多様な情報源を用いた隠れマルコフモデルによる医学/生物学論文の専門用語認識器

新福哲† 風間淳一‡ 宮尾祐介‡ 辻井潤一‡*
† 東京大学 大学院理学系研究科 情報科学専攻
‡ 東京大学 大学院情報理工学系研究科 コンピュータ科学専攻
* CREST, 科学技術振興事業団

1 はじめに

近年、医学/生物学論文からの情報抽出が生物学の新たな研究分野として注目を集めている。GENIA プロジェクト [5] は、自然言語処理の技術を応用し、医学/生物学の論文から多くの情報の獲得を行う事を目的としている。

本研究では GENIA プロジェクトの一環として、医学/生物学論文に**専門用語タグ** (専門用語をその種類により区別する為のタグ) の自動付与を行うシステムを提案する。医学/生物学論文の専門用語認識では、局所的な文脈だけからでは専門用語を判別する事が出来ない事が多い。そこで本システムでは、2 つの確率モデルからなる **Term hidden Markov model** を用い、より広範囲の文脈を考慮する事で、高精度を達成する。また、本システムの有効性を評価する為、GENIA プロジェクトで開発された GENIA コーパス [6] での実験結果を報告する。

本稿では、まず 2 節で専門用語認識において解かねばならない問題とその対処法について述べ、3 節で問題を解決する為の確率モデルとその確率モデルを用いたシステムを提案する。4 節では、GENIA コーパスを使用した専門用語タグ付与の実験結果について報告し、システムの有効性を評価する。

2 タスクと背景

本研究のタスクは、医学/生物学論文中に出現する専門用語を検出し、専門用語タグを付与する事である。本研究のタスクや、(第 6 回) Message Understanding Conference [1] における Named entity (NE) task では、一般に次の二つの問題を解決しなければならない。

ターム分割 文を専門用語とそれ以外の単語列に区分する

タグ選択 正しい種類の専門用語タグを付与する

このことから、専門用語の自動認識を行うシステムの構成としては、次の二つが考えられる。

対処 (1) 単語単位にタグ選択を行い、その後結果のタグ列を用いてターム分割を行う

対処 (2) ターム分割を行い、分割されたターム単位にタグ選択を行う

対処 (1) には、これまで研究されてきた品詞タグ付けや、チャンキング等の手法がそのまま適用できるという利点がある。同分野のテキストにおける先行研究である Collier の研究 [4] では、対処 (1) の手法でマルコフモデルを使用し、精度 (F-値)0.73 を達成していた。しかし対処 (1) の手法には広範囲の文脈を考慮するのが難しいという問題点がある。

一方、NE task において Nymble システム [3] では、対処 (2) の手法を用いて高精度 (F-値)0.93 を達成した。Nymble システムは、ターム分割については、任意の分割が全て等確率で起こると仮定し、状態をタームに対応させた特殊な HMM を用いて最尤ターム列を推定していた。

本研究は、Nymble システムをベースとし、医学/生物学論文の専門用語認識にて高精度を達成するシステムの構築を目的とする。医学/生物学の専門用語には、次のような事例が多い事から NE task よりも難易度が高いと考えられる。

- 普遍的な単語が専門用語を構成する事が多い。例えば、*disease* や *human* が専門用語を構成する事がある。
- 1 つの専門用語を構成する単語の数が多い。例えば、GENIA コーパスでは 17 単語からなる専門用語も存在する。
- 文脈等によって同じ単語列に違う種類のタグが付く事が多い。例えば、*NF-kappa B* は通常 PROTEIN タグが付与されるが、DNA タグが付与される場合もある。また、後ろに来る単語によっては OTHERS タグが付与される専門用語の一部となる場合もある。

この事から、本タスクではより広範囲の文脈、及び様々な単語の特性を考慮に入れたモデル化を行う必要がある。次節では、それら解決する確率モデルを提案する。

3 Term hidden Markov model

本研究では 2 つの確率モデルによって構成される Term hidden Markov model (Term-HMM) を提案する。Term-HMM に含まれる 2 つの確率モデルは、それぞれターム分割とタグ選択の各問題を解消する為のモデルとなっている。ターム分割を行うモデルでは、専門用語とそれ以外の単語列との区切り目の有無を決定する為に、条件部に前後の単語の情報を含ませた確率を使用する。タグ

選択を行うモデルでは、広範囲の文脈を考慮するため、Nymbleのモデルをトライグラムモデルに拡張し、さらに、条件部に多くの情報を加えた確率を使用する。本研究では、これらの結果として複雑になったモデルのパラメータ推定法として、データスパースネスに強いとされている最大エントロピー (ME) 法 [2] を用いる。

