系列ラベリングのための素性構造と文脈長を同時に考慮した素性抽出

石原 靖弘　　竹内 孔一
岡山大学 自然科学研究科
{ishihara, koichi}@cl.cs.okayama-u.ac.jp

1　はじめに

自然言語処理における固有表現抽出や品詞タグ付け，意味役割付与などのタスクは系列ラベリングでモデル化することができる．系列ラベリングには確率モデルの Conditional Random Fields(CRF) が代表的である．CRF は最大エントロピーモデルであるため任意の数の素性を追加することができるが，適切な識別を行うためには使用する素性による素性空間がラベルを分離できるものになっている必要がある．現在使っている素性で十分な性能が得られないとき，予測したいラベルの推測に用いるラベル周辺の情報の範囲を広げて，より遠くの文脈情報を素性に用いるようにするといった方法や，素性間の共起関係を新たな素性として追加するといった方法がある．しかし素性空間の次元を増やすことは必ずしも識別能力の向上にはつながらない．素性の数をむやみに増やしても過学習に陥り，汎化性能は低下する．例えば，ラベル周辺のある位置とある位置に出現した単語の共起情報を新たに素性として加えると最大で単語数×単語数の素性ができる．実際は学習データに出現したもののみを使用すると追加される素性の数は高々サンプル数にとどまるが，問題はこのように共起関係を考慮した素性の潜在的な素性空間の大きさが組み合わせの数を増やすにつれて指数関数的に大きくなることである．そのため複数の共起関係を考慮した素性は学習データ中に 1 度しか観測されないものがほとんどになり，そのような観測にもとづき推定されたパラメータは統計的に信頼できるものにならない．また使用する素性としては単語の表層だけでなく単語の基本形や品詞いった単語の表層より抽象的な素性も同時に使われる．そしてある位置にある単語の表層とある位置にある品詞の共起情報といった素性も使用することができる．そのような共起を考慮した素性の数は考慮する文脈の長さ，単語や品詞といった素性の種類数，組み合わせる数によって指数関数的に増加していき，過学習や学習にかかる時間の増加を招く．そのため素性は無差別に追加すればよいというわけでなく，ラベルの推定に有効な素性を選択することが重要になる．

Haruno ら [1] は単語の表層や品詞といった複数種類の素性を考慮した可変長オーダーのマルコフモデルを利用して品詞タグ付けを行っている．例えば,「彼女/は/東京/に/上京した」という文に品詞をタグ付けする場合を考えると,「に」の品詞に対する条件付確率は P(a—助詞, 固有名詞, に) のようになるかもしれない．また,「上京した」の品詞に対しては P(a—は，東京，助詞，上京する) になるかもしれない．このように文脈に応じて考慮する文脈の深さと，素性の抽象度を変えながら予測を行うモデルになっている．Haruno らは文脈を表現する接尾辞木を構築し，その各ノードにその文脈に対応する条件付確率分布を割り当てている．接尾辞木のエッジには素性が関連付けられていて文脈の長さに応じて根から葉までの長さが異なる．

Haruno らは接尾辞木の構築し，接尾辞木の各ノードに対応する確率分布を使って品詞の予測を行っている．一方，我々は Haruno らと同様に接尾辞木を構築するが，接尾辞木は素性抽出に用いる点が Haruno らと異なる．

以下ではまず，簡単のため素性の種類が 1 つの場合の可変長オーダーの文脈を表現する接尾辞木の構築の手続きについて説明したあと，複数の素性への対応を説明する．

2　接尾辞木の構築

2.1　可変長の文脈への対応

木の構築は根から順に貪欲法的に行う．Weinberger らの情報量基準 [2](以下，WIC と呼ぶ) を利用して枝

を生やすべきかどうかの判断を行う．

$$\begin{aligned}\Delta(sb) &= \sum_{a\subseteq A} n(a|sb) log \frac{P(a|sb)}{P(a|s)} \\ &= n(sb) \sum_{a\subseteq A} \frac{n(a|sb)}{n(sb)} log \frac{P(a|sb)}{P(a|s)} \\ &= n(sb) \sum_{a\subseteq A} P(a|sb) log \frac{P(a|sb)}{P(a|s)} \\ &= n(sb) D_{KL}(P(\cdot|sb)||P(\cdot|s)) \qquad (1)\end{aligned}$$

n(sb) は文脈 s に続いて b が出現した回数を表し，これは条件付確率 P(a—sb) の推定に使用したサンプル数に等しい．KL(P(a—sb),P(a—s)) は P(a—sb) と P(a—s) のカルバックライブラ情報量を表す．カルバックライブラ情報量は分布間の差異を表し，2 つの分布間の差が大きいほど大きい値をとり，2 つの分布が同一であれば 0 になる．

