自然言語処理 140−11
(2000. 11. 21)

単語の部分文字列を考慮した専門用語抽出と分類

山田 寛康, 工藤 拓, 松本 裕治
奈良先端科学技術大学院大学 情報科学研究科
{hiroya-y, taku-ku, matsu}@is.aist-nara.ac.jp

本稿では, 医学生物学分野の文献アブストラクトから, 単語の表層, 品詞情報だけでなく, 単語の部分文字列情報を考慮した専門用語の自動抽出・分類手法を提案する. 抽出・分類規則は, 人手により専門用語のタグ付けをしたデータから教師つき学習アルゴリズム Support Vector Machines を使用して自動的に学習する. 実験の結果, 医学生物学分野の文献に対する専門用語抽出に, 単語の部分文字列を考慮することが有効であることを確認した.

キーワード:情報検索, 情報抽出, 固有表現抽出, 専門用語抽出, サポートベクター学習

Using Substrings for Technical Term Exatraction and Classification

YAMADA Hiroyasu, KUDO Taku, MATSUMOTO Yuji
Graduate School of Infomation Science, Nara Institute Science and Technology
{hiroya-y, taku-ku, matsu}@is.aist-nara.ac.jp

We propose a method for extracting and classifying technical terms from MEDLINE abstracts in the molecular-biology domain. In our method, we use substrings of the terms as features in addition to the information of strings and parts-of-speech. Then we construct a human annotated corpus of technical terms , and apply Support Vector Machines to learn rules for automatic identification and classification of technical terms. Empirical results demonstrate the effectiveness of our method.

Keywords : Information Retrieval, Information Extraction, Named Entity, Technical Terms Extraction, Support Vector Learning

1 はじめに

人名・組織名といった固有表現を自動的に抽出する固有表現抽出 (Named Entity) の問題は, 情報検索, 情報抽出の基礎技術としてのみならず, 自然言語処理における形態素解析, 構文解析などの処理に大きな影響を及ぼすため, 重要な問題とされている.

一方, 医学生物学分野のような専門性の高い文献においては, 人名・組織名などだけでなく, 分野固有の言い回しや専門用語が頻繁に出現するため, このような分野に関する情報検索, 情報抽出をする場合, 専門用語を自動的に抽出する技術は必要不可欠である.

専門用語抽出を含め固有表現抽出の手法として, 大きく2つのアプローチがある. 一つは人手で作成した抽出規則を用いる方法であり, もう一つは人手により専門用語をにタグ付けした文献データから自動的に抽出規則を学習するアプローチ [10, 15] である.

専門用語抽出に関して, 人手作成規則を用いる研究として, 福田らの研究がある [16]. 福田らは医学生物学分野の文献アブストラクトを対象に, タンパク質名の自動抽出を行なった. 彼らはタンパク質名の候補となる core-term と呼ばれる単語を人手で作成した5つの処理を使用して抽出し, さらに表層的な手がかりや, 品詞情報などのを利用した9の規則により, タンパク質名を同定した.

しかしこれらの処理や規則は, 限られたタンパク質名に特化したもので, タンパク質以外の用語に適用するためには, 人手により新たな処理や規則を追加する必要がある. 福田らを含め人手作成規則を利用するアプローチの問題は, 次々と出現する新しい固有表現や専門用語に対して, 人手により新たに規則を追加しなければならない点である. 特に専門性の高い分野では, 別の分野の専門用語を抽出する場合, 適用分野に特化した規則を人手により追加しなければならず, その結果, 汎用性に欠けることが指摘される.

これに対して専門用語抽出規則を自動的に学習するアプローチに Collier ら [9] や合原ら [14] の研究がある. Collier らは, 医学生物学分野の文献から, タンパク質名を同定するために, 単語情報, 品詞情報, 文字種情報を素性とし, 隠れマルコフモデルにより抽出規則を学習する手法を提案した.

合原らは, 医学生物学分野の文献から, 13種類に分類した専門用語の自動抽出・分類規則を行なった. 彼らは単語の文字種, 品詞, さらに係り受け情報を素性とし, 決定木学習と Co-training [5] というブートストラップに基づく手法を適用し, 少量の訓練データと大量の未知データを相補的に利用し高精度の分類規則を学習する手法を提案した.