Term-HMM は、文 w_1^N が与えられた時に、考え得るタグ列 t_1^M の中で一番高い確率値 $P(t_1^M|w_1^N)$ を持つタグ列 $\hat{t}_1^M$ を検出する為の HMM である。Term-HMM では、$P(t_1^M|w_1^N)$ は次の近似式により与えられる。ここで、σ_n はターム分割を示す二値の確率変数で、文の区切り目がある場合に 1 を示す。

$$P(t_1^M|w_1^N) \cong P(\sigma_1^N|w_1^N)P(t_1^M, w_1^N|\sigma_1^N)$$

上式の $P(\sigma_1^N|w_1^N)$ を**ターム分割モデル**と呼び、$P(t_1^M, w_1^N|\sigma_1^N)$ を**タグ選択 HMM** と呼ぶ。

3.1 ターム分割モデル

ターム分割モデルは、文 w_1^N が与えられた時に $P(\sigma_1^N|w_1^N)$ を与える確率モデルである。本研究では、注目している単語 w_n と直後の単語 w_{n+1} から、専門用語とそれ以外の単語列の区切り目の有無に対して確率値 $Ps(\sigma_n|w_n, w_{n+1})$ を与える最大エントロピーモデルを使用した。そして、$P(\sigma_1^N|w_1^N)$ は以下のように Ps の積で表わされる。

$$P(\sigma_1^N|w_1^N) = \prod_{n=1}^{N} Ps(\sigma_n|w_n, w_{n+1})$$

3.2 タグ選択 HMM

タグ選択 HMM は、Nymble の確率モデルと同じタイプの HMM の生成確率によって、ターム単位のタグ列 t_1^M に対して確率値 $P(t_1^M, w_1^N|\sigma_1^N)$ を与える確率モデルである。本研究では、確率値 $P(t_1^M, w_1^N|\sigma_1^N)$ は次の式で与えられる。ここで、I_m は状態 t_m からの出力単語数を表わす。

$$P(t_1^M, w_1^N|\sigma_1^N) = \prod_{m=1}^{M}\{Pt(t_m|t_{m-1}, t_{m-2}, ht_m)\prod_{i=1}^{I_m} Pe(w_i|t_m, he_i)\}$$

Nymble システムはバイグラムの HMM を使用していたが、より広範囲の文脈を考慮する為、本研究ではタグ選択 HMM はトライグラムの HMM を使用した。これは専門用語同士の関係をより考慮した状態遷移を行う為である。Nymble で使われていた HMM では、状態がタームと対応しており、また、状態遷移を行う場合に必ず違う種類のタームと対応している状態へ遷移する。故に、一つもしくは二つ前の状態は、必ず専門用語と対応している状態となる。つまり、二つ前までの状態をみる事で、少なくとも 1 つ前の専門用語を考慮する事が出来る。

さらに本研究では、より多くの局所的な文脈も考慮する為に、状態遷移確率 Pt、および記号出力確率 Pe の条件部に直前の単語までのヒストリ (ht_m, he_i) を加えた。ヒストリの詳細については、3.4 節で述べる。

3.3 最尤タグ列の推定

以上をまとめると、文 w_1^N が与えられた時に、Term-HMM によって与えられる各タグ列 t_1^M の確率値 $P(t_1^M|w_1^N)$ は、次の式より計算される。

$$P(t_1^M|w_1^N) = \prod_{n=1}^{N} Ps(\sigma_n|w_n, w_{n+1})\times \prod_{m=1}^{M}\{Pt(t_m|t_{m-1}, t_{m-2}, ht_m)\prod_{i=1}^{I_m} Pe(w_i|t_m, he_i)\}$$

$\hat{m} = 1 + \sum_{i=1}^{n}\sigma_i$ とすると、上式は次のように変形される。

$$P(t_1^M|w_1^N) = \prod_{n=1}^{N}\begin{cases} Ps(\sigma_n|w_n, w_{n+1})\times & \\ Pe(w_n|t_{\hat{m}}, he_n) & \sigma_n = 0 \text{ の時} \\ Ps(\sigma_n|w_n, w_{n+1})\times & \\ Pt(t_{\hat{m}}|t_{\hat{m}-1}, t_{\hat{m}-2}, ht_{\hat{m}})\times & \\ Pe(w_n|t_{\hat{m}}, he_n) & \sigma_n = 1 \text{ の時} \end{cases}$$

