自然言語教示によるフレーズ抽出器の学習に関する研究

齊藤　亮将

フレーズ抽出とは，テキストデータ（自然言語文）から特定の単語やフレーズを抽出する情報抽出タスクであり，様々な技術に利用されている．例えば，固有表現抽出（Named Entity Recognition; NER）もフレーズ抽出の一つで，質疑応答システムや対話システムなどに利用されている．他にも専門用語抽出などが挙げられ，これは論文上の専門用語を専門用語辞書のようなコンピュータ上で扱いやすい形へと変換するために利用されている．

このようにテキストデータからフレーズを抽出することは重要なタスクであり，膨大なテキストデータを扱うにはフレーズ抽出の自動化が必須である．そこで，機械学習を用いて学習器を学習させようとすると訓練データに大量のフルアノテーションコーパスが必要になる．事前に用意出来る環境にあれば良いが，大量に用意しなくてはいけない点と一つのフルアノテーションコーパス作成にかかる時間やコストを考えると，アノテーションコーパスの作成がフレーズ抽出器モデルの学習を行う上で障壁となる可能性もある．

そこでアノテーションコストと訓練データ数を削減することにより，低コストで効率的なフレーズ抽出器の学習を行うことが本研究の目的である。提案手法では，自然言語による説明文「自然言語教示」を用いて，少ない訓練データから効率的なフレーズ抽出器の学習を行う．自然言語教示を用いることで，アノテータが単純にラベル付与をする以上の情報を獲得することができ，獲得した情報の関数化や単語辞書を検索する上でのキーワードとして使用することで，生コーパスデータに対して機械的にラベル付けを行い，擬似的に大量の訓練データを作成することができる．この手法により，必要な訓練データ数の削減とアノテータの一タスクあたりのコストを軽減させる．

提案手法の有用性を示すために，擬似的に作成した訓練データからフレーズ抽出器を学習させ，用意したテストデータでモデルの性能を示した．

（指導教員　若林　啓）