これらの研究は人手で抽出分類規則を作成するアプローチに比べて, 少量の訓練データを作成すれば, 訓練データから適用分野に対応した抽出規則を自動的に学習できるため, 様々な分野に適用可能であり, 汎用性が高く, 人手コストを大幅に削減できるという利点がある.

また専門用語抽出規則を学習するために Collier らや合原らが使用した素性は, ある専門用語の表層文字列, 品詞情報, さらに数字や特殊記号などが単語にどの程度含まれているかなど, 文字種に関するものを用いた.

文字種などの素性は専門知識を持たない者が専門用語であるか否かを判定する基準の一つとして直観的にも容易に予測でき, また彼らの実験によりその有効性が報告されている. しかし様々な分野に有効な文字種の規則を網羅的かつ一意的に決定することは難しい.

本稿では医学生物学分野の文献アブストラクト MEDLINE[1] から専門用語を自動的に抽出・分類することを目的とする. Collier らや合原ら同様に, 様々な分野や新しい専門用語にも対応するために, 専門家により専門用語タグが付けられたデータから, 分類・抽出規則を自動的に学習するアプローチを採用する. また Collier らや合原らにより, 専門用語の抽出分類規則には, 単語の文字種などが有効であることが報告されているため, これを考慮した抽出規則の学習を行なう. 文字種の規則を学習においては, Collier らや合原らのように特定の文字種やパターンに限定するのではなく, 単語の部分文字列を考慮することで対処する.

以下次節では, 専門用語抽出・分類のための分割モデルついて述べる. 3節では抽出・分類規則を学習するために使用した Support Vector Machines について説明し, 4節では専門用語抽出・分類実験の結果と考察について述べる. 最後に5節でまとめと今後の課題について述べる.

2 専門用語抽出と分類

2.1 BIO 分割モデル

Tjong Kim Sang らは chunking 問題に対して 分割する token に対して chunk の開始 (Begin), chunk 内 (In)

, chunk 以外 (Other) という 3 種類のタグを付与する分割モデルとを提案した [6, 8].

専門用語抽出 · 分類も, 専門用語の開始, 専門用語内, 専門用語以外の 3 つを決定する chunking タスクとしてみなすことができるため, 我々も Rawshow らと同様 BIO 分割モデルを採用した.

単語	抽出	分類
CcdB	B	protein
prevented	O	O
CcdA	B	reaction
degradation	I	reaction
by	O	O

図 1: BIO 分割モデル

図 1 は "CcdB prevented CcdA degradation by ···" という文に対して, BIO 分割モデルを適応した例を示す. この文で "CcdB", "CcdA degradation" が順に "protein" , "reaction" のクラスに属する専門用語であるとする. 専門用語抽出では, 出現順に B, O, B, I, O のタグを付与し, 専門用語分類では, B-protein, O, B-reaction, I-reaction, O のタグを付与することで, 専門用語となる単語列とそうでないものとを識別する. 以後本稿では単語に対して付与する B, I, O 分割モデルに基づくタグを総称して BIO タグと呼び, 専門用語抽出 · 分類タスクを, 出現する単語に関して B, I, O に分類する分類問題として扱う.

3 抽出 · 分類規則の学習

抽出 · 分類規則の学習は出現する単語を B, I, O に分類する規則の学習として扱う. 従って学習に使用する事例は単語毎に生成する. はじめに学習に使用する事例の素性について説明する.

3.1 訓練事例と素性

文中に i 番目に出現する単語に対して, その単語の BIO タグを Y_i , 学習する素性を要素とするベクトルを $\vec{x}_i$ で表現すると, 訓練事例は 分類すべき BIO タグとその素性ベクトルのペア (x_i, Y_i) で表現できる.