この式は、$Ps(\sigma_n|w_n, w_{n+1})$ を直前の状態と同じタグに対応した状態への状態遷移確率、$Ps(\sigma_n|w_n, w_{n+1})Pt(t_{\hat{m}}|t_{\hat{m}-1}, t_{\hat{m}-2}, ht_{\hat{m}})$ を直前の状態と違うタグに対応した状態への状態遷移確率、$Ps(w_n|t_{\hat{m}}, he_n)$ を記号出力確率として見ることで、以下に示す通常の HMM と同等の構造をしている。

$$\prod_{n=1}^{N}\{Pt(t_n|t_{n-1}, t_{n-2}, ht_n)Pe(w_n|t_n, he_n)\}$$

従って、HMM に対する最尤状態列探索アルゴリズムである Viterbi アルゴリズム [7] を応用する事が出来る。

3.4 パラメータ推定法

各確率の条件部に大量の条件を加えている為、本研究のモデルは、単純なパラメータを使用した確率モデルよりもデータスパースネスに弱くなってしまっている。そこで確率モデルのパラメータの推定法として、単純な相対頻度法よりもデータスパースネスに強いとされている、ME 法 [2] を使用した。

ME 法では、$P(x|y)$ を学習データ中の事象 (x, y) の相対頻度 $\tilde{p}(x, y)$、及び事象 (y) の相対頻度 $\tilde{p}(y)$ から推定する。具体的には $P(x|y)$ は以下の式でモデル化される。

$$P(x|y) = \frac{\prod_i \exp(\lambda_i f_i(x, y))}{\sum_x \prod_i \exp(\lambda_i f_i(x, y))}$$

ここで、f_i は素性関数と呼ばれ、事象 (x, y) がある特性 i を持つ時のみ、1 となる二値関数である。それに対する重み λ_i は、学習データの尤度を最大化するように推定される。

本システムでは単語を以下の三つの素性の組とする。

wd 頻度順に 2499 単語とその他を示す単語のクラスに分類

pos 品詞情報により 32 種類のクラスに分類

表 1: 状態遷移確率推定に使用した素性の組合わせ

$(t_m, t_{m-1}, t_{m-2}, wd_{n-1}, pos_{n-1}, wd_{n-2})$	$(t_m, t_{m-1}, t_{m-2}, fea_{n-1}, pos_{n-1})$	$(t_m, t_{m-1}, pos_{n-1}, wd_{n-2})$
$(t_m, t_{m-1}, t_{m-2}, cha_{n-1}, pos_{n-1}, wd_{n-2})$	$(t_m, t_{m-1}, t_{m-2}, wd_{n-1})$	$(t_m, t_{m-1}, wd_{n-1}, cha_{n-2})$
$(t_m, t_{m-1}, t_{m-2}, wd_{n-1}, pos_{n-1}, cha_{n-2})$	$(t_m, t_{m-1}, t_{m-2}, cha_{n-1})$	$(t_m, t_{m-1}, cha_{n-1}, cha_{n-2})$
$(t_m, t_{m-1}, t_{m-2}, cha_{n-1}, pos_{n-1}, cha_{n-2})$	$(t_m, t_{m-1}, t_{m-2}, pos_{n-1})$	$(t_m, t_{m-1}, pos_{n-1}, cha_{n-2})$
$(t_m, t_{m-1}, t_{m-2}, wd_{n-1}, wd_{n-2})$	(t_m, t_{m-1}, t_{m-2})	$(t_m, t_{m-1}, wd_{n-1}, pos_{n-1})$
$(t_m, t_{m-1}, t_{m-2}, cha_{n-1}, wd_{n-2})$	$(t_m, t_{m-1}, wd_{n-1}, pos_{n-1}, wd_{n-2})$	$(t_m, t_{m-1}, cha_{n-1}, pos_{n-1})$
$(t_m, t_{m-1}, t_{m-2}, pos_{n-1}, wd_{n-2})$	$(t_m, t_{m-1}, cha_{n-2}, pos_{n-1}, wd_{n-2})$	(t_m, t_{m-1}, wd_{n-1})
$(t_m, t_{m-1}, t_{m-2}, wd_{n-1}, cha_{n-2})$	$(t_m, t_{m-1}, wd_{n-1}, pos_{n-1}, cha_{n-2})$	$(t_m, t_{m-1}, cha_{n-1})$
$(t_m, t_{m-1}, t_{m-2}, cha_{n-1}, cha_{n-2})$	$(t_m, t_{m-1}, cha_{n-1}, pos_{n-1}, cha_{n-2})$	$(t_m, t_{m-1}, pos_{n-1})$
$(t_m, t_{m-1}, t_{m-2}, pos_{n-1}, cha_{n-2})$	$(t_m, t_{m-1}, wd_{n-1}, wd_{n-2})$	(t_m, t_{m-1})
$(t_m, t_{m-1}, t_{m-2}, wd_{n-1}, pos_{n-1})$	$(t_m, t_{m-1}, cha_{n-1}, wd_{n-2})$	(t_m)

