自然言語処理　150－27
(2002. 7. 16)

生物学文献からの専門用語抽出における機械学習モデルの検討

竹内孔一　コリアー・ナイジェル
国立情報学研究所
〒101-8430 東京都千代田区一ツ橋2－1－2
E-mail: {koichi, collier}@nii.ac.jp

本論文ではSVM(Support Vector Machine)による機械学習モデルを利用した専門用語抽出モデルの構築について議論する．対象とする分野は分子生物学でこの分野の文献から専門用語を認識し，あらかじめ用意する意味カテゴリーに分類する．固有抽出表現ではMUC-6に見られるような新聞記事を中心としたタスクが行なわれてきたが，認識すべき用語の困難さが大きく異なることが特徴である．本研究では比較のためにHMM (Hidden Markov Model)を用いた学習モデルを用意し，SVMによる学習モデルとの精度を分子生物学データ，MUC-6データの各々に対して適用しその精度を評価した．実験の結果，両データに対してSVMの抽出精度がHMMの精度を上回る結果を得た．この結果から最適な抽出モデルについて検討をおこなう．
キーワード　専門用語抽出，分子生物学，SVM，

Analysis of Machine Learning Model for Technical Term Extraction in Biological Science Papers

Koichi Takeuchi Nigel Collier
National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan
E-mail: {koichi,collier}@nii.ac.jp

This paper explores the use of Support Vector Machines (SVMs) for an extended named entity task. We investigate the identification and classification of technical terms in the molecular biology domain and contrast this to results obtained for traditional NE recognition on the MUC-6 data set. Furthermore we compare the performance of the SVM model to a standard HMM bigram model. Results show that the SVM utilizing a rich feature set of a ±3 context window and orthographic features had a significant performance advantage on both the MUC-6 and molecular biology data sets. From the results, the paper show what kind of parameter sets are important for constructing the best extraction model.
key words　Term extraction, Molecular biology, Support vector machine

1 はじめに

自然言語処理における固有表現抽出は，入力文書に対して浅いレベルの意味づけを行なう中心的技術として確立されている．固有表現抽出は 1995 年に DARPA によって開催された Message Understanding Conference (MUC)-6[7] に始まり，その後以下に示すように様々方向に発展して行った．

- アドホックな手法: 人手であらかじめ分析しておいたルールにより解析を行なう手法 [8]
- タグづけされたコーパスを利用した教師つき学習による手法 [1]
- 教師無し学習による手法 [5]

従来の固有表現抽出の内容は新聞記事を対象として，個人の名前や企業，組織名，日付，場所などの同定を目的として来た．こうした固有名詞の表現の多様性は大きく，さらに未知語が常にうまれることから，従来の自然言語処理でも十分に困難な課題であった．

しかし，本稿が取り上げている分子生物学における専門用語の抽出は個々の固有表現だけでなく概念的な意味クラスを取り出す必要性が生じるため，従来の新聞記事中心の固有表現抽出タスクを越えたところに位置づけできると考えられる．つまり，従来手法より高度な言語処理モデル，例えば，文字種解析，形態素解析，簡易な統語解析や文脈解析も必要となる．例をあげてみる．

- unliganded (apo)- and liganded (holo)-LBD

これは and という単語で LBD という 主辞に係ってるがこれ全体で専門用語である．こうした表現の解析には簡易なチャンク解析を越えた正確な統語解析が必要になる．

また，テキスト上の処理だけでなく知識処理も用語抽出には必要となる．分子生物学の専門用語は意味概念としてのクラスを持っており (タンパク質，遺伝子など) オントロジーのような深い知識表現の活用も必要となる．我々が抽出する対象としている用語は人名，タンパク質名，化学式なども含むことになる．

用語抽出というタスクを分類化と捉えると，隠れマルコフモデル (Hidden Markov Model (HMM))[11] や 誤り学習モデル (Transformation based error-driven learning (TBL)) [2] といった良く知られている学習モデルを適用することができる．最近，サポートベクターマシン (Support Vector Machine (SVM)) [14][6] と呼ばれる学習モデルがテキスト分類において用いられ大きな成果をあげている．これは SVM が大規模でスパースな入力データに対する処理が優れている点が特徴的である．さらに，SVM を利用してチャンク解析を含む品詞のタグづけにおいても良い成果が報告されている [10].