WIC がしきい値を超えた場合，ノードを追加する．これによりサンプル数とノードの追加による事後確率の変化を同時に考慮する．

2.2 階層構造を持つ素性への対応

自然言語における識別モデルの素性としては語の表層やその語の基本形，品詞など素性間に抽象度の違いからなる階層構造がある場合がある．これら階層構造を持つ素性を同時に考慮した接尾辞木を構築することで，異なる抽象度の素性からなる適切な抽象度をもった素性を抽出する．ここで，階層構造を持つ素性を同時に考慮した接尾辞木には構築する木に対する制約の強さの違いによりいくつかのバリエーションが考えられる．ここで 3 タイプの木について説明を行う．1 つ目の木は制約が最も強く，接尾辞木にノードを追加する際にある親ノードに対するすべての子ノードは同じ素性の階層構造に位置しているものになる．2 つ目の木は 1 つ目より制約は弱く，接尾辞木にノードを追加する際にある親ノードに対して異なる素性の階層に位置している素性が選ばれてもかまわないが，素性の階層構造において先祖と子孫の関係にあるもの同士が混在することは許されない．3 つ目の木は制約が最も弱く，接尾辞木にノードを追加する際に制限はなく，素性の階層構造の先祖と子孫の関係にある子ノードの追加も認める．これらの 3 つの接尾辞木を比較すると，1 つ目の木は最も制約が強く構築される接尾辞木の自由度が低くなる一方で，ある親ノードの下の子ノードはすべて同じ素性の階層に位置しているという制約から，与えられた素性の組から接尾辞木を辿り素性の抽出を行うことは容易になる．逆に 3 つ目の木は最も制約が弱く構築される接尾辞木の自由度は高くなる一方で，与えられた素性の組から接尾辞木を辿り素性を抽出するのはあらゆる階層の素性が混在しているために複雑になる．そして 2 つ目の木は 1 つ目の木と 2 つ目の木のちょうど中間の性質をもった木になっている．ここでは，2 つ目の木について説明する．学習データの素性の組の履歴にしたがって接尾辞木を辿り葉ノードに到達したら、そこからもう一つ履歴を遡った文脈に対応する頻度表の正解ラベルのカウントを 1 増やす．頻度表を更新する前までは同じだが，頻度表は素性の種類だけ作成する．そして情報量基準に基づいてノードの追加を判断する段階では，素性の階層の最上位にあるものは 1 つだけで通常の接尾辞木を構築するときと変わらないが，それ以下の階層にあるものは複数存在する．先祖と子孫の関係にある素性が混在しないようにするため，追加するノードはそれぞれ先祖と子孫の関係にある素性のうちどれか 1 つの階層から選択することになる．最上位の階層にある素性以外を新たに追加した文脈における情報量基準はそれぞれの階層に位置するものの平均値をとったものを使う．そして，その情報量基準が最も高い階層に属するすべての素性を新たに接尾辞木のノードとして追加する．

3 素性の抽出

素性は構築した接尾辞木を使って取り出す．推定したいラベルの文脈にしたがって接尾辞木を辿っていき，葉に到達するか辿る文脈がなくなった時点での根からそのノードまでのパスが使用する素性になる．CRF は最大エントロピーモデルであるため，素性の追加が容易である．そのため，最長パスだけでなく最長パスの一部を取り出して，最長パスと同時に素性として使うこともできる．また，Haruno らは予測する品詞の前の文脈のみを考慮したモデルを立てていたが，我々のモデルでは予測する意味役割の前後の文脈に対応した 2 つの接尾辞木を構築することで，前後の文脈に対する素性を使用することができる．

4 おわりに

本稿では Haruno らが提案した素性の階層構造と文脈の長さを同時に考慮したマルコフモデルで構築していた接尾辞木を使い，CRF などの識別モデルの素性を抽出するために用いることを提案した．本稿では手法の提案にとどまり，提案手法の有効性を確認するため

の実験はまだ行っていない．我々は今後本手法によって抽出した素性を使用した意味役割付与モデルの構築を行い，その有効性の確認を行う予定である．実験が間に合えば，本大会において結果を報告したい．

参考文献

[1] Masahiko Haruno and Yuji Matsumoto. Mistake-driven mixture of hierarchical tag context trees. In *in Proceedings of the 35th Annual Meeting of the Association of Computational Linguistics*, 1997.

[2] MJ. Weinberger, JJ. Rissanen, M. Feder. A universal finite memory source. In *IEEE Transaction on Information Theory*, 1995.