cha Collier の研究で使用されていた文字面に関するクラスに、頻度の高い語頭 3 文字 60 種にて単語を分類したクラスを加えたクラスに分類

これにより、ヒストリに入れる単語を次のように特性の組に変換して使用できる。

$$\begin{aligned} & f_i(z|ht_m) \\ &= f_i(z|w_{n-1}, w_{n-2}) \\ &= f_i(z|wd_{n-1}, cha_{n-1}, pos_{n-1}, \\ &\quad wd_{n-2}, cha_{n-2}, pos_{n-2}) \end{aligned}$$

しかし、これらの特性を利用した記号出力確率を推定する場合、以下の確率を使用しなくてはならなくなる。

$$Pe(wd_n, pos_n, cha_n|t_m, he_m)$$

ME 法では、条件付き確率 $P(x|y)$ の x の種類が多い場合、パラメータ推定のコストが大きくなるという欠点がある。よって本研究では以下のように記号出力確率を分割し、それぞれに対して ME 法による推定を行った。

$$\begin{aligned} & Pe(wd_n, pos_n, cha_n|t_m, he_m) \\ &= Pe_1(wd_n|pos_n, cha_n, t_m, he_m) \\ &\times Pe_2(pos_n|cha_n, t_m, he_m) \\ &\times Pe_3(cha_n|t_m, he_m) \end{aligned}$$

表 1 は、ヒストリに直前の 2 単語の情報を入れた状態遷移確率の推定の為に実際に使用した、素性関数 $f_i(t_m, t_{m-1}, t_{m-2}, ht_m)$ を列挙したものである。他の 4 つの確率を推定する為に用いた素性は、紙面の都合により省略する。

4 システム評価

4.1 実験環境

GENIA コーパス [6] に人手で品詞タグを付与した 480 アブストラクトを使用し、本システムの評価実験を行った。評価用コーパスには、品詞タグ付き GENIA コーパス中からランダムに選択した 80 アブストラクトを、学習用コーパスには残りの 400 アブストラクトを使用した。タグセットには、GENIA コーパスに付与されていたタグを GENIA オントロジー [5] に則って SOURCES、PROTEIN、DNA、RNA、OTHERS (前の四つ以外の専門用語) の 5 種類のタグに変換し、使用した。

表 2: 先行研究との比較 (F-値)

	MM	Term-HMM
完全正解	0.3867	0.4072
ターム分割	0.4949	0.5710
単語タグ	0.4969	0.5604
単語判定	0.6917	0.8297

システムの性能を様々な側面から考察する為に、評価基準として次の 4 つの基準を用いた。

完全正解 ターム分割、タグ選択の両方を正解しているもの

ターム分割 ターム分割のみ成功したもの

単語タグ 単語単位にタグ選択のみ正解しているもの

単語判定 単語単位に専門用語であるという判定のみ成功しているもの

実験に使用したシステムは次の通りである

MM Collier の研究 [4] におけるシステム

MM+ME Collier のシステムのパラメータ推定法を ME 法に置き換えたシステム

タグ選択 HMM タグ選択 HMM のみを用いたシステム

Term-HMM 今回提案したシステム

評価の値には、適合率と再現率から計算される F-値を用いる。適合率はシステムの出力のうち正解した割合、再現率は評価用コーパス中の正解のうちシステム出力が正解した割合、F-値は適合率を P と再現率を R として次式で計算される値である。

$$\frac{2 \times P \times R}{P + R}$$

4.2 結果と考察

まず、Collier の研究と我々の提案したシステムとの比較実験を行った (表 2)。この比較では、全ての側面にお

表 3: 既存の研究に ME 法を適用した場合 (F-値)

	MM	MM+ME
完全正解	0.3867	0.3934
ターム分割	0.4949	0.5885
単語タグ	0.4969	0.4438
単語判定	0.6917	0.7514

表 4: タグ選択 HMM の有効性 (F-値)

