES	FI	1.07	13	DE	ES	1.96
14	TW	ES	0.70	14	US	JP	0.63	14	US	JP	0.87	14	FI	GR	1.60
15	ES	JP	0.61	15	TH	FR	0.63	15	FI	JP	0.70	15	ES	GR	1.27
16	CH	ES	0.58	16	DE	GR	0.49	16	FR	JP	0.65	16	TW	TH	1.16
17	TW	GR	0.50	17	DE	FI	0.47	17	CH	FI	0.64	17	CH	TW	1.08
18	CH	GR	0.46	19	TW	US	0.40	18	TH	FI	0.62	18	TW	GR	0.86
21	GR	JP	0.40	20	CH	TW	0.38	19	CH	TW	0.59	19	DE	JP	0.81
22	US	JP	0.39	21	CH	US	0.38	20	TW	FR	0.40	20	ES	FI	0.70
23	DE	JP	0.34	24	CH	GR	0.28	22	TH	FR	0.36	21	TH	GR	0.62
26	CH	TW	0.29	25	TH	US	0.26	23	CH	GR	0.31	22	TW	ES	0.58
27	TH	GR	0.29	26	TH	GR	0.22	24	ES	GR	0.29	23	FR	JP	0.56
28	CH	DE	0.27	30	CH	JP	0.19	25	TW	FI	0.26	24	CH	GR	0.55
30	TH	ES	0.25	31	TH	FI	0.18	30	CH	JP	0.17	25	TH	FI	0.53
31	TW	DE	0.23	33	CH	FI	0.14	31	TH	GR	0.16	26	TW	FI	0.47
32	FI	JP	0.21	34	DE	JP	0.12	32	CH	ES	0.16	27	CH	TH	0.44
33	TH	DE	0.21	35	TH	JP	0.11	35	CH	TH	0.11	28	TH	JP	0.38
37	CH	JP	0.13	40	TW	GR	0.08	37	ES	JP	0.09	29	TW	JP	0.35
39	CH	FI	0.12	43	TW	JP	0.08	38	TH	JP	0.09	30	ES	JP	0.28
43	TH	FI	0.09	52	CH	DE	0.05	40	GR	JP	0.09	31	CH	ES	0.28
44	CH	TH	0.07	58	CH	TH	0.04	41	TW	GR	0.06	32	CH	FI	0.28
46	TW	FI	0.06	60	TW	FI	0.04	43	TH	ES	0.06	33	GR	JP	0.22
56	TH	JP	0.05	73	TW	DE	0.02	50	TW	ES	0.04	34	CH	JP	0.22
69	TW	JP	0.03	89	TH	DE	0.01	53	TW	TH	0.03	36	FI	JP	0.18
70	TW	TH	0.03	93	TW	TH	0.01	77	TW	JP	0.01	43	TH	ES	0.10

1 位，2 位，13 位に出現し，累計は 42.26% であった．スペインでは 5 位と 10 位に共通してドイツが現れ，一方ドイツでは 6 位と 9 位に共通してスペインが現れており，またアメリカで 13 位，フランスで 4 位となった 3 エリア組はどちらもドイツ，スペインの双方を含むものであることから，ドイツとスペインの間の類似性はグループ B の他のエリアと比較して低いものと考えられる．これらの結果から，グループ B の 4 エリアの関係は，お互いに対等な類似性を持つものではなく，フランスとアメリカが間に入る形でスペインとドイツが繋がっているような状態と推測できる．スペイン，フランス，ドイツは地理的にもフランスを間に入れて並んでいる形となっており，この推測は自然に感じられる．

グループ B の 4 エリアは同じ言語圏に属するものではないが，地理的，歴史的，文化的背景から，他のエリアと比べて類似した地名が多いことが予想され，これがグループの形成に繋がったと考えられる．特にフランスとスペインについては，n-gram 情報に類似が見られる (4.1.3 節) ことから，n-gram に基づく生成確率を用いた ACSc モジュールで類似エリアとして扱われることは自然である．

表 6.9: 3 エリア組の出現傾向 (グループ C)

日本				ギリシャ				フィンランド			
順位	JP 以外		%	順位	GR 以外		%	順位	FI 以外		%
1	CH	TW	13.71	1	US	FI	19.34	1	DE	US	16.81
2	FI	GR	13.39	2	FR	US	9.09	2	US	GR	13.28
3	CH	GR	9.80	3	FR	FI	8.89	3	DE	GR	8.75
4	CH	FI	7.10	4	FR	ES	8.18	4	FR	GR	6.16
5	CH	TH	5.78	5	TH	FI	6.51	5	DE	FR	5.86
6	CH	US	5.54	6	DE	FI	5.86	6	FR	ES	5.50
7	TH	US	4.06	7	ES	FI	4.86	7	FR	US	4.82
8	US	GR	3.76	8	ES	US	3.98	8	ES	GR	4.80
9	TH	GR	3.64	9	FI	JP	3.96	9	GR	JP	4.57
10	ES	GR	3.60	10	DE	US	3.79	10	ES	US	2.57
11	US	FI	3.30	11	TH	US	3.53	11	TH	GR	2.41
12	TH	FI	3.29	12	CH	FI	2.30	12	CH	DE	2.36
13	TW	TH	2.26	13	DE	FR	2.08	13	DE	JP	2.11
14	TW	US	2.18	14	US	JP	1.21	14	CH	GR	1.59
15	ES	FI	1.90	15	CH	US	1.11	15	CH	JP	1.58
16	CH	ES	1.89	16	TH	DE	0.98	16	TH	DE	1.37
17	ES	US	1.40	17	TH	FR	0.94	17	US	JP	1.36
18	TW	GR	1.29	18	ES	JP	0.93	18	CH	TW	1.32
19	TW	FI	1.22	19	CH	TW	0.91	19	DE	ES	1.30
20	DE	US	0.90	20	CH	TH	0.86	20	TH	JP	1.09
21	FR	ES	0.88	21	FR	JP	0.80	21	CH	US	0.99
22	DE	FI	0.86	22	CH	FR	0.80	22	TH	US	0.92
23	DE	GR	0.70	23	DE	ES	0.78	23	ES	JP	0.85
24	CH	DE	0.66	24	TW	FI	0.67	24	FR	JP	0.76
25	TH	ES	0.66	25	TW	US	0.65	25	CH	TH	0.60
26	FR	FI	0.61	26	CH	ES	0.59	26	TW	JP	0.42
27	FR	GR	0.57	27	CH	JP	0.57	27	TW	GR	0.38
28	FR	US	0.51	28	TW	FR	0.53	28	CH	ES	0.37
29	TW	ES	0.42	30	TH	ES	0.42	29	CH	FR	0.35
30	CH	FR	0.36	31	TH	JP	0.40	30	TH	ES	0.34
31	TH	DE	0.35	33	TW	ES	0.33	31	TW	DE	0.32
32	DE	ES	0.26	34	CH	DE	0.30	32	TW	US	0.31
33	TW	DE	0.25	35	DE	JP	0.30	36	TH	FR	0.26
34	DE	FR	0.24	43	TW	JP	0.10	38	TW	FR	0.23
35	TH	FR	0.17	44	TW	TH	0.10	41	TW	TH	0.15
36	TW	FR	0.16	46	TW	DE	0.09	45	TW	ES	0.13

同様に，表 6.6 のグループ C のエリアについて 3 エリア組の出現状況をまとめたものを表 6.9 に示す．グループ C としてまとめられたのは，日本，ギリシャ，フィンランドの 3 エリアである．グループ C の 3 エリアは，互いを含む 3 エリア組があまり上位に入っておらず，他のグループとは明らかに異なる出現傾向を示している (表 6.9)．グループ C の 3 エリア組の出現割合は，日本で 13.39% (2 位)，ギリシャで 3.96% (9 位)，フィンランドで 4.57% (9 位) と，少なくはないものの他のグループでの互いの出現傾向と比べて低い．また，1% 以上の出現割合を持つ 3 エリア組が 15 位から 20 位前後まで続く等，他のエリアに比べて，頻度の高い 3 エリア組の組合せの数が多い傾向がある．他の 2 グループで同様に 1% 以上の 3 エリア組が多いエリアはタイとアメリカで，他のエリアでは 1% 以上の割合を持つ 3 エリア組は 10 個前後に絞れている．図 6.4 の曲線の緩さ等も考え合わせると，これらの 3 エリアは積極的にグループを組んだものではなく，他のエリアとの類似性の低さから相対的にお互いとの類似性が高く評価されたものと考えられる．グループ C の 3 エリアは，母音 unigram の出現状況に類似性が見られ (4.1.3 節)，これがグループの形成に繋がった可能性がある．

グループ C のエリアのもう一つの特徴として，正解エリアを含む 3 エリア組の出現順位の高さが挙げられる．正解エリアを含む 3 エリア組の出現順位のうち，最も順位の低いものは，日本は 120 種類の 3 エリア組の中の 36 位，ギリシャは 46 位，フィンランドは 45 位であった．これと同程度の順位を示したのはアメリカの 43 位のみで，他のエリアはタイで 60 位，それ以外は 70 位以降となっている．グループ A，グループ B のエリアは，正解エリアを含む 3 エリア組であっても，他のグループのエリアが混在するものは出現頻度が低くなり，結果的に順位が低くなる傾向が見られた．これは，地名自身がグループの特徴に合っていなかったためと考えられ，例えばグループ A の地名がなんらかの理由でグループ B の特徴を示した場合には，正解エリアも含めてグループ A の可能性自体が低くなり，グループ B の 3 エリア組が優先されたものと思われる．それに対して，グループ C のエリアでは 3 エリア組の組合せに関わらず正解エリアを含む 3 エリア組が優先されたと考えられ，グループ C に属する地名は他のグループの地名と比べてグループとしての特徴よりもエリアの特徴が強く影響した結果，正解エリアを含む 3 エリア組の順位が相対的に高くなったものといえる．

推定グループの出現傾向と妥当性

表 6.6 で示したグループの妥当性の検討のため，実験対象エリア全体の地名を対象とした 3 エリア組の出現傾向を示す (図 6.5)．y 軸はその出現割合 (ratio) を表している．グループの存在の仮定および表 6.6 のグループの仮定が正しい場合，ACSc モジュールによる所属エリア候補推定が正しく行われれば，出力される 3 エリア組は理想的には同じグループのエリアのみから成る 3 エリア組となるはずである．本章で仮

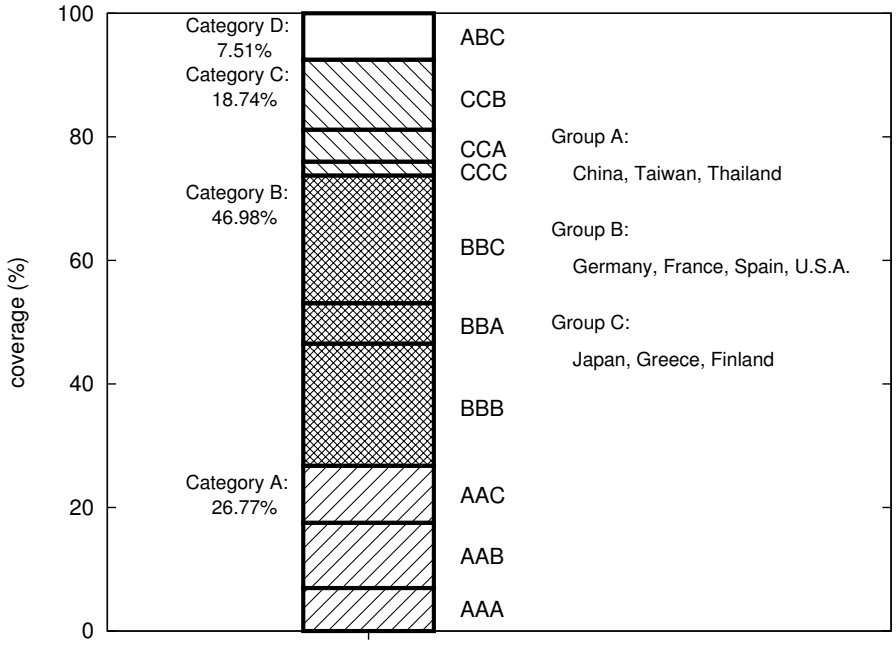


図 6.5: グループに着目した絞込み結果の分類

定しているグループは排他的なものではなく，均一なものでもないため，同一グループのエリアのみから成る 3 エリア組だけを正解とするのは条件が厳し過ぎる．そのため，図 6.5 では特定のグループに偏った 3 エリア組という意味で，3 エリア組をカテゴリに分けて集計した．ここでのカテゴリとは，出力された 3 エリア組の持つグループの偏りを示すものである．図 6.5 では，絞込み候補の 3 エリア組を，グループ A のエリアが 2 個以上出力された場合 (カテゴリ A)，グループ B のエリアが 2 個以上出力された場合 (カテゴリ B)，グループ C のエリアが 2 個以上出力された場合 (カテゴリ C)，すべてのグループから 1 エリアずつ出力された場合 (カテゴリ D) の 4 カテゴリに分けて示している．例えば図 6.5 の ‘AAB’ はグループ A に属するエリア 2 個とグループ B に属するエリア 1 個から成る 3 エリア組が絞込み結果として出力された地名の割合を示し，グループ A に属するエリアが 2 個以上であることから，グループ A に偏った出力としてカテゴリ A に入れる．