本稿では SVM と HMM にもとづく専門用語抽出モデルを作成し，分子生物学文献に対する専門用語抽出ならびに MUC-6 の固有表現抽出タスクに適用した．HMM と SVM のもっとも大きな違いの一つは SVM の出力は各カテゴリー間の距離であるのに対して，HMM は Viterbi アルゴリズム [15] を利用した全体のパスとして確率最大のものを出力する．この特性がどのように解析に反映されるのかを明かにして専門用語抽出における良いモデルの条件を考察する．

2 方法

2.1 SVM

本研究で使用する SVM は奈良先端大で開発されている Tiny SVM パッケージ [1] を利用する．これは Vapnik の SVM に最適化アルゴリズム [9] を組み合わせたものである．

SVM は学習データからの入力をもとにカテゴリー化する関数を作成し分類を行なう．特徴的なのは大きな特徴空間に対して頑健性が高い点である．この性質は専門用語抽出タスクにおいて理想的であると言える．マージンを最大化することができるのも大きな特徴である．

学習に必要な時間が大きくなる傾向があるが，解析には時間がかからない．なぜなら正例か負例かを保持する support vector を調べるだけで解析を行なうことができるからである．ただし，基本的に 2 値分類器なので多値分類が扱えるように仕掛けを用意する必要がある．

SVM は学習データとして $\chi \times \{\pm 1\}$ から分類関数 $f : \chi \to \{\pm 1\}$ を評価する．分類関数はテストセットがそのクラスに属していれば 1 そうでなければ-1 を返す．入力パターン χ に対して特徴空間 Γ に非線形関数を使って写像する ($\Phi : \chi \to \Gamma$) ことでより複雑な問題に対処可能になる点が重要な点で

[1] Tiny SVM は http:// cl.aist-nara.ac.jp/ taku_ku/ software/ TinySVM/ から入手可能

ある．SVM に関する詳細は先に挙げた文献にゆずるとして，我々が学習で用いた特徴量について説明する．

SVM に与える学習用データは語彙的な属性によるベクトルの集合である．特徴量として，表層の単語，orthographic 属性 [4][2] を利用した．また，予備実験において品詞情報を利用した学習を行なったが解析精度を下げてしまったので利用しなかった．これは，特に分子生物学に対して POS tagger がうまく機能しなかったことが原因であった[3]．

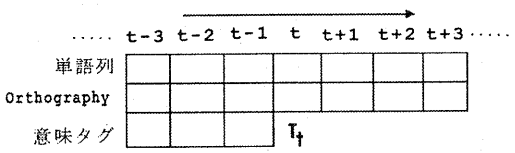

図 1: 窓幅 (特徴ベクトルを作る範囲)

これらの語彙属性をもとに学習用の特徴ベクトルを作成する．ここで図 1 に示すように窓幅を導入する．これはある意味タグ T_tのベクトルとして前後の単語，orthographic 属性さらに直前までの意味クラスを要素とすることを意味している[4]．窓幅は予備実験から±3 とした．ただし，窓幅内の語順は本実験では反映されておらず，集合として特徴ベクトルとなる．

また，専門用語は複数の単語からなるので I O B 表記を利用して意味タグの中に chunk の情報を組み入れた．意味タグ C_tの単語のはじまりを B_C_t それ以外を I_C_t，意味タグのない単語を O とする．

SVM は基本的に 2 分類器なので多値分類できるように複数の分類器を利用する．本手法では Tiny SVM を利用して one-against-one 法を用いた．

本稿では 2 種類の SVM を作成した．一つは窓幅を ±3 に設定したもの (SVM^1)，もう一つは 窓幅を -1 から 0 に設定したもの (SVM^2) である．これは HMM との比較に用いる．カーネル関数は Polynominal (poly) を用いた．実験では d 値を 1 から 4 まで変えて精度を調べた．

