文字分散表現に基づく辞書情報を活用した固有表現抽出器の学習に関する研究

Training Named Entity Recognizer by Character-based Word Classification using a Domain Specific Dictionary

学籍番号：201721690
氏名：平松　淳
Hiramatsu Makoto

固有表現抽出は自然言語処理の基盤的なタスクの 1 つであり，活発に研究が行われている．固有表現抽出とは，テキストの中に含まれる人名，組織名，場所名などの固有表現を自動的に抽出するタスクである．既存研究の多くは新聞記事のテキストからの固有表現抽出に取り組んでいる．しかしながら，現実世界への応用を考えると新聞記事以外のテキストに対する固有表現抽出の重要性は高い．例えば，バイオインフォマティクス分野では，遺伝子名やタンパク質名のデータベースの自動構築のニーズが高く，これはテキストからの自動用語抽出タスクとして解ける．また，料理テキストは食材名や調理器具，調理操作などの用語を自動で認識し，認識結果を用いて料理レシピを機械可読な形式へと変換する研究が行われているなど，言語の基礎解析として固有表現抽出の技術が活用されている．一般的に，特定のドメインのテキストに対して固有表現抽出を行うためには，対象ドメインのテキストに対して固有表現のアノテーションを付与したコーパスを構築する必要がある．しかし，特定のドメインのコーパス構築はドメイン固有の用語に対する専門的知識が必要であり，コストが高い．

本研究では，固有表現のアノテーションが付与されたコーパスに加え，ドメイン固有の辞書を活用して固有表現抽出器を学習する．我々は料理オントロジーと呼ばれる料理ドメインの用語の関係が記述されたデータを加工して辞書を構築した．得られた辞書は単語と単語が属するカテゴリのペアをエントリとしたデータである．辞書を教師データとして，単語が入力されたときにその単語が属するカテゴリを予測する分類器を学習する．この分類器は単語の文字レベル特徴量を用いてカテゴリを予測する．これにより，辞書には直接含まれない単語に対して，辞書情報を活用した単語特徴量を獲得できる．分類器は入力に対してカテゴリ数次元のベクトルを出力し，各次元の値は入力された単語が各カテゴリに属する確率を示す．この出力ベクトルを固有表現抽出器の特徴量に加えることで，提案手法は辞書情報を考慮できる．

提案手法の有用性を示すため，レシピドメインの固有表現抽出タスクを用いて既存手法との比較実験を行った．比較実験において提案手法が F 値において既存手法を上回る性能を発揮することを確認した．

研究指導教員：手塚　太郎
副研究指導教員：若林　啓