Web2.0時代の名寄せを実現する固有表現グラウンディング技術

ブログに代表されるユーザ生成コンテンツを的確に分析・活用するためには，多種多様に表記された情報を整理（名寄せ）することが必要です．本稿では，テキスト中の固有表現の意味を同定することで，テキストコンテンツと各種データベースの連携サービスを可能にする固有表現グラウンディング技術を紹介します．

松尾 義博（まつお よしひろ） /小林 のぞみ（こばやし）
平野 徹（ひらの とおる） /高橋 いづみ（たかはし）
NTTサイバースペース研究所

グラウンディング技術

「名寄せ」とは，データベースに複数登録されている同一人物や同一企業のデータをマージする作業を指します．名寄せが必要な代表例は金融機関の口座の管理で，例えばペイオフの限度金額の判定には名寄せは必須となります．

私たちが取り組んでいるグラウンディング技術は，同一事物を指す言語表現をまとめるという点でこの名寄せと類似しています．「グラウンディング」とは人工知能研究の用語で，機械と人間との対話中で言及される名称と，実世界での事物との関連付けをすることを意味します．これは，適切な対話が成立するためには，ユーザがしゃべる「Xさん」と機械が思っているXさんが同一である必要があるからです．

固有表現グラウンディング技術で名寄せするデータは，データベースのレコードではなく，Webページなどに記述されている人名や地名などの固有表現です．Webページでは，同一の対象を指す固有表現が多種多様に言及されます．特にCGM（Consumer Generated Media：消費者生成メディア）と呼ばれる一般消費者が書いた文書ではこの傾向が顕著で，略称，愛称，隠語まで含めて，実にさまざまな表現で記述されます．

これらのWebページを検索する場面を考えてみます．例えば，電電一郎首相について記述されたWebページを探しているとしましょう．ここで，検索窓に「電電一郎」と入力すれば十分でしょうか？　実際にgooウェブ検索などで検索してみると，「電電一郎」でヒットするページ数に比べ，「電電 AND 首相」でヒットする件数が大きく上回っていることが想像できます*．後者はその大部分が電電一郎首相に関する文書と推定されますので，単純にフルネームで検索したのでは相当量を取りこぼしていることが分かります．

私たちの取り組んでいるグラウンディング技術はこういった問題の解決を目指したもので，テキスト中に出てきたさまざまな表記に対して，ユニークなID（グラウンド）を付与する機能を提供します．先の例でいえば，「首相の電電さん」「電電内閣総理大臣」「電ちゃん」といったテキストに対して，単一のIDを与える機能です（**図1**）．

Webテキストで言及された名称に対してIDが付与できると，文書の名寄せが可能となります（**図2**）．さらには

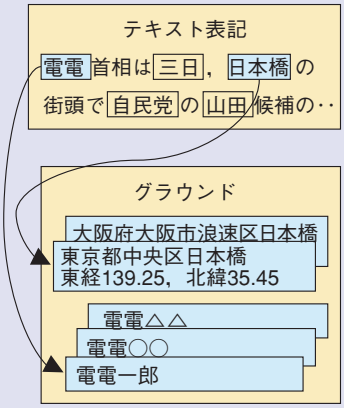

図1　表記とグラウンド

このIDをデータベースと連携させることによって，関連する情報を一元的に提示するなどの高度なWebサービスも可能となります．

本稿では，Webテキストの名寄せを実現するグラウンディング技術について解説します．

グラウンディング技術の難しさ

グラウンディングに必要な技術は大きく2つに分類できます．1つは，言及された名称が誰を指しているのかというあいまい性を解消する技術で，自

＊ この例は架空人物なので検索不能ですが，実在の首相名で検索すると後者が2～3倍程度ヒットすることが分かります．

然言語処理の多義解消技術が必要です．「福田さん」と書かれたとき，それが実際は誰のことなのかには多くの可能性があり，例えばWikipediaには福田姓の人物は70人以上が掲載されています．Wikipediaに載るということは，ある一定の知名度のある人物ですから，ブロガーの友人などまで含めると無数の可能性（多義）があるといえます．この多義を文脈などから解消するのが第1の技術です．

もう1つは，それぞれの固有物が，どういった表記で言及される可能性があるのかの知識を獲得する技術です．例えば，山田一郎さんは「ヤマさん」と呼ばれたり，佐藤花子さんは「さとっち」といった愛称を持っているかもしれません．これらの知識なしで「さとっち」が佐藤花子さんのことであると同定することは困難です．愛称や略称は日々新しいものが用いられるようになるため，これら同義語を自動的に獲得する技術が第2の技術です．