出現する i 番目の単語 (事例) の素性は $i-2$ から $i+2$ 番目の単語と品詞, $i-2$, $i-1$ 番目の BIO タグ, i 番目の単語の部分文字列とする. 部分文字列としては, 単語を構成するすべての部分文字列を考慮する substring と prefix と suffix のみ考慮する prefix·suffix の 2 つを扱う。

図 2 は "CcdB prevented CcdA degradation by ···" という文に対して, 単語 "CcdA" についての素性を示す.

図 2 で substr(CcdA) は単語 "CcdA" の部分文字列を表す. 即ち, Substr(CcdA) = {^C, c, d, A$, ^Cc, cd, dA$, ^Ccd,cdA$} である. prefix·suffix(CcdA) は単語 "CcdA" の prefix と suffix を表す. 即ち prefix·suffix(CcdA) = {^C, A$, ^Cc, dA$,^Ccd,cdA$} である. "^" , "$" はそれぞれ単語の始まりと終りを表す記号である. 今回は展開した部分文字列のうち長さが 3 文字以上のものを採用した.

	素性			
位置	単語	品詞	BIO	部分文字列
-2	CcdB	NNP	B	-
-1	prevented	VBD	O	-
0	CcdA	NNP	B	prefix·suffix(CcdA) or substr(CcdA)
+1	degradation	NN	-	-
+2	by	IN	-	-

図 2: 訓練事例の素性

3.2 Support Vector Machines

図 3 は Support Vector Machines (SVMs) の概要図を示す. SVMs は訓練事例 $(\vec{x}, y)$(正例 $(y = 1)$, 負例 $(y = -1)$) を正しく分離する超平面 $f(\vec{x}) = \vec{w} \cdot \vec{x} + b$ を margin 最大化に基づく戦略により求める教師あり学習法であり, 学習により得られた分類器は二値線形分類器である [12]. SVMs には次の 2 つの特長がある.

- SVMs の汎化能力は理論的に素性空間の次元に依存しないということが証明されており [12], これが高次元素性空間での学習においても過学習を軽減することができる.
- Kernel 関数を使用することで非線形分類問題を扱うことが可能であり, 特に Kernel 関数として多項式関数を用いることで素性の組合せ考慮した学習が容易に実現できる.

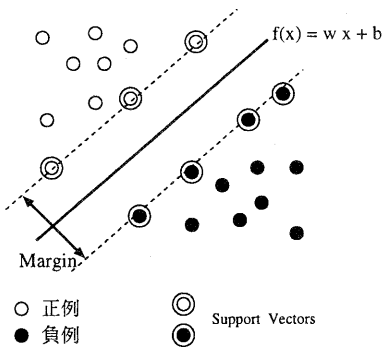

図 3: Support Vector Machines の概要

これら 2 つの特長により, SVMs は数千から数万にも及ぶ高次元空間を扱う文書分類 [7, 13] や chunking [11] の自然言語処理技術に適用され, 素性空間の次元に依存しない高い汎化能力によりその有効性が実証されている. 本稿でも単語や品詞, 部分文字列情報を使用することで高次元空間での分離問題を扱うため 学習アルゴリズムとして SVMs を採用する.

SVMs は二値分類器であるため, 3 クラス以上の分類問題に適用できるよう, 拡張する必要がある. 分類したい k 個のクラス数だけ SVMs を構築し, k 個のクラスそれぞれに対する 二値分類問題に分割することで対処する. これを K Class Classification と呼ぶ.

図 4 は BIO タグ分類問題に対して K Class Classification を適用した概要を示す. 訓練事例を B であれば正例そうでなければ負例, という二値問題に変換し, SVMs により学習し, 分離超平面 H_B を得る. 同様に I , O について事例を二値問題に変換することで I, O それぞれに対する分離超平面 H_I, H_O を得る.

3.3 SVMs による専門用語抽出 · 分類

SVMs により学習した超平面を用いて, テスト文の単語に対する専門用語抽出 · 分類, 即ち BIO タグの決定方法について述べる.

抽出 · 分類はテスト文の先頭から順に行なう. i 番目の単語の素性は, $i-2$ から $i+2$ までの単語と品詞情報, i 番目の部分文字列情報を使用する. $i-1, i-2$ の BIO タグは SVMs によって順次決定された結果から動的に決定される.

