階層的時系列モデルによる固有表現抽出

金子　悟士 †　林　朗 ‡　末松　伸朗 ‡　岩田　一貴 ‡
広島市立大学　広島市立大学　広島市立大学　広島市立大学

1　はじめに

情報抽出の対象になりやすい人名、組織名を文中で特定する技術を固有表現抽出といい、固有表現抽出システムは教師付き機械学習で実現する方法が主流である. 文献 [1] では、階層隠れマルコフモデル (HHMM) を使った固有表現抽出を提案している.

また、近年 HMM に代わる確率モデルとして CRF(条件付確率場) が提案され、様々な研究において HMM より性能が良いことが示されている.HHCRF は HMM を一般化した確率モデル HHMM に対応する識別モデルである.[2] 本研究では HHMM と HHCRF を使った固有表現抽出を提案し、比較する.

2　時系列モデル

2.1　HHMM

HHMM は HMM の状態内に HMM をもつ階層モデルである. 下層では短期的な部分時系列を、上層ではより長期的な部分時系列の表現を可能にする [3]. 図 1 の左側はダイナミックベイジアンネットワーク (DBN) で表した HHMM である.

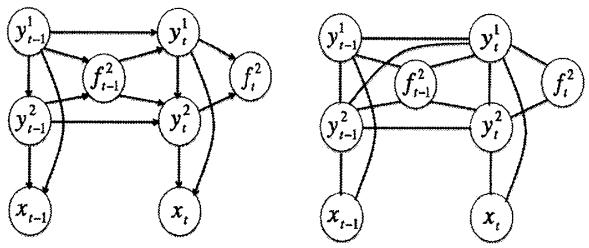

図 1　時刻 t-1 から t に関する HHMM の DBN と HHCRF の無向グラフ. 階層数 2(D=2)、$y_t^d(d1 \in, ..., D)$ は、時刻 t における深さ d の状態変数を表す. f_t^d は指標変数と呼ばれ、y_t^d が時刻 t において終了状態に遷移する場合に 1 を、それ以外の場合は 0 をとる.

2.1.1　HHMM のパラメータ推定

教師付き EM アルゴリズムにより HHMM のモデルパラメータを学習する.EM アルゴリズムとは、与えられた学習データに対するパラメータの期待値を計算する E ステップと、パラメータの更新を行う M ステップを繰り返す学習アルゴリズムである. 教師付き学習では、観測系列とそれに対応する状態系列によりモデルを学習する.

2.2　HHCRF

生成モデルである HMM に対し、CRF は識別モデルと呼ばれる. 生成モデル HHMM に対応する識別モデルは階層隠れ CRF(HHCRF) であり、無向グラフは図 1 右側のように表される. 生成モデルが入力と出力の同時確率をモデル化するのに対し、識別モデルは入力を条件とした出力の確率をモデル化する. 隠れ状態がないとき、ある観測系列 $x = \{x_1, ..., x_T\}$ に対する状態系列 $[y^{1:D} = \{y_1^{1:D}, ..., y_T^{1:D}\}, f^{2:D} = \{f_1^{2:D}, ..., f_T^{2:D}\}]$ の条件付き確率は以下のように表される.

$$p(y^{1:D}, f^{2:D}|x;\theta) = \frac{\exp \sum_{k=1}^{K} \sum_{t=1}^{T} \lambda_k \Phi_k(y_{t-1}^{1:D}, y_t^{1:D}, f_{t-1}^{2:D}, f_t^{2:D}, x_t)}{Z(x;\theta)} \quad (1)$$

ここで,$\Phi_k(y_{t-1}^{1:D}, y_t^{1:D}, f_{t-1}^{2:D}, f_t^{2:D}, x_t)$ は素性関数と呼ばれ、λ_k は素性関数の重み、$Z(x;\theta)$ は $p(y^{1:D}, f^{2:D}|x;\theta)$ の正規化関数である.

2.2.1　HHCRF のパラメータ推定

HHCRF では観測系列集合 $\{x^{(i)}|1 \leqq i \leqq N\}$ に対する、状態系列集合 $\{y^{1:D(i)}|1 \leqq i \leqq N\}$ の条件付き対数尤度 $l(\theta)$ を最大化することでパラメータ $\theta = \{\lambda_k|1 \leqq k \leqq K\}$ を推定する.

$$l(\theta) = \sum_{i=1}^{N} \log p(y^{1:D^{(i)}}, f^{2:D^{(i)}}|x^{(i)};\theta) \quad (2)$$

条件付き対数尤度の最大化には $l(\theta)$ の導関数 (式 (3)) を用いる.

$$\frac{\partial l(\theta)}{\partial \lambda_k} = \sum_{i=1}^{N} \sum_{t=1}^{T} \Phi(y_{t-1}^{1:D^{(i)}}, y_t^{1:D^{(i)}}, f_{t-1}^{2:D^{(i)}}, f_t^{2:D^{(i)}}, x_t^{(i)}) - \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{y_{t-1}} \sum_{y_t} \sum_{f_{t-1}} \sum_{f_t} \Phi(y_{t-1}^{1:D}, y_t^{1:D}, f_{t-1}^{2:D}, f_t^{2:D}, x_t^{(i)}) * p(y_{t-1}^{1:D}, y_t^{1:D}, f_{t-1}^{2:D}, f_t^{2:D}|x_t^{(i)}) \quad (3)$$

式 (3) の右辺第一項は経験分布に基づく Φ_k の期待値であり、これは Backward-Forward-Backward アルゴリズム [4] により求まる. 第二項はモデル分布に基づく Φ_k の期待値であり、これは Forward-Backward アルゴリズムにより求まる.

