SVMと３つ組／４つ組法による日本語係り受け解析に関する研究

石村 健二 (110011)

北陸先端科学技術大学院大学 情報科学研究科

2003年2月14日

キーワード: 日本語係り受け解析, 3つ組／4つ組モデル, Support Vector Machine, HPSG.

本研究は, 日本語係り受け解析における３つ組／４つ組モデルにSupport Vector Machine(SVM)を導入し，構文解析器の解析精度の向上を図ることである．近年，文法フォーマリズムが発達し，新聞や雑誌などの文章に対して構文木を出力できる文法(SLUNGなど)が研究されているが，そこで扱われる文法とはこれらを対象に可能な限り構文を列挙したものであり, 曖昧性が高い．また, 金山らが提案した３つ組／４つ組法は，この曖昧性を取り除くために開発された統計的手法で，高い精度が報告されている．しかしながら，この３つ組／４つ組モデルでは，他の類似した手法よりもさらに多くの素性を取り扱っているため，ここで用いられている最大エントロピー法(ME法)による係り受け解析では，扱う素性を慎重に選択する必要があった．そこで，高次元の素性集合を用いても過学習を起こしにくいSVMを係り受け解析に導入する．

日本語の構文解析において, 係り受け解析は重要な基本技術の一つとして認識されている．一般的に係り受け解析とは，２文節間の係りやすさを数値化した係り受け行列を作成し，動的計画法などを用いて文全体が最適な係り受け関係となるように, 各文節の係り受け関係を求めることである．従来の係り受け解析では，ある係り元文節の複数ある係り先文節候補に対して, それぞれ独立な確率で係り先文節を決定していた．これに対して，３つ組／４つ組モデルでは，各係り先文節候補に対する係り受け確率推定のために, 係り元文節と複数の係り先文節候補の素性を同時に考慮する確率推定モデルを定義する．この確率推定モデルにより, ある係り元文節に対する各係り先文節候補は，相対的・合理的に係り元文節に対する係り先文節を決定することができる．

本手法のベースである３つ組／４つ組モデルは，HPSGに基づいた文法およびヒューリスティクスによって，係り先文節候補を３文節以下に絞り，係り元文節と複数の係り先文節候補の素性を, 同時に考慮する確率推定モデルにより係り受け確率を推定している．ここで言われているヒューリスティクスとは，日本語の文節の係り先文節を観察した結果,

文法的に許されている係り受け候補の中で，係り元から最も近い文節，２番目に近い文節，最も遠い文節のいずれかに係り受け関係が成立する確率が 98.6%と高いこと利用している．このモデルは，文法とヒューリスティクスによる係り先候補の絞り込みにより EDR コーパスに対して 88.6 %という高い文節正解率を達成している．しかしながら, 金山らがモデルの推定に使用している ME 法には, 人手により学習に用いる素性を適切に選択する必要があり，素性が適切に選択されなかった場合, 過学習を起こすなどの問題が指摘されている.

SVM は統計的機械学習法であり，学習サンプルと分類境界の間隔を最大化するような戦略に基づく手法の線形二値分類器である．また，従来の学習モデルと比較しても極めて汎化能力が高く，高次元の学習データを用いても過学習しにくい．さらに，Kernel 関数と呼ばれる計算技術を用いる事により，計算量をほとんど増やすことなく，線形分離できない空間を高次元に写像することができ，これまで線形分離することができなかった空間を分離することができる．また，Kernel 関数を変更することで，線形・非線形などの様々なモデル空間を自由に仮定することができる. さらに, ME 法とは違い人手により素性の組み合わせを考慮しなくても, Kernel 関数によって複数の素性の組み合わせを考慮した学習モデルを製作することが可能になるなど，多くの利点がある．この利点を用い, 工藤らは日本語係り受け解析に SVM を用いて京大コーパスで評価実験を行った, その結果，非常に少ない学習データで文節正解率が 89.09 %という高い精度が報告されている.

本研究で用いる３つ組／４つ組モデルとは, ある係り元文節に対して, 文法的にその文節が,“ 係り元文節に一番近い文節 ”,“ 係り元文節に二番目に近い文節 ”又は,“ 係り元文節から最も遠い文節 ”のいずれと係り受け関係が生じるかを判別する多値分類器である. そこで, 既存の SVM を多値分類器に拡張するため one vs. rest 法を用いた. また，金山らが ME 法を用いて３つ組／４つ組モデルを構築したときには、データの過疎性の問題が生じるため素性に一部しか入れられなかった各文節の主辞の単語を素性にすべていれ，さらなる性能の向上を図った.

最後に以上に述べた統計モデルを使った構文解析実験を行った．実験には, 金山らが３つ組／４つ組モデルを構築したときに使用した素性のうち, 明示的に素性の組み合わせを与えられているものを除いた全ての素性と, 各文節（係り元文節, 係り元文節に一番近い文節, 二番目に近い文節, 最も遠い文節）毎に、主辞として学習コーパス中に現われた単語全てをリスト化し, SVM の素性として与えた. 学習およびテストには EDR コーパスを用い, それぞれ 101,540 文, 2,603 文を与えたが, 得られた精度は最大 87.7%であり，金山らが ME 法を使って達成した性能には及ばなかった.