[2] 概して言うと，英語の固有表現によくあるパターンを特徴量にしたものである．例えば "NE" は capitalized というように大文字化や記号の構成パターンを特徴化する．

[3] Brill tagger[3] を利用したが分子生物学では十分な精度が得られなかった．

[4] 松本研で作成されている Yamcha を利用した．

2.2 HMM

本実験で使用する HMM は文献 [4] に記載されているものを利用する．この HMM は文献 [4] で定義した orthographic と表層の単語を入力シンボルとする bigram モデルである．なお，品詞情報は用いていない．

表層の単語 W と orthographic 属性 F の列 $\langle W, F\rangle$ に対してモデルをたてる．まず，単語列の先頭の意味クラス C_0の生成確率を定義する．

$$\begin{aligned} & Pr(C_0| < W_0, F_0 >) \\ &= \sigma_0 f(C_0| < W_0, F_0 >) \\ &+ \sigma_1 f(C_0| < \_, F_0 >) \\ &+ \sigma_2 f(C_0) \end{aligned} \quad (1)$$

それ以外の意味クラス C_tは以下のように定義する．

$$\begin{aligned} & Pr(C_t| < W_t, F_t >, < W_{t-1}, F_{t-1} >, C_{t-1}) \\ &= \lambda_0 f(C_t| < W_t, F_t >, < W_{t-1}, F_{t-1} >, C_{t-1}) \\ &+ \lambda_1 f(C_t| < \_, F_t >, < W_{t-1}, F_{t-1} >, C_{t-1}) \\ &+ \lambda_2 f(C_t| < W_t, F_t >, < \_, F_{t-1} >, C_{t-1}) \\ &+ \lambda_3 f(C_t| < \_, F_t >, < \_, F_{t-1} >, C_{t-1}) \\ &+ \lambda_4 f(C_t|C_{t-1}) \\ &+ \lambda_5 f(C_t) \end{aligned} \quad (2)$$

これらの関数 f(·) は学習データから獲得する．また，定数 λ_iと σ_i は$\sum \sigma_i = 1.0$ で $\sum \lambda_i = 1.0$ である．$\sigma_0 \geq \sigma_1 \geq \sigma 2$，$\lambda_0 \geq \lambda_1 ... \geq \lambda_5$ という範囲で人手により決定している．

HMM は上述の 2 つの式で確率が計算されてから Viterbi アルゴリズム [15] により最適なパスを出力する．

2.3 データ

比較のために 2 種類のデータを用意する．1 つは分子生物学文献 (MEDLINE) の 100 タイトル分の概要に遺伝子と遺伝子の生成物の専門用語に対して意味タグを付与したデータ [12] (23586 単語)．もう 1 つは MUC-6 で使用された 60 個の連続したテキストデータ (24617 単語) である．各々の意味タグ

の統計量を表1と表2に提示する．内容を見て明らかなように，生物学文献における専門用語の意味タグの方がMUC-6の意味タグに比べてより複雑で頻度のばらつきも大きいことが特徴である．

表 1: 分子生物文献中の意味クラスの統計量

意味クラス	出現回数	説明
PROTEIN	2125	proteins, protein groups, families, complexes and substructures
DNA	358	DNAs, DNA groups, regions and genes
RNA	30	RNAs, RNA groups, regions and genes
SOURCE.cl	93	cell line
SOURCE.ct	417	cell type
SOURCE.mo	21	mono-organism
SOURCE.mu	64	multi-celled organism
SOURCE.vi	90	viruses
SOURCE.sl	77	sublocation
SOURCE.ti	37	tissue

