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

長田 誠也　末永 圭吾　善積 正伍　庄司 和正　吉田 享晴　橋本 恭明
ヤフー株式会社
{sosada, ksuenaga, syoshidu, kashouji, tayoshid, yahashim}@yahoo-corp.jp

1. はじめに

自然言語テキストの意味を計算機に理解させるという課題は、自然言語処理技術における大きな課題の１つである。このテキストの意味理解というタスクの一部として、入力されたテキスト中に出現する実世界上のモノやコト等の実態（エンティティ）を認定し、それを知識ベースのエンティティのエントリにリンクするというエンティティリンキングの技術に近年関心が高まっている[1][2]。この中でも入力テキスト中に含まれる地名や住所に特化したエンティティリンキングの研究が行われている[3]。入力テキストに地点情報が付与できるようになると、例えばオンラインニュース記事や電子メールに地点情報を付与して、そのテキスト情報と同時にその地点情報を含む地図を表示することで、人間にとってより直感的な情報を提示することができるようになる。

本稿では、入力テキストとしてオンラインニュース記事をエンティティリンキングのシステムに入力して、結果として得られるエンティティの中から、特に場所、組織、人のエンティティに注目し、これらのエンティティから得られる所在地や出身地の情報を用いて、入力テキストに地点情報を付与する方法を提案する。

また、この提案手法を用いた応用例として、ユーザーが興味のある位置情報と組み合わせることで、この位置情報に関連するオンラインニュース記事を配信するアプリケーションを提案する。

2. 関連研究

日本語テキストを入力とした地点情報に関するエンティティリンキングシステムにGeoNLP[3][4] の地名テキスト解析システムが存在する。このシステムはさまざまな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、オンラインニュース記事等の入力テキストから非常に多くの地名を抽出することができている。

3. 提案手法

入力テキストに特徴的な地点情報を付与するためにエンティティリンキングシステムを構築し、このシステムから得られたエンティティを用いて特徴的な地点情報を付与する。

次に、このエンティティリンキングシステムの構成と、エンティティリンキングシステムから得られた複数のエンティティを元に特徴的な地点情報を付与する方法を示す。

3.1. エンティティリンキングシステムの構成

エンティティリンキングシステムを次の a)から d)の 4 ステップで構築する。

a) **知識ベースを形態素解析ユーザー辞書に追加**
人、組織、場所等の内部で収集したエンティティの辞書を従来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器で形態素に分割**
形態素解析用の辞書に a)で追加したユーザー辞書を含めて入力テキストを形態素解析する。

c) **エンティティを含む文字列を抽出**
形態素解析の結果から a)で追加した辞書にマッチした部分を抽出する。

d) **エンティティ曖昧性解消**
複数のエンティティを持つ文字列に対してはエンティティの曖昧性解消をしてエンティティを１つに決める。

3.2. エンティティを用いた特徴的な地点情報付与

「3.1. エンティティリンキングシステムの構成」で追加した人、組織、場所のエンティティにそれぞれ表１で示した地点情報を事前に付与しておく。

地点情報は、一般的な住所情報だけではなく、都道府県や市区町村までの住所情報、緯度経度等の内容で構成される。

表１ エンティティのタイプ別の地点情報

エンティティのタイプ	地点情報
人	出身地
組織	(企業名等の)所在地
場所	所在地

１つの入力テキストをエンティティリンキングシステムで解析すると一般的には複数のエンティティが取得でき、各エンティティから地点情報が得られるが、応用時には１つもしくは数個に絞り込んだ地点情報が求められることも多い。よって、1 つの入力テキストから特徴的な数個の地点情報を重み付きで付与することを考える。

１つの入力テキストは１つの話題を扱っていることを仮定し、1 つの話題は少数の特徴的なある範囲の地点情報が付与できることを仮定する。例えば、この範囲を都道府県の単位とすると、1 つの入力テキストから取得した複数のエンティティを、エンティティの数よりも少ない都道府県にマッピングすることで実現する。

そこで、取得したエンティティに付与された都道府県の情報を用いて、以下の式(1)で都道府県別にスコアを求めて、このスコアが高い都道府県を入力テキストに付与する都道府県とする。

$$Score_i = \sum_j \delta_{ij}\, e_j \qquad \cdots (1)$$

$Score_i$: i番目の都道府県のスコア
e_j: 入力テキストに出現するj番目のエンティティのタイプ別の重み
δ_{ij}: i番目の都道府県とj番目のエンティティの都道府県の距離に応じた重み

4. 評価

入力テキストから最も特徴的な都道府県を１つ出力する提案手法と、入力テキスト中の最初に出現した都道府県名を出力するベースライン手法の２つの手法に対する評価結果と分析結果を以下で述べる。