本実験では各エリアの地名コーパスに含まれる地名数は同じであるため，絞込みが完全に成功すればカテゴリ A に 30%，カテゴリ B に 40%，カテゴリ C に 30% の割合で 3 エリア組が出現するはずである．カテゴリ D はグループを絞り込めない 3 エリア組を示し，グループの存在が認められる場合には理想的には 0% となる．

実験結果をカテゴリ毎に集計した結果，カテゴリ A の 3 エリア組が 26.77%，カテゴリ B の 3 エリア組が 46.98% 出現していた (図 6.5)．グループ A およびグループ B については理論値に近い出現率が得られたことから，この 2 個のグループは意味のあるグループと推定できる．

カテゴリ D の 3 エリア組の出現率は全体の 7.51% であった．このことから，90% 以上の地名についてどれかのグループに偏った 3 エリア組が出力されたことがわかる．このことから，同時に候補に残りやすい組合せの存在がわかり，本フェーズが所属エリア候補の絞込み処理として妥当かつ有効であることがいえる．

カテゴリ C の 3 エリア組は 30% の出現率が期待されるが，実際には 18.74% となった (図 6.5)．前述のとおりグループ C は他の 2 グループに比べて互いの類似性が低い．表 6.9 でも，同じグループに属するエリアを含む 3 エリア組が必ずしも高い順位を得るとは限らず，他のグループと混同されやすいとの結果が出ていた．3 個のグループが混在した 3 エリア組を示すカテゴリ D の出現率は 7.51% であり，これとカテゴリ C の出現率 18.74% を合わせると 26.25% となる．これらから，カテゴリ C の 3 エリア組の代わりにカテゴリ D の 3 エリア組が出力された例が多いものと予想される．

6.4.4 グループに着目した所属エリア推定

6.4.2 節では，第 1 フェーズでの絞込みの妥当性を実験結果から説明し，第 1 フェーズで 3 エリアを出力することにより 98% 程度の再現率で所属エリア候補の絞込みができることを示した．また 6.4.3 節では，第 1 フェーズでの絞込み結果を検証し，エリア間の類似性に基づくグループの存在と，第 1 フェーズでの絞込みがグループの絞込みをある程度実現していることを示した．このことから，第 2 フェーズは，グループが絞り込まれた状態でグループ内のエリア間の識別を行うことで，地名の所属エリア

候補を正しく推定し，可能性の低い所属エリア候補を排除する処理と位置付けられる．

本節では，グループの特定が行えた場合を想定し，グループ内のエリアを対象とした機械学習による所属エリア推定の実験結果を示す．表 6.10 に，グループ A の 3 エリア (中国，台湾，タイ) について，この 3 エリアを対象に機械学習器を用いて所属エリア推定を行った結果を示す．グループ A に属する 3 エリアのみを対象として機械学

表 6.10: 機械学習器による所属エリア推定の結果 (グループ A)

エリア	F 値	適合率 (%)	再現率 (%)	正解率 (%)
中国	0.83	70.96	100.00	86.36
台湾	0.73	59.87	93.71	76.97
タイ	0.95	90.71	99.75	96.51

習を行った場合，表 6.10 に示したとおり，F 値は最低で 0.73，最大で 0.95 となった．正解が台湾の場合の適合率が 59.87% と低いが，この場合でも再現率は 93.71% であり，正解が候補として出力される頻度は十分に高い．中国では再現率 100%，タイでも再現率は 99.75% と非常に高い値を示している．この傾向は他のグループでも同様で，再現率では台湾の 93.71% が最も低い値であった．このことから，グループに着目して所属エリア推定を行うことで高い再現率が期待できる．

表 6.7 の結果を見ると，中国と台湾では上位 8 個までの頻出 3 エリア組の出現順位はほぼ同じである (7 位と 8 位のみ順序が入れ替わっている)．このことから，中国と台湾はお互いの類似性が非常に高く，中国のみ高い類似性を示すエリアや台湾のみ高い類似性を示すエリアが存在するわけではないことがわかる．このことがグループ A 内での識別でこの 2 エリアの F 値がタイと比べて低い理由である．それに対してタイは同じグループに属するが性質が異なり，中国および台湾との間の識別ではその特徴の差異を捉えることに成功したため，高い適合率を達成している．

表 6.11 は，グループ B に属する 4 エリアを対象として所属エリア推定を行った結果である．グループ B は 4 エリアから成るため，ここではこの 4 エリアから成る 4 種類の 3 エリア組それぞれについて機械学習を行っている．4 エリアの中ではフランスの地名に対する結果が悪いが，これはフランスの適合率が他と比べて低いためである．再現率については，フランスも含め，どのエリアでも 95% 以上であり十分な値を達成している．このことから，フランスは他のエリアと混同される割合が高いことがうかがえる．フランスも台湾もそれぞれ他のエリアとの混同により精度が低下しているが，混同されるエリアがすべて精度の低下を起こしているわけではないことから，類似性に方向性があることが確認できる．フランスは n-gram に基づくパープレキシティ (6.1.3 節) や長さ情報等でスペインとドイツの双方と類似性が高い．それに対してスペインとドイツの間では類似性は高くはない．このようにグループ内の複数のエリアとの間で高い類似性を示すエリアはグループ内での識別の精度が低下する傾向がある．台湾の精度が低いことも，台湾が中国との間で n-gram 情報で高い類似性を示すと同時に，タイとの間では n-gram 情報の類似性に加えて長さ情報で高い類似性を

表 6.11: 機械学習器による所属エリア推定の結果 (グループ B)

エリア	F 値	適合率 (%)	再現率 (%)	正解率 (%)
フランス	0.75	60.91	98.80	78.46
スペイン	0.86	77.50	96.36	89.46
ドイツ	0.86	75.29	99.64	88.98
フランス	0.79	66.03	98.16	82.55
スペイン	0.85	75.69	95.93	88.37
アメリカ	0.87	78.94	97.73	90.55
フランス	0.80	67.48	98.60	83.69
ドイツ	0.84	72.76	99.82	87.48
アメリカ	0.86	78.03	95.55	89.55
スペイン	0.90	83.32	98.42	92.91
ドイツ	0.89	80.72	99.99	92.04
アメリカ	0.87	79.45	95.86	90.35

示し，中国とタイの双方との間で類似性が高いことが要因と考えられる．グループ A 内でも，中国とタイの間の類似性は高くなく，そのためこの 2 エリアは精度を落としていない．

表 6.12 はグループ C に属する 3 エリアを対象として所属エリア推定を行った結果である．このグループでも再現率はすべて 95% 以上となっており，正解が排除されずに出力する可能性は十分に高い．このグループは他の 2 グループに比べてエリア間の

表 6.12: 機械学習器による所属エリア推定の結果 (グループ C)

エリア	F 値	適合率 (%)	再現率 (%)	正解率 (%)
日本	0.85	75.02	99.02	88.68
ギリシャ	0.75	61.45	95.93	78.59
フィンランド	0.80	67.15	97.86	83.33

類似性が低いことが推定される．しかし，グループ内のみを対象とした所属エリア推定結果を見る限り，このグループの結果は他の 2 グループの結果と比べて大きな差異はない．このことから，グループ内のエリアのみを対象として所属エリア推定を行うことで，類似したエリアであっても十分に識別することが可能であるといえる．グループ内の互いの類似性の高さに関わらずグループ内全体の識別精度が劣化する傾向は低く，グループによらず安定した識別が行われることがわかった．また台湾やフランスの例から，グループ内の複数のエリアとの間で特に類似性が高いエリアのみ差別化が十分にできずに精度が落ちる傾向が見て取れる．

これらの結果から，同じグループに属するエリアの間での分類であっても，十分な正解率が得られることが示された．実験対象とした 10 エリア全体を対象とした学習では，類似性の低いもの同士の間の差異が強調され，類似性の高いもの同士の間の差異の学習は進み辛い．そのため手法*1*は類似エリアの存在に弱く，類似エリアを含む実験ではよい結果が得られなかった (4.4.3 節)．それに対して，手法*3*では 2 段階処理を導入し，第 1 フェーズで絞込みを行った後の所属エリア候補群のみを対象として第 2 フェーズで局所的な所属エリア推定を行うことで，類似性の高いエリアの間での所属エリア推定の精度の向上を実現した．

6.4.5 機械学習器の個数

第 1 フェーズでの絞込み結果に応じて第 2 フェーズでの推定処理対象が変化する手法*3*では，第 2 フェーズで機械学習器を用いる手法を導入する場合，実験対象エリアの数の増加に応じて必要な機械学習器の数が爆発的に増加する可能性がある．実験対象エリア数を 10，第 1 フェーズでの絞込み数を 3 とした本実験の場合，理論上可能な 3 エリア組の種類は ${}_{10}C_3 = 120$ 個，学習および判定するために必要な機械学習器の数は ${}_{10}C_3 \times 3 = 360$ 個となる．数多くの機械学習器を適切に学習させるためには，相当なデータ量と時間とが必要となり，現実性が低くなる．このように，実験対象エリア数が増えると実現が難しくなる可能性があることが手法*3*の問題として挙げられる．

手法*3*ではグループの存在を仮定しているため，実際には，すべての組合せの機械学習器を準備する必要はない．図 6.4 に示したとおり，第 1 フェーズで出力される所属エリア候補の組合せは出現頻度に大きな偏りを持っている．これは手法*3*で仮定したグループの存在によるものであり，実際に出力される所属エリア候補の組合せはグループの特徴を意識したものとなる．理想的には，第 1 フェーズは，グループの特定を行う処理と位置付けられ，その出力は特定のグループに属する類似エリアの集合となることが期待される．第 1 フェーズでグループを特定できるのであれば，第 2 フェーズで必要となる機械学習器を減少させることができる．

6.5 手法*3*のまとめ

地名を対象とした所属エリア推定処理では，言語的な類似性や地理的，歴史的要因等さまざまな理由によって所属エリアの判別が困難な場合がある．手法*1*，手法*2*，手法*1+2*では，類似エリアの識別が十分に機能しているとはいえなかった．これは，抽出された地名の表層情報を利用する際に，対象エリア全体を所属エリア推定対象として評価していたためである．

本章では，類似した特徴を持つエリアがグループを成すものと仮定し，グループの特定を行う第 1 フェーズで大域的な所属エリア推定を行い，グループ内でのエリアの識別を行う第 2 フェーズで局所的な所属エリア推定を行うような 2 段階の処理を導入することで，類似エリア識別の問題を解決した (手法*3*)．この方法を用いた 10 エリア

を対象とした評価実験では，平均 95.03% の再現率，74.08% の適合率を得ることに成功し，その有効性を確認した．

手法*3*はきめ細かい識別処理を実現した反面，そのために必要な学習の量が増加しており，実験対象エリア数の増加によって実現が困難になるという点を考慮する必要がある．このため，全世界的なレベルの処理の実現へ向け，学習量を抑える改良が必須である．

第7章　表層情報に基づく地名の所属エリア推定手法の提案 (手法*4*)

本研究では，文字レベルの表層情報の所属エリア推定での利用 (手法*1*: 第 4 章) について調べた結果を基に，所属エリア推定に適した処理単位の検討 (手法*2*，手法*1+2*: 第 5 章)，類似エリア識別手法の検討 (手法*3*: 第 6 章) を行ってきた．本章では，これらを統合し，候補エリア数が増加した場合にも対応可能な，頑健性，汎用性に優れた手法を提案する (手法*4*)．手法*4*は単純にこれらを繋ぎ合わせたものではなく，手法*1*，手法*2*，手法*1+2*，手法*3*の研究から得られたデータを基に，これらのアプローチを再構築したものである．

7.1　頑健な所属エリア推定手法の提案

地名は含有する文字列の長さが短く，また固有表現のため一般の語句を用いた言語推定に比べて頻出文字列が少ない等，地名の文字レベルの表層情報は適切な利用が難しい．第 4 章では再現率を重視した手法として，文字レベルの表層情報を利用した機械学習器による所属エリア候補の削減を行う手法を提案し (手法*1*)，文字レベルの表層情報を用いて地名の所属エリア推定を行うことは十分可能であることを示した．第 5 章では，第 4 章で課題となった適合率の低さの解決を図り，適合率を重視した手法として，地名に適した処理単位としてブロックの概念を提案し，ブロックレベルの表層情報を用いることで，所属エリア候補としての確度を測る新しい指標を提案した (手法*2*)．これは，文字レベルの表層情報での課題であった適合率の低さを解消するだけでなく，単語レベルの表層情報で発生しやすいデータスパースネスの問題も解消するもので，これにより地名の所属エリア推定処理の適合率を向上させることに成功した．第 6 章では，同一の地名や非常に類似した地名が複数のエリアに存在するという地名の特殊性を考慮し，類似エリア間の識別の問題を検討し，類似エリアが所属エリア推定の対象エリアに含まれるような場合においても，所属エリア推定の精度を確保する方法を提案した (手法*3*)．しかし，手法*3*では，実験対象エリアの数の増大に伴って処理に必要な学習量が大幅に増大することがわかっており，実用に向けた対策が課題として残った．

これらの成果を受けて，本章では，これまで提案してきた各手法を統合すると同時にそれぞれの手法の持つ問題点の解決を図ることで，地名の所属エリア推定手法の確立を目指し，より頑健性と汎用性の高い手法を提案する．

7.1.1 所属エリア推定手法の統合

表層情報を用いた所属エリア推定を行う上で重要な性質は，既に述べたように，高い再現率と高い適合率，さらに，高い汎用性である．本章では，ここまでに検討してきた手法*1*，手法*1+2*，手法*3*の統合によって，これらの性質を同時に実現する (図 3.9).