次に，人名と地名に対するあいまい性解消技術と，同義固有表現獲得技術について述べます．

人名のあいまい性解消

CGMで言及・検索されることの多い人名のグラウンディングは重要な課題です．例えば，政治評論家の「電電花子」さんとサッカー選手の「電電太郎」さんを考えます．人はどうやって「電電さん」が花子さんなのか太郎さんなのかを見分けているのでしょうか？何の脈絡もなく出現した「電電さん」をどちらか判断することは困難です．見分けることができるのは，脈絡，すなわち文脈があるからといえます．例えば，政治面に出現していれば花子さんである可能性が高く，スポーツ面に出現していれば太郎さんの可能性が高いでしょう．また，○○党の「□□幹事長」と同時に言及される可能性が高いのは花子さんで，Jリーグの「△△選手」と一緒に書かれるのは太郎さんと推定できます．

この文脈を機械的に推定して判定するのがグラウンディング技術です（**図3**）．本技術では，事前に大量のテキストから，それぞれの人名の周辺に出現した語の分布を計算しておきます．花子さんの場合には，「自民党」や「国会」といった語が高い頻度で収集されることが期待できます．

入力されたテキストの「電電さん」が，実際には誰のことであったかを判

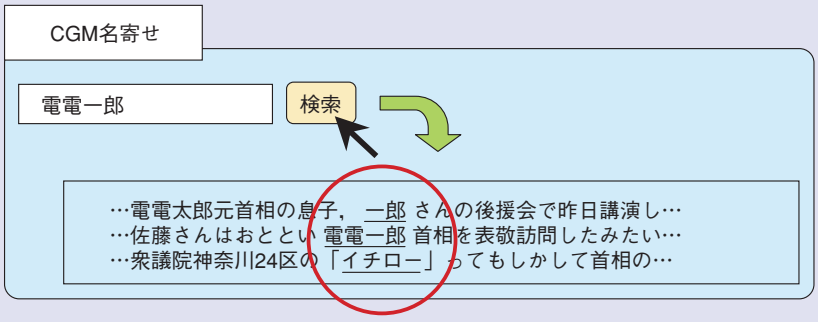

図2　CGMの名寄せ

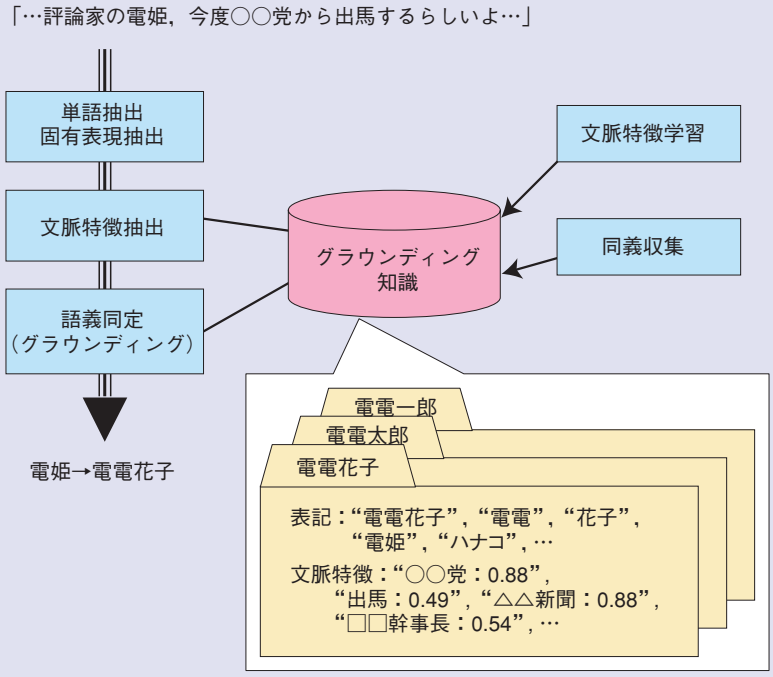

図3　固有表現のグラウンディング処理

定する際には，「電電さん」の周辺に出現した語の分布と，事前に収集していた分布を比較し，どちらの分布に似ているかで，実際の人物を推定します．

技術的には，周辺に出現する語のうち推定に役立つ語とそうでない語をうまく重み付けすることが課題となります．例えば，どちらの人物でも周辺に「テレビ」といった語が出現しているかもしれません．本技術では，ある人で多く出現し，他の人であまり出現しない語を統計的に推定することで，効率的な推定を可能にします．