表 2: MUC-6 中の意味クラスの統計量

意味クラス	出現回数
DATE	542
LOCATION	390
ORGANIZATION	1783
MONEY	423
PERCENT	108
PERSON	838
TIME	3

3 実験結果と分析

学習モデルの結果を評価するにはF値[13]を利用する．この値はCoNLLで使われた評価用スクリプト[5]から計算する．F値とは $F = (2PR)/(P+R)$ で定義される．ここで P は適合率 (Precision) R は再現率 (Recall) を表す．それぞれを説明すると以下のようになる．

$$(\text{適合率}) = \frac{\text{正しく検出された専門用語数}}{\text{検出された全ての専門用語数}} \quad (3)$$

$$(\text{再現率}) = \frac{\text{正しく検出された専門用語数}}{\text{真に正しい専門用語数}} \quad (4)$$

表3はHMMと2種類のSVMについて，生物学文献ならびにMUC-6に対する用語抽出結果を示している．全ての結果は全データを使って10 holdの交差検定を行なった．

表3の結果から，SVM^1の能力はHMMに対して生物学文献とMUC-6の両方のコーパスに対して優位な結果を得た．また，HMMと同じ特徴量だけで学習させたSVM^2ではHMMに対して低い結果を得た．このことから，SVMを用いるには多数の特徴量を扱うことができる利点を最大限に利用しないと優位な効果が得られないことを示している．

生物学文献の専門用語抽出とMUC-6のような新聞記事ベースの固有表現抽出のタスクの違いを観測すると，SVMの全ての結果で生物学文献の方が低い解析精度を得たが反対にHMMでは生物学文献の方が高い結果が得られた．これが何に起因しているのか明らかな要因がつかめていないが，おそらくHMM側のモデルが最初から生物学文献を対象に計画されてきたため何らかのモデルの特性の良さが反映された可能性が考えられる．

各意味クラスごとの精度を表4に示す．この表からSVM^1ではHMMに全体の精度では優ったものの，低い出現頻度であるRNAやSOURCE.mo, SOURCE.tiなどに対してはHMMの精度の方が優っている．これはSVMは窓幅の中の属性値の順序が考慮されていないのに対して，HMMは扱っている特徴量は一つ前の情報だけであるが，Viterbiアルゴリズムによって全体のパスの中で最適なものを獲得するという戦略がここで効果を発揮しているためと思われる．

4 結論

SVMを利用した専門用語抽出モデルならびにHMMを利用したモデルを作成し，分子生物学文献と新聞記事ベースのMUC-6に対して抽出実験を行ない精度を比較した．本稿でのSVMはパラメータ構築において強力な最適化は行なわなかったがHMMに対して2つの分野に対して精度が優るこ

[5] http://lcg-www.uia.ac.be/conll2002/ner/bin から入手可能

表 3: HMM と SVM の意味タグ付与能力の評価 (F 値)

コーパス	学習モデル					
	HMM	$SVM^1(poly)$ degree d=				$SVM^2(poly)$ degree d=
		1	2	3	4	2
生物	**70.97**	71.33	**71.78**	68.54	65.09	**65.63**
MUC-6	**70.38**	72.86	**73.21**	69.22	65.12	**65.94**

とが実験から明かになった．この時，SVM のパラメータは±3 の広い窓幅で多くの特徴量を使ったことが精度向上の大きなポイントであった．

しかしながら，ここでの SVM 学習の結果は明らかに最適化されていない．例えば窓幅内の属性パラメータ内に語順を反映することなどである．また，統語的要素として parser を使い構文的な情報 (主辞情報など) を入れることも考えられる．現時点ではコーパス量が限られていることから用いないが今後の課題としたい．また，現時点では実時間で学習がおわるようにプログラミングするのが簡易ではないが写像関数も polynominal 以外の他のものについて試すことを計画している．

5 謝辞

この研究の一部は文部科学省の補助金 (番号 14701020) に補助を受けました．また，辻井潤一教授には意味タグが付与された分子生物学コーパスの使用を許して下さったことに感謝いたします．

参考文献

[1] Bikel, D., Miller, S., Schwartz, R. and Wesichedel, R.: Nymble: a high-performance learning name finder, *In Proceedings of the Fifth Conference on Applied Natural Language Processing (ANLP'97)*, pp. 194–201 (1997).

[2] Brill, E.: Transformation-based error-driven learning and natural language processing:A case study in part-of-speech tagging, *Computational Linguistics 21*, pp. 543–565 (1995).