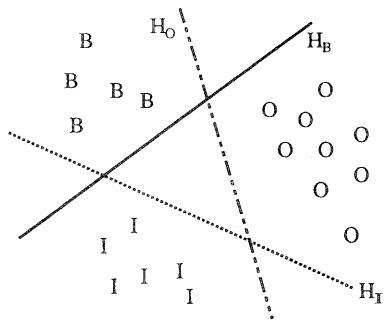

図 4: SVMs による複数クラスの分類

位置	素性			
	単語	品詞	BIO	部分文字列
-2	CcdB	NNP	動的に決定	-
-1	prevented	VBD	動的に決定	-
0	CcdA	NNP	?	prefix·suffix(CcdA) or substr(CcdA)
+1	degradation	NN	-	-
+2	by	IN	-	-

図 5: テスト事例の素性

SVMs より学習した k 個のクラス C_k に関する分離超平面 を $H_k(\vec{x}) = \vec{w} \cdot \vec{x} + b$ とすると, テスト事例 $\vec{x}_{test}$ のクラス C_i は 分類超平面からの距離 $H_k(\vec{x}_{test})$ を使用して式 1 のように書ける.

$$C_i = arg \max_i H_i(\vec{x}_{test}) \quad (1)$$

即ち, 各クラスの分離平面との距離を計算し, 最大の距離を得る分離超平面のクラスをテスト事例のクラスと決定する.

図 5 にテスト文として "CcdB prevented CcdA degradation by ···" が入力されたとき, 単語 "CcdA" についての素性を示す. 単語 "CcdA", "prevented" の BIO タグは式 1 により順次決定され, 後の単語の BIO タグ素性として利用される.

4 実験

4.1 データ

データは医学生物学分野の文献アブストラクト MEDLINE[1] を対象にした. MEDLINE に登録され

表 1: 専門用語の種類とその頻度

専門用語クラス	被験者 A	被験者 B
condition	199	141
function	172	201
gene	117	133
material	54	87
ethod	168	159
others	23	1
protein	365	396
reaction	355	307
research field	150	64
sequence	292	247
site	90	136
species	169	135
structure	114	150
合計	2268	2157

ている文献に対して専門家二人によって専門用語をタグ付けし実験データとした. 収集した文献アブストラクト数は 77 件, 総文数は 525 文, 総単語数は 14990 語であった. 単語の品詞付与には英語版「茶筌」[2] を使用した. 表 1 に使用したデータに対して二人の被験者がタグ付けした専門用語の種類とその出現頻度を示す.

実験は抽出・分類共に, 実験データを文を単位に 5 等分に分割し, 交差検定することで行なった. 評価はタグ正解率, 専門用語抽出・分類の適合率, 再現率, F 値を使用した. タグ正解率, 専門用語抽出・分類の適合率, 再現率, F 値 の詳細を以下に記す.

- タグ正解率 $= \frac{\text{正解 } BIO \text{ タグ数}}{BIO \text{ タグ総数}} \times 100$
- 適合率 $= \frac{\text{専門用語正解数}}{\text{システムが出力した専門用語数}} \times 100$
- 再現率 $= \frac{\text{専門用語正解数}}{\text{被験者がタグ付けした専門用語数}} \times 100$
- F 値 $= \frac{2 \times \text{適合率} \times \text{再現率}}{\text{適合率} + \text{再現率}}$

4.2 実験結果

被験者間の一致

まず最初に専門用語抽出・分類の難しさを測るために, 被験者 A をシステムの出力, 被験者 B を正解として評価した. 表 2 に被験者間における抽出・分類精度を示す. 表 2 より, 抽出は被験者間でも揺れが少なく, 高精

表 2: 被験者間の精度

抽出			
タグ正解率	適合率	再現率	F 値
89.60	78.62	82.92	80.71
分類			
タグ正解率	適合率	再現率	F 値
72.43	60.54	63.84	62.15
クラス			
condition	46.23	65.25	54.12
function	53.49	45.77	49.33
gene	78.63	69.17	73.60
material	51.85	32.18	39.72
method	55.49	60.38	57.83
others	0.00	0.00	0.00
protein	84.15	77.78	80.84
reaction	56.30	65.47	60.54
research field	36.67	85.94	51.40
sequence	58.22	68.83	63.08
site	75.56	50.00	60.18
species	57.89	71.74	64.08
structure	68.42	52.00	59.09

