Closed class に着目した教師なし品詞タグ推定精度向上の検討

柴田　尚樹

自然言語処理のタスクに、品詞タグ推定がある。品詞タグ推定とは、文書中の単語の品詞を推定することである。品詞タグ推定は、自然言語処理の基盤的な技術であるため、テキスト解析、機械翻訳や質問応答といった様々な応用タスクの性能向上に関わっている。

品詞タグ推定は、教師あり品詞タグ付けと教師なし品詞タグ推定に分類される。現状では、教師あり品詞タグ付けの正解率が教師なし品詞タグ推定の正解率を上回っている。しかし、教師あり品詞タグ付けでは、ウェブ上のテキストといったコーパスや辞書データに存在しないスラングや顔文字を含むテキストへの対応が困難である。一方、教師なし品詞タグ推定は辞書データに存在しない単語にも対応しうる推定手法である。このような応用範囲の点から、教師なし品詞タグ推定が教師あり学習と比べて精度が大きく劣っている現状の改善は重要な課題であると言える。

本研究は、教師なし品詞タグ推定精度の向上を目的としている。

そのために教師なし品詞タグ推定でしばしば用いられる隠れマルコフモデル(HMM)に対して、オープンクラス(Open class)とクローズドクラス(Closed class)に着目した改良を加えた。

オープンクラス、クローズドクラスとは、統語論における品詞の分類方法のひとつである。 オープンクラスには 名詞や動詞など、新たに言葉が増える可能性が高い品詞が属し、クローズドクラスには、冠詞や前置詞など、新たに言葉が増える可能性が低い品詞が属している。

HMM に改良を加え、状態に属す単語の出現回数の多寡によって、クローズドクラスに属す単語が大部分を占めている状態を推定し、その状態においてオープンクラスに属すと推定される単語を別の隠れ状態に遷移させる、という方法を検討した。

いくつかの検討を行ったが、その状態において出現回数第(t+1)位の単語の出現回数を第 t 位の単語の出現回数で割った値が 0.20 以下の時、第(t+1)位以下の出現回数の単語を別の状態に遷移させる手法が今回の実験では最も有効であった。

評価手法に Many-to-One、V-Measure に加え、独自に定義した ClosedBasedMany-to-One を用いたところ、若干の精度向上を確認することができた。ClosedBasedMany-to-One は、単語ごとの正解率の和を、語彙数で割ったものとして定義した。

今後の展望としては、クローズドクラスに属す単語の推定方法および、オープンクラスに属す単語の推定方法のさらなる模索が挙げられる。

（指導教員　若林 啓）