Entity Linkingを用いたニュース記事に対する市区町村単位の地域情報の付与

井上 裁都　　末永 圭吾　　長田 誠也　　立石 健二
ヤフー株式会社
{tatinoue,ksuenaga,sosada,ktateish}@yahoo-corp.jp

1. はじめに

筆者らは現在，地域に根ざしたニュース記事を適切な利用者に配信するサービスの品質向上に向け，記事に対する地域情報付与の研究開発を進めている．GPS受信機を搭載したデバイスの普及により，端末所有者の場所に紐付く情報の入手は容易になった．ニュース記事にそれが言及する地域情報を付与できれば，端末所有者にマッチした地域の記事が配信できる．端末所有者は関心がある記事を効率的に閲覧できる．

ニュース記事の地域情報付与に関する従来研究として，Entity Linkingを応用したD'Ignazioら[1]，長田ら[2]の研究がある．これらの研究では，まずニュース記事から場所に関するEntityを抽出し，それらの曖昧性解消を行う．その後，Entityを重み付けして集計し，記事全体の地域情報を付与する．

従来研究の問題は，細かい粒度での地域情報付与に対する評価がされていないことである．D'Ignazioら[1]は国単位で，長田ら[2]は都道府県単位でニュース記事に地域情報を付与している．しかし，より粒度の細かい市区町村単位については，同様の方法を用いて十分な精度が得られるか報告されていない．

筆者らの研究の目的は二つある．一つは，ニュース記事に対する市区町村単位での地域情報付与について，難易度を明らかにすることである．もう一つは，市区町村単位の地域情報を高精度に付与することである．本稿では，まず地域情報付与の従来手法を概説し，適合率を向上に有用な改良手法を提案する．次に，従来手法による都道府県単位と市区町村単位の付与精度について，比較評価をする．最後に，従来手法と提案手法を比較評価し，提案手法が適合率向上に有効であることを示す．

2. 関連研究

近年，KBP Entity Linking Track[3]や，NEEL Challenge[4]といった評価型ワークショップを通して，Entity Linkingに関する技術開発が盛んに行われている．Entity Linkingではニュース記事やツイートに対して人物・場所・組織等に関するEntityの出現位置を特定し，ナレッジ(Wikipedia, DBpedia)へのリンクを付与する．

場所に関するEntity Linkingシステムとしては，GeoNLP[5]が存在する．このシステムはLOD(Linked Open Data)の地名辞書を持ち，この辞書を形態素解析ソフトウェア(MeCab)で利用できるようにすることで，ニュース記事から非常に多くの地名を抽出できる．

テキストに地域情報を付与する研究として，Webテキストを対象としたAmitaらのWeb-a-Where[6]，LiebermanらのSTEWARD[7]に関する研究がある．これらは地名の階層構造[6]や地名同士の共起関係[7]を利用して，Webテキストに地域情報を付与する．

ニュース記事に地域情報を付与する研究としてD'Ignazioら[1]，長田ら[2]の研究がある．D'Ignazioら[1]は，CLAVINと呼ばれる既存のオープンソースをベースとして，曖昧性解消に改良を加え，Entity出現の頻度で地域情報を付与した結果，国単位の精度が約90%まで向上したと報告している．長田ら[2]は，同様なEntity Linkingを用いた手法で，都道府県単位の精度が約88%であったと報告している．

3. 地域情報付与の従来手法

長田らが報告したEntity Linkingを用いた地域情報の付与手法について概略を述べる．本稿ではこの長田らの手法を従来法と呼ぶ．

3.1. Entity Linkingシステム

ニュース記事を対象にEntity Linkingをすることで，記事中の場所Entityを抽出し，これを地域情報の付与に利用する．Entity Linkingシステムは次の4ステップで構成される．

Copyright(C) 2016 The Association for Natural Language Processing.
All Rights Reserved.

