地名情報処理環境 GeoNLP の紹介と歴史的な地名に関する課題

北本 朝展（国立情報学研究所）

1．はじめに

地名は地理情報における重要な一要素であるにもかかわらず、従来の地理情報システム（Geographic Information Systems / GIS）では扱いづらい存在であった。地理情報システムは、座標系という幾何学的な空間で、多様な地物を一様に扱えることが大きな特徴である。しかし地名は概念であって、そこで問題となるのは位相学的な空間である。例えば「○○の一部」といった関係性も重要であるし、特に歴史地名の場合はその概念が指し示す中心点や境界も不明確である。こうした状況は GIS のモデルと整合しないにもかかわらず、従来の歴史地理情報処理では GIS を無理やり使うという方法論が主流であった。そこで、我々が推進するプロジェクト GeoNLP は、地名を対象とするオープンな地名情報システム（Toponym Information Systems / TIS）を構築し、GIS とは異なる体系で地理情報を扱える情報基盤を創生することを目指す。

２．地名の曖昧性

そもそも「地名とは何か？」という問題に統一された定義はないが、GeoNLP プロジェクトでは「位置に関する属性をもつ固有名（named entity）」と定義する。狭義の地名は、土地に付与された名前のみを扱うが、GeoNLP では広義の地名として、住所などの行政的な位置表現だけでなく施設などの人工物に付与された名前も対象とし、こうした名前を固有名と位置（緯度・経度）との対応表である地名辞書にあらかじめ登録する。この問題設定において地名情報処理の課題とは、1) 自然言語文中に出現する地名を見つけ、2) これに対応する地名辞書のエントリを検索し、3) その位置を結果として示す、と整理できる。

ここで技術的に困難な挑戦課題となるのが、地名という概念の「曖昧さ」の処理である。上記に示した 3 段階の処理では、それぞれ対処すべき曖昧さの課題が異なっている。まず 1)では、自然言語文の中から地名文字列を切り出さねばならないが、日本語の場合は単語が空白文字等で区切られていないため、形態素解析等を用いて単語に分割せねばならず、ここに曖昧さが生じる。次に 2)では、ある地名文字列に対して地名辞書に複数のエントリが存在する場合が多々あり、そのどれを選ぶかという点に曖昧さが生じる。この曖昧さを多義性（ambiguity）と呼ぶ。最後に 3)では、地名に対する位置をどう定義するかという点に曖昧さが生じる。代表点および境界を明示的に定義できる場合は問題ないが、日常的

に使う地名や歴史的な地名の場合はどちらも曖昧な場合が多々ある。この曖昧さを漠然性（vagueness）と呼ぶ。こうした曖昧さを解決できる地名情報基盤として GeoNLP を誰でも簡単に使えるようにすることが、プロジェクトの最終的な目標である。

現在のところ GeoNLP は多義性の問題を中心的に扱っており、曖昧性解消のために自然言語文に地名が出現する際の文脈情報を利用する方式を研究している。特に地名の共起に注目し、同じ種類の地名や距離的に近い地名が出現しやすいとの仮定のもと、共起の起こりやすさを評価する関数を用いて共起しやすい候補を選んでいる。ただし評価関数は改善の余地が大きく、文脈情報が不足している場合に適切な候補が選ばれにくいという問題がある。さらに根本的な問題として、precision と recall のジレンマという問題は今後の重要な研究課題である。地名辞書のエントリを増やしていけばいくほど、確かに抽出可能な地名は増える（recall は上昇する）ものの、地名ではない単語を誤って地名として抽出してしまう場合も増える（precision が低下する）ことが避けられない。これはこの種の処理において原理的に避けられない問題であり、特に歴史地名のように一般名詞的な地名が多い地名辞書を使えば、より深刻な問題となることも予想される。実用的な観点から precision と recall のバランスをコントロール可能な手法が必要である。

３．歴史地名に関する課題

GeoNLP を歴史地名に適用する際に生じる重要な問題として、1) 古文用形態素解析器の導入、2) 地名の変遷のデータベース化、3) 地名の有効期間の活用、を論じておきたい。1)は GeoNLP が用いる形態素解析器（MeCab）が現代文に最適化されており、古文には必ずしも適用できないという問題を指すが、これは形態素解析器の差し替えを可能にすることで部分的には対処可能である。2)は地名の変遷をどう辞書に登録するかという問題を指すが、これは地名辞書の本体に登録するのではなく、我々が開発する GeoLOD に地名の関係を Linked Data として登録する方法で解決したい。最後に 3)は、地名の変遷を登録するだけでなく、地名が主に出現する期間を属性として定義することで、期間を指定すると特定の地名が選ばれやすくなる評価関数を定義する方法を考えている。いずれも、実際の歴史的なテキストで試しながら改善していく必要があり、共同研究などを進めていきたい。

GeoNLP はオープンソースとオープンデータという、オープンな方針のもとに地名情報処理環境を実現することを目標とする。プロジェクトで開発したソースコードやデータは、ウェブサイト https://geonlp.ex.nii.ac.jp/ から入手できる。しかしこの問題は我々だけで解決できるものではなく、ソフトウェアとデータの改善はコミュニティベースで進めていきたいと考えている。関心のある方々は、ぜひご協力いただきたい。