IPSJ Transactions on Bioinformatics Vol. 3 82–90 (Oct. 2010)

Original Paper

A Method to Extract Sentences Containing Protein Function Information with Training Data Extension Based on User's Feedback

Kazunori Miyanishi,[†1] Tomonobu Ozaki[†2]
and Takenao Ohkawa[†3]

A protein expresses various functions by interacting with chemical compounds. Protein function is clarified by protein structure analysis and the obtained knowledge has been stated in a number of documents. Extracting the function information and constructing the database are useful for various application fields such as drug discovery, understanding of life phenomenon, and so on. However, it is impractical to extract the function information manually from a number of documents for constructing the database, which strongly provide motivation to study automatic extraction of the function information. Extraction of protein function information is considered as a classification problem, namely, whether each sentence from the target document includes the function information or not is determined. Typically, in the case of addressing such a classification problem, a classifier is learned using the training data previously given. However, the accuracy is not high when the training data is not large enough. In such a case, we attempt to improve the accuracy of classification by extending the training data. Effective sentences for getting high accuracy are selected from the reference data aside from the training data set, and added to the training data. In order to select such effective sentences, we introduce the reliability of temporary labels assigned to sentences in the reference data. Sentences with low reliability temporary labels are presented to users, assigned true labels as users' feedback, and added to the training data. Additionally, a classifier is learned by the training data with sentences with high reliability temporary labels. By iterating this process, we attempt to improve the accuracy steadily. In the experiment, compared with the related approach, the accuracy is higher when the iteration steps of feedbacks and the number of sentences returned by users' feedback are small. Thus, it is confirmed that the training data is appropriately extended based on users' feedback by the proposed method. In addition, this result serves a purpose of reducing users' load.

1. Introduction

A protein expresses various functions by interacting with other chemical compounds, and plays important roles in organism activity[1]. Protein function is clarified by protein structure analysis and the obtained knowledge has been stated in a number of documents. In order to make the knowledge available readily, it is required to make a database of protein function information. Many protein-related databases have been developed [e.g., PIR (Protein Information Resource)[2], PDB (Protein Data Bank)[3], and Swiss-Prot (Swiss Protein Database)[4]]. However, the useful information that has not been registered in such databases is still contained in huge volumes of documents.

Recently, there have been many attempts to extract significant information from biomedical documents. For example, Tsai, et al.[5] and Sun, et al.[6] proposed an approach to biomedical named entities recognition using Conditional Random Fields (CRF)[7] based on orthographical features or words conjunctions. On the other hand, researches to extract protein interactions information from biomedical documents have been also conducted. Bunescu, et al.[8] attempted to identify human protein names and extract protein interactions using various information extraction methods [for example, dictionary-based extraction, Rapier and BWI (a rule learning algorithm), Hidden Markov Models (HMMs)[9], Support Vector Machine (SVM)[10], and existing protein name identification systems (KEX[11] and ABGene[12])]. Cooper, et al.[13] proposed a method for the dictionary of protein-protein interactions using a combination of linguistic information (e.g., verbs used to describe protein interactions) and graphical relations between proteins. Hao, et al.[14] proposed an approach to discover English expression patterns, optimizing them and extracting protein-protein interactions using them.

While these researches aim to extract biomedical named entities or protein interactions, we have focused on documents about protein structure analyses, and proposed a framework which assists users in extracting protein function information interactively. In our scheme, a concept of extracting sentences containing the protein function information by iterative learning[15] has been introduced. Extraction of the sentences can be considered as to classify sentences based on whether they contain the function information or not. The SVM is used as a