(a) **Entity 辞書を形態素解析辞書に追加**
場所などのEntityを収集した辞書を事前に用意し，この辞書データを既存の形態素解析器のユーザー辞書に追加する．

(b) **入力テキストを形態素解析**
(a)のユーザー辞書を使い形態素解析する．

(c) **Entity とマッチする形態素列を抽出**
(b)の結果と Entity 辞書をマッチングさせ， マッチした Entity を抽出する．

(d) **Entity 曖昧性解消**
1 形態素列に対しマッチする Entity が複数あれば，曖昧性解消して 1 Entity に定める．

Entity収集の情報源としてはWikipediaなどを利用する．曖昧性解消の手法は石川ら[8]の報告が詳しいため参照されたい．

3.2. 地域との関連度スコア算出

都道府県毎または市区町村毎に記事との関連度スコアを求める．スコアが事前に与える閾値を超えたものを記事の地域情報として付与する．

3.1 節で抽出した Entity には場所・組織・人などのカテゴリ情報があらかじめ付与されている．場所 Entity であれば, Entity が存在する都道府県や市区町村などの情報も付与されている．カテゴリはツリー構造の体系になっており, 場所 Entity は行政区画, 自然地名，建造物，道路などのより細かいカテゴリ情報を持つ．

本手法では Entity のカテゴリ毎に重みを付け, 式(1)により各地域のスコアを求める．

$$w(x) = \sum_{e \in E_x} \theta(c_e) \qquad \cdots (1)$$

ここで， $w(x)$ は地域 x のスコア，E_x は x と紐付く記事中の Entity の集合，$\theta(c_e)$ は e のカテゴリ c_e の重みである．$\theta(c_e)$は 5.1 節で述べる開発用データを使って評価しつつ人手で調整する．例えば，建造物など所在地が一意に定まるカテゴリは大きく，道路など範囲があるカテゴリは小さく設定する．

4. スコア補正による適合率改善

サービス利用者へのニュース記事配信という課題においては，地域情報付与の再現率よりも適合率の高さが重視される．これは適切な記事の配信機会を損なうよりも，不適切な記事を誤配信する方がサービス利用者にとって利便性を損なうと考えるためである．したがって，再現率よりも適合率を改善することには十分な意義がある．

一方，3 節で述べた従来法を用いたとき，付与する地域情報の粒度が細かくなるほど，課題の難易度は高くなる．このため，記事への都道府県単位の情報付与は，市区町村単位の付与よりも易しく，その結果を信頼できる．そこで，都道府県単位と市区町村単位の両者に対して式(1)のスコアを算出し，後者のスコアを式(2)で補正することで，適合率を改善することを提案する．

$$w'(d) = \frac{w(p_d)}{\sum_{p \in P} w(p)} w(d) \qquad \cdots (2)$$

ここで，d は市区町村，p_d は d の都道府県，P は都道府県の集合， $w(x)$ は地域 x の補正前スコア(式(1)の値)，$w'(x)$ は x の補正後スコアである．式(2)は，都道府県単位のスコアを総和が 1 になるよう正規化し，その上で都道府県単位と市区町村単位のスコアを乗算することを意味する．

従来法のシステムが記事の主題ではない都道府県の Entity を誤抽出し, かつ記事の主題である都道府県の正規化したスコアが 1 に近いとき, 式(2)は有効に働く．このとき，主題でない都道府県に属する市区町村は，主題の市区町村と比べ大幅にスコアを下げられる．最終的にはスコア閾値の調整により，適合率の向上が可能になる．

5. 評価

3 節で述べた従来法による地域情報付与について, 都道府県単位と市区町村単位の付与精度を比較評価する．また，市区町村単位の付与精度について，従来法と 4 節で述べた提案法を比較評価する．

5.1. データセット

Yahoo!ニュースに掲載された 2014 年のニュース記事から 300 件をサンプリングし，150 記事を開発用データ，残り 150 記事を評価用データとしてデータセットを作成した．各記事を対象に，記事と関連性が高い都道府県と市区町村を正解として，正解を人手で付与した．正解が複数あれば，全て付与対象とした．開発用データは 3.2 節で述べた$\theta(c_e)$の調整に利用した．

正解は関連度に応じて“GOOD”，“FAIR”の 2 種類を付与した. GOOD は各記事の主題となるシステムが必ず付与すべき地域を対象に付与した．FAIR は

Copyright(C) 2016 The Association for Natural Language Processing.
All Rights Reserved.

