Analysis and Improved Recognition of Protein Names Using Transductive SVM

Masaki Murata
National Institute of Information and Communications Technology
Email: murata@nict.go.jp

Tomohiro Mitsumori
Miyazono Patent Office
Email: mitsumor01@yahoo.co.jp

Kouichi Doi
Pharma Security Consulting Inc.
Email: doy@pharmasecurity.jp

***Abstract*— We first analyzed protein names using various dictionaries and databases and found five problems with protein names; i.e., the treatment of special characters, the treatment of homonyms, cases where the protein-name string may be a substring of a different protein-name string, cases where one protein exists in different organisms, and the treatment of modifiers. We confirmed that we could use a machine-learning approach to recognizing protein names to solve these problems. Thus, machine-learning methods have recently been used in research to recognize protein names. A classifier trained in a specific domain, however, can cause overfitting and be so inflexible that it can only be used in that domain. We therefore developed a new corpus on breast cancer and investigated the flexibility of classifiers trained on the GENIA [1] or the breast-cancer corpora. We used a transductive support vector machine (SVM) to avoid overfitting, and we evaluated the effect of transductive learning. We found that transductive SVM prevented overfitting in experiments and yielded higher accuracies than were obtained from the conventional SVM. The transductive SVM increased the F-scores (70.46 to 79.64 and 70.63 to 74.61) in our two experiments for the criterion of "Sub" that we define in this paper.**

***Index Terms*— overfitting, protein name recognition, biomedical literature, SVM, transductive SVM, different domain**

I. Introduction

Many studies using the technology of natural language processing (NLP) to extract information from the biomedical literature have recently been reported. Typical studies have explored the extraction of terms related to protein-protein interactions or the function of genes [2]–[4], and the first step in this extraction is the recognition of technical terms such as "protein," "gene," "RNA," "disease," and "cure" [5]–[9].

Various techniques for recognizing technical terms have been proposed. Fukuda et al. [5] and Franzen et al. [9] analyzed words that were used as protein names in abstracts and prepared rules that allow protein names to be recognized automatically. When rule-based methods are used, new rules must be created when new protein names appear, and this complicates rule management.

Protein names and corresponding gene names are stored in databases such as SWISS-PROT [10], and several researchers have extracted protein names from such databases, compiled dictionaries, and undertaken protein-name recognition based on pattern matching or dynamic programming [11]. Dictionary-based methods, however, result in many false positives, which are instances in which word sequences that are not protein names are mistakenly recognized as protein names. This occurs because the protein name may be identical to a word that is often used in ordinary text. The maltose-binding protein, for example, is abbreviated to MAP, which is also the word for a graphical representation of a portion of the surface of the earth. Also, "white," the word for the color of objects that reflect nearly all visible wavelengths, is used as a gene name. Another shortcoming of dictionary-based methods is that they cannot recognize a protein name that is not stored in a dictionary.

The use of machine-learning methods to recognize biomedical terms has also been explored [6], [12]–[14]. Correct answers in the training data (corpus) need to be annotated when a machine-learning method is used. Collier et al. [6] annotated 100 abstracts, and recognized technical terms by using hidden Markov models (HMMs). Kazama et al. [12] used the GENIA corpus as training data, and extracted technical terms by using support vector machines (SVMs). Task 1 of the Biocreative workshop[1] was a competition to recognize gene or protein names, and some of the machine-learning methods used

This paper is based on "Overfitting in protein-name recognition in biomedical literature and method of preventing it through use of transductive SVM," by M. Murata, T. Mitsumori, and K. Doi, which appeared in the Proceedings of the 4th International Conference on Information Technology: New Generations (ITNG 2007), Las Vegas, USA, April 2007. © 2007 IEEE.

[1] http://www.pdg.cnb.uam.es/BioLINK/BioCreative.eval.html

```
<protein*>  <name>protein name</name>
</protein> <gene> <name>gene name</name>
</gene>.
```

Figure 1. Tagging rule in XML format

```
<TreeNumber>D12.776.+
</TreeNumber>  <String>protein
name</String>  <TreeNumber>D08.586.+
</TreeNumber>  <String>enxyme
name</String>.
```

Figure 2. Tagging rule in XML format

in this workshop performed well. Although machine-learning methods are known to be efficient, it is expensive to produce the training data they require [15]. The GENIA corpus is a collection of abstracts containing the terms *human*, *blood cell*, and *transcription factor*, and the machine-learning method used to recognize protein names in this corpus yielded an F-score of about 80%. Some researchers extracting information from the biomedical literature, however, do not want to collect information about a specific domain but about a broad domain, and producing the training data for such a domain is very expensive. A great deal of information should therefore be obtained using only a small amount of training data. The purpose of the research reported here was to investigate the flexibility of a classifier trained in a specific field. We evaluated the accuracy with which a classifier trained to recognize protein names in one specific field could recognize protein names in a different specific field. Further, because classifiers trained in one specific domain are known to cause overfitting [16], [17] — i.e., to be inaccurate in a different specific field — we also explain how overfitting can be avoided by using a transductive method.

This paper is organized as follows. Section II describes our analysis of protein names. We analyzed protein names using public protein databases and documents. We also summarized the problems related to protein names. Section III explains SVMs and the transductive SVM we used in our studies on the recognition of protein names. Section IV describes the experiments we conducted in these studies. We confirmed that a conventional SVM caused overfitting in the experiments and yielded low accuracies when trained in a different domain. We also confirmed that the transductive SVM avoided overfitting and obtained higher accuracies than the ordinary SVM when trained in a different domain. We discuss how effective our method of using the transductive SVM was in Section V and we provide concluding remarks in Section VI.

II. Analysis of Protein Names

Before we investigated the use of machine learning for recognizing protein names, we examined some problems related to protein-name recognition. We first extracted protein names from three public databases — Swiss-Prot,

TABLE I.
NUMBERS OF PROTEIN NAMES IN THREE DIFFERENT DATABASES.

Database name	Class	Proper
Swiss-Prot (protein)	-	96,195
Swiss-Prot (gene)	-	115,663
MeSH (protein)	3,246	9,831
MeSH (enzyme)	1,745	6,973
LIGAND	206	15,087

TABLE II.
NUMBERS OF CO-OCCURRENCES OF PROTEIN NAMES IN DATABASES.

Database	Co-occurrences
Swiss-Prot and MeSH	620
Swiss-Prot and LIGAND	560

Medical Subject Heading (MeSH)[2], and LIGAND[3] — and used the obtained list of names to search for protein names in Medline abstracts. Here, we describe results from the process of information retrieval of protein names, such as the treatment of special characters, the treatment of homonyms, cases where the protein-name string may be a substring of a different protein-name string, cases where one protein exists in different organisms, and the treatment of modifiers.

