点予測と能動学習を用いた固有表現抽出の提案

小林　滉河

固有表現抽出（Named Entity Recognition; NER）とは，文章の中から人名・地名・組織名といった固有名詞や時間表現，金額表現等の語句を抽出する自然言語処理の技術のことであり，情報抽出タスクの一つとして知られている．抽出された固有表現は情報検索や対話システムといった様々なタスクに応用される．近年では食材名や調理方法などの固有表現を持つ料理ドメインや，物質材料ドメイン，バイオドメインといった様々な専門分野に対して固有表現抽出を行う研究が増えている．

既存の研究では大量のアノテーションコーパスが利用できる状況において高精度で固有表現抽出を行うことを目的としている．しかし，専門ドメインにおけるアノテーションコーパスは一般ドメインのアノテーションコーパスに比べて非常に少ない．その上，専門ドメインのアノテーションには専門的知識が要求される為，一般ドメインに比べてコーパスの作成に大きなコストがかかる．本研究は一部の箇所に対してタグを付与出来る部分的アノテーションコーパスを利用可能な点予測を用いることで，アノテーションが少ない状況においても高い性能を実現できる固有表現抽出手法と能動学習への適用を提案する．能動学習は，ラベルなしデータ集合の中から情報量が多いデータのラベルをアノテータに問い合わせることで少ないコストで効率よく学習を進める手法である．提案手法では部分的アノテーションを教師データとして利用可能である点予測のメリットを最大化するために能動学習を適用した．

提案手法の有用性を示すため，提案手法と既存手法である条件付き確率場と深層学習モデルに能動学習を適用させ，同一のテストデータに対しての性能を計測した．比較実験の結果，アノテーションの数が少ない場合において提案手法が既存手法を上回った．

（指導教員　若林啓）