[3] Brill, E.: A simple rule-based part of speech tagger, *In Third Conference on Applied Natural Language Processing - Association for Computational Linguistics*, pp. 152–155 (2000).

[4] Collier, N., Nobata, C. and Tsujii, J.: Extracting the names of genes and gene products with a hidden Markov model, *In Proceedings of the 18th International Conference on Computational Linguistics (COLING'2000)* (2000).

[5] Collins, M. and Singer, Y.: Unsupervised models for named entity classification, *In Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in Natural LanguageProcessing and Very Large Corpora* (1999).

[6] Cortes, C. and Vapnik, V.: Support-vector networks, *Machine Learning, 20*, pp. 273–297 (1995).

[7] DARPA: *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, Morgan Kaufmann (1995).

[8] Fukuda, K., Tsunoda, T., Tamura, A. and Takagi, T.: Toward information extraction: identifying protein names from biological papers, *In Proceedings of the Pacific Symposium on Biocomputing'98(PSB'98)*, pp. 707–718 (1998).

[9] Joachims, T.: Making large-scale SVM learning practical, *Advances in Kernel Methods - Support Vector Learning* (Scholkopf, B., Burges, C. and Smola, A.(eds.)), MIT Press (1999).

[10] Kudo, T. and Matsumoto, Y.: Use of support vector learning for chunk identification, *In Proceedings of the Fourth Conference on Natural Language Learning (CoNLL-2000)*,

表 4: 各意味クラスに対する HMM と SVM の精度 (F 値)

意味クラス	学習モデル					
	$SVM^1(poly)^\dagger$ d=2			HMM		
	p	r	$F_{\beta=1}$	p	r	$F_{\beta=1}$
分子生物学文献						
PROTEIN	75.99	78.87	77.40	78.84	78.81	**78.82**
DNA	69.32	48.88	**57.33**	48.80	45.79	47.25
RNA	80.00	13.79	23.53	41.67	16.67	**23.81**
SOURCE.cl	75.00	48.91	**59.21**	50.00	38.71	43.64
SOURCE.ct	78.89	54.68	64.59	70.60	64.51	**67.42**
SOURCE.mo	0.00	0.00	0.00	77.78	35.00	**48.28**
SOURCE.mu	72.50	45.31	**55.77**	60.00	32.81	42.42
SOURCE.vi	91.80	60.87	**73.20**	78.08	62.64	69.51
SOURCE.sl	72.41	54.55	**62.22**	65.67	57.14	61.11
SOURCE.ti	50.00	2.70	5.13	47.06	21.62	**29.63**
All	75.89	68.09	**71.78**	73.10	58.95	70.97
MUC-6 のデータ						
DATE	84.52	62.56	**71.90**	74.47	61.67	67.47
LOCATION	67.78	48.22	**56.35**	66.23	40.32	50.12
ORGANIZATION	70.17	81.19	**75.28**	65.33	66.01	65.67
MONEY	83.11	76.88	**79.87**	77.71	80.62	79.14
PERCENT	75.68	51.85	61.54	85.48	98.15	**91.38**
PERSON	88.62	76.70	82.23	87.06	80.46	**83.63**
TIME	0.00	0.00	0.00	0.00	0.00	0.00
All	76.13	73.24	**74.66**	73.09	67.87	70.38

pp. 142–144 (2000).

[11] rabiner, L. and Juang, B.: An introduction to hidden Markov models, *IEEE ASSP Magazine*, pp. 4–16 (1986).

[12] Tateishi, Y., Ohta, T., Collier, N., Nobata, C., Ibushi, K. and Tsujii, J.: Building an annotated corpus n the molecular-biology domain, *COLING'2000 Workshop on Semantic Annotation and Intelligent Content* (2000).

[13] van Rijsbergen, C.: *Information Retrieval*, Butterworths (1979).

[14] Vapnik, V.: *The Nature of Statistical Learning Theory*, Springer-Verlag (1995).

[15] Viterbi, A. J.: Error bounds for convolutions codes and an asymptotically optimum decoding algorithm, *IEEE Transactions on Information Theory,IT-13(2)*, pp. 260–269 (1967).