The protein names were collected from three databases: Swiss-Prot, which is a database of protein sequences and function information; MeSH, which is a vocabulary thesaurus used for indexing articles in the PubMed database concerned with life science; and the LIGAND database, which provides information on chemical compounds and reactions in biological pathways.

The Swiss-Prot database, when released on 30 April, 2003, contained 125,744 entries. We extracted the protein and gene names from the provided XML format source as shown in Fig. 1. The definition of protein names begins with `<protein>` and finishes with `</protein>`. The expressions of protein names and synonyms are recorded between `<name>` and `</name>`. This rule also applies to gene names. We extracted 254,462 protein names, including synonyms, and 197,597 gene names. 96,195 protein names and 115,663 gene names remained after duplicate entries were removed.

The MeSH database covers all keywords in life science, and these are ordered in a tree structure. After this, we will refer to the lowest layer (leaf) entries of the tree structure by their *proper name* and all the upper layer (node) entries by their *class name*. We extracted names classifying proteins and enzymes from the XML format source shown in Fig. 2. The protein/enzyme names are defined between `<string>` and `</string>`. We obtained 3,246 (protein) class names and 1,745 (enzyme) class names; 9,831 (protein) proper names and 6,873 (enzyme) proper names were found in the lowest layer.

We extracted enzyme names, their respective syn-

[2]http://www.nlm.nih.gov/mesh/meshhome.html
[3]http://www.genome.jp/ligand/

TABLE III.
TOTAL NUMBER OF EXTRACTED PROTEIN AND GENE NAMES.

	Proteins	Genes
Total number	132,153	115,663

TABLE V.
TWENTY EXAMPLES OF *Escherichia coli* PROTEIN NAMES EXTRACTED FROM MEDLINE ABSTRACTS.

Protein names	Frequencies
MAP	89015
PHI	37697
PTH	37694
ThiS protein	26438
UK	21615
Lysozyme	20747
Reverse transcriptase	20656
CHD	19898
TF	19302
Beta galactosidase	17489
HSP70	13213
GPI	11971
PAD	11171
Glutathione S transferase	11100
DBP	10825
NOD	9146
Beta glucuronidase	8801
PDT	7547
Thymidine kinase	7394
Enolase	7001

onyms, and class names from the LIGAND database. We obtained 206 class names and 15,087 proper names. The results are listed in Table I. Table II lists the frequency of duplicated protein names extracted from the three databases and the total numbers of extracted protein and gene names are listed in Table III.

We unexpectedly found that the frequency of co-occurrence of protein names was low. We attributed this to (1) the different concepts on which the three databases were built on, and (2) trivial differences in the expressions of protein names as listed in Table IV.

We next estimated the frequencies of protein names obtained from Medline abstracts. Medline is a bibliographic database that lists 12 million biomedical texts and has been managed by the National Library of Medicine since the mid-1960s.

We investigated five typical problems that we considered may have had consequences regarding the application of information extraction to protein names and consequently to protein interactions in the biomedical literature.

1) Treatment of special characters
2) Treatment of homonyms
3) Cases where the protein-name string may be a substring of a different protein-name string
4) Cases where one protein exists in different organisms.
5) Treatment of modifiers

First, we detected problems with the pattern matching of protein names; e.g., “TNFalpha1R”, “TNF alpha-1 R”, “TNFalpha-1(R)”, and “TNF alpha1R”. (also see Ariadne, 2003). For example, “12alpha-hydroxysteroid” and “12-alpha-hydroxysteroid” represent the same protein in the first entry of Table IV; however, neither of them will produce a regular match of expressions.

TABLE VI.
TWENTY EXAMPLES OF *Bacillus subtilis* PROTEIN NAMES EXTRACTED FROM MEDLINE ABSTRACTS.

Protein names	Frequencies
MAP	89015
PHI	37697
PTH	37694
CAP	27910
H1	25214
UK	21615
PL	18340
Esterase	17619
TK	16678
BK	15394
SK	13227
HSP70	13213
ALS	12381
GPI	11971
PAD	11171
H3	11100
PDT	7547
Thymidine kinase	7394
Enolase	7001
G6PD	6208

Second, we noted that some abbreviations may occasionally have homonyms (i.e., words that are pronounced or spelled the same way but have different meanings, such as “map” and “Map”). We can find *Escherichia coli* and *Bacillus subtilis* protein-name frequencies in decreasing order in Medline abstracts in Tables V and VI by ignoring symbols; e.g., (,), /). Table VII lists examples of sentences extracted from the Medline abstracts, which include the protein names in Tables V and VI, and highlighted homonym expressions.

Third, a protein-name string may just be a substring of a different protein-name string. For example, the string corresponding to protein ”ACP” is a substring of the string corresponding to protein name ”ACP phosphodiesterace 1”. The expression ”ACP” will match both expressions ”ACP” and ”ACP phosphodiesterace 1” in pattern matching. Other examples are listed in Table VIII. Note that these cases occurred often when protein names were extracted.

Fourth, the same protein may exist in several different organisms. For example, 23OHBP oxygenase belongs to *Burkholderia cepacia* and *Pseudomonas paucimobilis*. This may present a problem when one would like to classify biological information regarding a specific organism. Some examples are given in Table IX.

Fifth, we examined words modifying protein names. Both “myosin proteins” and “myosin” were regarded as protein names in the phrase“... of the myosin proteins to ...”. This meant there was ambiguity about whether the “proteins” were part of the protein names. These words were not canonical words in the noun phrase containing the protein names. They were often found on the boundary of protein names. Some words occurred to the left of

TABLE IV.
EXAMPLES OF TRIVIAL DIFFERENCES IN PROTEIN NAMES.

EC nomenclature	Swiss-Prot	LIGAND
EC 1.1.1.176	12alpha-hydroxysteroid dehydrogenase	12-alpha-hydroxysteroid dehydrogenase
EC 3.5.99.7	Putative 1-aminocyclopropane -1-carboxylate deaminase	1-aminocyclopropane -1-carboxylate deaminase
EC 1.1.1.53	20beta-hydroxysteroid dehydrogenase	20-beta-hydroxysteroid dehydrogenase
EC 1.13.11.32	2-nitropropane dioxygenase precursor	2-nitropropane dioxygenase

TABLE VII.
EXAMPLES OF HOMONYMOUS EXTRACTED SENTENCES.