手法*1+2*は，ACR モジュールによる所属エリア候補の削減の処理として手法*1*を，ACEb モジュールによるブロックを用いた所属エリア候補の検証の処理として手法*2*を順に実行する 2 段階処理の形で統合したものだった (図 5.5)．手法*1+2*では，適合率の向上のために，最終的な所属エリア候補の出力を 1 個に絞り込んでおり，このアプローチでのさらなる精度の向上は困難である．より精度の高い手法を実現するためには，ブロックレベルの表層情報を利用し適合率を確保しつつも，ブロックレベルの情報だけでは判断が難しい場合を考慮するため複数の所属エリア候補の出力を許し，さらにそこから所属エリア候補を絞り込む必要がある．所属エリアの可能性が高いと判断される複数の所属エリア候補からの絞込み処理は，手法*3*が目的とした類似エリア識別処理に帰着できる．手法*3*は，第 1 フェーズで所属エリア候補の絞込みを行った結果に対して，第 2 フェーズとして類似エリア識別処理を行う形である．そこで，第 1 フェーズの ACS モジュールに手法*2*の成果を導入することで第 1 フェーズの絞込みの精度を向上させ，手法*3*の第 2 フェーズを活かすことを検討する．

本章で提案する手法*4*のシステムの概要を図 7.1 に示す．手法*4*は，所属エリア候補の絞込みを行う area candidate selection (ACSb) モジュールによる処理と，所属エリア候補の削減を行う area candidate reduction (ACR) モジュールによる処理の 2 フェーズで構成する．ACSb モジュールでは，入力地名について，所属エリア候補の絞込み

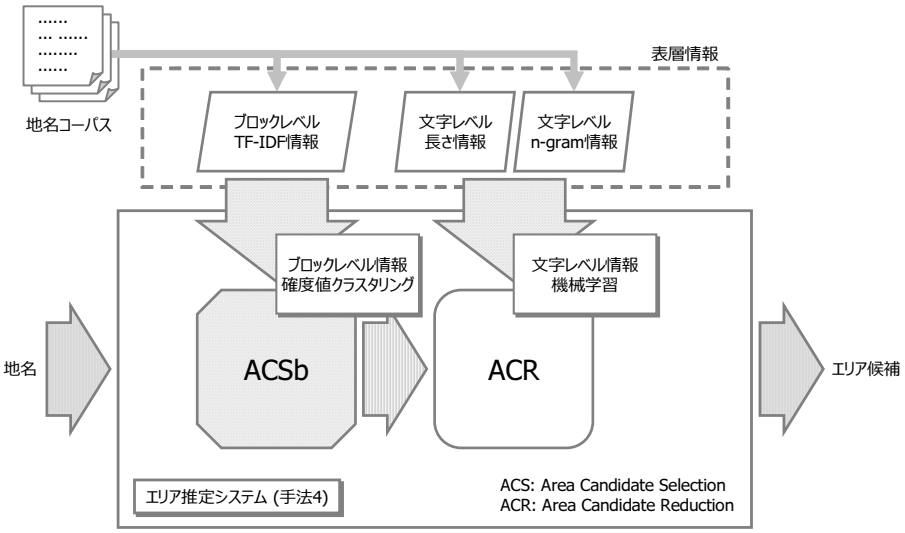


図 7.1: 手法*4*の処理の流れ

を行い実験対象エリア全体からあらかじめ設定した候補数まで所属エリア候補を絞り込む．これは，ブロックレベルの TF-IDF 情報をベースとした階層的クラスタリングによる絞込み処理である．次の ACR モジュールでは，絞込み後の所属エリア候補群

それぞれに対して所属の可能性を推定し，可能性があると推定されたものを所属エリア推定の結果として出力する．ACR モジュールでは，文字レベルの n-gram 情報と長さ情報を用いた所属エリア推定を行う．本システムの出力は，入力地名文字列が所属する可能性のあるエリアであり，推定結果に応じて 0 個 (第 2 フェーズで候補のすべてが排除された場合) から第 1 フェーズでの絞込み個数まで (第 2 フェーズで候補のすべての候補が排除されなかった場合) となる．

7.1.2 ブロックレベル情報を用いた所属エリア候補の絞込み

第 6 章で述べたとおり，特定のエリア同士，あるいは特定の複数のエリア間において，表層的な特徴を共有する部分があることがわかっている．このような表層情報の共有はエリア間の類似性を生む．エリア間の類似性は，エリア間の 1 対 1 の関係とは限らず，複数のエリアの集合としてのグループの固有の特徴として表層情報を共有する等，階層的な関係を持つことが示唆される．このことから，地名のエリアに対する所属確度についても階層的な関係を持っていると考えるのが自然である．第 6 章では，類似エリア間のグループの存在を仮定し，グループを同時に所属エリア候補として残りやすい 3 エリア組の集合と定義した．6.4.3 節で述べたとおり，この定義では同じグループ内のエリア間の類似性の差を反映することができないため，グループ C のようにエリア間の類似性の低いグループでは十分に効果を発揮できない可能性があった．このことから，エリア間の類似性を適切に反映するためには，エリアを固定したグループとしてまとめたうえでグループ内か否かとの尺度で処理することは効果的でない場合が考えられる．そこで本章では，類似エリア群の中での類似性の差を考慮する手法として，階層的クラスタリングの概念を導入する．

エリア間の階層的な関係を図 7.2 に例示する．図 7.2 はフランスのある地名に対してエリアのクラスタリングを行った例である．図 7.2 では，地名に対する各エリアの

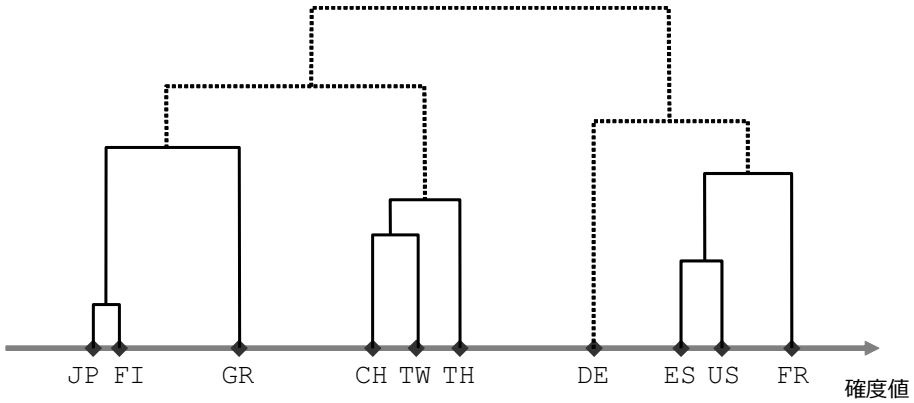


図 7.2: 階層的クラスタリング

確度値に基づいて実験対象エリア 10 エリアを x 軸上に並べている．階層的クラスタリングでは，初期段階では 1 個のエリアのみを含むクラスタを構成し，順々に，距離の近いクラスタをまとめていく．階層的クラスタリングは，類似度が高いエリアの間

では地名に対する確度値は互いに似た値となりやすく，そのため同一クラスタにまとまりやすいという特徴を利用するものである．階層的クラスタリングでは，すべての要素が 1 個にまとまるまで処理を続けることができる．そのため，どこまでをクラスタと認めるかの定義が必要である．

手法*4*の ACSb モジュールでは，適合率が期待できるブロックレベルの表層情報をベースとし，地名の確度値に基づいてエリアの階層的クラスタリングを行うことで所属エリア候補の絞込みを行う．この地名の確度値の算出法は，手法*2*の ACEb モジュールと同じである．階層的クラスタリングは，ウォード法 [73] を用いる．ウォード法では，クラスタ内の各値から，そのクラスタの重心までの距離を最小化するように，クラスタを統合する手法であり，クラスタ間の距離 $D(C_1, C_2)$ は式 (7.1) で定義される．

$$D(C_1, C_2) = E(C_1 \cup C_2) - E(C_1) - E(C_2) \tag{7.1}$$

ただし，$E(C_i)$ は，クラスタ内の各値 x とクラスタの重心 c_i の距離 $D(x, c_i)$ の 2 乗和である (式 (7.2))．

$$E(C_i) = \sum_{x \in C_i} (D(x, c_i))^2 \tag{7.2}$$

階層的クラスタリングでは，図 7.2 のように，類似性の高いエリアがより早くクラスタとして統合され，類似性の低いエリアが統合されるのは遅くなる．この特徴を利用してグループの概念を表現し，高い確度値を持つエリアの近辺を所属エリア候補として絞り込む．

手法*4*では，第 1 フェーズに当たる ACSb モジュールの階層的クラスタリングが終了した時点で，最大の確度値を持つエリアと同一のクラスタ内にあるエリアを，第 1 フェーズでの所属エリア候補とする．階層的クラスタリングでは階層的に要素が繋がり，自明なクラスタ境界が存在するわけではない．そのため，なんらかの方法でクラスタ境界を指定する必要がある．クラスタ境界の指定の方法としては，(1) 閾値による指定，(2) 要素数による指定，等が一般的である．第 6 章では，第 1 フェーズで適切に候補を絞り込むことによって，再現率 98% を確保するために必要な所属エリア候補の数は 3 個程度に抑えられることが示された．第 1 フェーズで十分な絞込みを行うことが第 2 フェーズの識別精度の向上に欠かせないため，第 2 フェーズで手法*3*の成果を活かせるよう，手法*4*でも第 1 フェーズで手法*3*と同等の絞込みを行うこととし，第 1 フェーズでの絞込み個数を最大 3 個とする．このため，手法*4*の ACSb モジュールでは，階層的クラスタリングの実行に際してクラスタ境界を要素数で指定することとし，最大の確度値を持つエリアを含むクラスタのサイズは 3 を超えないものとしてクラスタの統合を行う．この結果，第 1 フェーズでの出力エリア数は 1 個から 3 個と幅を持つ．

第 5 章で述べたとおり，地名は固有表現であるため，ブロックレベルでは十分な表層情報が引き出せない場合がある．このような場合には適切な確度値の算出ができず，極端な場合にはすべてのエリアに対して同一の確度値が算出される．第 1 フェーズで正解エリアが排除されることは望ましくなく，第 1 フェーズでは再現率の確保を保証する必要がある．そこで，すべてのエリアに対する確度値が同一の値であった場合には，

手法*4*の第 1 フェーズとして，ACSb モジュールの代わりに手法*3*の ACSc モジュールを用いて n-gram 情報を用いた生成確率による確度値の算出を行い，上位 3 個を第 1 フェーズの所属エリア候補とする．これにより，ブロックレベルの情報では判断しきれないような場合についても再現率を確保する．

7.1.3 類似エリアを対象とした所属エリア候補の削減

手法*3*では第 1 フェーズで出力される所属エリア候補数を 3 に固定した．それに対して，手法*4*の第 1 フェーズに当たる ACSb モジュールによって出力される所属エリア候補の数は，1 個から 3 個と幅を持つことになる．第 2 フェーズに当たる ACR モジュールは，手法*3*の ACR モジュールと同等のものであり，第 1 フェーズで出力される所属エリア候補群を対象とした機械学習器を用いて各エリアへの所属の可能性の判定を行う．手法*3*では ACR モジュールへの入力エリア数は 3 に固定されていたため，必要な機械学習器は 3 エリアを対象としたもののみであったが，手法*4*では入力エリア数が 1 から 3 とさまざまであり，それぞれに応じた対応が必要となる．

第 1 フェーズで出力される所属エリア候補数が 3 個の場合の処理は，手法*3*の ACR モジュールと同様である．第 1 フェーズで出力される所属エリア候補数が 2 個の場合，第 2 フェーズの ACR モジュールは，この 2 個のエリアだけを対象とした機械学習器を用いて所属エリア候補の削減処理を行う．機械学習器の学習対象が異なる[1]だけで，処理は手法*3*の ACR モジュールと同様である．第 1 フェーズで出力される所属エリア候補数が 1 個の場合，第 2 フェーズでの絞込みは不要と判断し，ACR モジュールによる判定を行わずにこの所属エリア候補をそのまま第 2 フェーズの出力とする．