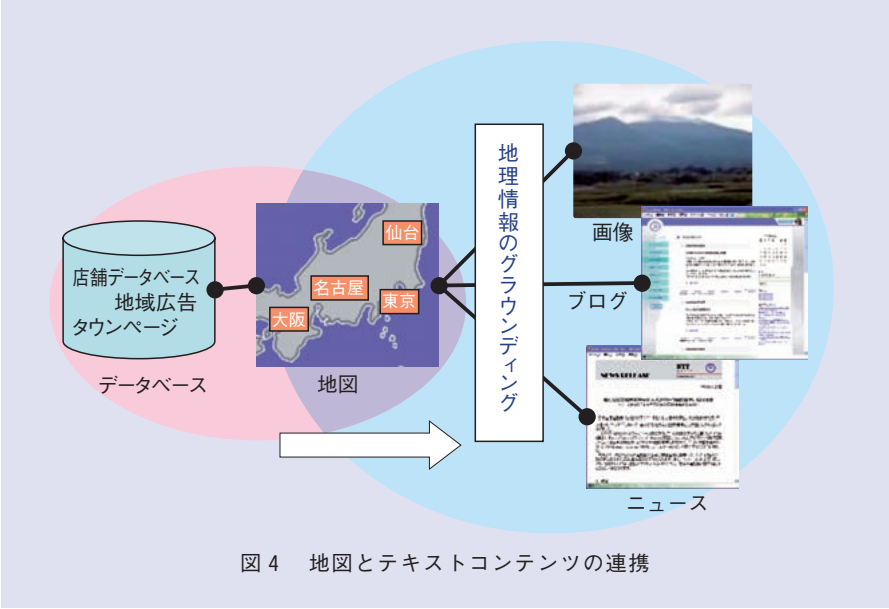

図4　地図とテキストコンテンツの連携

地名のあいまい性解消

昨今のGIS（Geographic Information System：地理情報システム）の発展に伴い，地理的な事物の実世界での位置を同定することの重要性が高まっています．特に，goo地図API [(1)] などのWebでの地図APIの登場により，各種データベース情報を地図上で整理して提示するサービスが急速に普及しました．テキスト中の地名の実世界での位置を推定することが可能になれば，GISで整理する対象を，データベースから各種テキスト情報に拡大することが可能になります（**図4**）．

テキストに書かれた地名の位置を推定するためには，あいまいな表現で記述された地名の実際の指示対象を同定する必要があります．例えば「日本橋には難波から行くのが便利」という文章を考えます．国土交通省の街区レベル位置参照情報によると，日本全国の地名で「日本橋」を含む地名は2カ所，「難波」を含む地名は18カ所あります．地名のあいまい性を解消するとは，これらの候補の中から適切なものを選択するという問題といえます．

あいまい性を解消するために，本技術では主に2つの観点を用います．1つは，テキスト中に出現する地名は相互の距離が近いことが多い，という観点です．ここで，距離とはテキスト内での出現位置ではなく実世界での距離を指します．テキスト中の地名と，それぞれの地名の緯度経度の候補をすべて列挙し，すべての組み合わせの中から相互に近い位置候補をその地名の実世界での位置と決定すれば，先の例の「日本橋」と「難波」は，それぞれ，「大阪府大阪市浪速区日本橋（北緯＝34.39，東経＝135.30）」「大阪府大阪市中央区難波（北緯＝34.40，東経＝135.31）」と決定できます．

もう1つの観点は，テキスト中に前置きなく出現する地名は「有名な」ものであるという観点です．全国には多数の「銀座」がありますが，前置きなく述べられた銀座は，そのほとんどが東京都中央区のものと考えられます．東京以外の「銀座」に言及する場合には「X県Y市銀座町」といった言及をすることが一般的といえます．

では，地名の「有名度」をどうやって推定すればいいでしょうか？　いくつかの手法が提案されていますが，1つの手法としてそれぞれの地域に存在する店舗の数から推定することが可能です [(2)]．商業施設が多数存在しているということはその地域に多数の人が訪れるということですので，店舗数と，その地をテキストの読者が知っている割合とは一定の相関があると考えられます．これらのあいまい性解消技術を組み合わせ，ブログ記事を対象にした評価実験では，92.7%の精度で実世界での位置が正しく推定できることが確認できています．