Protein names	Homonymous sentences
MAP	butanol parametric **map** was created from any subset
ThiS protein	a high homology of **this protein** to the third domain
PAD	the patellar tendon and within Hoffa's fat **pad** of the knee
MAT	This study used a pressure-sensing **mat** to investigate saddle fit
LAP	A specific history of **lap** belt injury, bicycle handlebar injury,
protein A	per mg total **protein. A** purification scheme has
BS A	orally either bosentan **(BS), a** nonselective ET receptor antagonist

TABLE VIII.
EXAMPLES OF PROTEIN NAME STRINGS INCLUDED AS SUBSTRINGS OF DIFFERENT PROTEIN NAMES.

Substrings	Protein names
50S ribosomal protein L1	50S ribosomal protein L10
	50S ribosomal protein L11
	50S ribosomal protein L13
	50S ribosomal protein L14
	50S ribosomal protein L15
	50S ribosomal protein L16
	50S ribosomal protein L17
	50S ribosomal protein L18
	50S ribosomal protein L19
ACP	ACP phosphodiesterase 1
	ACP phosphodiesterase 2
Adenylosuccinase	Adenylosuccinate lyase
	Adenylosuccinate synthetase
Aspartokinase I	Aspartokinase I alpha subunit
	Aspartokinase I beta subunit
BL1	BL10
	BL11
	BL15
	BL17

TABLE IX.
EXAMPLES OF PROTEINS BELONGING TO DIFFERENT ORGANISMS.

Protein names	Organism names
23OHBP oxygenase	*Burkholderia cepacia (Pseudomonas cepacia)*
	Pseudomonas paucimobilis (Sphingomonas paucimobilis)
	Pseudomonas pseudoalcaligenes
	Pseudomonas sp. (strain KKS102)
	Rhodococcus sp.
Carbonic anhydrase 1	*Flaveria linearis*
	Equus caballus (Horse)
	Homo sapiens (Human)
	Macaca mulatta (Rhesus macaque)
	Macaca nemestrina (Pig-tailed macaque)
ADP-ribosyl cyclase 1	*Mus musculus (Mouse)*
	Oryctolagus cuniculus (Rabbit)
	Rattus norvegicus (Rat)

protein names while others occurred to their right. We called these modifiers and examined their appearance in the Medline abstracts. We undertook protein name recognition in a distribution of the MEDLINE database, November 1, 2002 NLM MEDLINE DTD, by using SVM and the same features (i.e., bag-of-words (BoW), parts-of-speech (POS), orthographic, prefix, suffix, and preceding class features) mentioned in Section IV-B. There were a total of 6,229,339 abstracts. We used the GENIA corpus [1] as training data. As a result, a total of 10,819,286 protein names were recognized. We counted the frequency of each word that appeared as the word to the left or to the

TABLE X.
ONE-HUNDRED MOST FREQUENT CANDIDATES FOR THE LEFT MODIFIER WORD AND THEIR FREQUENCIES OF APPEARING ON THE RIGHT OF PROTEIN NAMES IN MEDLINE ABSTRACTS

Modifiers	Freqs.	Modifiers	Freqs.
the	2477927	interleukin-1	3637
protein	83600	coagulation	3556
monoclonal	81025	mature	3477
human	60544	identified	2845
serum	57371	homologous	2834
recombinant	48071	intermediate	2798
two	47530	pituitary	2758
major	35103	**early**	2679
plasma	31997	Src	2526
extracellular	27800	kinase	2437
cytoplasmic	25760	blood	2362
purified	23573	**eukaryotic**	2297
endogenous	21292	urinary	2286
nuclear	20372	Drosophila	2095
surface	20101	metabotropic	1796
N-terminal	19846	HIV-1	1779
class	19645	HIV	1286
C-terminal	16026	goat	1179
wild-type	15099	microbial	1140
viral	13437	chromosomal	1123
enzyme	13127	microsatellite	1116
bovine	12648	**adenovirus**	1096
platelet	12054	subunit	1093
rat	11924	antigen-specific	1068
active	11230	single-stranded	1049
mutant	10801	upstream	940
cytosolic	10464	non-specific	925
mitochondrial	10349	late	908
mouse	9780	3’	853
murine	9686	long	848
liver	9117	avian	844
mammalian	7944	equine	838
vascular	7246	salmon	820
lipoprotein	6379	tyrosine-phosphorylated	794
yeast	6206	clotting	790
glycoprotein	5876	tobacco	712
microsomal	5693	phosphodiesterase	712
secreted	5304	MAPK	631
truncated	5145	p21	610
carboxyl-terminal	4736	CSF	593
complete	4677	alpha-	593
bacterial	4669	basolateral	588
immunoglobulin	4617	90-kDa	562
intact	4362	CMV	559
muscle	4268	heterogeneous	557
phosphorylated	4203	30-kDa	543
gene	4187	varicella-zoster	529
chicken	4057	Xenopus	477
core	3754	type-specific	461
full-length	3646	43-kDa	452

TABLE XI.
ONE-HUNDRED MOST FREQUENT CANDIDATES FOR THE RIGHT MODIFIER WORD AND THEIR FREQUENCIES OF APPEARING ON THE RIGHT OF PROTEIN NAMES IN MEDLINE ABSTRACTS

Modifiers	Freqs.	Modifiers	Freqs.
in	491184	**component**	6444
protein	419821	superfamily	5806
proteins	282634	fragment	5549
receptors	230770	promoter	5296
antibodies	172016	mutations	5268
factors	160784	precursor	5244
factor	150189	**members**	5231
complex	128542	polypeptide	4730
antibody	118387	**ligand**	4590
domain	111918	fragments	4133
kinase	92671	**transcript**	3992
molecules	70015	**transcripts**	3991
alpha	64639	**families**	3887
gene	55142	**mutants**	3836
subunit	52018	fusion	3749
domains	46534	genes	3727
sequence	46287	peptide	3448
molecule	46116	repressor	3406
I	43892	segment	3149
enzyme	39515	transporter	3051
mRNA	39237	homologue	3040
complexes	38768	segments	2903
1	36343	**motif**	2839
region	33283	mutation	2730
family	33006	particles	2662
products	31243	alleles	2448
2	30518	endonuclease	2372
product	28791	cDNA	2106
chain	23994	isoenzymes	1981
regions	23799	**heterodimer**	1959
beta	23156	mRNAs	1868
subunits	20387	isoenzyme	1835
system	20362	pathways	1754
kinases	20117	cascade	1667
residues	18574	nuclease	1659
A	18142	combination	1649
3	16276	repeat	1643
protease	13233	MAPK	1598
site	11475	**heterodimers**	1590
isoforms	10782	cofactor	1489
glycoprotein	10780	**member**	1394
gamma	10365	delta	1373
4	10186	**variants**	1343
mutant	10136	isomer	1176
pathway	9888	message	1151
sites	9395	locus	1106
mAb	8044	core	1045
RNA	7678	s	954
sequences	7608	**motifs**	947
isoform	6700	repeats	923

right of protein names, and we also collected one word on the left and on the right boundary of protein names. In the next sample fragment,

> “`human I/PRO kappa/PRO B/PRO alpha/PRO protein`”,

protein names regarded by our system were tagged using `/PRO`. `I kappa B alpha` was recognized as a protein name. We regarded `human` and `protein` as candidates for modifiers. We also regarded `I` and `alpha` as candidates for modifiers. Table X lists the 100 most frequent candidates for left-hand modifiers. Table XI lists the 100 most frequent candidates for right-hand modifiers. The frequencies of appearance on the left or right of protein names in the Medline abstracts are also listed. The bold font in the tables means the modifiers appeared both in the BioCreAtIvE evaluation data and in the GENIA corpus. The modifiers played the following roles related to protein names.