4.1. 提案手法の詳細

提案手法のエンティティのタイプ別の重みe_jを表２のようにし、距離に応じた重みδ_{ij}をi番目の都道府県とj番目のエンティティの都道府県が一致したときに 1、それ以外のときに 0 となるようにした。

表２：提案手法におけるe_jの値

詳細タイプ	e_j	詳細タイプ	e_j
場所: 市区町村	1.0	組織: インフラ企業	0.5
場所: その他	0.8	人: スポーツ選手	0.5
組織: 学校	0.8	人: その他	0.1

また、得られた $Score_i$の中で最も高くかつ閾値が 0.8 を超えた都道府県を１つ出力する。また最高のスコアが閾値を超えない記事は地点情報なしとする。

4.2. ベースライン手法の詳細

ベースライン手法は、入力テキストに対して形態素解析を行い、解析結果から最初に出現した都道府県名を出力結果とした。ただし、都道府県名は「東京都」のように都道府県の接尾辞を含んだものと「東京」のように都道府県の接尾辞を含まないもののどちらでもよいことにする。

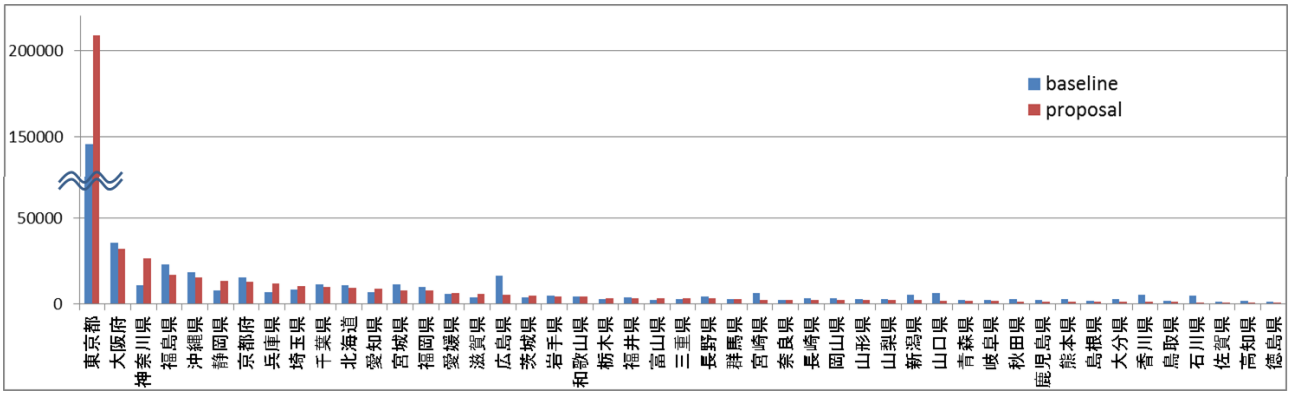

図１：都道府県別の記事数

4.3. カバレッジ評価

Yahoo!ニュースに掲載された 2013 年の約 100 万件のニュース記事に対して、上記の提案手法とベースライン手法で出力した結果を図１に示す。どちらの手法でも、すべての都道府県に対する記事が出力できていることがわかり、特に東京都のような都市部の都道府県や、2013 年に話題の多かった福島県や沖縄県の記事が多く出力できていることがわかる。

4.4. 提案手法とベースライン手法の差の分析

提案手法とベースライン手法で特に差が顕著な東京都、神奈川県、山口県、広島県について、結果に差があった記事について分析した。

- **東京都**

  提案手法では、地域情報として企業名や POI と呼ばれる場所の所在地情報も使用しているため、ニュースで頻出するような企業や POI が多く所在する「東京都」が出力されている。

> 『2014NHK ソチオリンピック・パラリンピック』放送テーマソングとなるコブクロの楽曲披露発表会が、10 月 23 日 NHK ホールで行われた。1 本の花の大輪をイメージし、作り上げたというテーマソング・・・
> **Billboard Japan［コブクロ 2014NHK ソチオリンピック・パラリンピックテーマ曲披露］**

- **神奈川県**

  神奈川県に関する記事では「神奈川県」の文字列はなく「横浜市」のように市名から記述されていることが多いが、提案手法で正しく「神奈川県」が出力されている。

> 災害時に迅速な被害状況の把握と的確な初動体制を確立するため、海老名市は１１月から・・・
> **カナロコ［被害状況素早く把握、海老名市防災カメラ導入へ／神奈川］**

- **山口県**

  ベースライン手法で、人名の「山口」から「山口県」と出力されることがある。

> ・・・麻生財務相は国会出席のため、山口俊一財務副大臣があいさつを代読した。
> **時事通信［景気、緩やかに持ち直し＝1～3 月期の地域経済―財務局長会議］**