度の抽出規則を学習することが期待できる．しかし分類では，各専門用語クラスで被験者間の揺れが大きく，専門用語を抽出しさらにある専門用語クラスに分類するタスクの難しさがわかる．

4.3 SVMs による抽出・分類結果

次に SVMs により学習した抽出・分類規則による実験について報告する．表 3 に 3 種類の実験の名称と使用した素性を示す．baseline では単語と品詞情報のみを素性として使用し学習した．単語表層と品詞情報だけでなく部分文字列を考慮した学習をするために，substring と prefix·suffix の 2 種類の実験をした．substring は 3 文字以上のすべての部分文字列を素性とし，prefix·suffix 3 文字以上の prefix suffix を素性として使用した．また SVMs による学習では Kernel 関数として二次の多項式関数を使用した．

表 3: 実験の種類

	単語	品詞	部分文字列
baseline	○	○	-
substring	○	○	部分文字列
prefix·suffix	○	○	prefix と suffix

baseline, substring, prefix·suffix について，被験者 A 被験者 B のそれぞれに対し交差検定を行なった．表 4 に訓練データ数とテストデータ数の比が 4:1 の場合の抽出・分類精度を示す．

図 6, 7 はそれぞれ被験者 A, 被験者 B の訓練データサイズによる抽出精度の変化を示す．図 8, 9 はそれぞれ被験者 A, 被験者 B の訓練データサイズの変化による専門用語分類精度を示す．図 6 から図 9 で縦軸は F 値を表し，横軸はテストデータ数 1(105 文) に対する訓練データ数の比を表す．

抽出

表 4 より専門用語抽出・分類共に，単語の表層，品詞情報だけでなく，部分文字列を考慮して学習した規則により精度の向上が見られた．特に部分文字列として，prefix·suffix を考慮した場合に F 値で 被験者 A, 被験者 B に対して それぞれ 82.79,78.82 という高い精度を得た．

図 6, 7 からわかるように，少量の訓練データで学習した場合にも，prefix·suffix が専門用語抽出に有効であることがわかる．

分類

専門用語クラス分類は表 4, 図 8, 9 より，多量の訓練データを使用し prefix·suffix を考慮した場合でも，F 値はそれぞれ 43.91,46.91 と専門用語抽出に比べ大幅に低い結果しか得られなかった．

精度低下の原因としては，分類タスクは専門用語抽出タスクよりも難しいタスクであるうえに，表 2 より被験者間でも専門用語クラスの分類に関して揺れが大きいことが考えられる．

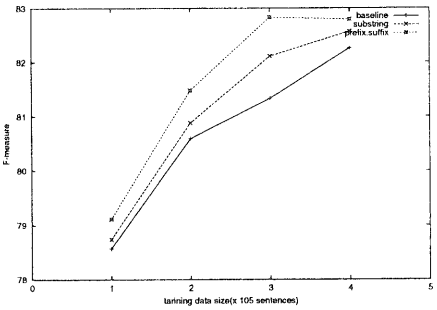

図 6: 被験者 A の訓練データ数の変化と抽出精度

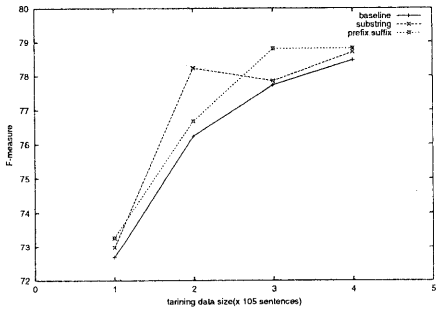

図 7: 被験者 B の訓練データ数の変化と抽出精度

表 4: SVM の精度

抽出 (被験者 A/被験者 B)				
	タグ正解率	適合率	再現率	F 値
baseline	92.19/89.69	81.30/76.95	83.23/80.05	82.26/78.47
substring	92.31/89.73	81.55/77.12	83.63/80.37	82.57/78.71
prefix·suffix	92.34/89.79	81.85/77.08	83.76/80.65	82.79/78.82
分類 (被験者 A/被験者 B)				
	タグ正解率	適合率	再現率	F 値
baseline	59.59/62.38	45.39/48.21	41.26/42.92	43.23/45.41
substring	59.58/62.82	45.41/48.42	42.32/44.12	43.81/46.17
prefix·suffix	59.50/63.03	45.82/49.44	42.14/44.63	43.91/46.91