- They contributed to detection of the boundary of noun phrases including protein names in the sentence. Some protein names were a part of a phrase. In the next example phrase:

  “... of the myosin proteins to ...”,

  the “myosin protein” is a noun phrase. “Myosin” is not a whole noun phrase. Complementing modifiers

contributes to parsing.

- Some modifiers restricted the meaning of protein names such as species, structures, and conditions (e.g., *human*, *rat*, *complex*, *subunit*, and *active*). For example, "AML1", "human AML1", and "mouse AML1" were found in the Medline abstract.
- Some modifiers change the gist of protein names to another meaning. For example, if *gene* occurs on the right of a protein name, the meaning tends to be a gene name. In the next sample phrase,

  "... activation of the IL-2 gene ...",

  "IL-2" is a protein name and "IL-2 gene" is a gene name. We are focusing on modifiers concerned with protein names for this reason.
- There were ambiguities in distinguishing the boundary of protein names. Ambiguous boundaries in evaluations with BioCreAtIvE are listed in `Correct.Data` to assess the gene or protein names. We reported in our previous study [18] that there were some ambiguous words with regard to determining the boundary of protein names in the GENIA corpus. Some ambiguous words are modifiers. For example, there is the case where "myosin" is considered to be a protein name and there is also the case where "myosin proteins" are considered to be protein names in the phrase"... of the myosin proteins to ...".

The problem where one protein exists in different organisms involves the process after protein recognition. We did not address this problem in this study. However, this is based on protein recognition and greatly influences it. As our study is useful for improving the accuracy of protein recognition, it would be useful for solving the problem where one protein exists in different organisms.

The remaining problems involve the difficulty of treating protein names. Since protein names are complicated and difficult to handle through dictionary-based or handmade-rule-based approaches, we should use machine-learning methods for recognizing proteins. We used machine-learning methods such as an SVM or a transductive SVM in this research.

However, a classifier trained in a specific domain can cause overfitting and yield very low accuracy in the recognition of protein names in different domains. We used a transductive SVM in this study to prevent the problem of overfitting. We expected the transductive SVM to prevent the problem of overfitting because it used test data in addition to training data in the learning process.

III. METHODS

A. SVM

Data consisting of two categories is classified by dividing space with a hyperplane with this method. When the two categories are positive and negative and the margin between the positive and negative examples in the training data is larger (see Fig. 3[4]), the possibility of incorrectly choosing categories in the open data is considered to be smaller. The hyperplane that maximizes the margin is determined, and classification is done using that hyperplane. Although the basics of this method are as described above, the inner region of the margin in the training data can include a small number of examples for extended versions of the method, and the linearity of the hyperplane can be changed to non-linear by using kernel functions. Classification with the extended method is equivalent to classification carried out using the following discernment function, and the two categories can be classified on the basis of whether the output value of the function is positive or negative [19]–[21]:

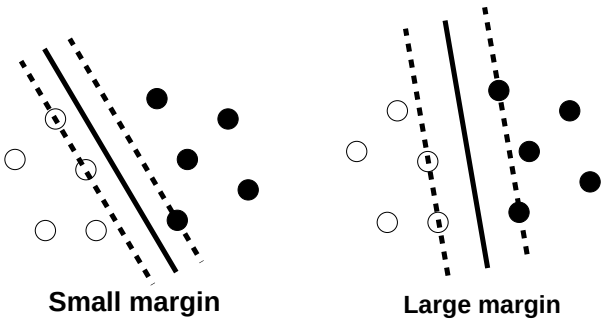

Figure 3. Maximizing margin

$$
\begin{aligned}
f(\mathbf{x}) &= sgn\left(\sum_{i=1}^{l}\alpha_i y_i K(\mathbf{x}_i,\mathbf{x})+b\right) \qquad (1)\\
b &= -\frac{max_{i,y_i=-1}b_i + min_{i,y_i=1}b_i}{2}\\
b_i &= \sum_{j=1}^{l}\alpha_j y_j K(\mathbf{x}_j,\mathbf{x}_i),
\end{aligned}
$$

where $\mathbf{x}$ is the context (a set of features) of an input example, $\mathbf{x}_i$ and y_i $(i = 1, ..., l, y_i \in \{1, -1\})$ indicate the context of the training data and its category, and function sgn is

$$
sgn(x) = \begin{array}{ll} 1 & (x \geq 0), \\ -1 & (otherwise). \end{array} \qquad (2)
$$

Each α_i $(i = 1, 2...)$ is fixed at the value of α_i when the value of $L(\alpha)$ in Eq. (3) is maximum under the conditions in Eqs. (4) and (5).

$$
L(\alpha) = \sum_{i=1}^{l}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{l}\alpha_i\alpha_j y_i y_j K(\mathbf{x_i},\mathbf{x_j}) \qquad (3)
$$

$$
0 \leq \alpha_i \leq C \ (i = 1, ..., l) \qquad (4)
$$

[4]The open circles in the figure indicate positive examples and the closed ones indicate negative ones. The solid line indicates the hyperplane that divides the space, and the broken lines indicate the planes at the boundaries of the margins.

$$\sum_{i=1}^{l} \alpha_i y_i = 0 \quad (5)$$

Although function K is called a kernel function and various types of kernel functions are used, we used the following polynomial function:

$$K(\mathbf{x}, \mathbf{y}) \quad = (\mathbf{x} \cdot \mathbf{y} + 1)^d. \quad (6)$$

C and d are constants set by experimentation, and in this paper C is fixed at 1 for all experiments. Three values are used for d: $d = 1$, $d = 2$, and $d = 3$. A set of $\mathbf{x}_i$ that satisfies $\alpha_i > 0$ is called a support vector, and the portion totaling the sum in Eq. (1) is calculated using only examples that are support vectors.