7.1.4 対象エリア数の増加への対応

機械学習器を用いる手法では，処理対象とするエリアの組合せ毎に機械学習器を作成する必要がある．手法*3*では，第 1 フェーズでの所属エリア候補の出力数を固定し，対象エリア数が $|A|$，絞込み数が p の場合には，${}_{|A|}C_p \times p$ とおりの機械学習器を作成した．手法*4*では第 1 フェーズで出力される所属エリア候補数の上限のみを固定するため，この上限値を下回る個数の所属エリア候補群を対象とした機械学習器も作成することとなり，処理に必要な機械学習器の数はさらに増える可能性がある ($\sum_{p=2}^{p_{\max}}({}_{|A|}C_p \times p)$)．

処理に必要となる機械学習器の数は，実験対象エリア数 $|A|$ や 絞込み個数 p の値が大きくなるにつれて増加する．しかし，対象エリア数 $|A|$ に対して制限を設けることは所属エリア推定タスクの大前提となる全世界的な処理の実現に対する制約となり，提案手法の汎用性の面で大きな問題である．また，絞込み数 p に対して制限を設けることは，所属エリア推定手法の精度の低下をもたらす要因となりえる．この機械学習器を用いた手法は，スケーラビリティという観点で大きな制約を抱えているといえ，必要な機械学習器の数を抑止することが課題となる．

[1]この場合には，出力される 2 個の所属エリア候補それぞれについて，対象エリアの地名を正事例，もう 1 個のエリアの地名を負事例として学習を行った機械学習器を用いた判定を行う．

6.4.3 節で述べたとおり，第 1 フェーズでの絞込みの結果出力される所属エリア候補群の組合せには大きな偏りがあり，理論上可能なすべての組合せが出現するわけではない．そこで，出現の可能性が十分に高い組合せに着目し，これらについてのみ機械学習器を作成することにすれば，対象エリア数を減らさずに大幅に機械学習器の数を減少させることが期待できる．前述のように，手法*4*の ACR モジュールは機械学習器の学習対象エリア数に幅があるため，理論上必要になる機械学習器の数は手法*3*の ACR モジュールよりも増える場合がある．しかし，手法*3*では第 1 フェーズで出力される所属エリア候補数が 3 個に固定されているのに対して，手法*4*では第 1 フェーズで出力される所属エリア候補数はたかだか 3 個であり，必ず 3 個出力されるわけではない．手法*3*の第 1 フェーズでの絞込み処理は，再現率を保ったまま，類似すると思われる所属エリア候補を 3 個に絞り込むことを目的としていた．これは，再現率のみを考慮したアプローチであり，適合率を考慮したものでない．6.4.1 節では手法*3*の第 1 フェーズで所属エリア候補が 1 個に絞れる可能性が 85.96% に上ることが示されており，また 6.4.2 節では上位 2 個までの絞込みで 95.22% がカバーできるとの結果が得られていることから，出力する所属エリア候補数を 3 個に固定した手法*3*では，第 1 フェーズで本来排除できていたはずのエリアを候補に含めて所属エリア候補群を出力していたことがわかる．

判定対象とする所属エリア候補の数が減らせれば機械学習器の学習に必要な時間とデータの量の削減に繋がるほか，適合率の向上を見込めることが 6.4.4 節で示されている．したがって，適合率を重視するアプローチを導入することで第 1 フェーズで出力する所属エリア候補の数を適切に調整することが可能となれば，第 2 フェーズでの機械学習器の数を削減することができる．

手法*4*では，ACR モジュールでの処理の際，処理の対象とするエリアの組合せをあらかじめ限定することで，機械学習器の数を削減する．手法*4*では，第 2 フェーズの実行に先立ち，第 1 フェーズで出力する所属エリア候補の各組合せに対して，その出現割合を第 1 フェーズの出力結果を基に算出する．これを基に所属エリア候補の組合せを順位付けし，出現割合の順位が 1 位の組合せから累積出現割合があらかじめ設定したカバー率を満たすまでに出現した組合せを抽出する．これらの組合せについては，手法*3*の ACR モジュールと同様に，機械学習器による所属エリア候補の削減処理を行い，その結果を第 2 フェーズで出力する所属エリア候補とする．これにより，カバー率を調整することで，機械学習器の数の制御が可能となる．抽出されなかった組合せは，第 1 フェーズで出力される出現割合が十分に低いことを意味し，意味のない組合せで出力されたものと考えられる．そこで，このような組合せで出力された所属エリア候補群については，判定不能と判断し，第 2 フェーズでの判定処理を行わずにそのまま第 2 フェーズの出力とする．例えば，実験対象エリア数を 10，第 1 フェーズでの出力エリア数の範囲を 1 から 3 とした場合，理論上可能な組合せの数は 175 となり，またこの学習に必要な機械学習器の数は 450 となる． これらの 175 の組合せについてそれぞれ第 1 フェーズでの出現割合を算出し，この値の高い順に並べた場合，30 位までの組合せで累積出現割合があらかじめ設定したカバー率 95% を超えたとすれば，第 2 フェーズで用いる機械学習器はこの 30 の組合せに対してのみ作成する．その場合，

必要な機械学習器の数は最大でも 90 (30 個の組合せがすべて 3 エリア組の場合) であり，最低で 40 (30 個のうち 10 個が 1 エリアのみの出力で，残り 20 個が 2 エリア組の場合) となる．またこの場合，他の 145 の組合せで出力された場合は判定不能との扱いとして，第 1 フェーズの絞込み結果をそのまま第 2 フェーズの出力とする．第 1 フェーズで出力する所属エリア候補数はたかだか 3 個であるため，そのまま所属エリア推定を出力する場合にも妥当と判断できる．

7.2 実験結果

7.2.1 実験環境

地名コーパス

本実験の対象エリアは 6.1.1 節と同じ，中国，台湾，タイ，日本，ギリシャ，フィンランド，フランス，スペイン，ドイツ，アメリカの 10 エリアである．

ACSb モジュールの設定 (第 1 フェーズ)

ACSb モジュールの所属エリア候補の絞込み数は 1 個から 3 個とする．

ACR モジュールの設定 (第 2 フェーズ)

本章の実験では，実験対象エリア数を 10 とする．第 1 フェーズでの絞込み数を 1 個から 3 個としたため，ACR モジュールで必要となる機械学習器の数は最大で 450 種類 ($\sum_{p=2}^{3}({}_{10}C_p \times p)$) である．

7.2.2 実験結果

手法*4*での所属エリア推定実験の結果を，手法*1*，手法*2*，手法*1+2*，手法*3*との比較の形で示す (表 7.1)．結果として，手法*4*は全体の 95% のカバー率の場合で適合率

表 7.1: 手法*4*とその関連手法での所属エリア推定結果

手法	F 値	適合率 (%)	再現率 (%)
手法*1*	0.74	61.58	94.05
手法*2*	0.87	90.51	83.84
手法*1+2*	0.89	85.54	91.92
手法*3*	0.83	74.08	95.03
手法*4*	0.91	88.30	94.03

88.30%，再現率 94.03% を達成した．表 7.1 で示されるとおり，再現率，適合率を個々に見た場合にはそれぞれ手法*4*(カバー率 95%) の結果を上回る値を得た手法が見られるが，手法*4*では双方について同時に各手法での最高レベルの値を示している．手法*4*(カバー率 95%) の F 値は 0.91 となり，この手法の有効性が確認できた．F 値が 0.9 を超えたのは手法*4*のみであり，この手法が再現率，適合率共に十分高いことが示された．

図 7.3 に，さまざまなカバー率での所属エリア推定の処理の適合率 (precision) と再現率 (recall) の結果を，手法*4*(菱形) と手法*1*(丸印)，手法*1+2*(四角)，手法*3*(三角) の比較の形で記載する．手法*1*は文字レベル表層情報のみを用いた処理 (ACR)，手法*1+2*は

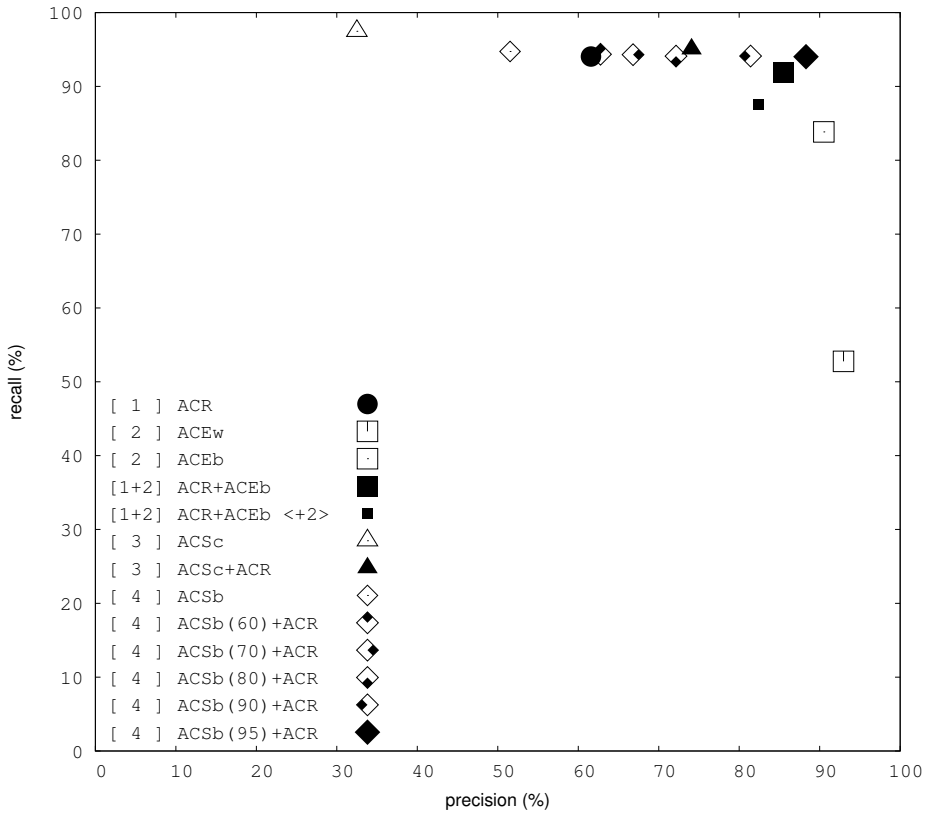


図 7.3: 手法*4*とその関連手法での所属エリア推定結果

ブロックを導入し再現率と適合率の双方を考慮した 2 段階処理 (ACR+ACEb)，手法*3*は類似エリア問題を考慮して文字レベル生成確率によるグループの絞込みを行う 2 段階処理 (ACSc+ACR) である．図 7.3 では，手法*4*の結果は左から，60%，70%，80%，90%，95% のカバー率の結果が並んだ．図 7.3 から，手法*4*はカバー率 60% で手法*1*と同等の結果を示し，カバー率 80% で手法*3*と同等の結果を示すことがわかる．また手法*4*では，カバー率を 95% にすることで，手法*1+2*の結果を上回る精度を実現している．類似エリア間では表層的な特徴を共有していることが多く，これらの分布には偏りがある．したがって，十分なカバー率を確保することで，多くの類似パターンを網羅することが可能となり，頻度の少ない組合せを枝刈りすることができる．

手法*4*では，第 2 フェーズで用いる機械学習器の数の削減も課題となっていた．実験結果から，全体の 95% のカバー率の場合で第 2 フェーズでの処理対象となった組合せ

数は 78 (理論値の 44.6%) であり，機械学習器の数は 182 個 (理論値の 40.4%) まで削減することができた．

次に，手法*4*でのエリア毎の実験結果を図 7.4 に示す．x 軸は適合率 (precision)，y 軸は再現率 (recall) を表している．最も高い F 値を得たエリアはタイ (0.93)，最も F 値が低くなったエリアは中国とスペイン (0.89) である．他の手法がエリア毎に F 値にばらつきが見られたのに対して，手法*4*ではエリア毎の F 値の値の差が小さい．図 7.4 によると，すべてのエリアで，再現率が 90% を越えていると同時に適合率も 85% を越えている．このように，すべてのエリアについて再現率，適合率の双方で安定して高い値を得ることに成功したことが，F 値の差に繋がっている．このことは，手法*4*がエリアに依らず十分な精度を期待できることを示しており，手法*4*は有効性，頑健性の双方を兼ね備えた所属エリア推定手法であるといえる．