本技術を応用したサービスとしてgoo画像検索 [(3)] が挙げられます．このサービスではスクロールした画像のキャプションなどに含まれる地名を利用して，地図インタフェース上での画像検索機能を提供しています．

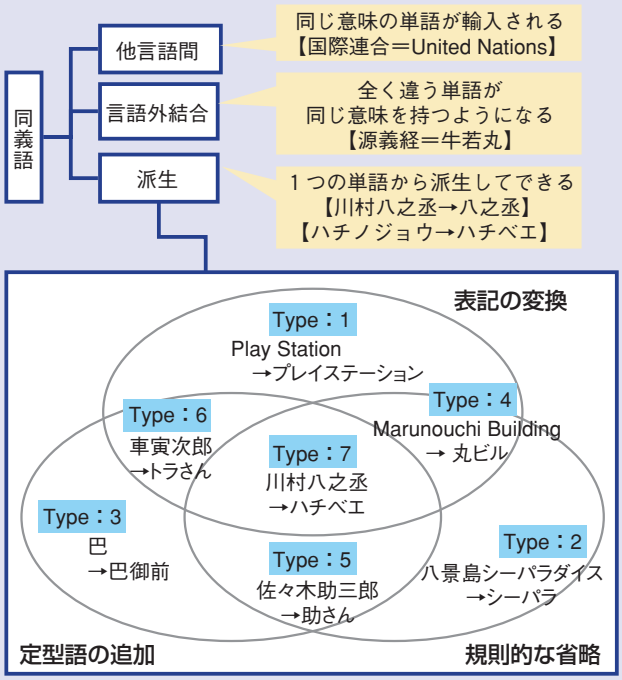

図5　固有表現の同義語生成モデル

グラウンディングに必要な同義語知識の自動獲得

固有表現の多義解消を実現するには，まずは，それぞれの表記がどの語義であり得るかの候補リストが必要となります．候補リスト中に「電ちゃん→電電一郎，山田電子」とあれば，多義解消とはそのどちらであるかを選択する問題となります．この候補リストは，ある固有物がどのような表記で表現されるかの同義関係のリストから生成することが可能です．

固有表現の同義語にはさまざまな種類があります．長音記号の省略のような単純な表記揺らぎのものから，想像もつかないような愛称まで，その発生過程によって多くのタイプのものがあります．私たちはこれらの発生過程を分類し，派生型の固有表現同義語を自動獲得する方法を考案しました[(4)]．

派生型の固有表現同義語とは，音の類似や表記の類似，規則的な省略，定型的な愛称語の追加（例えば，“八之丞”→“ハチベエ”）などによって発生した同義語です．これらの発生過程を組み合わせた分類図を**図5**に示します．テキストコーパスの分析により，同義語の約90％はこれらの組み合わせで発生が説明できることが分かっています．

ある語のペアが同義関係にあるかどうかを判定するには，上記の発生過程を計算機で再現して判定します．音の類似を判定するためには発音間の類似を編集距離などで判定し，定型的な愛称語は辞書で判定します．また規則的な省略は，省略関係にあるかどうかを識別学習によって判定します．

これらの判定規則を組み合わせた手法によるコーパスからの同義ペア抽出実験では，約70％の精度で抽出が可能なことが分かりました．

今後の取り組み

本稿では，固有表現の語義を推定することによって，Webコンテンツ等の名寄せを実現し，的確な分析・活用を可能にする，固有表現グラウンディング技術について述べました．今後は，処理対象の表現の拡大に取り組むとともに，新たなWebマイニングサービスへの適用を進める予定です．

■参考文献

(1) http://map.api.goo.ne.jp/icc/
(2) 平野・松尾・菊井：“地理的距離と有名度を用いた地名の曖昧性解消,”情報処理学会第70回全国大会, Vol.2, pp.85-86, 2008.
(3) http://bsearch.goo.ne.jp/maptop/
(4) 高橋・浅野・松尾・菊井：“単語正規化による固有表現の同義性判定,”言語処理学会第14回年次大会, pp.821-824, 2008.

（後列左から）平野　徹/ 松尾　義博
（前列左から）高橋　いづみ/ 小林　のぞみ

書き手の意図を正確に理解する，必要な情報を的確に伝える，そんな技術を目指して研究開発に取り組んでいきます．

◆問い合わせ先
NTTサイバースペース研究所
音声言語メディア処理プロジェクト
TEL　046-859-2670
FAX　046-855-1054
E-mail　matsuo.yoshihiro@lab.ntt.co.jp