B. Transductive SVM

A transductive SVM [22] is a machine-learning method where induction and deduction processes are simultaneously conducted.

The transductive SVM algorithm can be outlined as follows.

1) A boundary hyperplane is built using labeled data.
2) Unlabeled data are classified according to that hyperplane.
3) A data pair near the boundary hyperplane is chosen in Step 2 from data determined to be a positive or a negative example.
4) The labels of the pair chosen in Step 3 are exchanged if this exchange does not reduce the margin.
5) If the conditions for ending the algorithm are fulfilled, the algorithm is terminated; otherwise, the algorithm returns to Step 2.

Because a transductive SVM uses unlabeled data in addition to labeled data, it produces good performance even when the amount of labeled data is not large. Some papers report good results from the use of a transductive SVM [22]–[24]. We expected a transductive SVM to prevent the problem of overfitting that we faced in this paper, because it uses test data (unlabeled data) in addition to training data (labeled data) in the learning process. Therefore, we used a transductive SVM in this study.

IV. Experiments

A. Tagged Corpus

We used two corpora in this study: the GENIA and a corpus on breast cancer.

The GENIA corpus of 2,000 abstracts collected from Medline was assembled to support the use of NLP technology to extract information from biomedical texts. Semantic annotations of biomedical domains are given in these abstracts, which have *human*, *blood cell*, and *transcription factor* as MeSH terms [25].

Technical terms in 36 kinds of biomedical fields defined by GENIA ontology are tagged in the GENIA corpus (e.g., DNA, RNA, and proteins). A class of proteins has subclasses — such as protein_complex and protein_family_or_group — and in the research reported here all subclasses of proteins were defined as protein names. We constructed the corpus on breast cancer by using 1,000 abstracts that were related to breast cancer and that included numerous protein names. Protein names in this corpus were tagged by a biology specialist we supervised. The definition of protein names follows that in the GENIA corpus.

TABLE XII.
FEATURES USED IN SVM LEARNING.

Features	Descriptions
BoW	Words appearing in training data
POS	POSs tagged by Brill tagger
Orthographic	(explained in Table XIII)
Prefix of words	uni-, bi-, and tri-gram of the word
Suffix of words	uni-, bi-, and tri-gram of the word

TABLE XIII.
DETAILS OF ORTHOGRAPHIC FEATURES.

Features	Examples	Features	Examples
DigitNumber	15	CloseBracket	]
Greek	alpha	Colon	:
SingleCap	M	SemiColon	;
CapsAndDigits	I2	Percent	%
TwoCaps	RalGDS	OpenParen	(
LettersAndDigits	p52	CloseParen	)
InitCaps	Interleukin	Comma	,
LowCaps	kappaB	FullStop	.
Lowercase	kinases	Determiner	the
Hyphen	-	Conjunction	and
Backslash	/	Other	* + #
OpenBracket	[		

B. Features, parameters, and evaluation methods

The features used in this numerical experiment are listed in Table XII. POSs were tagged by the Brill tagger [26]. These features are usually used for biomedical named entities [6] [27]. The BoW and POS features are used in convention NLP. The orthographic features are effective for extracting protein names because some names contain capital letters, Greek letters, or numbers, and the suffix feature is effective because some protein names have characteristic endings such as ∼ase or ∼in. The prefix feature is also effective for extraction although its effect is slight [27]. It is useful for named entity recognition in NLP, but cannot be used in a transductive SVM. There is an example of feature extraction in Fig. 4, where the phrase is written vertically in the "Word" column and the extracted features are listed in the other columns (e.g., the orthographic features are listed in the column labeled "Ortho". The class T or O for each word (details are explained below) is listed in the "Class" column. Each feature is separated by a space. The elements of the feature vectors for the current word (i.e., "NF-kappaB") are shown in the shaded area in Fig. 4. The information from the two preceding and two following tokens (words) is used for each vector.

The parameters used in our experiment are listed in Table XIV. Linear, polynomial, RBF, and sigmoid kernels

	Word	POS	Ortho	Prefix	Suffix	Class
Position -3	such	JJ	Lowercase	s su suc	h ch uch	O
Position -2	as	IN	Lowercase	a as --	s as --	O
Position -1	NF-kappa	NNP	Greek	N NF NF-	a pa ppa	T
Position 0	B	NNP	SingleCap	B -- --	B-- --	T
Position +1	that	IN	Lowercase	t th tha	t at hat	O
Position +2	are	VBP	Lowercase	a ar are	e re are	O
Position +3	constitutively	RB	Lowercase	c co con	y ly ely	O

Figure 4. Feature extraction example using sample fragment "*... such as NF-kappa B that are constitutively ...*."

TABLE XIV.
PARAMETERS USED IN SVM LEARNING.

Parameters	Descriptions
Kernel	Linear, polynomial with degree two and three
Context window	-2, -1, 0, 1, 2
Preceding class (only ordinary SVM)	-2, -1
Direction of parsing	Forward

have been proposed, and we compared the experimental results obtained using polynomial and linear kernels. The direction of parsing is forward. The IOB2 tag is often used in the named entity of NLP. IOB2 means that the B tag is attached to the first word of the protein name, the I tag is attached to the remaining words in the protein name, and the O tag is attached to words that are not parts of protein names. As adapting to multiclass labeling like IOB2 is difficult in a transductive SVM, we used a TO tag in the work reported here. The TO tag means that the T tag is attached to words in protein names and the O tag is attached to words not in protein names.

The evaluation criteria used in this experiment are listed in Table XV. Some protein names are compounds. If a protein name is one word, there are two evaluation values: correct or incorrect. Evaluation criteria taking partial matching into account were added for protein names consisting of more than one word. "Exact" is the strictest criterion, "Left" and "Right" are looser criteria, and "Sub" is the loosest criterion.

C. Intradomain accuracy of protein-name recognition

The experimental concept underlying a closed corpus is outlined in Fig. 5, and the efficiency of protein-name-recognition measures — precision, recall, and F-score — for the GENIA corpus are listed in Table XVI. Ten-fold cross validation (10CV) was used to evaluate these numerical results. The data were divided into ten parts, nine of which were used as training data and one of which was used as test data. Ten kinds of training and test data were prepared. The SVM learned the ten parts of training data and classified the ten parts of test data. The results listed in Table XVI are the average values for the ten SVM classifications. A polynomial kernel of degree two was used in these experiments, and the other parameters used are listed in Table XIV. These parameters were based

TABLE XV.
EVALUATION CRITERIA IN THIS EXPERIMENT.

