自然言語処理における系列ラベリング問題のための高速で厳密な漸次的復号化アルゴリズム

鍛治伸裕　藤原靖宏　吉永直樹　喜連川優

東京大学 生産技術研究所
{kaji, fujiwara, ynaga, kitsure}@tkl.iis.u-tokyo.ac.jp

1 はじめに

品詞タグ付与など，自然言語処理における基礎解析の多くは系列ラベリング問題 [1] として定式化することが可能である．従来，系列ラベリングにおける復号化には，ビタビアルゴリズムが適用されてきた．しかし，ビタビアルゴリズムはラベル数の2乗に比例した計算時間を要するため，ラベル数が大きなタスクに適用された場合，極めて非効率となる．この問題に対処するため，ビタビアルゴリズムよりも高速で，なおかつ厳密解を保証できる復号化アルゴリズムを提案する．

2 提案アルゴリズム

系列ラベリングとは，入力トークンの系列 $\mathbf{x} = \{x_1, x_2 \ldots x_N\}$ が与えられたとき，それに最適なラベル列 $\mathbf{y} = \{y_1, y_2 \ldots y_N\}$ を予測する問題である．系列ラベリング問題は，モデル学習と復号化 (最適なラベル系列の探索) に分けて考えることが出来るが，本論文では後者を議論の対象とする．

以下本節では，系列ラベリングのモデルとして隠れマルコフモデル (HMM) を考える．ただし提案手法自体は HMM に依存するものではなく，他のモデルに対しても同様に適用することが可能である．

まず，提案アルゴリズムの核となる縮退ラティスという概念を導入する．系列ラベリングにおける復号化は，ラティスの最適経路を探索する問題と捉えることが出来る (図 1(a))．ここで，ラティスの同一列上の複数ノードを1つにまとめ上げることによって，元のラティスをより簡潔な構造に変換することが出来る (図 1(b))．これを縮退ラティスと呼ぶ．縮退ラティス構築の際に，まとめ上げられずに残ったラベル (=ノード) を活性ラベル，まとめ上げられたラベルを非活性ラベルと呼ぶ．そして，非活性ラベルをまとめ上げて新しく生成されたラベルを縮退ラベルと呼ぶ．

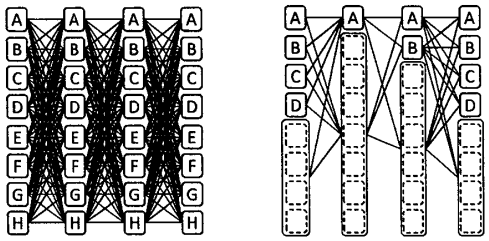

図 1: (a) ラティスの例．アルファベットはラベルを表す．(b) 縮退ラティス．

縮退ラベルの出力確率と遷移確率は下に示すように設定する．

$$
\begin{aligned}
p(x|z) &= \max_{y' \in I(z)} p(x|y'), \\
p(z|y) &= \max_{y' \in I(z)} p(y'|y), \\
p(y|z) &= \max_{y' \in I(z)} p(y|y'), \\
p(z|z') &= \max_{y \in I(z), y' \in I(z')} p(y|y'),
\end{aligned}
$$

ただし x は入力トークン，y は通常のラベル，z と z' は縮退ラベルを表す．また $I(z)$ は縮退ラベル z に対応する非活性ラベルの集合を表す．

このように確率を設定することにより，縮退ラベルを通る経路のスコアは，それに対応する非活性ラベルを通る経路のスコアに対する上限となる．したがって，縮退ラティスの最適経路が縮退ラベルを全く含まなければ，それは元のラティスの最適経路と一致する．