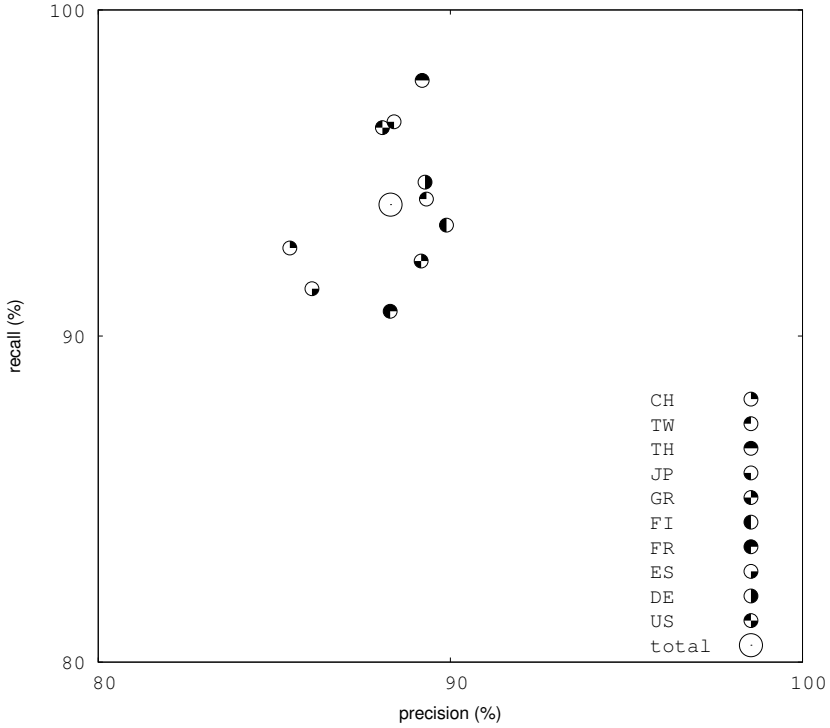


図 7.4: 手法*4*でのエリア毎の実験結果

7.3 考察

7.3.1 ブロックレベル情報を用いた所属エリア候補の絞込みの有効性

手法*4*の第 1 フェーズに当たる ACSb モジュールでの階層的クラスタリングによる絞込み結果の組合せの出現割合を表 7.2 に示す．表 7.2 は ACSb モジュールの出力をその出現割合で順位付けして記載したものであり，出現割合の高い組合せを上位 20 個まで記載した．上位 20 個までの組合せの累積出現割合は 72.46% であり，理論上可

表 7.2: ACSb モジュールによる絞込み結果

順位	出現割合 (%)	絞込み結果		
1	9.25	中国	台湾	
2	6.67	フランス	スペイン	
3	6.58	中国	台湾	タイ
4	5.56	アメリカ		
5	5.43	日本		
6	5.40	フィンランド		
7	5.16	ギリシャ		
8	4.52	タイ		
9	4.27	ドイツ		
10	3.28	フランス		
11	3.06	スペイン		
12	1.94	フランス	スペイン	アメリカ
13	1.72	フランス	アメリカ	
14	1.67	ドイツ	フランス	
15	1.50	タイ	フランス	
16	1.43	ドイツ	アメリカ	
17	1.32	中国		
18	1.26	ドイツ	フィンランド	
19	1.26	ギリシャ	フィンランド	
20	1.18	台湾		

能な 175 の組合せのうちこの 20 個の組合せで 第 1 フェーズの出力の 70% 以上をカバーしている．表 7.2 の 20 個の組合せのうち，出力が 1 エリアに絞られていたものが 10 個，出力が 2 エリアまで絞られていたものが 8 個で，3 エリア同時に候補として出力された組合せは 2 個のみであった．本実験の実験対象エリアは 10 エリアであり，表 7.2 の結果はこの 10 エリアすべてについて 1 個のエリアに絞り込むことに成功する割合が多いことを示している．ACSb モジュールによって，10 エリア中 8 エリアが上位 11 位までに 1 個のエリアに絞り込むことができている．また，10 個すべてのエリアが上位 20 位までに 1 個のエリアに絞り込むことができている．

ここでの実験は 10 エリアを対象としているため，正しく推定ができていると仮定した場合，中国または台湾と推定される出現割合は 20% となるはずだが，表 7.2 中，中国または台湾が含まれた組合せ 4 個の累計出現割合は 18.33% であり，このすべてが正しい出力と仮定すれば，上位 20 個までの組合せで約 90% の地名をカバーしたことになる．同様に，特定できた割合が比較的低いフランスとスペインも，この 2 エリアの組合せが 6.67% で 2 位となっており，この 2 エリア間の類似性が所属エリア推定に影響していることがわかる．

また，2 エリアまで絞り込んだ組合せでは，第 6 章で示したグループ分けに合わない組合せは，15 位のタイとフランスの組合せと，18 位のドイツとフィンランドの組合せのみである．3 エリア同時に候補として出力された組合せは，片方はグループ A のエリアをすべて含む組合せであり，もう一方はグループ B の 4 エリアのうち互いの類似性が低いと判断されたドイツとスペインのうちドイツが入らない組合せであった．これらのことから，ほとんどの地名について第 1 フェーズで類似する 2 エリア程度まで絞り込むことが可能であることが推測され，第 2 フェーズを類似エリア間の識別フェーズとする手法*4*のアプローチが適切に働いていることがわかる．

7.3.2 類似エリアを対象とした所属エリア候補の削減の妥当性

この機械学習器を用いる手法では処理対象とする組合せの数だけ機械学習器を作成する必要があり，組合せの数の増加に伴って必要な機械学習器の数が増加するという課題があった．本節では，手法*4*を手法*3*と比較する形で，実験結果を基に機械学習器の数の削減のための組合せの絞込みについて考察する．

表 7.2 の上位 20 個の組合せの出現割合の合計は 72.46% である．第 1 フェーズでエリアが 1 個に絞り込まれた場合，第 2 フェーズの処理は行われないため，絞込み結果として 2 個以上のエリアが残っている場合にのみ，第 2 フェーズの機械学習器を用いた ACR モジュールによる処理が行われる．機械学習器の数を抑制することを目的に，例えば，上位 20 個の組合せでカバーされる 72.46% の地名に対して第 2 フェーズの処理が行われるようにする場合，1 個に絞り込まれた場合を除いた 22 個の機械学習器を準備することになる．これは，すべての組合せについて機械学習器を作成した場合の 5% 程度に削減することができることを示している． 手法*3*ではすべての 3 個のエリアの組合せに対して機械学習器を作成する必要があったが，表 7.2 から，第 1 フェーズで所属エリア候補を 3 個以下に絞り込めるケースは十分多く，手法*3*よりも機械学習器の数を削減できる可能性が示された．

図 7.5 に，絞込み後の組合せ数に対して必要となる機械学習器の数の変化を示す．x 軸は絞込み後に機械学習器を準備する対象となる組合せの数 (number of combinations)，y 軸はカバー率 (coverage) と必要となる機械学習器の数 (number of classifiers) を表している．水平の線は手法*3*での機械学習器の数 (360) を示し，実線は組合せ数毎の手法*4*での機械学習器の数を示している．また点線は，組合せ数に対する手法*4*での絞込み結果のカバー率を表す．手法*4*の実験結果から，出現割合の高い組合せは 1 エリアと 2 エリアの組合せの割合が高いことがわかっており，ある程度以降の出現割合で出現する組合せはほぼ 3 エリア組のみである．そのため，図 7.5 の機械学習器の数は組合せ数の増加にしたがってほぼ線形に増加する．

手法*4*の実験結果では，第 1 フェーズの出力全体の 60% をカバーするように組合せを選択する場合には上位 12 個の組合せのみ抽出すればよく，この場合必要となる機械学習器の数は 18 である．この値はすべての組合せについて機械学習器を作成する場合の 4% に過ぎず，対象エリア数の増加に伴う機械学習器の数の大幅な増加の問題は解消している．同様に，95% のカバー率を確保するには，上位 78 個の組合せを選択す

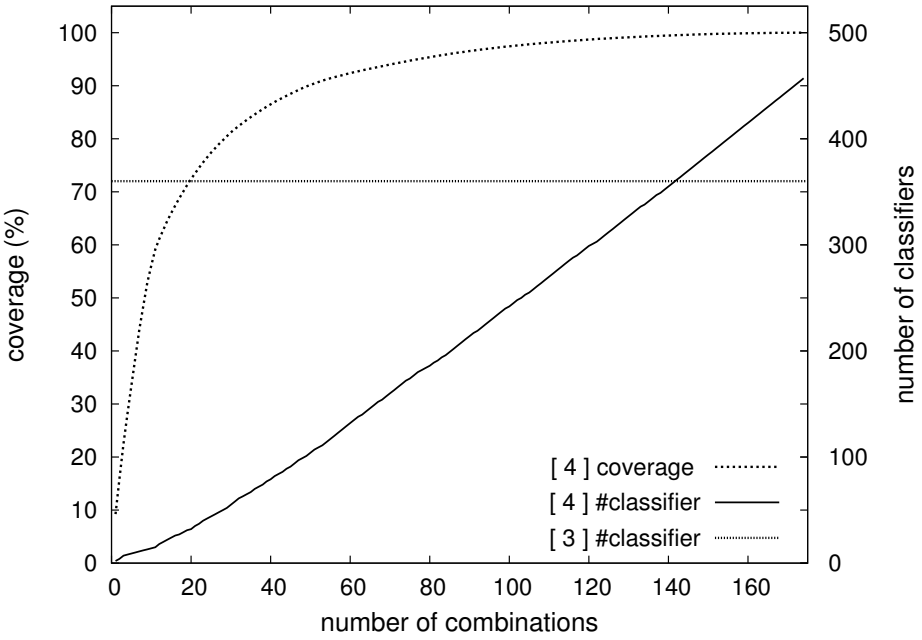


図 7.5: 機械学習器の数と組合せカバー率

ればよく，このとき必要となる機械学習器の数は 182 であり，この場合は機械学習器の最大数の 40.4% を必要とすることになる．この場合でも，機械学習器の数は手法*3*で必要とした数の 50.6% に過ぎず (図 7.5)，カバー率を十分に上げた場合でも手法*4*の絞込み手法で十分に対応できることがわかる．このことから，手法*4*は第 1 フェーズでの所属エリア候補の絞込みを適切に行っており，その結果第 2 フェーズで用いる機械学習器を削減することに成功しているといえる．

7.3.3 対象エリア数の増加の影響

地名の所属エリア推定タスクでは，理想的には全世界的な所属エリア推定が期待されるため，実験対象エリアの数の増大が課題となる．本節では，実験対象エリア数を増やして手法*4*を適用した場合の実験結果を示し，この問題について検討する．

7.2 節の実験の対象エリアに，新たに 10 のエリアを加え，20 エリアで実験を行った結果を図 7.6 に示す．x 軸は適合率 (precision)，y 軸は再現率 (recall) を表している．本節での所属エリア推定の実験対象エリアは，中国，チリ，エジプト，フィンランド，フランス，ドイツ，ギリシャ，インド，イタリア，日本，韓国，マレーシア，オランダ，ポルトガル，ロシア，スペイン，タイ，台湾，イギリス，アメリカの 20 エリアである (表 3.1)．この実験対象エリア群には，中国と台湾，スペインとチリ，イギリスとアメリカの 3 組の同一言語圏エリアを含めたほか，地理的，歴史的要因から類似エリアと考えられるエリア群を含めている．7.2 節では手法*3*の結果も比較対象としているが，実験対象エリア数を 20 に増やした本実験では手法*3*で必要とする機械学習器の数が大きくなり実行が難しいため，図 7.6 の比較には手法*3*は含まない．図 7.6 では，これまでの結果と同様に，再現率は手法*1*(丸印) がよい値を示すこと，適合率では手法*2*およ