Criteria	Descriptions
Exact	SVM predictions are correct if the left and right boundaries of the answer and the SVM predictions are the same.
Left	SVM predictions are correct if the left-hand side of the boundary of the answer and the SVM prediction are the same.
Right	SVM predictions are correct if the right-hand side of the boundary of the answer and the SVM prediction are the same.
Sub	SVM predictions are correct if part of the SVM prediction agrees with the answer.

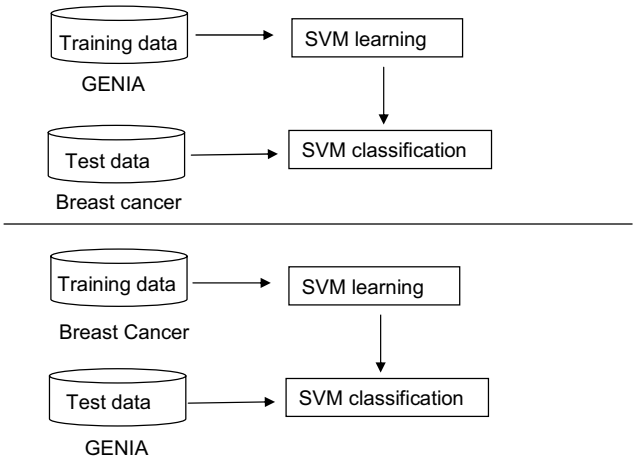

Figure 5. Intradomain experimental concept.

on the best set described in [27]. The accuracy measures for protein-name recognition in the breast-cancer corpus are listed in Table XVII. These values are also average values for 10CV.

Precision is the fraction of all positive predictions that are true positives:

$$Precision = \frac{tp}{tp + fp}. \tag{7}$$

Recall is the fraction of positive examples correctly predicted as positive:

$$Recall = \frac{tp}{tp + fn}. \tag{8}$$

The F-score is a harmonic average of precision and recall:

$$\textit{F-score} = \frac{2.0 \times Precision \times Recall}{Precision + Recall}. \tag{9}$$

The terms tp, fp, and fn are the numbers of true positives, false positives, and false negatives.

Precision was higher than recall in these experiments, and the accuracy of protein-name recognition in the breast-cancer corpus was better than that in the GENIA corpus. We think these differences in protein-name recognition are the result of there being fewer variations in protein names that appeared in the breast-cancer corpus than in those that appeared in the GENIA corpus.

TABLE XVI.
CONVENTIONAL SVM ACCURACY IN GENIA CORPUS (AVERAGE VALUES FOR 10-FOLD CROSS VALIDATION (10CV)).

Criteria	Precision	Recall	F-scores
Exact	79.08	72.57	75.68
Left	83.96	77.05	80.34
Right	85.89	78.82	82.19
Sub	89.42	83.41	86.30

TABLE XVII.
CONVENTIONAL SVM ACCURACY IN BREAST-CANCER CORPUS (AVERAGE VALUES FOR 10CV).

	Precision	Recall	F-scores
Exact	85.54	80.18	82.74
Left	88.85	83.29	85.95
Right	88.08	82.57	85.21
Sub	90.90	85.53	88.10

D. Cross-domain accuracy of protein-name recognition

We wanted to see whether a classifier trained in a specific domain could be used for recognizing names in other domains. The experimental concept underlying recognition of cross-domain names is outlined in Fig. 6.

E. Protein-name recognition accuracy of a conventional SVM

The protein-name-recognition accuracy measures obtained when the SVM learned the entire GENIA corpus and classified the entire breast-cancer corpus are listed in Table XVIII (see the upper diagram in Fig. 5), and the protein-name recognition accuracy measures obtained when the SVM learned the entire breast-cancer corpus and classified the entire GENIA corpus are listed in Table XIX (see the lower diagram in Fig. 5). We used three types of kernels in these experiments: linear, a polynomial of degree two, and a polynomial of degree three. The accuracies of recognition measures listed in these tables are significantly lower than those listed in Tables XVI and XVII.

F. Protein-name recognition accuracy of a transductive SVM

The recognition of protein names obtained when the transductive SVM learned the entire GENIA corpus and classified the entire breast-cancer corpus is indicated by the values listed in Table XX (see the upper diagram in Fig. 5), and the recognition of protein names obtained when the transductive SVM learned the entire breast-cancer corpus and classified the entire GENIA corpus is indicated by the values listed in Table XXI (see the lower diagram in Fig. 5). We also used three types of kernels in these experiments: linear, a polynomial of degree two, and a polynomial of degree three. The accuracy of protein-name recognition was better than that obtained when the conventional SVM was used (see Tables XVIII and XIX). When the conventional SVM was used the precision was higher than the recall, whereas when the transductive SVM was used the precision and recall were about the same or the recall was higher. The F-scores obtained when using the transductive SVM were better than those when using the conventional SVM.

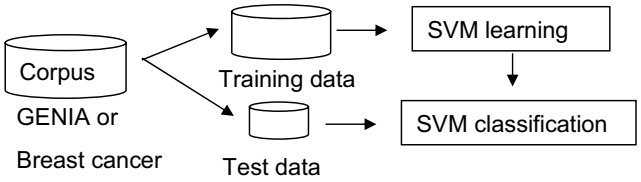

Figure 6. Cross-domain experimental concept.

TABLE XVIII.
ACCURACY OF PROTEIN-NAME RECOGNITION OF CONVENTIONAL SVM WHEN THE GENIA CORPUS IS USED AS TRAINING DATA AND THE BREAST-CANCER CORPUS IS USED AS TEST DATA (AVERAGE VALUES FOR 10CV).

Kernels	Evaluation	Precision	Recall	F-scores
Linear	Exact	66.30	53.26	59.07
	Left	78.63	63.16	70.05
	Right	71.38	57.33	63.59
	Sub	82.98	68.86	**75.26**
Polynomial (degree 2)	Exact	69.08	53.80	**60.49**
	Left	80.50	62.69	**70.49**
	Right	73.54	57.27	**64.39**
	Sub	84.78	67.12	74.92
Polynomial (degree 3)	Exact	68.87	47.10	55.94
	Left	80.63	55.14	65.49
	Right	73.09	49.98	59.36
	Sub	84.87	58.70	69.40

V. DISCUSSION

We evaluated the flexibility of classifiers that learned to recognize protein names in a specific domain. We used two corpora: the GENIA corpus and a corpus on breast cancer in this study. In the 10CV used to evaluate classifier accuracy in both corpora, the F-score obtained using the "Exact" criterion was 75.68 in the GENIA corpus and 82.74 in the breast-cancer corpus.

