自然言語文章からの上位下位関係自動抽出手法に関する研究

平松　淳

情報処理技術の飛躍的な進歩に伴い，ウェブ上に存在する文書の数は日々増大している．これら大量の文書は現実世界に関する多種多様な知識を含んでおり，質問応答をはじめとする自然言語処理のタスクのための情報源としての価値が高い．しかし，ウェブ上に存在する文書の多くは自然言語で記述された非構造化データであり，現在のところ計算機は自然言語を人間のように理解することはできない．そのため，自然言語で記述されたデータを計算機で活用するためには，データをなんらかの機械可読な形式で表現する必要がある．

本研究では，現実世界に関する知識の中でも基盤的な役割を担う上位下位関係について，これを自動的に抽出する手法の提案を行う．既存の研究では，与えられた単語ペアが上位下位関係にあるかどうかを高精度に判定することを目指している．しかし，上位概念・下位概念は一般に単語ではなくフレーズで表されることから，自然言語文章から候補となるフレーズペアを全列挙して上位下位関係判定を行うことは困難である．このため，本研究では，系列ラベリング問題の枠組みを利用することにより，フレーズペアの抽出と上位下位関係の判定をシームレスに行う手法を提案する．

提案手法の有用性を示すため，提案手法と既存手法を同一のテストデータに対して適用し，上位下位関係の抽出を行なった．性能の比較実験の結果，提案手法の上位下位関係抽出の性能が，精度で既存手法を上回り，F1 値は既存手法とほぼ同等という結果になった．一方で，再現率が既存手法と比較して低下した結果となり，今後考慮すべき課題である．

提案手法では，上位概念・下位概念の同定を系列ラベリング問題として定式化しているため，従来の手法では抽出することができなかった，形態素解析によって複数の単語にわかれてしまう複合語を抽出すること可能となった．このため，提案手法は上位下位関係の自動抽出というタスクに関する新しいアプローチを提供することができたと結論づける．

（指導教員　若林啓）