E-032

地名辞書を利用した地名特定方式

Place Name Identification Using Place Name Dictionary

金木 雄太† 山田 剛一† 絹川 博之† 中川 裕志‡
Yuta Kaneki Koichi Yamada Hiroshi Kinukawa Hiroshi Nakagawa

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]の住所情報を地域区分単位に分割して、地名情報を取得する。その地名情報を元に地名辞書を作成し、 新聞記事に対する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。図1にその例を示す。

(1) 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。
(2) 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。
(3) 該当地域階層より下位の地域階層は0で表現する。

地域番号

東京都	**001**	000	000	000	000
新宿区	001	**004**	000	000	000
西新宿	001	004	**006**	000	000

図1 地域番号の例

2.2 地名辞書の構成

地名辞書は２つのテーブルから構成される。

(1) 地域情報テーブル

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2）

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

図2 地域情報テーブル

地名表現文字列(STRING)	地域番号(LONG)

図3 文字インデックステーブル

(2) 文字インデックステーブル

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）

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。本研究で使用した形態素解析器は「茶筌 Ver2.0」である。

地名候補を漏れなく抽出するために、地域、人名、組織名と認識されたものはすべて抽出する。また、地域の接尾辞を利用し、地域を特定する際に有用な情報かどうかを判別する。もし、無用な情報ならその地名候補は利用しない。

無用な情報となる接尾辞は主に、“川名”や“道名”など地域を横断するものの名称である。

3.2 地域距離

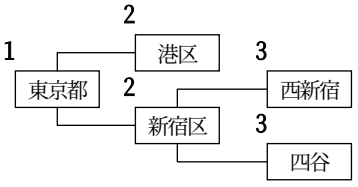
文書中の地名表現文字列から検索された地域候補が多数ある場合、その文書の地名特定はできない。しかし、文書中にそれとは別の地名表現文字列があった場合、これらの地域候補の関係を調べれば地域候補を絞ることが可能になる。

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。（図4）

地域距離の計算式は次のとおりとする。

$$D_{ij} = \frac{L_{ij}^{\ 2}}{L_i \times L_j} \qquad (i \neq j)$$

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。



例) 四谷と西新宿の距離

$$D = \frac{2^2}{3 \times 3} = \frac{4}{9}$$

図4 地域階層と地域距離の例

3.3 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、取得した全ての地域と地域の関係から得点を計算し、その得点からその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。

† 東京電機大学大学院 工学研究科 情報メディア学専攻
‡ 東京大学 情報基盤センター

(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。

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

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。

$\frac{Lsub_{jN}}{Lcomp_{jN}} \ge 2$ の場合

完全一致地域候補 $L_{jip} = \frac{LS_{jp}}{2} \times \frac{1}{Lcomp_{jN}}$

部分一致地域候補 $L_{jip} = \frac{LS_{jp}}{2} \times \frac{1}{Lsub_{jN}}$

$\frac{Lsub_{jN}}{Lcomp_{jN}} < 2$ の場合

完全一致地域候補 $L_{jip} = \frac{LS_{jp}}{\left(Lcomp_{jN} + \frac{Lsub_{jN}}{2}\right)}$

部分一致地域候補 $L_{jip} = \frac{LS_{jp}}{\left(Lcomp_{jN} + \frac{Lsub_{jN}}{2}\right)} \times \frac{1}{2}$

L_{jip} が LS_j から取得した地域候補の得点、$Lcomp_{jN}$ 、$Lsub_{jN}$ は地域候補数のうち、地域名と地名表現文字列が、それぞれ完全に一致したものの数と、部分的に一致したものの数を表す。

(3) 地域間の関連を得点に反映させる。

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。

$$L_{jis} = \left(L_{jip} + \sum_{g=1}^{n} \sum_{h=1}^{m} (L_{ghp} \times D_{jigh}) \right) \times D_{jigk} \quad (g \ne j)$$

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、D_{jigk} は L_{ji} と際近接地域候補 L_{gk} の地域距離を表す。

(4) 下位地域候補との関連を得点に反映させる。

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。

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

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。

(5) 地域候補 L_{ji} の得点は、(1)～(4)により得た L_{jit} とする。

4. 実験

4.1 実験方法

本研究の評価に用いた実験データは、毎日新聞社（1998 年度）の新聞記事から無作為に選択した 300 件の記事文書である。新聞記事は掲載面の種類ごとに分かれており、実験は社会面、経済面について行った。

新聞記事から抽出される地名情報の多くは、例えば「東京都新宿区西新宿」のように、本研究の特定方法を利用する必要のないものである。そこで、評価する記事をこの観点から分類した。

まず、抽出した新聞記事のなかから、地名情報が全く入っていないものを除外する。残った記事の内、本研究の特定方法を利用しなくても良いものと、そうでないものを分類し、それぞれの割合を求める。

最後に、本研究の特定方法を利用しなければならない記事を処理対象として、平均精度を求める。このとき、平均精度は地域の階層ごとに求める。

4.2 実験結果

表 1 本特定処理対象の割合（毎日新聞 1998）

	処理対象（%）	処理対象外（%）
社会面	24.3	75.7
経済面	12.8	87.2

表 2　平均精度（%）（毎日新聞 1998）

階層	1	2	3	4	5
社会面	0.862	0.753	0.774	0.653	0.526
経済面	0.843	0.685	0.623		

4.3 実験結果の考察

平均精度評価が下位階層において下がっているのは、下位階層の地域名の曖昧性が高いことが原因となっている。また、経済面での階層４以下の評価は、新聞記事に４階層以下の地名に関する記事がないためである。

本研究の手法では、地名の表現が曖昧な文書に対して処理を行うことによって、その文書の地域を特定することを目的としている。本特定方式の有効な適用分野について、実験をする必要がある。

5. おわりに

現在、文書の地域特定をする際に文書中の地名情報のみを利用しているが、地域を特定するようなランドマークや施設なども利用して特定を行えば、より曖昧な文書も特定が可能になるので、それを実現させるために研究を進めていく予定である。

また、本方式は、地域情報を検索する際に目的の地域とは異なった検索結果を除外する場面で有効であり、その効果を確認する予定である。

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/（2003）

[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/