When we evaluated the recognition of protein names of the classifier on the breast-cancer corpus, trained on the entire GENIA corpus, the F-score obtained using the "Exact" criterion decreased to 60.49. When we evaluated the recognition of protein names of the classifier on the GENIA corpus, trained on the entire breast-cancer corpus, the F-score obtained using the "Exact" criterion decreased to 58.06. These values were obtained using the polynomial kernel with degree two. These results indicate that a classifier trained in a specific domain does not recognize protein names as well in other domains.

We used the transductive SVM to improve the recognition of the classifier in a different domain. When we evaluated the recognition of protein names of a classifier on the breast cancer corpus, trained on the entire GENIA corpus, the F-score obtained using the "Exact" criterion improved from 60.49 to 64.43.

TABLE XIX.
CONVENTIONAL SVM ACCURACY OF PROTEIN-NAME RECOGNITION WHEN THE BREAST-CANCER CORPUS WAS USED AS TRAINING DATA AND THE GENIA CORPUS WAS USED AS TEST DATA (AVERAGE VALUES FOR 10CV).

Kernels	Evaluations	Precision	Recall	F-scores
Linear	Exact	62.11	53.47	57.47
	Left	70.46	60.65	65.19
	Right	66.45	57.20	61.48
	Sub	76.43	65.65	**70.63**
Polynomial (degree 2)	Exact	64.03	53.12	**58.06**
	Left	72.05	59.77	**65.34**
	Right	67.93	56.35	**61.60**
	Sub	77.67	64.13	70.25
Polynomial (degree 3)	Exact	63.66	46.88	54.00
	Left	72.39	53.31	61.40
	Right	67.14	49.44	56.95
	Sub	77.76	56.99	65.77

TABLE XX.
ACCURACY OF PROTEIN-NAME RECOGNITION OF A TRANSDUCTIVE SVM WHEN THE GENIA CORPUS WAS USED AS TRAINING DATA AND THE BREAST-CANCER CORPUS WAS USED AS TEST DATA.

Kernels	Evaluations	Precision	Recall	F-scores
Linear	Exact	60.65	60.20	60.42
	Left	72.50	71.95	72.22
	Right	65.34	64.86	65.10
	Sub	76.86	78.34	77.59
Polynomial (degree 2)	Exact	56.32	75.26	**64.43**
	Left	64.09	85.64	**73.31**
	Right	61.80	82.57	**70.69**
	Sub	69.11	93.96	**79.64**
Polynomial (degree 3)	Exact	55.54	75.04	63.83
	Left	63.11	85.28	72.53
	Right	60.67	81.98	69.73
	Sub	68.50	93.24	78.98

When protein name were recognized on the GENIA corpus using the classifier trained on the entire breast-cancer corpus, the recognition obtained using the "Exact" criterion was almost the same regardless of whether the conventional or the transductive SVM was used: the F-score was 58.06 with the conventional SVM and 57.22 with the transductive SVM. The advantages of using the transductive SVM were confirmed, however, when the other criteria were used. The F-score improved from 65.34 to 66.63 using the "Left" criterion, from 61.60 to 62.38 using the "Right" criterion, and from 70.25 to 74.25 using the "Sub" criterion. The boundary of a protein name varies according to the annotator [18]. We also discussed ambiguous boundaries for protein names related to their modifiers in Section II. For example, the following variations on the protein name "human NF-kappa B protein" can be considered: "NF-kappa B," "human NF-kappa B," "NF-kappa B protein," and "human NK-kappa B protein." Each variation has almost the same meaning. When ambiguous boundaries for protein names are taken into consideration, the transductive SVM effect that appears when using the criteria of "Left", "Right", or "Sub" is meaningful.

TABLE XXI.
ACCURACY OF PROTEIN-NAME RECOGNITION OF A TRANSDUCTIVE SVM WHEN THE BREAST-CANCER CORPUS WAS USED AS TRAINING DATA AND THE GENIA CORPUS WAS USED AS TEST DATA.

Kernels	Evaluations	Precision	Recall	F-scores
Linear	Exact	54.74	58.86	56.73
	Left	63.98	68.80	66.31
	Right	60.25	64.78	**62.43**
	Sub	72.63	76.71	**74.61**
Polynomial (degree 2)	Exact	56.97	57.47	**57.22**
	Left	66.34	66.92	**66.63**
	Right	62.11	62.66	62.38
	Sub	74.62	73.89	74.25
Polynomial (degree 3)	Exact	54.80	55.42	55.11
	Left	64.94	65.68	65.31
	Right	59.59	60.26	59.93
	Sub	72.85	72.27	72.56

Precision was higher than recall when the conventional SVM was used, and precision and recall were almost the same (or recall was higher) when the transductive SVM was used. We have demonstrated that a classifier trained on a specific domain causes overfitting and that this can be avoided by using the transductive SVM.

We demonstrated that the transductive SVM increased the flexibility of a classifier trained on a specific domain in this research. Using the transductive SVM rather than the conventional SVM, however, yielded lower precision, higher recall, and higher F-scores. If precision is considered to be important, the conventional SVM should be used. New proteins continue to be discovered, however, and the transductive SVM is more suitable for extracting new protein names from the literature when high recall and high F-score are required.

VI. CONCLUSION

We first analyzed protein names using various dictionaries and databases and found five problems with protein names; i.e., the treatment of special characters, the treatment of homonyms, cases where the protein-name string may be a substring of a different protein-name string, cases where one protein exists in different organisms, and the treatment of modifiers. We found that we should use a machine-learning approach to solve these problems in protein-name recognition. Machine-learning methods have recently been used in research on protein-name recognition. However, a classifier trained in a specific domain can cause overfitting and be so inflexible that it is useful in only that domain. We therefore developed a new corpus on breast cancer and investigated the flexibility of a classifier trained on the GENIA [1] corpus or the breast-cancer corpus. The conventional SVM in our experiments caused overfitting and obtained low accuracies when trained in a different domain, as we had expected. We used a transductive SVM to avoid overfitting, and we evaluated what effect transductive learning had. We confirmed through two experiments that the transductive SVM prevented overfitting and yielded higher accuracies than the conventional SVM did. For the criterion "Sub",

the transductive SVM increased the F-scores (70.46 to 79.64 and 70.63 to 74.61) in both experiments.

We would like to use our method on different corpora or diverse domain abstracts in the future. We would also like to use our method to solve different problems in applications other than protein-name recognition as there are also overfitting problems in other information processing areas. We hope to solve these using transductive SVMs.

REFERENCES

[1] T. Ohta, Y. Tateisi, and J.-D. Kim, "The genia corpus: an annotated research abstract corpus in molecular biology domain," *In Proceedings of he Human Language Technology Conference (HLT 2002)*, 2002.

