固有表現抽出における機械学習モデルの効率的な教示方法に関する研究

Efficient teaching methods of machine learning models in Named Entity Recognition

学籍番号：201921631
氏名：小林　滉河
Kobayashi Koga

固有表現抽出とは人名や地名といった固有表現を文章から自動的に取り出すことを目的とした自然言語処理のタスクである．取り出された固有表現は検索や情報推薦，形態素解析といった様々なタスクに応用され，応用先の精度を大きく左右させる要因になるため，非常に重要な技術として知られている．近年，固有表現抽出は一般的な分野だけではなく，物質材料やバイオ，薬学といったより専門性の高いドメインに対する適用を目指す研究が増加している．固有表現抽出は系列ラベリングタスクの一つとして知られており，教師データには通常アノテーションコーパスといわれるデータセットを利用することが多い．しかし，アノテーションコーパスは文章中に含まれる単語に一対一で対応する固有表現タグを付与する必要があり，作成コストが高い．専門分野における教師データの作成には，専門家によるアノテーションが必要不可欠である，そのため低コストで高い抽出性能を持つ固有表現抽出手法の提案が求められている．

本研究では，教師データの作成とそのデータを用いてモデルを学習する一連の流れのことを教示と呼び，固有表現抽出における教示方法をテーマに次の 3 つの課題に取り組んだ．1 つ目は教師データの作成のため，大量のアノテーションを必要とすること．2 つ目は辞書や生コーパスといった学習資源を有効に活用できていないこと．最後の課題はアノテータが間違ったラベルを付与する可能性があること．それぞれの課題に対して能動学習，遠距離教師あり学習，クラウドソーシングにおけるタスクの集約による解決・調査を試みた．第 2 章ではアノテーションコストの削減のため，点予測による単語単位をクエリに用いた能動学習を提案し，有効性を示した．第 3 章では固有表現抽出における遠距離教師あり学習について，辞書を用いた生コーパスへの間違ったアノテーションの付与を問題として取り上げ，ラベリング誤りを考慮するモデルの提案と既存手法の比較を行った．実験の結果，辞書の整備が不十分で多くの誤りが発生する場合において，再現率の向上が見られた．第 4 章にて固有表現抽出におけるクラウドソーシングでは，系列ラベリングタスク固有の集約で起きうる問題について述べ，生成モデルの複雑性と集約性能の関係について調査を行った．調査の結果，複雑なモデルを単語の出力分布に仮定すると出力確率が支配的になり遷移確率を無視したラベリングが増え，出力分布を導入しない場合には遷移確率の影響が大きすぎるためアノテーションを無視することが分かった．

研究指導教員：若林　啓
副研究指導教員：佐藤　哲司