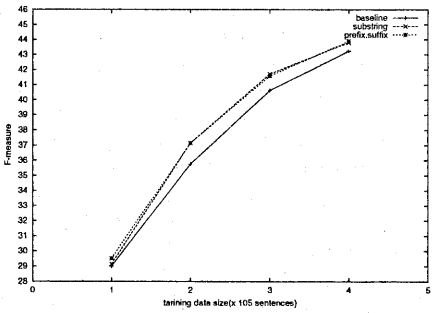

図 8: 被験者 A の訓練データ数の変化と分類精度

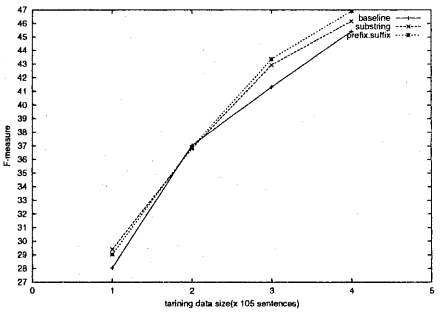

図 9: 被験者 B の訓練データ数の変化と分類精度

素性の組合せについて

素性間の組合せを考慮することで, 抽出精度に及ぼす影響をみるために学習に使用した多項式 Kernel 関数の次数の違いによる精度を比較した. 多項式 Kernel 関数の次数 d が 1 の場合, 素性を独立に考慮する線形分類器となる。多項式 Kernel 関数の次数 d が 2 の場合、2 つの素性の組合せを考慮する非線形分類器となる。表 5 は多項式 Kernel 関数の次数の違いによる F 値をしめす.

表 5 より 多項式 Kernel 関数の次元を 1 次から 2 次にあげることで baseline では 被験者 A, 被験者 B のそれぞれに対し 1.34, 9.9 の F 値の向上が見られた. それに対して部分文字列を考慮した substring , prefix·suffix は 1.93, 1.92 と 1.96, 1.62 という F 値の向上が見られた.

これにより単語, 品詞情報と部分文字列を独立に考慮して学習するよりも, 品詞の素性と部分文字列の組合せを考慮して学習することで抽出精度に貢献することがわかった.

5 まとめ

本稿では医学生物学分野の文献アブストラクト MEDLINE から単語の表層, 品詞情報に加えて, 単語の部分文字列を考慮した専門用語の自動抽出・分類手法を提案した. 分類・抽出規則は Support Vector Machines により単語の表層, 品詞情報, 部分文字列を考慮し自動的に学習した. 実験により部分文字列まで考慮して学習した規則が, 専門性の高い分野の専門用語抽出・分類に有効であることを示した.

今後の課題として以下 3 点があげられる.

表 5: kernel 関数の違いによる抽出結果

	F 値 (被験者 A/被験者 B)	
	$d=1$	$d=2$
baseline	79.35/75.39	80.69/76.29
substring	79.15/74.71	81.08/76.63
prefix-suffix	80.83/79.23	82.79/80.85

- (1) 専門用語分類の精度改善
- (2) 部分文字列を考慮した場合の学習の改善
- (3) 固有表現抽出タスクへ応用

(1) に関しては, 分類に関する被験者間の揺れが大きな原因と考えられるため, 分類する専門用語クラスの検討をするとともにコーパスの整備をする予定である.

