論文の内容の要旨

中近世スペイン語古文書の統計的年代推定・場所推定

Statistical Methods for Dating and Geolocation of Medieval and Modern Spanish Documents

川崎　義史

本研究の目的は，中近世スペイン語古文書の作成年代と作成場所を言語的特徴に基づき統計的に推定する方法を開発することである。現存する文献史料に基づく歴史学や通時言語学の研究において，作成年代と作成場所が不明の文献がいつ・どこで作成されたかを推定（特定）すること，及び文書の真贋を判定することは最重要課題である。正確な歴史の記述には，信頼できる文献資料が不可欠だからである。本研究では，文献史料を文書（document），文書の作成年代の推定を年代推定（dating），文書の作成場所の推定を場所推定（geolocation）と呼ぶことにする。データセットとして中近世スペイン語古文書コーパス CODEA（Corpus de Documentos Españoles Anteriores a 1700）を用いた。

本論文の貢献は，以下の四点である。

一つ目は，作成年代と作成場所を同時に推定する方法の提案である。管見の限り，同時推定を提案した研究は存在しない。先行研究では，年代推定・場所推定は個別に行われている。年代推定の研究では言語の空間的変異を無視している。同様に，場所推定の研究では言語の時間的変異を無視している。しかし，同年代における空間的変異や同地域における時間的変異が存在するので，言語の変異を扱う際には，時間軸と空間軸を同時に考慮する必要がある。実験により，作成年代と作成場所の個別推定に比べ，同時推定の方が常に予測精度が高くなることを示した。ただし，トレードオフとして，同時推定では計算量が増加する。

二つ目は，時空間カーネル平滑化の応用である。カーネル平滑化とは，カーネル関数を用いて，ある関数からより滑らかな関数を推定する方法である。本研究では，素性の出現頻度に対して，カーネル平滑化を適用した。カーネル平滑化では，関数に関して線形性やS字カーブなど特殊な性質を仮定する必要がない。カーネル平滑化により，データセットに点在する欠損値補間と頑健な推定が可能になる。時間軸と空間軸の各々においてカーネル平滑化を行う先行研究は存在するが，両者を組み合わせた時空間カーネル平滑化を文書分類のタスクに応用した研究は，管見の限り，存在しない。ただし，上述の利点がある一方で，本研究の実験では，カーネル平滑化の年代推定・場所推定への効果は限定的だった。

三つ目は，言語に依存しない年代推定法・場所推定法の提案である。素性として文字 *n*-gram を用いることで，単語毎に分かち書きされない言語（日本語や中国語など）の文書や，正書法が確立されていない時代の文書もそのまま扱うことができる。素性として単語 *n*-gram を用いる場合は単語分割やステミングなどの技術開発が必要となる。また文字 *n*-gram は，単語 *n*-gram に比べ，素性数を大幅に削減できるという利点がある。

四つ目は，文献学的特徴に基づく計量的な年代推定法・場所推定法の提案である。スペイン語で書かれた文書の年代推定・場所推定は，スペイン語文献学の大きな目標の一つである。今日まで，記述的研究には大きな蓄積があるが，年代推定・場所推定を正面から扱った研究は存在しない。先行研究により，各年代や地点に特有の言語的特徴は，ある程度，判明している。しかし，どれだけ細かな記述をしても，記述は記述に過ぎない。スペイン語史の記述から年代推定・場所推定という予測に移るには，計量的なアプローチが必要となる。重要性の異なる複数の証拠から総合的に判断するには，専門家の「勘」よりも，計量的手法の方が信頼性・実証性に勝るからである。「塵も積もれば山となる」というように，小さな証拠でも複数集まれば，大きな差異を生むことになる。本研究では，各々の証拠，つまり文献学的特徴の重みをデータから決定した。多くの文書に現れる特徴ほど，大きな重みが与えられる。この重みをプロットすることで，各特徴の年代推移・地理的変異を可視化することができる。これは，年代推定・場所推定の副産物として，スペイン語文献学への大きな貢献となる。

本論文の構成は以下の通りである。第 1 章では，研究背景の説明と問題定義を行った。第 2 章では，年代推定・場所推定に関連する先行研究を紹介した。第 3 章では，本研究で用いるコーパスの概要と記述的統計を示した。第 4 章では，素性として用いる文字 *n*-gram と文献学的特徴について説明した。第 5 章では，欠損値補間と頑健な推定を可能にするカーネル平滑化について説明した。第 6 章では *n*-gram 言語モデルによる年代推定法・場所推定法を，第 7 章では JS 情報量に基づく年代推定法・場所推定法を，第 8 章ではナイーブベイズ多変数ベルヌーイモデルによる年代推定法・場所推定法を説明した。第 9 章では，年代推定・場所推定の実験結果を示した。第 10 章で，結論と今後の課題を述べた。

以上。