	MM+ME	タグ選択 HMM
完全正解	0.3934	0.4068
ターム分割	0.5885	0.5717
単語タグ	0.4438	0.5604
単語判定	0.7514	0.8297

いて、我々のシステムの方が優位である[1] という結果が示された。

次に、この結果がパラメータ推定に ME 法を用いたことにのみに依存したものか調査する為に、Collier のシステムと、Collier のシステムのパラメータ推定を ME 法でおこなったシステムとの比較実験を行った (表 3)。完全正解、ターム分割、単語判定では精度は向上したが、単語タグでは精度が低下する事が分かる。よって、ME 法を適用した場合の方が良い結果を出すシステムとなるが、本研究での精度向上はそれのみに依るものではないことが示された。

また、タグ選択 HMM とターム分割モデルによる効果を調査する為に、Collier の研究のモデルとタグ選択 HMM のみを使用したシステムの比較 (表 4) と、タグ選択 HMM のみを使用したシステムと Term-HMM を使用したシステムの比較 (表 5) を行った。この実験では、パラメータ推定法は ME 法に統一している。Collier の研究とタグ選択 HMM のみとの比較では、ターム分割ではわずかに精度が下がっているが、全体的には精度が向上した。特に単語単位での精度に関しては大きく向上し、タグ選択 HMM の有効性が確かめられた。しかし、タグ選択 HMM のみと Term-HMM との比較では、わずかに精度が向上したものの、期待した程の精度の向上は見られなかった。この事は、本タスクにおけるターム分割を行うには、今回用いたターム分割モデルのような単純なモデルでは、大きな精度向上は得られないことを示している。

最後に、学習量による精度変化を見る為に、学習量を 80、160、240、320、400 アブストラクトと変化させた場合の精度を図 1 に示す。このグラフによると、学習量を増やす事で、更なる精度の向上が期待される。

5 おわりに

本研究は、医学/生物学論文に現れる専門用語に専門用語タグを付与するシステムの為の確率モデルの提案、および実装を行った。GENIA コーパスを使用した比較

[1] 2 節で述べた 0.73 という精度は、本研究で用いた GENIA コーパスの公開前のアルファバージョンでの精度である。

表 5: ターム分割モデルの有効性 (F-値)

	タグ選択 HMM	Term-HMM
完全正解	0.4068	0.4072
ターム分割	0.5717	0.5710
単語タグ	0.5594	0.5604
単語判定	0.8295	0.8297

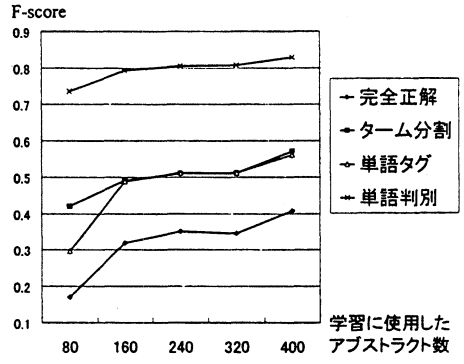

図 1: 学習量と精度の関係

実験を行い、精度比較を行った結果、同分野のテキストを対象とした Collier による先行研究のシステムより全ての側面で精度が向上している事が確かめられた。

今後の課題としては、確率モデルの変更や使用素性の変更といったターム分割モデルの改良と、現在作成中である GENIA コーパス 1000 アブストラクトを使用した大量のコーパスを学習に使用した場合の実験と分析が考えられる。

参考文献

[1] *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, San Francisco, November 1995. M. Kaufmann.

[2] A. L. Berger, S. A. Della Pietra, and V. J. Della Pietra. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71, 1996.

[3] D.M. Bikel, S. Miller, R. Schwartz, and R. Weischedel. Nymble: A high-performance learning name-finder. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, pages 194–200, March 1997.

[4] N. Collier, C. Nobata, and J. Tsujii. Extracting the Names of Genes and Gene Products with a Hidden Markov Model. In *Proc. COLING 2000*, pages 201–207, 2000.

[5] T. Ohta, Y. Tateisi, N. Collier, C. Nobata, and J. Tsujii. Building an annotated corpus from biology research papers. In *Proc. COLING 2000 Workshop on Semantic Annotation and Intelligent Content*, pages 28–34, 2000.

[6] T. Ohta, Y. Tateisi, N. Collier, C. Nobata, and J. Tsujii. GENIA project home page. http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/, 2000.

[7] A. J. Viterbi. Error bounds for convolutional codes and an asymptotically optimal decoding algorithm. *IEEE Transactions on Information Theory*, IT-13:260–269, 1967.