[†1] Graduate School of Science and Technology, Kobe University
[†2] Cybermedia Center, Osaka University
[†3] Graduate School of System Informatics, Kobe University

classifier, where one sentence corresponds to one instance, and characteristics (keywords, patterns, etc.) of each sentence corresponds to the features of the sentence. In this approach, however, if not enough training data set is given, the accuracy of classification is decreased extremely. Additionally, because it is necessary for experts in the related fields to read documents, and determine whether each sentence should be extracted, it is troublesome and almost impractical to create large training data. To address this problem, in Ref. 15), all classified sentences are assigned correct labels by users' feedback to augment the training data. On the other hand, in this paper, only effective sentences for improving the accuracy are selected and presented to users in order to reduce users' load. Concretely, besides a small amount of the training data, non-labeled available data (called "reference data") is used. While each sentence in the training data is assigned a label indicating whether the sentence contains the protein function information ('positive') or not ('negative'), one in the reference data is not assigned a label. A label of a sentence in the reference data can be predicted based on the training data. Such a predicted label is called a "temporary label". On the other hand, a label assigned to a sentence in the training data is called a "true label". A label assigned by feedback is also a true label, because it is assumed that users' feedback is correct. However, temporary labels may be mistaken, and if all sentences with temporary labels are added to the training data, the accuracy of classification is affected negatively. Then, in this paper, we attempt to improve the accuracy using two ways of extending the training data. The first is the extension method using users' feedback. In the second extension, the reliability of each temporary label is introduced, and sentences with reliable temporary labels only are added to the training data. In the first method, sentences with true labels obtained by users' feedback are added to the training data because the feedback is very reliable. However, taking into account users' cost, it cannot be expected to get a lot of feedback. Therefore, measure of the reliability of temporary labels is introduced, and the training data is extended using sentences with high reliability temporary labels, because the training data cannot be extended drastically by only users' feedback.

In the fields of machine learning, there is the approach called active learning, that the learner actively selects the examples to be labeled. The proposed method is considered as a kind of active learning with pool-based sampling[16)–18)], which selects most informative examples from a large pool of unlabeled data. The proposed method is also considered as a variety of semi-supervised learning in terms of utilizing unlabel data. There are several studies about active and semi-supervised learning. In terms of the way of selecting examples for presenting to users, the methods for active and semi-supervised learning are divided into two categories, committee-based methods[16),19),20)] and certainty-based methods[21)–24)]. In committee-based methods, the multiple learners are learned based on the labeled data, and the unlabeled examples whose labeling results are inconsistent each other when the learners are applied to them are presented to users. In certainty-based methods, the learner labels each example in the unlabeled data with a degree of certainty. The examples with the lowest degrees of certainty are presented to users. We employ the certainty-based method in the proposed method. Various kinds of criteria to select examples from unlabeled data have been proposed. Zhu, et al.[21)] selected examples to minimize the estimated classification errors over unlabeled data. Tur, et al.[22)] used probability estimates based on the logistic function as confidence scores. Yu, et al.[23)] focus on active learning in terms of experimental design, and select examples to minimize the predictive variance of the target data. For sequence labeling tasks, Tomanek and Hahn[24)] employed the utility function based on the marginal and conditional probability of the most likely label sequence as the degree of certainty. In the proposed method, the reliability based on the distance between examples on the feature space is used as the criterion. More informative examples are selected by relearning the distance metric using feedbacks, and it is aimed to improve the classification accuracy by iterative learning. In the proposed method, sentences that it is hard to correctly classify until the previous step are preferentially selected, and presented to users. Therefore, it is considered that the accuracy becomes high with fewer feedbacks that is with less users' load.

The aim of the proposed method is to reduce users' load. In other words, it is required to achieve a high accuracy with fewer feedbacks. Therefore, we evaluate the effectiveness of the proposed method in the situation where both the iteration steps of feedbacks and the number of sentences returned by the users' feedback are restricted to the small values.