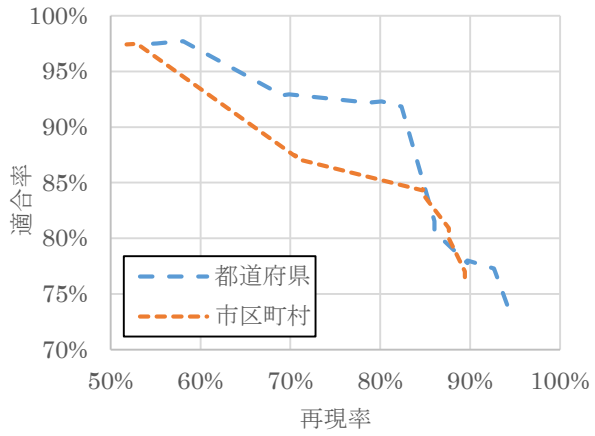

図1：粒度別の比較評価結果

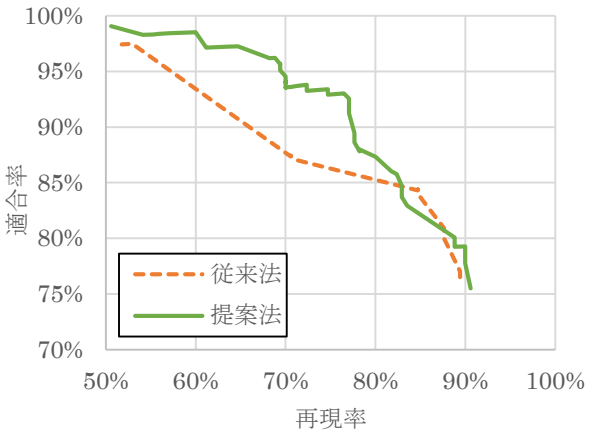

図2：従来法と提案法の比較評価結果

各記事と関連性はあるが付与は任意で良い地域を対象に付与した．例えば，記事中にスポーツ大会の開催地の記述があれば，その地域を GOOD として付与した．また，スポーツ選手の出身地や出身校の記述があれば，その地域は FAIR として付与した．

5.2. 評価指標

5.1 節で定義した正解に対するシステム出力の適合率と再現率で評価する．ただし，適合率は GOOD と FAIR の両方を正解とし，再現率は，GOOD のみを正解として算出した．

適合率算出において GOOD のみを正解とすると，GOOD ではないが記事と関連がある地域(すなわち FAIR)を付与しても，記事と無関係な地域を付与しても，共に同じ誤りとして集計されてしまう．一方，再現率算出において GOOD と FAIR を共に正解とすると，必ず付与すべき GOOD とそうではない FAIR が，未付与のときに同じ誤りとして集計されてしまう．

適合率と再現率の算出において GOOD と FAIR の扱いを変えると前述の課題が避けられ，直感的な評価値を得ることができる．データセット中の記事により GOOD の個数に大きな差があることから，適合率・再現率は記事毎のマクロ平均ではなく，マイクロ平均で評価する．

5.3. 評価結果

5.1 節で述べた評価用データ 150 記事を使い，3 節の長田らの手法(従来法)による都道府県付与，市区町村付与，ならびに 4 節で提案したスコア補正を適応したときの市区町村付与の適合率・再現率を算出した．

従来法でのスコア閾値による適合率・再現率の変化を、都道府県単位と市区町村単位で比較した結果を図 1 に示す．都道府県単位と比較するとやはり若干低いと言えるが，市区町村単位でも比較的高い精度が得られていることがわかる．

続いて，市区町村単位での適合率・再現率の変化曲線を、従来法と提案法で比較した結果を図 2 に示す．再現率が 80%以下のとき，提案法の適合率は従来法よりも有意に高くなることがわかる．

4 節で述べた通り，ニュース記事配信という課題では再現率より適合率の方が重視される． 5.4, 5.5 節では，適合率が 90%を超え，再現率は 75%を確保できるポイントの閾値を用い，適合率を低下させた誤りを中心に分析結果を述べる．

5.4. 従来法と提案法の比較分析

提案法では付与されないが，従来法では誤付与される市区町村の例を挙げる．

- **比喩表現由来の誤付与**

以下の例では，Entity Linking システムが「東京ドーム」を抽出するが，この Entity の場所(東京都文京区)は記事の主題の地域(神奈川県横浜市)とは関係がない．従来法では，横浜市よりは低いが文京区にも比較的高いスコアが付く．

……「山下公園に埋まっている、関東大震災で出たがれきは東京ドーム何個分」などのクイズを出し、……
神奈川新聞［「3・11」に歴史学ぶ、Y校生徒ら慰霊碑など遺構訪問／横浜 ］

この例文は記事の一部だが，記事全体を解析すると神奈川県に属する Entity を多数抽出でき，神奈川

Copyright(C) 2016 The Association for Natural Language Processing.
All Rights Reserved.