- **広島県**

  野球チームやサッカーチーム名が「広島」と表記されることが多く、ベースライン手法では誤って「広島県」と出力されることがある。

> ・・・◆中日３―２巨人（１８日・ナゴヤドーム） 巨人が中日に競り負けた。７回、高橋由の１０号ソロで試合を振り出しに戻したが、その裏に３番手の沢村が、森野に決勝ソロを浴びた。２位の阪神が広島に完勝し・・・
> **スポーツ報知［【巨人】沢村３年連続１０敗！ボール先行し中継ぎ初失点］**

4.5. 提案手法の精度評価

「4.3. カバレッジ評価」で評価した約 100 万件のニュース記事から提案手法を用いて得られた都道府県別の記事集合からランダムに 10 件ずつ計 470 件のニュース記事を抽出し、得られた都道府県がふさわしいかを人手で判定した結果を表３に示す。

表３：抽出都道府県の正解率

	記事数	割合
正解数	415	88.3%
不正解数	55	11.7%

4.6. 提案手法の誤り分析

「4.5. 精度評価」で評価した提案手法の結果がふさわしくない都道府県と判定された 55 記事に対して誤りの傾向を調査した。

- **記事長の問題**

  正解記事と不正解記事の記事本文のバイト長を測定したところ、正解記事は平均 1549byte、不正解記事は平均 2778byte と約 1.8 倍の差があった。式(1)のスコアは記事長に依存して単調増加になるにも関わらず、閾値は固定しているため、長い記事で誤判定されやすい。

- **エンティティの抽出誤り**

  エンティティリンキングシステムから得られたエンティティが誤っているため、誤った都道府県を出力している。以下の例では「大手」を「長野県松本市大手」として取得している。

> 大手銀行の住宅ローン金利引き下げ競争が転機を迎えた。三菱東京 UFJ 銀行、みずほ銀行、三井住友銀行、りそな銀行の大手 4 行は 31 日、相次いで 10 年固定型の最優遇金利を 0．2%引き上げ・・・
> **時事通信社［金利下げ競争に転機＝住宅ローン差別化急ぐ―大手行］**

- **多数エンティティ（特にスポーツ記事）**

  スポーツに関連する記事の中には、多くのチーム名、場所名や人名が記述され、たまたま多く出現した都道府県が出力されている。

> ・・・下北沢成徳（東京）が準々決勝で昨年の国体を制した九州文化学園（長崎）に２－１で競り勝ち、全国高校総体優勝の橘（神奈川）は３回戦敗退。男子は総体覇者の星城（愛知）が準々決勝で鎮西（熊本）を２－０で下し、前回優勝の大村工（長崎）も順当勝ちした。
> **京都新聞［京都橘 ４強逃す　全日本高校バレー］**

5. 応用システムの提案

スマートフォン等のモバイル端末が普及し、端末の現在位置を端末自体が持つ GPS 等の機能を用いて取得できるようになっている。

一方、今回の提案手法を用いることで、大量のオンラインニュース記事に特定の地点や地域を関連付けることができる。これにより、端末の現在位置に関わるオンラインニュース記事等をモバイル端末に配信することが可能になり、端末所有者は「探す手間」を省きながら効率的に現在位置等に関わるオンラインニュース記事を入手できるようになる。

なお、モバイル端末がバックグラウンドで発する位置情報の利用に抵抗を感じる端末利用者は、地域や地点を明示的に示すことで同様の記事配信が可能と考えられる。

6. おわりに

本稿では、オンラインニュース記事に対して、エンティティリンキングシステムを用いた地点情報の付与手法を提案し、この手法で精度よく地点情報を付与できることを示した。また、本稿の提案手法で地点情報に関連付けた大量のオンラインニュース記事を、ユーザーに関連する位置情報と組み合わせて配信する応用システムを提案した。このような構成のシステムが普及することで、都市圏以外の地域・地方に即した効率的なコンテンツ配信が可能になり、地域・地方ニュースといったコンテンツ制作が都市圏以外で促進され、都市圏と地方における情報流通の格差が解消していくことも期待される。

参考文献

[1] TAC KBP 2013 Entity Linking Track
http://en.wikipedia.org/wiki/Entity_linking

[2] R. Mihalcea and A. Csomai, “Wikify!: linking documents to encyclopedic knowledge” in *Proceedings of the 16th ACM CIKM*, 2007, pp. 233–242.

[3] 北本, 相良, 有川, "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 年 11 月

[4] GeoNLP
https://geonlp.ex.nii.ac.jp/