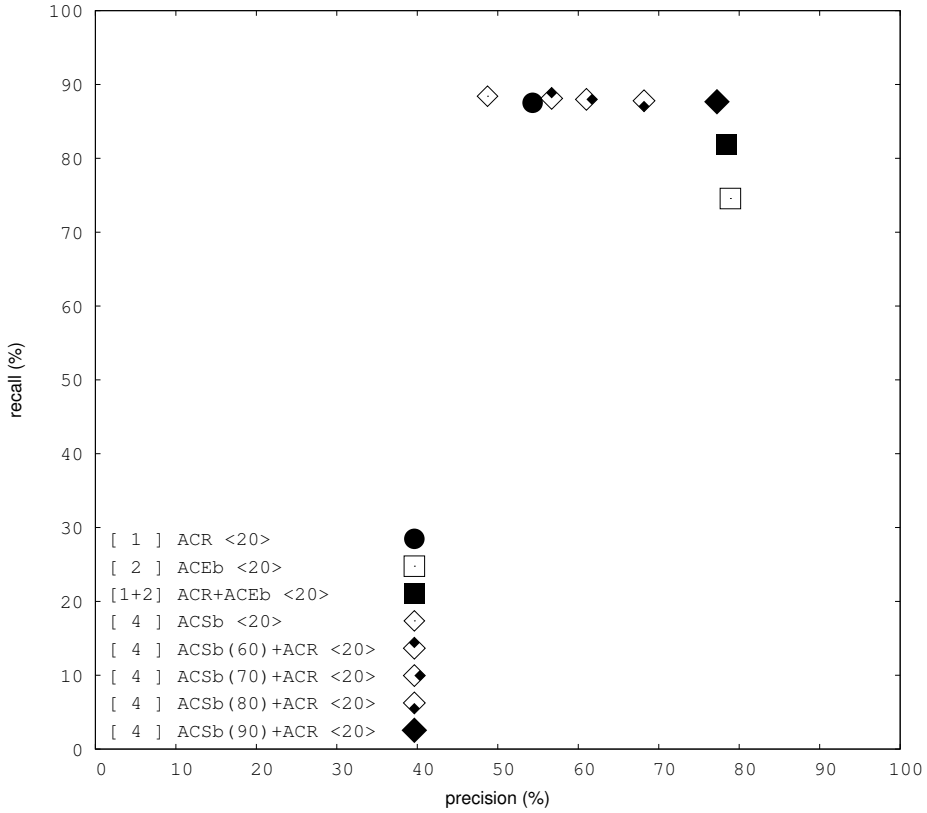


図 7.6: 実験対象エリアを増やした場合の実験結果

び手法*1+2*(四角) がよい値を示すこと，そして手法*4*(菱形) が再現率をほぼ保ったまま高い適合率を示すことがわかる．このように，実験対象エリア数が倍増しても各手法の傾向は変わらない．

実験対象エリア数が 10 の場合と 20 の場合での実験結果の変化を図 7.7 に示す．x 軸は適合率 (precision)，y 軸は再現率 (recall) を表している．10 エリアを対象した実験ではカバー率 90% で F 値 0.87，再現率 94.11%，適合率 81.41% だったのに対して，実験対象エリア数を 20 エリアに増やした場合では，カバー率 90% で全体の F 値は 0.82 であり，10 エリアでの F 値より劣るものの，十分な結果を示した．20 エリアに増やした場合では，再現率は全体で 87.66%，適合率は 77.23% となっており，この双方で値を落としたことが F 値の低下の原因である．

実験対象エリア数を 20 に増やした場合のエリア毎の実験結果を図 7.8 に示す．この実験では，実験対象エリアそれぞれについて他の実験と同様に各エリア 10,000 個の地名からなる地名コーパスを作成し，この地名コーパス群から抽出した表層情報を用いて実験を行っている．同一エリア内の地名には重複は許さないが，エリア間での同一地名の共有は許している．

実験対象エリア数を 20 エリアに増やした場合，10 エリアの場合 (図 7.4) と比較してエリア間の結果の差が大きくなる傾向が見られる．しかし，再現率についてはすべてのエリアで 70% を超えており，適合率もすべてのエリアで 60% 以上を得ることができた．

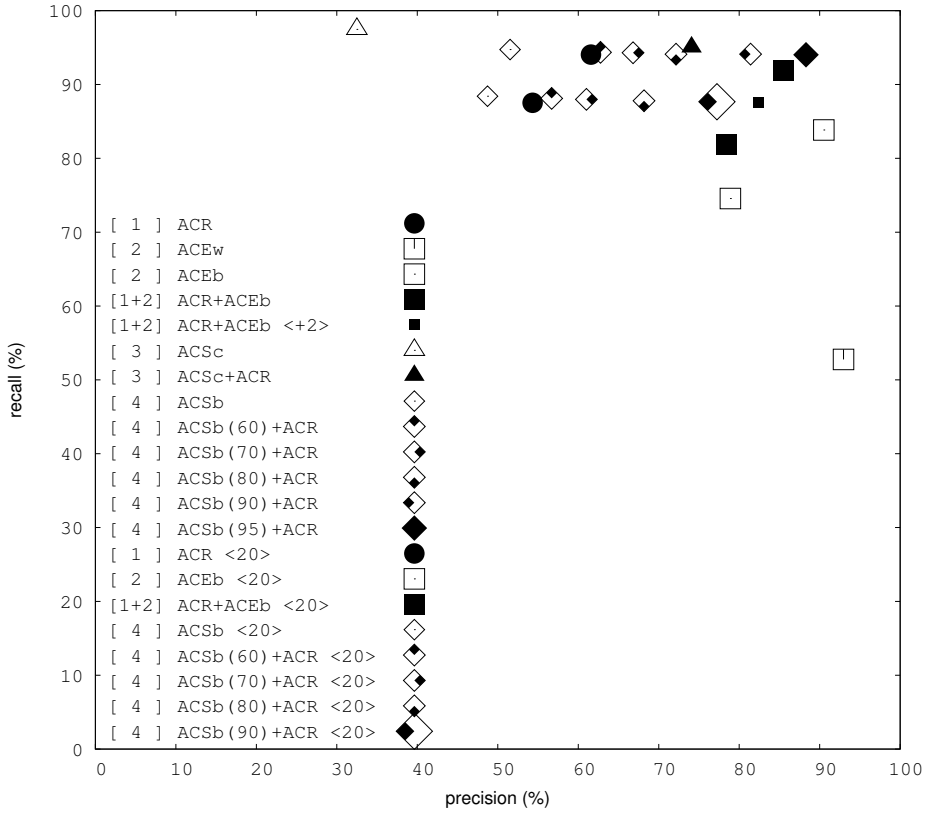


図 7.7: 実験対象エリア数の変化による実験結果の変化

同一言語圏の 3 組のエリア対のうち，中国と台湾は，それぞれ F 値 0.86 と 0.89 を得ており，他のエリアと比べて比較的よい結果となった．この 2 エリアは他のエリアからの類似性が相対的に低く，ほとんどの地名について第 1 フェーズでは正解エリアが特定されるかこの 2 エリア組と判定されるかの出力結果となった (7.3.1 節) ために，第 2 フェーズでこの 2 エリアに特化した識別が行われたことがよい結果をもたらしたといえる．これは，手法*4*の主張と合致する結果であり，手法*4*のアプローチが適切に働いた好例といえる．同様に，イギリスとアメリカは，それぞれ F 値 0.81 と 0.85 を示し，他のエリアと同等の精度を達成した．

それに対して，スペインとチリはそれぞれ F 値 0.70 と 0.81 を示し，チリは他のエリアと同等の結果を得たものの，スペインについては他のエリアと比べて低い値となった．特にスペインの F 値は 20 エリア中最低だが，これはスペインの適合率が 62.81% と 20 エリアの中で特に低い値を示したこと，再現率も 79.29% と相対的に低い値だったことが影響している．これに対してチリは再現率 87.16%，適合率 75.87% と他のエリアと遜色ない結果を示している．スペインについては，10 エリアを対象としたこれまでの実験でもフランスとの混同による F 値の低下が指摘されていた．さらに，20 エリアを対象とした実験ではフランスおよびスペインに言語的，地理的に近いエリアとしてイタリアが追加されている．フランスの再現率は 79.14%，適合率は 73.29% と他のエリアに比べて低い値となっており，イタリアの実験結果はフランスとほぼ同等 (再現率 81.19%，適合率 72.37%，F 値 0.77) であった．このことから，スペイン，フラ

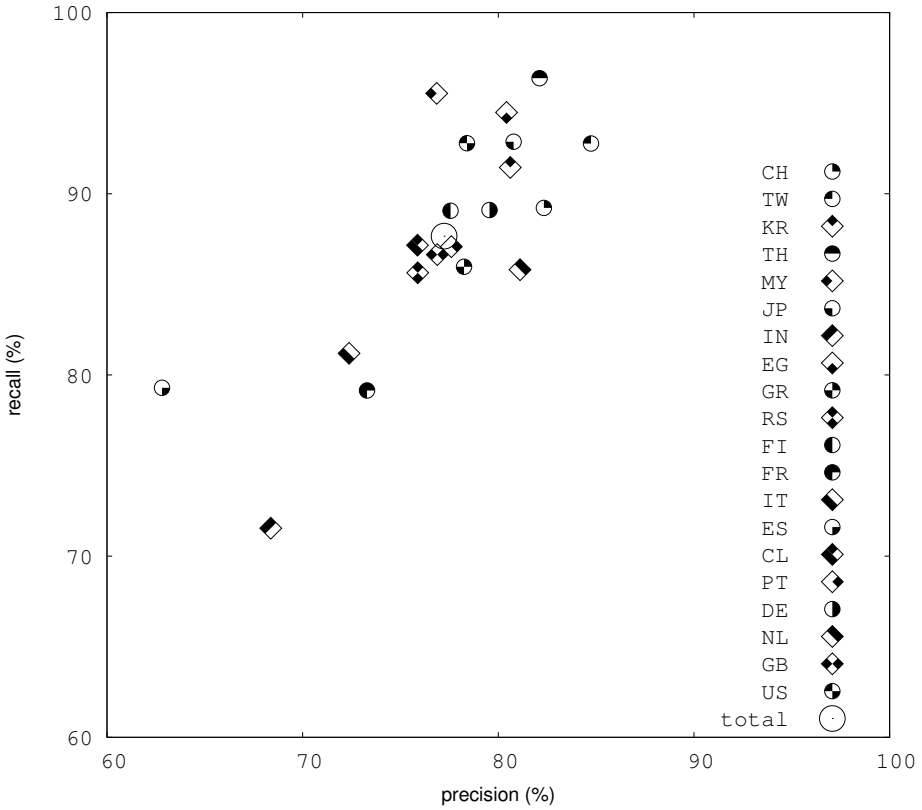


図 7.8: 実験対象エリア数を増やした場合の手法4のエリア毎の実験結果

ンス，イタリアの互いの識別が不十分であり，この 3 エリアの所属エリア推定の精度が低下したものと考えられる．特にスペインについては，地理的に近いエリアとの混同と言語的に近いエリアとの混同の双方が影響したため，他のエリアに比べて大きく F 値が低下した．これらから，類似エリアの数や性質によっては F 値の低下が見られ，さらなる改善の余地があることが示される．

このほか，低い F 値となったエリアとしてインドがある (F 値 0.70)．インドは再現率 71.54%，適合率 68.37% であり，他のエリアと比べて再現率が大幅に劣ることが F 値の低下の要因である．インドは広い国土を持ち，多くの言語が利用されている．また，インドの地名は，地名あたりの文字数が約 8.7 であり，他のエリアと比較して小さなものであった．これらの要因から，インドというひとつのエリアとしての特徴を地名コーパスから抽出することが困難であったことが，低い再現率につながったものと考えられる．

これに対して，結果がよかったのは台湾 (F 値 0.89)，エジプト (F 値 0.87) であった．中国，韓国，日本，マレーシア，アメリカもそれぞれ 0.85 以上の F 値を得ており，アジア圏のエリアが比較的よい結果を得た．これらのエリアに共通する要因は，高い再現率である．これらのエリアでは，中国の再現率が 89.22% と 90% に届かなかったほかはすべて 90% 以上となっており，このことが高い F 値に繋がっている．反面，適合率については他のエリアと比べて比較的高い傾向がある程度で，必ずしも高い値を得ているわけではなかった．このことから，再現率の向上が所属エリア推定処理の精

度の向上に重要な要因であることが推測できる．全体として，低い F 値 (0.70) となった 2 エリア (インド，スペイン) 以外の 18 エリアは F 値の範囲は 0.76 から 0.89 の間に収まり，全体での F 値 (0.82) との差はたかだか 0.07 との結果となった．このことから，手法*4*は実験対象エリア数の増加にも十分対応できるものといえる．

今回の実験では，実験対象エリアの増加に伴い類似エリアの割合が大きく増えたが，表層情報の取得に用いる地名コーパスのサイズ (地名コーパスを構成する地名の数) はこれまでの実験と同じである．地名コーパスのサイズを大きくすることで，さらにきめ細かい表層情報の利用が可能となる可能性がある．本実験ではコーパスサイズの妥当性や地名コーパスのサイズのエリア毎のばらつきの影響についての検討により，処理精度のさらなる改善が考えられ，これらは今後の課題として挙げられる．

