能動学習による複合語を考慮した専門用語抽出

小田倉　史麿

専門用語は，専門分野における特定の概念や事物を表す言葉である．新たな研究論文の公開に伴っては，日々数多くの専門用語が登場している．このため，専門用語を収集および整理することは，今後の研究に資する活動である．しかし，従来では，専門家が人手で研究論文から専門用語を収集することが多く，その作業にかかるコストが大きいために，用語情報を最新の状態に保つことが困難である．このため，コーパス中から自動で専門用語を抽出する技術の高度化が求められている．

既存の研究には，専門家によって与えられた語と並列関係にあたる語をコーパス中から探し出す専門用語抽出のアプローチが存在する．これまでには，パターンマッチングによって並列関係を得る手法や，語の分散表現によって意味合いの近い語を獲得する手法などが提案されている．しかし，ひとつの言葉は概ね複数の属性を持ち，文脈によって意味が変化する．このため，抽出された専門用語が，専門家が念頭におく文脈にしたがうとは限らず，専門家の意図に沿わない用語群が抽出されてしまう懸念がある．

本研究では，専門用語の候補を段階的に専門家に提示し，念頭におく文脈にしたがう用語を繰り返し選択してもらうことで，専門用語を帰納的に類推しながら抽出する枠組みを提案する．提案手法では，コンピュータとラベル付けを行う専門家が対話を繰り返し，効率の良い学習を目指す「能動学習」を用いることで，人間による教師ラベル作成のコストを抑えつつ，文脈を考慮した専門用語抽出を行う．提案手法では，既存の言語モデルを用いた品詞マッチングによる候補の絞り込みを行い，共起頻度の特徴量ベクトルを SVM（Support Vector Machine）にかけることで，獲得する候補ランキングの並べ替えを行う．また，用語抽出における課題のひとつである「複合語の考慮」を可能にするために，事前に転置インデックスを作成することで，複合語の出現位置の動的な獲得を行う．

提案手法の有効性を示すため，比較手法として「分散表現獲得」および「固有表現抽出」を挙げ，提案手法と同様に能動学習を適用したうえで，一定の文脈によって分類された既知の専門用語群をテストデータとし，抽出性能を比較した．

（指導教員　若林啓）