GeoNLP：自然言語文を対象とした高度なジオタギングに向けて

北本 朝展[1], 相良 毅[2], 有川 正俊[3]
[1]国立情報学研究所, [2]株式会社 ピコラボ, [3]東京大学 空間情報科学研究センター
連絡先: <kitamoto@nii.ac.jp> Web: <http://agora.ex.nii.ac.jp/GeoNLP/>

(1) 目的: 自然言語テキストに含まれる地名や住所にタグ付けを行い，対応する経緯度の情報などを埋め込む処理を「ジオタギング（geo-tagging）」と呼ぶ．ジオタギングに対するニーズは大きく，たとえばオンラインニュース記事や電子メール，twitter などに含まれる「場所」を自動的に抽出して地図に表示できれば，災害時などに有用なことは明らかである．

ジオタギングを行うシステムは既に存在するが，地名語辞書に新しい語を追加することができないため，抽出できない地名が残ってしまうという問題があった．そこで GeoNLP プロジェクトでは，地名語辞書の共同構築環境を用意し，抽出できなかった地名語を利用者が自分で登録できる手段を提供することで，地名語辞書の持続的な更新とジオタギング処理の精度向上を同時に実現する．また，ジオタギング処理を WebAPI として一般に公開し，自然言語文を対象としたジオタギングを誰でも手の届く処理にすることを目標とする．

(2) アプローチ: 現在は主に GeoNLP の Web API 版を開発中である．その動作方法の概略を以下に示す．まず地名に関する公開情報源を収集，統合して地名語辞書を構築する．現在は国土数値情報（国土地理院）や駅データ.jp（株式会社コードプラス）などを利用している．次に，この地名語辞書を形態素解析ソフトウェア（MeCab）で利用できるよう，それぞれの語に適切なコストを自動的に算出する．そして自然言語テキストに形態素解析を適用し，地名語の候補となる単語列を，品詞パターンなどにより抽出する．その後，独自の評価アルゴリズムにより，「港区」が東京都か大阪市か，「東京」が東京都か東京駅かという確率を高速に算出し，同綴地名を解決する．最後に，地名語に対応する緯度経度等のメタデータを，元のテキストに付与して出力する．

(3) 結果: 東日本大震災に関連するオンラインニュース（Yahoo! News）記事約 14 万件を対象に，GeoNLP の開発版を用いてジオタギング処理を実行した結果，全記事をジオタギングすることに成功した．その精度については現在検証中であるが，結果の概略は以下の通りである．1) 地名語の抽出はおおむね成功した，2)同綴地名の解決にはまだ失敗が多く，特に鉄道駅に関しては異なる場所にジオタギングされることが多かった．今後の課題としては，GeoNLP の Web API 版の精度問題に関して，同綴地名の問題を中心に対応することで，早い段階で Web API を一般に公開したいと考えている．

(4) その他:

・ 本研究は，科学技術振興機構(JST)さきがけ「知の創生と情報社会」および国立情報学研究所共同研究費（いずれも代表：北本朝展）の支援を受けた．

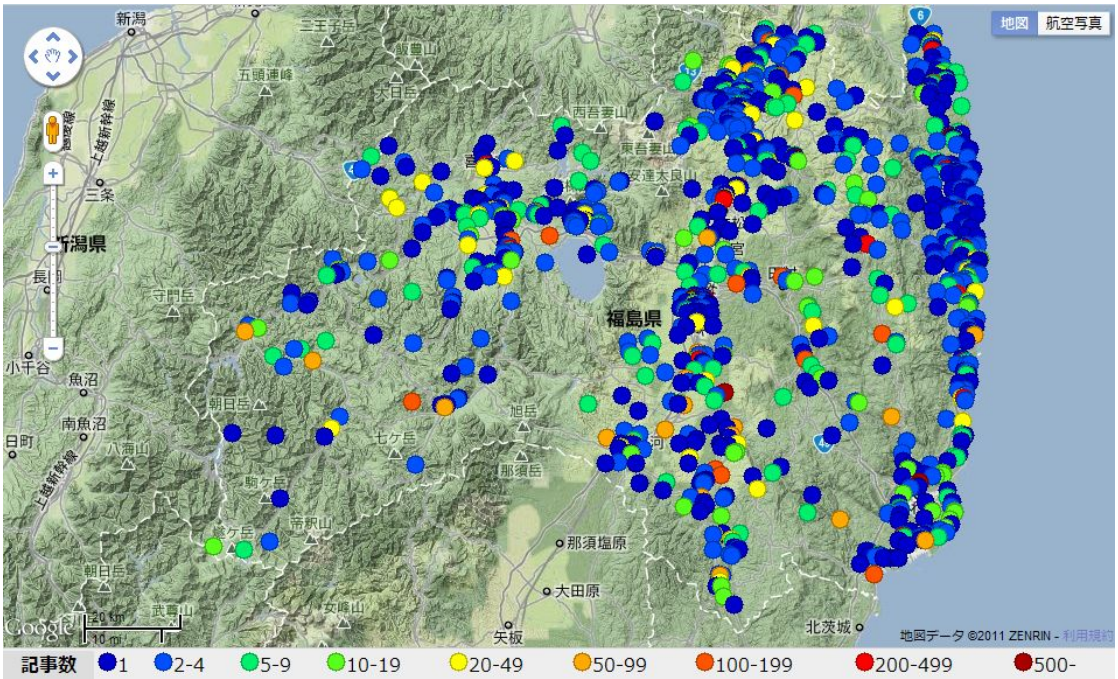

図 1: オンラインニュース記事を対象としたジオタギングの例
東日本大震災関連オンラインニュース記事を地名ごとに整理した結果を示す．福島県では全部で 1,070 個の地名を抽出した．点の色は地点ごとの記事数を表す．