2E1-03

地名辞書を利用した地名の曖昧性解消と文書の地域分類

Place Name Discrimination using Place Name Dictionary and Document Classification about Place.

金木雄太[*1] Yuta Kaneki　山田剛一[*1] Koichi Yamada　絹川博之[*1] Hiroshi Kinukawa　中川裕志[*2] Hiroshi Nakagawa

[*1] 東京電機大学大学院 Graduate School, Tokyo Denki University

[*2] 東京大学 情報基盤センター Information Technology Center, The University of Tokyo

A place name is very important to get accurate information. But in the Web information different places of the same name often appear. Then we propose a place name discrimination method using place name dictionary, co-applarane of place names and a landmark information which equally expresses a place name. The proposed method is evaluated.

1. はじめに

最近では必要な情報を得るためにネット上で検索することが多い。しかし必要な情報を効率よく得ることはなかなかできない。地名は地域のニュースや情報を得るのにとても重要な要素である。そこで本研究では、毎年保守管理されている全国住所辞書[1]の住所情報を利用し、文書の地域特定に使用するものとする。

2. 地名辞書

2.1 地域番号

日本全国には、同名の地名は多数存在し、地名だけで地域を判別することはできない。そこで、地名辞書を作成する際に、日本全国の地名それぞれに対して識別可能な番号を割り当てる必要がある。本研究では、その番号を地域番号と呼ぶ。

地域番号の割り当て規則は次のとおりである。

- 現在、日本の地域は最大 5 階層で表現される。そこで本研究でも地域を5つの階層で表現する。
- 地域は、ひとつの地域階層を 3 桁で表した 15 桁の整数値で表現する。
- 該当地域階層より下位の地域階層は“000”で表現する。

2.2 地名辞書の構成

地名辞書は 3 つのテーブルから構成される。

(1) 地域情報

主に、地域番号が与えられた場合、その番号に対応する地域情報（地域名、読み仮名、郵便番号）を取得するときに利用するものである。（図 1）

地域番号(LONG)	地域名(STRING)	読み(STRING)	郵便番号(INT)

図 1 地域情報テーブル

(2) ランドマーク情報

ランドマークとは、地域を特定する際に、地名と同等の意味があるものを表す。与えられたランドマークに対応する地域番号と、関連度を取得するのに利用するものである。（図 2）

関連度とはランドマークと地域の一致する割合のことであり（3 章で定義）、あるランドマークに対して候補となる地域が一つしかない場合、関連度は最大の 1 となる。

地名表現文字列<ランドマーク>(STRING)	地域番号(LONG)	関連度(DOUBLE)

図 2 ランドマークテーブル

金木雄太　東京電機大学大学院工学研究科情報メディア学専攻
〒101-8457　東京都千代田区神田錦町 2-2　Tel:03(5280)3631
E-mail:kaneki@cll.im.dendai.ac.jp

(3) 文字インデックス

主に、文書中から抽出された地名表現文字列が与えられた場合、その文字列を含む地域の地域番号を取得するときに利用するものである。（図 3）

地名表現文字列＜地域名＞(STRING)	地域番号(LONG)

図 3 文字インデックステーブル

2.3 ランドマーク登録方法

ランドマークを登録することにより、以降の特定の際に地域を判別する要素が増えるのでより正確に分類することが可能となる。そのため登録するランドマークは一定の水準以上の関連度が必要となる。

ランドマークの登録方法は次のとおりである。

- 登録するランドマークを含む、適切な情報を Web から 100 件取得する。
- 取得した文書情報から、構文解析器「Cabocha」を利用して、固有表現を抽出し本地名特定方式を利用する。
- 本特定方式を利用する際、登録ランドマークと共起する地名があった場合、地域距離のかわりにランドマーク距離を利用して計算する。
- 最終的に、関連度が一定の水準以上であるランドマークをランドマーク情報へ登録する。

3. 地名特定方式

3.1 地名抽出方法

文書の地域を特定するには、文書中の地名を全て抽出する必要がある。そのため、まず形態素解析を行い、単語情報を取得する。次に構文解析を行い、地域を表している単語を全て取得した。本研究で使用した形態素解析器は「茶筌 Ver2.0」、構文解析器は「Cabocha」である。

3.2 地域距離

文書中の地名表現文字列から検索された地域候補が多数ある場合、その文書の地名特定はできない。しかし、文書中にそれとは別の地名表現文字列があった場合、これらの地域候補の関係を調べれば地域候補を絞ることが可能になる。

そこで、地名特定をする際に、共出する2つの地域間の距離を表す値を計算する。本研究ではその距離を地域距離と呼び、地域距離を計算する際に地域階層の値を利用する。地域階層には最上位を”1”とし、順に”5”までの整数値で表現した階層値を与える。

地域距離の計算式は次のとおりとする。

$$D_{ij} = \frac{L_{ij}^{2}}{L_i \times L_j} \times \left(LMsim_{\ i} \times LMsim_{\ j} \right) \quad (i \neq j) \qquad \text{(式 1)}$$

L_i、L_j は同一文書中に出現した地域で、異なる地名表現文字列から取得された地域候補、L_{ij} は L_i、L_j 両方に共通な上位階層で一番階層の低い地域、D_{ij} は L_i、L_j の間の地域距離を表す。$LMsim_i$ と $LMsim_j$ はそれぞれ、ランドマークと地域の関連度を表し、地域名そのものの場合は 1 となる。