手法*3*のように実験対象エリア数の増加に伴って処理に用いる機械学習器の数が増大する手法では，実験対象エリア数の増加への対応は難しい．それに対して，第 1 フェーズでの絞込み結果にしたがって必要な機械学習器の数を調整する手法*4*では実験対象エリア数が増加しても機械学習器の数が爆発的に増加することは抑えられる．カバー率を 90% とした場合，手法*4*の第 2 フェーズで処理対象とする組合せの数は 205 種類となり，機械学習器の数は 471 となった．実験対象エリア数を 20，第 1 フェーズでの出力エリア数を 1 個から 3 個とした今回の実験では，すべての組合せに対して機械学習器を作成すると 3,800 個となり，実際に作成した機械学習器の割合はその 12.4% に抑えられた． 手法*3*を利用する場合の機械学習器の数と比較してもその 13.8% に過ぎず，手法*3*では実行が困難だった 20 エリアを対象とした実験が手法*4*では問題なく行えたのはこのためである．実験対象エリア数が 10 の場合にはカバー率 90% では，機械学習器が作成された組合せ数は 50 であった．これは全組合せ数の 11% に相当する．実験対象エリア数が 20 の場合では，カバー率 90% での組合せ数は 205 で，全組合せ数の 15% に相当する． 実験対象エリア数が増えることで類似エリアの割合は増加することが予想されるため，特定のカバー率を達成するために必要な組合せの全組合せに対する割合は少しずつ増えることが予想されるが，手法*4*では十分な絞込みを実現しているためその増加幅はかなり抑えられる．手法*4*では，カバー率を高くすることで，これまでの手法の特徴を損なうことなく，高い再現率と高い適合率を確保できる．また，実験対象エリア数が増加し第 2 フェーズでの処理対象となる出力候補群の組合せの増加が予想される場合には，カバー率を抑えることで現実的な時間での処理が可能である．

7.4 手法*4*のまとめ

地名の表層情報のみを用いた頑健な所属エリア推定を目的として，第 4 章では表層情報の利用による再現率の高い所属エリア推定手法について検討し，第 5 章では高い適合率を実現するためブロックという新しい処理単位を提案した．また第 6 章では，類似エリアの問題の解決方法を示した．本章で提案した手法*4*は，これらの研究結果を統合すると同時に，実験対象エリア数についてのスケーラビリティの問題の解決を図るものである．

手法*4*では，第 1 フェーズで所属エリア候補を類似性の高いエリア群のみに絞り込み，第 2 フェーズで類似性の高いエリア間の識別を行うことで地名の所属エリア推定を実現した．第 1 フェーズでは再現率が重視されるが，このフェーズで十分妥当と判断される所属エリア候補のみを出力することが第 2 フェーズの負荷に大きく影響するため，第 1 フェーズに確度の高いブロックレベルの表層情報の利用を導入し，再現率と適合率の双方を備えた絞込みを実現した．この結果に対して第 2 フェーズで機械学習器による所属エリア推定を行うことで，再現率，適合率共に高い値を達成し，エリアによらず安定した推定を行うことに成功した．さらに，第 2 フェーズでは第 1 フェーズでの出力の妥当性の判定を行っており，十分な妥当性を認めた結果に対して処理を行う形とすることで，負荷の軽減と高い精度とを実現した．

実験の結果，10 エリアを対象とした所属エリア推定実験では，F 値 0.91，適合率 88.30%，再現率 94.03% を達成した．また，実験対象エリア数の増加の影響を調べるため，同じ条件でエリア数 10 の場合と 20 の場合を比較する．10 エリアを対象とした実験で F 値 0.87，適合率 81.41%，再現率 94.11% だったのに対して，20 エリアを対象とした実験でも F 値 0.82，適合率 77.23%，再現率 87.66% を達成し，実験対象エリア数の増大により類似エリアの含有率が高くなった場合でも十分な結果が得られることを示した．手法*3*で問題となった実験対象エリア数の増加に伴う負荷の増大については，手法*4*では第 1 フェーズでの適切な絞込みにより十分に負荷を抑えることに成功し，スケーラビリティに関する課題も解決した．

これらの実験結果より，手法*4*がエリアの類似性やエリア数の変化に強い汎用的な所属エリア推定の手法であることが示された．手法*4*は地名辞書やヒューリスティクスを用いることのない高い頑健性を持つ手法であり，地名の所属エリア推定の手法として現実的かつ有効な手法である．

第8章　おわりに

8.1　結論

本研究は，自然言語処理の課題である未知語処理の積極的な解決を目的とし，固有表現の属性推定手法の提案およびその有効性の評価を行うものである．

固有表現等の外国語句を含む文書は数多く存在するが，現在の自然言語処理では一般に単一言語を前提として処理を行うため，文書中に埋め込まれた外国語句は未知語として扱われてきた．自然言語処理では精度の面から辞書ベースの手法が主流となっている．外国語句が未知語して扱われるのはこのためであるが，その解決方法として外国語句すべてを辞書に登録することは現実的ではなく，辞書ベースの手法には限界がある．

文書に埋め込まれた外国語句は固有表現の場合が多く，言語推定よりも所属エリア推定が必要とされる．従来の所属エリア推定は，辞書引きによる検索処理と文脈情報を用いた曖昧性解消処理の組合せであり，辞書やヒューリスティクスへの依存度が高く，頑健性や汎用性の面で問題があった．それに対して本研究は，固有表現を構成する文字の並びに着目し，そこから簡単に抽出することが可能である表層的な情報のみを用いることで，言語や文脈に依存しない，頑健性と汎用性の高い所属エリア推定の手法を提案するものである．

本研究では，第一に，地名の表層的な情報を用いた所属エリア推定の基本的な手法を提案した．第二に，効果的な表層情報の利用を目的として，言語的な特徴を有しつつ十分な量の情報を抽出できるブロックと呼ぶ新たな単位を導入し，低い適合率を改善することに成功した．第三に，類似した地名を有するエリアが複数存在するという地名特有の問題に取り組み，類似エリアを含む所属エリア推定においても再現率を下げることなく適合率を向上させることに成功した．最終的に，これらの手法を効果的に組み合わせることで，辞書情報や文脈情報に依存しない頑健性とエリア数の変化や類似エリアの有無に対応できる汎用性を兼ね備えた所属エリア推定手法を実現した．類似エリアを含む 10 個のエリアを対象とした実験で適合率 88.30%，再現率 94.03%，F 値 0.91 を達成し，表層情報のみに基づく頑健な地名の所属エリア推定を実現した．語句の短さや類似エリアの存在等の地名の特殊性を考慮すると，この結果は，辞書やヒューリスティクスに頼らない手法にもかかわらず十分よい結果であると同時に，有効性も十分に有するものであると結論付ける．

8.2 今後の展望

本研究では，これまで未知語として扱われてきた固有表現に着目し，これに意味的な属性を自動的に付与する手法を提案した．本研究は外国語句からの有用な情報の抽出を可能にするものであり，属性情報を取得することが困難な場合が多い固有表現から自動的に属性を引き出すことで，機械翻訳や情報抽出等への適用等が期待できる．これにより，自然言語理解や意味解析等の高次の自然言語処理を行う際に，固有表現の属性を積極的に活用することが可能となり，その精度の向上や可能性の拡充が期待できる．地名に限らず固有表現はさまざまな情報とリンクしており，これらを積極的に活用することで，固有表現を単に字面で処理するだけでなく，その意味も補うという高レベルの処理が可能となる．本研究では固有名詞の中で特に地名に着目したが，表層情報のみを利用する本手法は固有名詞に限らず一般的な語句も含めて広く応用することが可能である．本研究は，固有表現を含めた語句の属性推定を可能とすることで，高レベルで実用的な自然言語処理の実現に向けて，その一助を成すものであると考える．

謝辞

本研究は，著者が慶應義塾大学大学院理工学研究科後期博士課程在学中に，同大学理工学部斎藤博昭准教授の御指導の下で行ったものです．斎藤先生には，卒業論文，修士論文に引き続き，博士論文の御指導を頂き，研究内容の相談，学会での発表，論文の書き方，推敲等，研究生活すべてにおいてたいへんお世話になりました．これまでの研究を博士論文の形でまとめることができたのは，熱心で親身な御指導のおかげであり，感謝の念にたえません．また，故 中西正和教授には，残念ながら博士論文の御指導をいただくことは叶いませんでしたが，卒業論文，修士論文でいただいた御指導は，本論文にも生きているものと信じています．

本論文の審査の過程において，櫻井彰人教授，高田眞吾准教授，小原京子准教授には数々の御助言と御指導を賜りました．お忙しい中，副査を引き受けて頂いた先生方に深謝申し上げます．

斎藤研究室の皆様にも，本論文をまとめるにあたり，たいへんお世話になりました．特に，岡本紘幸君，鈴木宏哉君，松原正樹君，菊地弘晶君，延澤志保さんは，勉強会のメンバとして積極的に支えてくださり，心より感謝いたします．本研究の内容を深めることができたのは，勉強会での継続的な議論のおかげであり，たいへん有益なものであったと確信しています．

最後に，この研究室で博士論文をまとめ上げることができたことを幸せに思います．ありがとうございました．

2010 年 3 月

佐 野　智 久

参考文献

[1] Einat Amitay, Nadav Har'El, Ron Sivan, and Aya Soffer. Web-a-Where: Geotagging Web Content. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '04)*, pp. 273–280, New York, NY, USA, 2004. ACM.

[2] Sophia Ananiadou. A Methodology for Automatic Term Recognition. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING '94)*, Vol. 2, pp. 1034–1038, 1994.

[3] Shlomo Argamon, Ido Dagan, and Yuval Krymolowski. A Memory-Based Approach to Learning Shallow Natural Language Patterns. In *Proceedings of the 17th International Conference on Computational Linguistics (COLING-ACL '98)*, pp. 67–73, August 1998.

[4] Kay M. Berkling and Etienne Barnard. Language Identification of Six Languages Based on a Common Set of Broad Phonemes. In *Proceedings of the 3rd International Conference on Spoken Language Processing (ICSLP '94)*, pp. 1891–1894, September 1994.

[5] John D. Burger, John C. Henderson, and William T. Morgan. Statistical Named Entity Recognizer Adaptation. In *Proceedings of the 6th Conference on Natural Language Learning*, pp. 1–4, Morristown, NJ, USA, 2002. Association for Computational Linguistics.

[6] Hsin-Hsi Chen, Sheng-Jie Huang, Yung-Wei Ding, and Shih-Chung Tsai. Proper Name Translation in Cross-Language Information Retrieval. In *Proceedings of the 17th International Conference on Computational Linguistics (COLING-ACL '98)*, pp. 232–236, August 1998.

[7] Hsin-Hsi Chen and Jen-Chang Lee. Identification and Classification of Proper Nouns in Chinese Text. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING '96)*, pp. 222–229, August 1996.

[8] Stanley F. Chen and Joshua Goodman. An Empirical Study of Smoothing Techniques for Language Modeling. In *Proceedings of the 34th Annual Meeting on Association for Computational Linguistics*, pp. 310–318, Morristown, NJ, USA, 1996.

[9] Paul Clough. Extracting Metadata for Spatially-aware Information Retrieval on the Internet. In *Proceedings of the 2005 ACM Workshop on Geographic Information retrieval (GIR '05)*, pp. 25–30, 2005.

[10] Silviu Cucerzan and David Yarowsky. Language Independent Named Entity Recognition Combining Morphological and Contextual Evidence. In *Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pp. 90–99, 1999.

[11] Silviu Cucerzan and David Yarowsky. Language Independent NER Using a Unified Model of Internal and Contextual Evidence. In *Proceedings of the 6th Conference on Natural Language Learning*, pp. 1–4, Morristown, NJ, USA, 2002. Association for Computational Linguistics.

[12] James R. Curran and Stephen Clark. Language Independent NER Using a Maximum Entropy Tagger. In *Proceedings of the 7th Conference on Natural Language Learning*, pp. 164–167, 2003.

[13] Susan Dumais, John Platt, David Heckerman, and Mehran Sahami. Inductive learning algorithms and representations for text categorization. In *Proceedings of the 7th International Conference on Information and Knowledge Management (CIKM '98)*, pp. 148–155, New York, NY, USA, 1998. ACM.

[14] Ted Dunning. Statistical Identification of Language. Technical Report MCCS 94-273, Computer Research Laboratory Technical Report, March 1994.

[15] Katerina T. Frantzi and Sophia Ananiadou. Automatic Term Recognition Using Contextual Cues. In *Proceedings of the 3rd DELOS Workshop*, 1997.

[16] 福本淳一, 下畑光夫, 桝井文人. 固有名詞抽出における日本語と英語の比較. 情報処理学会研究報告, Vol. 98-NL-126, pp. 107–114, 1998.

[17] Pascale Fung. Extracting Key Terms from Chinese and Japanese Texts. *The International Journal on Computer Processing of Oriental Language, Special Issue on Information Retrieval on Oriental Languages*, Vol. 12, No. 1, pp. 99–121, 1998.

[18] Eric Garbin and Inderjeet Mani. Disambiguating Toponyms in News. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT '05)*, pp. 363–370, 2005.

[19] Gregory Grefenstette and Julien Nioche. Estimation of English and non-English Language Use on the WWW. In *Proceedings of Recherche d'Information Assistée per Ordinateur (RIAO 2000)*, pp. 237–246, 2000.

[20] Alexander G. Hauptmann and Andreas M. Olligschlaeger. Using Location Information from Speech Recognition of Television News Broadcasts. In *Proceedings of the ESCA ETRW Workshop on Accessing Information in Spoken Audio*, pp. 102–106, 1999.

[21] 久光徹, 丹羽芳樹. 辞書と共起情報を用いた新聞記事からの人名獲得. 情報処理学会研究報告, Vol. 97-NL-118, pp. 1–6, March 1997.