[2] T. Ono, H. Hishigaki, A. Tanigami, and T. Takagi, "Automatic extraction of information on protein-protein interactions from biomedical literature," *Bioinformatics*, vol. 17, no. 2, pp. 155–161, 2001.

[3] J. Thomas, D. Milward, C. Ouzounis, S. Pulman, and M. Carroll, "Automatic extraction of protein interactions from scientific abstracts," *Proceedings of the Pacific Symposium on Biocomputing*, vol. 5, pp. 538–549, 2000.

[4] C. Blaschke and A. Valencia, "Can bibliographic pointers for known biological data be found automatically? Protein interactions as a case study," *Comparative and Functional Genomics*, vol. 2, pp. 196–206, 2001.

[5] K. Fukuda, A. Tamura, T. Tsunoda, and T. Takagi, "Toward information extraction: Identifying protein names from biological papers," *Proceedings of the Pacific Symposium on Biocomputing*, pp. 705–716, 1999.

[6] N. Collier, C. Nobata, and J. Tsujii, "Extracting the names of genes and gene production with a hidden Markov model," *Proceedings of the 18th International Conference on Computational Linguistics (COLING'2000)*, pp. 201–207, 2000.

[7] K. Takeuchi and N. Collier, "Bio-medical entity extraction using support vector machine," *Proceedings of the ACL 2003 Workshop on NLP in Biomedicine*, pp. 57–64, 2003.

[8] L. Tanabe and W. J. Wilbur, "Tagging gene and protein names in biomedical text," *Bioinformatics*, vol. 18, no. 8, pp. 1124–1132, 2002.

[9] K. Franzén, G. Eriksson, F. O. L. Asker, P. Lidén, and J. Cöster, "Protein names and how to find them," *International Journal of Medical Informatics*, vol. 67, pp. 49–61, 2002.

[10] B. Boeckmann, A. Bairoch, R. Apweiler, M.-C. Blatter, A. Estreicher, E. Gasteiger, M. J. Martin, K. Michoud, C. O'Donovan, I. Phan, S. Pilbout, and M. Schneider, "The SWISS-PROT protein knowledgebase and its supplement TeEMBL in 2003," *Nucleic Acids Research*, vol. 31, no. 1, pp. 365–370, 2003.

[11] Y. Tsuruoka and J. Tsujii, "Boosting precision and recall of dictionary-based protein name recognition," *Proceedings of the ACL 2003 Workshop on NLP in Biomedicine*, pp. 41–48, 2003.

[12] J. Kazama, T. Makino, Y. Ohta, and J. Tsujii, "Tuning support vector machines for biomedical named entity recognition," *In proceedings of the Natural Language Processing in the Biomedical Domain (ACL2002)*, pp. 1–8, 2002.

[13] K. Yamamoto, T. Kudo, A. Konagaya, and Y. Matsumoto, "Protein name tagging for biomedical annotation in text," *Proceedings of the ACL 2003 Workshop on NLP in Biomedicine*, pp. 65–72, 2003.

[14] P. Chen and H. Al-Mubaid, "Context-based term disambiguation in biomedical literature," *In proceedings of the 19th International FLAIRS Conference*, 2006.

[15] J. Baldridge and M. Osborne, "Active Learning and the Total Cost of Annotation," *Proceedings of Empirical Methods for Natural Language Processing*, pp. 9–16, 2004.

[16] T. Masuyama and H. Nakagawa, "Two step POS selection for SVM based text categorization," *IEICE Transactions on Information and Systems*, vol. E87–D, no. 2, pp. 373–379, 2004.

[17] J. Drish, "Obtaining calibrated probability estimates from Support Vector Machines," *Final Project for CSE 254: Seminar on Learning Algorithms*, 2001.

[18] T. Mitsumori, S. Fation, M. Murata, K. Doi, and H. Doi, "Boundary correction of protein names adapting heuristic rules," *Proceedings of Computational Linguistics and Intelligent Text Processing (CICLing 2004)*, pp. 172–175, 2004.

[19] V. N. Vapnik, *The Nature of Statistical Learning Theory*. Springer., 1995.

[20] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, 2000.

[21] T. Kudoh, "TinySVM: Support Vector Machines," http://cl.aist-nara.ac.jp/ taku-ku// software/TinySVM/ index.html, 2000.

[22] T. Joachims, "Transductive inference for text classification using support vector machines," *International Conference on Machine Learning (ICML)*, 1999.

[23] K. Shimada, K. Hayashi, and T. Endo, "Product specification extraction using SVM and transductive SVM," *Journal of Natural Language Processing*, vol. 12, no. 3, pp. 43–66, 2005, (in Japanese).

[24] K. Duh and K. Kirchhoff, "Lexicon acquisition for dialectal Arabic using transductive learning," *Proceedings of Empirical Methods in Natural Language Processing*, pp. 399–407, 2006.

[25] National Library of Medicine, "Medical subject headings, mesh," http://www.nlm.nih.gov/mesh/, 2003.

[26] E. Brill, "Some advances in transformation based part of speech tagging," *National Conference on Artificial Intelligence AAAI Press*, pp. 722–727, 1994.

[27] T. Mitsumori, S. Fation, M. Murata, K. Doi, and H. Doi, "Gene/protein name recognition based on support vector machine using dictionary as features," *BMC Bioinformatics*, vol. 5, no. Suppl 1, p. S8, 2005.

Masaki Murata received his Bachelor's, Master's, and Doctorate degrees in engineering from Kyoto University in 1993, 1995, and 1997.

He is currently a senior researcher at the National Institute of Information and Communications Technology, Japan, which is an independent administrative institution. His research interests include natural language processing, machine translation, information retrieval, and question answering.

He is a member of the Information Processing Society of Japan, the Japanese Society for Artificial Intelligence, the Institute of Electronics, Information, and Communication Engineers, the Mathematical Linguistic Society of Japan, and the Association for Computational Linguistics.

Tomohiro Mitsumori received his Bachelor's, Master's, and Doctorate degrees in science from Saga University in 1992, 1994, and 1997.

He is currently a member at Miyazono Patent Office, Japan. His research interests include natural language processing, machine learning and information retrieval.

Kouichi Doi Kouichi Doi received his Bachelor's, Master's, and Doctorate degrees in engineering from The University of Tokyo in 1985, 1988, and 1991. His research interests include natural language processing, software engineering, cognitive science, and bioinformatics.

He is currently a technical adviser at Pharma Security Consulting Inc., Japan.

He is a member of the Information Processing Society of Japan, the Japanese Society for Artificial Intelligence, and the Japanese Cognitive Science Society.