Named Entity Extraction by Hierarchical Time Series Models

† Satoshi Kaneko: Faculty of Information Sciences,Hiroshima City University
‡ Akira Hayashi, Nobuo Suematsu, Kazunori Iwata: Graduate School of Information Sciences, Hiroshima City University
Email: kaneko@prl.info.hiroshima-cu.ac.jp

3 固有表現抽出

今回行う固有表現抽出では文章中から関連のある英単語のペアを情報抽出する. 固有表現抽出をするモデルは、HHMM、HHCRF どちらも 2 層階層構造をもつ. それぞれのモデルに対応する観測系列は文章中の英単語の列である. 状態系列は、上層ではフレーズ (句) の系列からなり、下層では対応する観測値が固有表現抽出の対象となる英単語であるかないかが系列として表されている.

与えられた文章が固有表現を含むか否かはポジティブ/ネガティブモデルの尤度を比較することにより推定する. 固有表現を含む場合、固有表現の単語のペアはポジティブモデルの最適状態系列を Viterbi アルゴリズムにより求めることで得られる.

4 実験

4.1 時系列モデルによる固有表現抽出

文献 [1] での実験で用いられたデータセットである、生物医学記事から『ある (不特定の) タンパク質の細胞内の場所』を表している箇所を抽出する. 今回使用したデータセットは 600 の文から構成されており、そのうち『あるタンパク質』と『そのタンパク質の細胞内の場所』を表している固有表現のペアを含む文は 300 存在する. データセットについて 5 分割交差検定を行い、全てのデータについてテストする.

HHMM と HHCRF を比較するとき、各モデルの評価には再現率適合率 (Recall-Precision:PR) 曲線を用いる. 正しく抽出できた固有表現を含む文の数を H_{it}、テストデータ中の固有表現を含む文の数を T_{est}、Viterbi アルゴリズムにより推定された固有表現を含む文の数を V_{iterbi}、とすると、再現率と適合率はそれぞれ $Recall = H_{it}/T_{est}$ および $Precision = H_{it}/V_{iterbi}$ で表される.

また、それぞれの最適状態系列について信頼度を計算する. 最適状態系列 $[\hat{y}^{1:D}, \hat{f}^{2:D}]$ の尤度

$$\delta(i) = p(\hat{y}^{1:D}, \hat{f}^{2:D}|x^{(i)}; \theta) \tag{4}$$

と、観測系列に対するモデルの尤度

$$\alpha(i) = \sum_{y^{1:D}} \sum_{f^{2:D}} p(y^{1:D}, f^{2:D}|x^{(i)}; \theta) \tag{5}$$

を計算することにより信頼度 $C(i) = \delta(i)/\alpha(i)$ を求める. 信頼度について閾値を変化させてゆくことにより PR 曲線を作成する.

図 2 では、ほぼ全体的に HHCRF の方が HHMM より高い Precision を示していることから、このデータセットについては HHCRF の方が精度が高い. 一方 Recall では、HHMM の方が高い値を示しているところから、HHMM は HHCRF と比べると正しく推定する数が多い.

また、HHMM、HHCRF どちらのモデルも Recall が 0 付近で Precision が急激に下がっている. これはどちらのモデルでも、誤推定した系列が最も高い信頼度をもったためである. これについては学習データセットを増やす等、モデルの精度を上げることにより解決する必要がある.

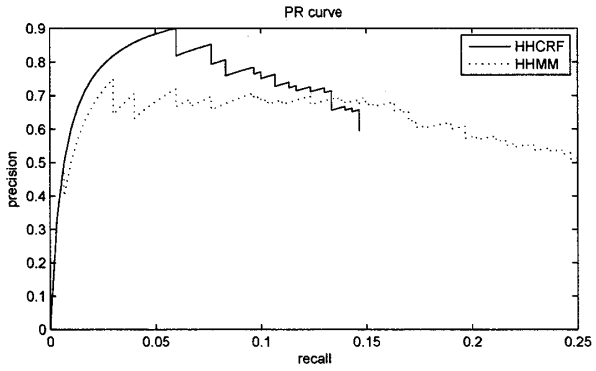

図 2 Recall-Precision 曲線.

5 まとめ

HHMM と HHCRF を用いて生物医学記事からの固有表現抽出を行い、2 つの手法の性能を比較した. 今回取り扱ったデータセットに対しては適合率では HHCRF、再現率では HHMM が高い性能を示した. 今後は 2 つの手法を使い、違うデータセットについても実験して性能を比較する必要がある.

参考文献

[1] M.Skounakis: "Hierarchical Hidden Marcov Models for Information Extraction", Proc.18th Int. Joint Conf. Artificial Intelligence (IJCAI 2003),2003

[2] T.Sugiura, N.Gotoh and A.Hayashi: "A discriminative model corresponding to hierarchical hmms", Proc.14th Int. Conf. Intelligent Data Engineering and Automated Learning (IDEAL 2007), 2007

[3] K.P.Murphy: "Linear Time Inference in Hierarchical HMMs", Advances in Neural Information Processing Systems ,2001

[4] T.Scheffer and S.Wrobel: "Active hidden Marcov models for information extraction", Lecture Notes in Computer Science, vol.2189,2001