[22] 池原悟, 白井諭, 河岡司. 大規模日本語コーパスからの連鎖型および離散型の共起表現の自動抽出法. 情報処理学会論文誌, Vol. 36, No. 11, pp. 2584–2596, November 1995.

[23] 磯崎秀樹, 賀沢秀人. 自然言語処理のための SVM 高速化. 電子情報通信学会技術研究報告. PRMU, パターン認識・メディア理解, Vol. 102, No. 317, pp. 17–22, September 2002.

[24] 北研二. 言語と計算 (4) 確率的言語モデル. 東京大学出版会, November 1999.

[25] R. Kneser and H. Ney. Improved Backing-Off for M-gram Language Modeling. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Vol. 1, pp. 181–184, May 1995.

[26] 工藤拓, 松本裕治. Support Vector Machine を用いた Chunk 同定. 情報処理学会研究報告, Vol. 2000-NL-140, No. 170, pp. 9–16, November 2000.

[27] Susana Ladra, Miguel R. Luaces, Oscar Pedreira, and Diego Seco. A Toponym Resolution Service Following the OGC WPS Standard. In *Proceedings of the 8th International Symposium on Web and Wireless Geographical Information Systems (W2GIS '08)*, pp. 75–85, Berlin, Heidelberg, 2008. Springer-Verlag.

[28] Jochen L. Leidner. Toponym Resolution: A First Large-Scale Comparative Evaluation. Research Report EDI–INF–RR–0839, School of Informatics, University of Edinburgh, Edinburgh, Scotland, UK, July 2006.

[29] Jochen L. Leidner. Toponym Resolution in Text: Annotation, Evaluation and Applications of Spatial Grounding. *SIGIR Forum*, Vol. 41, No. 2, pp. 124–126, 2007.

[30] Jochen L. Leidner. *Toponym Resolution in Text: Annotation, Evaluation and Applications of Spatial Grounding of Place Names.* PhD thesis, University of Edinburgh, Scotland, UK, 2007.

[31] Huifeng Li, Rohini K. Srihari, Cheng Niu, and Wei Li. Location Normalization for Information Extraction. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING '02)*, pp. 1–7, 2002.

[32] Huifeng Li, Rohini K. Srihari, Cheng Niu, and Wei Li. InfoXtract Location Normalization: A Hybrid Approach to Geographic References in Information Extraction. In *Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References*, pp. 39–44, 2003.

[33] Rafael Dueire Lins and Paulo Goncalves. Automatic Language Identification of Written Texts. In *Proceedings of the 2004 ACM Symposium on Applied Computing (SAC '04)*, pp. 1128–1133, 2004.

[34] Kim Teng Lua. Experiments on the Use of Bigram Mutual Information for Chinese Natural Language Processing. In *Proceedings of International Conference on Computer Processing of Oriental Languages (ICCPOL)*, pp. 23–25, February 1995.

[35] Robert Malouf. Markov Models for Language-Independent Named Entity Recognition. In *Proceedings of the 6th Conference on Natural Language Learning*, pp. 1–4, Morristown, NJ, USA, 2002. Association for Computational Linguistics.

[36] Christopher D. Manning and Hinrich Schuetze. *Foundations of Statistical Natural Language Processing*. The MIT Press, June 1999.

[37] Bruno Martins and Mário J. Silva. Language Identification in Web Pages. In *Proceedings of the 2005 ACM Symposium on Applied Computing (SAC '05)*, pp. 764–768, 2005.

[38] D. Matrouf, M. Adda-Decker, L. F. Lamel, and J. L. Gauvain. Language Identification Incorporating Lexical Information. In *Proceedings of the 5th International Conference on Spoken Language Processsing (ICSLP '98)*, pp. 181–184, December 1998.

[39] Paul McNamee and James Mayfield. Entity Extraction without Language-Specific Resources. In *Proceedings of the 6th Conference on Natural Language Learning*, pp. 1–4, Morristown, NJ, USA, 2002. Association for Computational Linguistics.

[40] Andrei Mikheev, Marc Moens, and Claire Grover. Named Entity Recognition without Gazetteers. In *Proceedings of the 9th Conference on European Chapter of the Association for Computational Linguistics*, pp. 1–8, Morristown, NJ, USA, 1999. Association for Computational Linguistics.

[41] 永田昌明. 未知語の確率モデルと単語出現頻度の期待値に基づくテキストからの語彙獲得. 情報処理学会論文誌, Vol. 40, No. 9, pp. 3373–3386, 1999.

[42] 中渡瀬秀一, 木本晴夫, 中川優. 統計的手法による辞書未登録語の獲得法. 電子通信学会論文誌, Vol. J81-D-II, No. 2, pp. 238–248, February 1998.

[43] 延澤志保, 斎藤博昭, 中西正和. 文字の共起情報のみを利用した文字列抽出. 情報処理学会研究報告, Vol. 98-NL-127, pp. 141–148, September 1998.

[44] Shiho Nobesawa, Hiroaki Saito, and Masakazu Nakanishi. String Extraction Based Only on Statistic Linkability. In *Proceedings of the 18th International Conference on Computer Processing of Oriental Languages (ICCPOL'99)*, pp. 23–28, March 1999.

[45] Shiho Nobesawa, Hiroaki Saito, and Masakazu Nakanishi. Automatic Semantic Sequence Extraction from Unrestricted Non-Tagged Texts. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000)*, pp. 579–585, July 2000.

[46] 延澤志保, 斎藤博昭, 中西正和. 文字間統計情報に基づく口語文字列の自動抽出. 言語処理学会 自然言語処理, Vol. 8, No. 3, pp. 39–57, July 2001.

[47] Shiho Nobesawa, Tomohisa Sano, Hiroaki Saito, and Masakazu Nakanishi. Incremental Learning of D-bigram Data against New Domains. *Journal of Chinese Language and Computer (formerly: Journal of the Chinese and Oriental Language Information Processing Society)*, Vol. 12, No. 1, pp. 1–22, June 2002.

[48] 延澤志保, 佐藤健吾, 斎藤博昭. ドメイン固有の文字列情報の組み込みによる形態素解析処理の精度の向上. 言語処理学会 自然言語処理, Vol. 9, No. 3, pp. 21–40, July 2002.

[49] Shiho Nobesawa and Ikuo Tahara. Language Identification for Person Names Based on Statistical Information. In *Proceedings of the 19th Pacific Asia Conference on Language, Information and Computation (PACLIC 19)*, pp. 289–296, December 2005.

[50] Shiho Nobesawa and Ikuo Tahara. Area Identification of English Person Names Based on Statistical Information. In *Proceedings of the 19th IEEE Canadian Conference on Electrical and Computer Engineering (CCECE '06)*, pp. 1688–1691, May 2006.

[51] Shiho Nobesawa, Junya Tsutsumi, Da Jiang Sun, Tomohisa Sano, Kengo Sato, and Masakazu Nakanishi. Segmenting Sentences into Linky Strings Using D-bigram Statistics. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING '96)*, pp. 586–591, August 1996.

[52] 奥野陽, 萩原将文. インターネットを用いた日本語入力システム. 情報処理学会研究報告, Vol. 2009-NL-190, No. 36, pp. 1–6, March 2009.

[53] Andreas M. Olligschlaeger and Alexander G. Hauptmann. Multimodal Information Systems and GIS: the Informedia Digital Video Library. In *Proceedings of the 1999 ESRI User Conference*, pp. 102–106, 1999.

[54] David D. Palmer. A Trainable Rule-Based Algorithm for Word Segmentation. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 321–328, 1997.

[55] Jon Patrick, Casey Whitelaw, and Robert Munro. SLINERC: the Sydney Language-Independent Named Entity Recogniser and Classifier. In *Proceedings of the 6th Conference on Natural Language Learning*, pp. 1–4, Morristown, NJ, USA, 2002. Association for Computational Linguistics.

[56] Bruno Pouliquen, Marco Kimler, Ralf Steinberger, Camelia Ignat, Tamara Oellinger, Ken Blackler, Flavio Fuart, Wajdi Zaghouani, Anna Widiger, Ann-Charlotte Forslund, and Clive Best. Geocoding Multilingual Texts: Recognition, Disambiguation and Visualisation. *CoRR*, Vol. abs/cs/0609065, pp. 53–58, 2006.

[57] Bruno Pouliquen, Ralf Steinberger, Camelia Ignat, and Tom De Groeve. Geographical Information Recognition and Visualization in Texts Written in Various Languages. In *Proceedings of the 2004 ACM Symposium on Applied Computing (SAC '04)*, pp. 1051–1058, New York, NY, USA, 2004. ACM.

[58] Erik Rauch, Michael Bukatin, and Kenneth Baker. A Confidence-Based Framework for Disambiguating Geographic Terms. In *Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References*, pp. 50–54, 2003.

[59] Ellen Riloff. Automatically Constructing a Dictonary for Information Extraction Tasks. In *Proceedings of the 11th National Conference on Artificial Intelligence*, pp. 811–816, 1993.

[60] 佐野智久. 動的な統計情報を用いた自然言語処理. 修士論文, 慶應義塾大学大学院, 1997.

[61] 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, January 2010.

[62] Tomohisa Sano, Shiho Hoshi Nobesawa, Hiroyuki Okamoto, Hiroya Susuki, Masaki Matsubara, and Hiroaki Saito. Robust Toponym Resolution Based on Surface Statistics. *IEICE Transactions on Information and Systems*, Vol. E92-D, No. 12, pp. 2313–2320, December 2009.

[63] Tomohisa Sano, Shiho Hoshi Nobesawa, and Hiroaki Saito. Automatic Country Identification of Area Names Based on Surface Features. In *Proceedings of the 7th International Symposium on Natural Language Processing (SNLP '07)*, pp. 13–18, 2007.

[64] 佐野智久, 堤純也, 孫大江, 延澤志保, 佐藤健吾, 中西正和. d-bigram を用いた自然言語文評価に関する実験. 情報処理学会第 51 回全国大会 Vol.3, pp. 3–4, 1995.

[65] 佐野智久, 堤純也, 孫大江, 延澤志保, 佐藤健吾, 大森久美子, 中西正和. 自然言語文評価における d-bigram 情報の活用方法に関する実験. 言語処理学会 第 2 回年次大会, pp. 185–188, 1996.

[66] Penelope Sibun and A. Lawrence Spitz. Language Determination: Natural Language Processing from Scanned Document Images. In *Proceedings of the 4th Conference on Applied Natural Language Processing*, pp. 15–21, 1994.

[67] David A. Smith and Gregory Crane. Disambiguating Geographic Names in a Historical Digital Library. In *Proceedings of the 5th European Conference on Research and Advanced Technology for Digital Libraries (ECDL '01)*, pp. 127–136, London, UK, 2001. Springer-Verlag.

[68] Rohini Srihari, Cheng Niu, and Wei Li. A Hybrid Approach for Named Entity and Sub-Type Tagging. In *Proceedings of the 6th Conference on Applied Natural Language Processing*, pp. 247–254, Morristown, NJ, USA, 2000. Association for Computational Linguistics.

[69] 堤純也, 孫大江, 延澤志保, 佐藤健吾, 佐野智久, 中西正和. d-bigram と trigram の相関に関する実験. 情報処理学会第 51 回全国大会 Vol.3, pp. 5–6, 1995.

[70] 宇津呂武仁, 松本裕治. コーパスを用いた言語知識の獲得. 人工知能学会誌, 第 10 巻, 第 2 号, pp. 197–204, March 1995.

[71] 宇津呂武仁, 颯々野学, 内元清貴. 正誤判別規則学習を用いた複数の日本語固有表現抽出システムの出力の混合. 言語処理学会 自然言語処理, Vol. 9, No. 1, pp. 65–100, January 2002.

[72] Takahiro Wakao, Robert Gaizauskas, and Yorick Wilks. Evaluation of an Algorithm for the Recognition and Classification of Proper Names. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING '96)*, pp. 418–423, August 1996.

[73] Joe H. Ward. Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, Vol. 58, pp. 236–244, March 1963.

[74] Dekai Wu, Grace Ngai, Marine Carpuat, Jeppe Larsen, and Yongsheng Yang. Boosting for Named Entity Recognition. In *Proceedings of the 6th Conference on Natural Language Learning*, pp. 1–4, Morristown, NJ, USA, 2002. Association for Computational Linguistics.

[75] 湯本紘彰, 森辰則, 中川裕志. 出現頻度と連接頻度に基づく専門用語抽出. 情報処理学会研究報告, Vol. 2001-NL-145, pp. 111–118, September 2001.

[76] GuoDong Zhou and Jian Su. Named Entity Recognition using an HMM-based Chunk Tagger. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL '02)*, pp. 473–480, 2001.

[77] Wenbo Zong, Dan Wu, Aixin Sun, Ee-Peng Lim, and Dion Hoe-Lian Goh. On Assigning Place Names to Geography Related Web Pages. In *Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL '05)*, pp. 354–362, 2005.