$LMsim_i$、$LMsim_j$ は、後述の(式6)～(式8)の D_{jigk} を DM_{ij}、DM_{ji} で置き換えることで得られる。

3.3 ランドマーク距離

ランドマークを登録する際、登録するランドマークと共起する地名が重要となってくる。その共起する地名に重点を置いて地域候補同士の関係を数値化したものである。

ランドマーク距離 DM の計算式は次のとおりとする

$$DM_{ij} = D_{ij} \times D_{je} \quad (i \neq j) \qquad (式\ 2)$$

$$DM_{ji} = D_{ji} \times D_{ie} \quad (i \neq j) \qquad (式\ 3)$$

D_{je}、D_{ie} はそれぞれ、共起する最も近い地域候補 L_e との地域距離である。DM_{ij} は地域候補 L_i、L_j の地域距離を表している。

3.4 地名特定方法

本研究では、文書中に出現する全ての地名を取得し、その全ての地域と地域の関係から得点を計算し、文書を表す地名を特定することを目的としている。得点を計算する手順を説明する。

(1) 地名表現文字列に得点を与える

文書に地名表現文字列がn種類出現した場合、各地名表現文字列 LS_i の得点は次の式で表される。

$$LS_{ip} = LS_{iN} \times 100 \quad (1 \leq i \leq n) \quad (式4)$$

LS_{ip} は LS_i の得点、LS_{iN} は LS_i の出現回数を表す。

(2) 地域候補に得点を与える

地域候補は地名表現文字列から地名辞書を用いることによって取得することができ、文書と関連のある地域の候補のことである。地名表現文字列 $LS_j(1 \leqq j \leqq n)$ から m_j 個の地域候補が取得された場合、地域候補 $L_{ji}(1 \leqq i \leqq m_j)$ の得点を次の式で表す。

$$L_{jip} = \frac{LS_{jp}}{m_j} \times SIM \qquad (式\ 5)$$

Ljip が LSj から取得した地域候補の得点を表す。SIM は地名表現文字列 LSj と地域候補 Lij の文字的類似度を表す。

(3) 関連のない地域を省く

全ての地域候補のなかで最も地域距離の近い候補を探す。地域距離が近い要素がほかにある地域は、ない地域より重要という点から最近接候補 L_{gk} との地域距離 D_{jigk} を得点に乗算する。

$$L_{jip} = L_{jip} \times D_{jigk} \quad (g \neq j) \qquad (式\ 6)$$

(4) 地域間の関連を得点に反映

すべての地域候補に対して地域距離を計算し、その距離に応じた点数を加算する。そして、最も距離の近い地域候補との地域距離を点数に反映させる。このとき、同じ地名表現文字列から取得された地域候補同士での地域距離の計算は行わない。計算式は次のとおりである。

$$L_{jis} = L_{jip} + \sum_{g=1}^{n} \sum_{h=1}^{m} (L_{ghp} \times D_{jigh}) \quad (g \neq j) \qquad (式\ 7)$$

L_{jip}、L_{ghp} は地域候補の得点、D_{jigh} は L_{ji} と L_{gh} の地域距離を表す。

(5) 下位地域候補との関連を得点に反映

下位階層の地域候補が存在する場合は、すべての下位地域候補との地域距離を計算し、その値に応じた得点を加算する。

$$L_{jit} = L_{jis} + \sum_{g=1}^{n} \sum_{r=1}^{m} (L_{grp} \times D_{jigr}{}^2) \qquad (式\ 8)$$

L_{grp} は L_{ji} の下位地域候補 L_{gr} の得点、D_{jigr} は L_{ji} と L_{gr} の地域距離を表す。L_{gr} は L_{gh} に含まれるうち、L_{ji} の下位地域の地域候補である。

(6) 最後

地域候補 L_{ji} の得点は、(1)～(5)により得た L_{jit} とする。

4. 実験

4.1 実験方法

登録するランドマークは地名特定に大きくかかわってくる。そこで登録されるランドマークの正解率を調べる。このときランドマーク距離の有効性を示すために、ランドマーク距離を使う場合(手法Ⅰ)と使わない場合(手法Ⅱ)の2通り実験する。

次に、ランドマークが実際に地名特定に貢献しているか調べるために、ランドマークを利用した場合(手法 A)としない場合(手法 B)の 2 通り調査し、その正解率を求める。手法 B はランドマーク距離を使うか否かで B-Ⅰと B-Ⅱの二通り実験する。

4.2 実験結果

表 1 ランドマークと地域の関連度

	手法Ⅰ(%)	手法Ⅱ(%)
東京タワー	100	100
味の素スタジアム	31	46
東京電機大学	55	58

表 2 地域特定の正解率

	手法 A(%)	手法 B-Ⅰ(%)	手法 B-Ⅱ(%)
港区	38	70	75
調布市	40	60	64
府中市	42	55	62

4.3 実験結果の考察

手法 A よりも手法 B のほうが地域特定の正解率が高いことから、地域特定におけるランドマークの有効性を確かめることが出来た。しかし、ランドマークと地域の関連度が低いものも多いので、その点を改善していけば地域特定の正解率も上がる可能性がある。また、手法 B-Ⅰより手法 B-Ⅱの精度が高いことからも同様のことが言える。

5. おわりに

今回の実験でランドマークの有効性が確かめられたので、今後はランドマーク自体の正解率を上げていく予定である。

・参考文献

[1] 株式会社レムトス「REM-DIC 住所名マスタファイル」http://www.remtoss.co.jp/(2004)
[2] 奈良先端科学技術大学院大学 松本研究室「茶筌」http://chasen.naist.jp
[3] 金木 雄太 “地名辞書を利用した地名特定方式”FIT2004 第 3 回情報科学技術フォーラム 第 2 分冊 p181