半教師ありトピックモデルを利用した Twitter ユーザの生活に関わる地域の推定 Estimation of Twitter User's Life-Area Using Semi-Supervised Topic Model

学籍番号：201221592
氏名：堂前友貴
Yuki DOUMAE

Twitter において，ユーザの生活に関わる地域は，社会行動の分析において重要な属性の一つであるが，プロファイルに明示的に記述されていることは少ない．そこで，本研究では，Twitter ユーザを対象として，半教師ありトピックモデルを利用した地域特徴語の選択に基づく，生活に関わる地域属性の推定手法を提案する．ここで，ユーザの生活に関わる地域とは，居住地や勤務地など，日常生活で関わることの多い地域とする．ユーザの生活に関わる地域を推定することで，ユーザ支援や分析への応用が期待できる．

本研究では，ツイート中のトピックには，地域に特徴があるものと，共通で出現するものがあるという仮定のもと，地域に特徴的なトピックを生成し，地域特徴語を選択することで，ユーザの生活に関わる地域を推定する．具体的には，地域情報サイトから収集した地域語を含むツイートを教師データとした，半教師ありトピックモデルにより，地域に特徴的なトピックを抽出する．そして，トピックから選定した地域特徴語を使用し，ツイートごとに地域ラベルを付与する．各ユーザの生活に関わる地域は，ユーザの ツイートに割り当てられた地域ラベルに基づき推定する．

提案手法に基づき，都道府県を，生活に関わる地域の単位とし，16 の都道府県を対象として，ユーザの生活に関わる地域を推定する実験を行った．まず，教師なしトピックモデルを適用して，地域特徴語の選択を行い，課題について検討を行った．次に，その課題を改善するために，半教師ありトピックモデルを適用した地域特徴語の選択の有効性について検証した．実験では，2012 年の日本語ツイートを対象とし，半教師ありトピックモデルを一カ月ごとに適用した．地域特徴語の選択に用いたデータは，各月を単位として，ラベルが付与された 1,600,000 件のツイートと，ラベルが付与されていない最小 4,640,000 件，最大 4,960,000 件のツイートである．人手によって判定した 1,600 人のユーザについて評価を行ったところ，精度 0.65，再現率 0.67，F 値 0.66 の評価値が得られた．

本研究の貢献を以下にまとめる．

1. 半教師ありトピックモデルを用いることで，教師なしトピックモデルよりも適切に地域に特徴的なトピックを生成し，地域特徴語の選択が行えることを示した．

2. ツイートごとにトピックを付与し，トピックに付与されたラベルに基づいて地域を推定することが有効であることを明らかにした．

研究指導教員：佐藤 哲司
副研究指導教員：関 洋平