(2) の問題は, 部分文字列を考慮にして学習する時, すべての単語に対して部分文字列を展開しているため過剰に高次元な素性空間となる. その結果 kernl 関数を使用しない学習時に精度の低下の原因となっている. この問題に対処するために, 部分文字列を考慮せず学習し, そこで得られた 分類の難しい Support Vector となった事例に対してのみ部分文字列展開を行ない再学習する方法が考えられる. 今後この方法を適用し実験により検証する.

(3) は今回提案した手法を MUC[3] や IREX [4] などの固有表現抽出タスクに適用し本手法の汎用性を検証する予定である.

なおこの研究は文部省科学研究費重点領域研究「ゲノムサイエンス」の援助を受けている.

参考文献

[1] *Internet Grateful Med Development Team, National Libarary of Medline: MEDLINE(1996).*

[2] 英語版「茶筌」: ChaSen Home Page: http://cl.aist-nara.ac.jp/lab/nlt/chasen/.

[3] *MUC:Proceedings of the 7th Message Undestanding Conference (MUC-7)*, 1998.

[4] IREX 実行委員会: IREX homepage, http://cs.nyn.edu/cs/projects/proteus/irex, 1999.

[5] Avrim Blum , Tom Mitchell. Combining Labeled and Unlableled Data with Co-Training. In *Annual Conference on Computational Learning Theory(COLT-98)*, pp. 92–100, 1998.

[6] Erik F. Tjong Kim Sang and Jorn Veenstra. Representing text chunks. In *In Proceedings of EACL'99*, pp. 173–179, 1999.

[7] Hirotoshi Taira "," Masahiko Haruno. Feature Selection in SVM Text Categorization. In *AAAI-99/IAAI-99 Proceedings *Sixteenth National Conference on Artificial Intelligence / Eleventh Conference on Innovative Applications of Artificial Intelligence, Orlando, Florida*, pp. 480–486, 18-22 Jul 1999.

[8] Lance A. Ramshaw and Mitchell P. Marcu. Text Chunking Using Transformation-Based Learning. In *In Proceedings of the Third ACL Workshop on Very Large Corpora*, pp. 82–94, 1995.

[9] Nigel Collier and Hyun Seok Park and Norihiro Ogata and Yuka Tateishi and Chikashi nobata and Tomoko Ohta and Tateshi Sekimizu and Hisao Imai and Katsutoshi Ibushi and Jun-ichi Tsujii. The GENIA Project: Corpus-based Knowledge Acquisition and Information Extraction from Geonome Research Papers. In *Ninth Conference of the European Chapter of the Association for Computational Linguistics *Proceedings of EACL '99, Bergen, Norway*, pp. 271–272, 8 Jun 1999.

[10] Satoshi Sekine and Ralph Grishman and Hiroyuki Shinnou. A Decision Tree Method for Finding and Classifying Names in Japanese Texts. In *the Sixth Workshop on Very Large Corpora*, pp. 171–178, 1998.

[11] Taku Kudoh and Yuji Matsumoto. Use of Support Vector Learning for Chunk Identification. In *Computational Natural Language Learning (CoNLL-2000)*, pp. 142–144, 2000.

[12] Vladimir N. Vapnik. *The Nature of Statistical Learning Theory.* New York, 1995.

[13] Yiming Yang "," Xin Liu. A Re-examination of Text Categorization Methods. In *SIGIR '99 *Proceedings of the 22nd International Conference on Research and Development in Information Retrieval, University of California, Berkeley*, pp. 42–49, 9 Aug 1999.

[14] 合原 博 and 宮田 高志 and 松本裕治. 医学生物学分野からの専門用語抽出分類. 情報処理学会研究会報告, 第 2000 巻, pp. 41–48, 2000.

[15] 内元 清貴 and 馬 青 and 村田 真樹 and 小作 浩美 and 内山 将夫 and 井佐原 均. 最大エントロピーモデルと書き換え規則に基づく固有表現抽出. 自然言語処理, 第 7 巻, pp. 63–90, 2000.

[16] 福田 賢一郎"," 角田 達彦"," 田村 あゆち"," 高木利久. 医学生物学文献からの専門用語抽出にむけて: タンパク質名の自動抽出. 情報処理学会論文誌, 第 39 巻, pp. 2421–2429, 8 1998.