県のスコアが東京都より有意に高くなる．提案法ではこれを利用し，横浜市と文京区の差を拡大するようにスコアを補正する．結果，スコア閾値により誤付与を防ぐことが可能になる．

- **都道府県間の曖昧性解消の誤り**

次の例では，Entity Linking システムが「佐倉」を千葉県佐倉市と誤判定する．正しくは福島県福島市の佐倉地区を意味する．この記事では千葉県より福島県のスコアが高くなるため，提案法の補正により佐倉市のスコアが低くなり，誤付与を防げる．

……同校によると、教職員が校庭に入り込んだニホンカモシカを発見し、福島署佐倉駐在所に通報した。……
福島民報［ 校庭に珍客　福島の荒井小にニホンカモシカ ］

5.5. 提案法の誤り分析

従来法でも提案法でも，誤付与してしまう市区町村の例を挙げる．なお，今回作成したデータセットにおいては，従来法で付与されず，提案法でのみ誤付与される市区町村の例は見当たらなかった．

- **作品名由来の誤付与**

この例では，Entity Linking システムが作品名に含まれる「乃木坂」を誤抽出してしまう．記事自体は地域と無関係なため地域情報の付与は不要である．しかし，スコア補正に意味がないためにこの誤りは除去できず，提案法は効果がない．

……付属の DVD 特典映像「T.M.Revolution｜SCANDAL 平成ガチ BATTLE ～乃木坂の戦い～」の SPOT 映像がオフィシャル YouTube チャンネルにて公開となり……
CD ジャーナル［ T.M.Revolution｜SCANDAL、1 対 1 のガチンコトークバトル期間限定独占公開！ ］

- **都道府県内の曖昧性解消の誤り**

こちらの例では，Entity Linking システムが「元町」を横浜市中区元町と誤判定する．横浜市と茅ヶ崎市は同じ神奈川県内のため，提案法のスコア補正に意味はなく，この誤りも除去できない．

……「ｓｔｕｄｉｏ　ＣＯＯＣＡ（スタジオクーカ）」（平塚市平塚４丁目）の展示即売会が茅ケ崎市元町の茅ケ崎ラスカで開かれ……
神奈川新聞［ 独創アートで魅了　障害者福祉施設が展示即売会／茅ケ崎 ］

6. おわりに

本稿では，長田らが報告したニュース記事への地域情報の付与手法が都道府県単位だけでなく市区町村単位でも有効であることを示した．また，市区町村単位の付与で都道府県単位のスコアを利用することで，適合率を向上できることも示した．

提案手法により十分高い適合率を達成できたため，再現率も向上させることが今後の課題である．再現率が十分でない理由の一つに，ニュース記事中の学校名に比較的多い，略称への対応が難しいことが挙げられる．これを解決するため，Entity Linking 用の辞書の拡充や Entity 間の曖昧性解消精度の改善を進めていきたい．

参考文献

[1] Catherine D'Ignazio, Rahul Bhargava, Ethan Zuckerman, Luisa Beck, CLIFF-CLAVIN: Determining Geographic Focus for News Articles, NewsKDD, 2014

[2] 長田誠也，末永圭吾，善積正伍，庄司和正，吉田享晴，橋本恭明，エンティティリンキングを用いたドキュメントに対する地点情報の付与とその応用，言語処理学会第 21 回年次大会, 2015.

[3] Heng Ji, Joel Nothman and Ben Hachey, Overview of TAC-KBP2014 Entity Discovery and Linking Tasks, TAC2014, 2014.

[4] Giuseppe Rizzo, Bianca Pereira, Amparo E. Cano, Andrea Varga, Making Sense of Microposts (#Microposts2015) Named Entity rEcognition & Linking Challenge, Microposts2015, 2015.

[5] 北本 朝展，相良 毅，有川 正俊, GeoNLP：自然言語文を対象とした高度なジオタキングに向けて, CSIS Days 2011, No. D10, 2011.

[6] Einat Amitay, Nadav Har'El, Ron Sivan, Aya Soffer, Web-a-Where: Geotagging Web Content, SIGIR'04, 2004.

[7] Michael D. Lieberman, Hanan Samet, Jagan Sankaranarayanan, Jon Sperling, STEWARD: Architecture of a Spatio-Textual Search Engine, ACMGIWS'07, 2007.

[8] 石川裕貴，小林健，長田誠也，ウェブ検索ログと Wikipedia 内部リンクを用いたエンティティの曖昧性解消，言語処理学会第 21 回年次大会, 2015.

Copyright(C) 2016 The Association for Natural Language Processing.
All Rights Reserved.