提案アルゴリズムは以下の 3 ステップから構成される．

初期化ステップ 入力 $\mathbf{x} = \{x_1, x_2 \ldots x_N\}$ に対して，

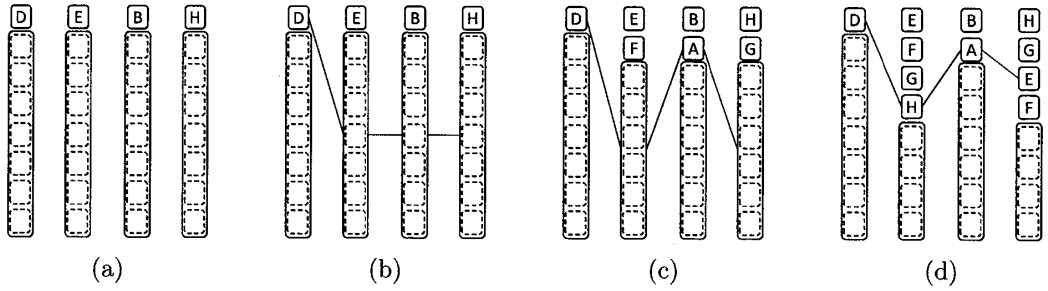

図 2: (a) 初期化ステップで構築された縮退ラティス．(b) ビタビアルゴリズムによって発見された最適経路．(c) 拡張ステップの実行後に，新たに発見された最適経路．(d) 非活性ラベルを含まない最適経路．

全ての列に活性ラベルが 1 つだけ存在するような縮退ラティスを作成する (図 2(a))．n 行目の活性ラベルには $p(y|x_n)$ を最大化するラベルを選択する．$p(y|x)$ の値は訓練事例を用いて推定する．

探索ステップ ビタビアルゴリズムを用いて最適経路を探索する (図 2(b))．もし最適経路が縮退ラベルを含まなければ，それは元のラティスの最適経路に等しいので処理を終了する．そうでなければ，次のステップへ進む．なお，探索の際には枝刈りを行うことが可能であるが誌面の都合上省略する．

拡張ステップ 探索ステップで得た最適経路の n 番目のラベルが縮退ラベルであれば，その列の活性ラベルの数を倍にする．新しい活性ラベルは $p(y|x_n)$ の大きな順に選択する．活性ラベルを増やした後，再び探索ステップを実行する (図 2(c))．この処理は，探索ステップで縮退ラベルを含まない最適経路が発見されるまで続けられる (図 2(d))．

3 実験結果

提案アルゴリズムの有効性を，品詞タグ付け，品詞タグ付けと基本句同定の結合タスク (結合タグ付与と呼ぶ)，スーパータギングの 3 タスクにおいて検証した．結合タグ付与においては，品詞タグと基本句タグ (BIO 形式) を結合したものをラベルとした．データはそれぞれ，Penn Treebank(PTB) コーパス，CoNLL2000 コーパス，PTB コーパスを HPSG 形式に変換したものを用いた [2]．

表 1 に，HMM とパーセプトロンにおける提案手法の復号化速度 (文数/秒)，及びビタビアルゴリズムとの結果を示す．いずれの問題設定においても復号化速度が大きく向上しているが．特にラベル数の多いタスクほど速度向上が顕著であった．

表 1: HMM での復号化速度 (文数/秒)．

	品詞タグ付与	結合タグ付与	スーパータギング
ビタビ	4600	100	1.3
提案法	**39,000**	**4200**	**570**

表 2: パーセプトロンでの復号化速度 (文数/秒)．

	品詞タグ付与	結合タグ付与	スーパータギング
ビタビ	4000	100	1.3
提案法	**23,000**	**5100**	**450**

4 まとめ

本論文では，系列ラベリングのための高速で厳密な復号化アルゴリズムを提案した．実験の結果，提案手法は従来主流であったビタビアルゴリズムよりも最大 400 倍高速であることを確認した．今後は，系列ラベリング以外の問題に対する拡張や，自然言語処理ドメイン以外での有効性の検証などに取り組みたい．

参考文献

[1] John Lafferty, Andrew McCallum, and Fernand Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of ICML*, pp. 282–289, 2001.

[2] Takuya Matsuzaki, Yusuke Miyao, and Jun'ichi Tsujii. Efficient HPSG parsing with supertagging and CFG-filtering. In *Proceedings of IJCAI*, pp. 1671–1676, 2007.