2. Extracting Protein Function Information as Classification

In this paper, extracting protein function information is treated as classifying whether each sentence in the target documents contains such information. That is to say, the classifier is learned based on the training data, is applied to new documents, and distinguishes whether each sentence in the documents contains the protein function information. The classifier is learned using features of sentences as follows,

(1) atomic distance between interacting substances
If names of residues or atoms are written in one sentence, their physical characteristics on three-dimensional structures can be a clue to determine whether the sentence contains protein function information. Concretely, when a residue interacts a substance, an atom or a part of atoms in the residue gets close to the substance. Therefore, if the distance in three dimensions between a residue and a substance written in a sentence is shorter than a certain pre-defined threshold, it is considered that the residue interacts the substance and "1" is assigned, otherwise "0" as one feature of the sentence.

(2) keywords
Frequently occurring words in sentences containing protein function information are significant as a hint for classification. Thus, if each of the keywords, for example "interact", "bind", "hydrogen bond", and so on, is included in a sentence, "1" is assigned to the sentence, otherwise "0" is assigned.

(3) patterns
Frequently occurring sequences of words in sentences containing protein function information are also the hint for classification. These sequences are defined as patterns with wild-card characters, and used as features. For example, "<residue> (.)* play (.)* <function>", "<protein> (.)* contain(.)* <residue>", where "<residue>" means the name of residue, for example "Arg21" and "His23". Similarly, "<function>" and "<protein>" mean respectively the name of function and protein. If a sentence matches each of the patterns, "1" is assigned to the sentence, otherwise "0" is assigned.

3. Extension of the Training Data

3.1 The Outline of Extending the Training Data

If enough instances as training data are not given in advance, an accurate classifier cannot be built based on only the training data. Then, we propose a method of extending training data based on the user's feedback. The outline of the proposed method is shown in **Fig. 1**. In the initial stage, the classifier is learned based on only existing training data, and applied to the reference data whose sentences are not labeled. As a result, a temporary label is assigned to each sentence in the reference data. Here, we introduce the reliability of each temporary label that is the degree of convincing that the temporary label is true, and its value of each sentence in the documents is calculated. Sentences with high reliability are added to the training data without any conditions (Extension I). Sentences with low reliability are presented to users in order to check whether each of sentences contains protein function information, true labels of the sentences are returned as feedback, and the sentences with the true labels are added to the training data (Extension Ⅱ). In the iterative process, the selection of the previous step is discarded, and sentences are afresh selected based on the newly calculated reliability. In this framework, it is necessary to present effective sentences to improve the accuracy of classification. If effective sentences are properly selected, the accuracy increases with a few steps of feedback, which contributes to reducing users' load.

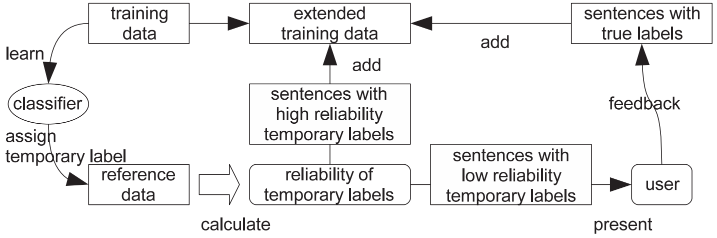

training data
extended training data
sentences with true labels
learn
add
add
classifier
sentences with high reliability temporary labels
feedback
assign temporary label
reference data
reliability of temporary labels
sentences with low reliability temporary labels
user
calculate
present

Fig. 1 The outline of the proposed method.

3.2 Extension of the Training Data

3.2.1 The Reliability of the Temporary Label

It is conceivable that instances that the same label is assigned to have similar features, therefore they are close to each other in the feature space described in Section 2. Thus, it seems highly possible that the temporary label, which is assigned to one sentence (called a target sentence), is correct if many sentences with true and the same labels (as the target sentence) are located around the target sentence in the feature space. In contrast, it is possible that a temporary label of a sentence is false if many sentences with true and the different labels are located around the target sentence. Therefore, the reliability of the temporary label is calculated focusing on the distribution of sentences with true labels around the target sentence with the temporary label.

From the distribution of sentences with true labels in the feature space, the distance measure between sentences is redefined as sentences with the same true label come close each other by reconstructing the space. In the field of semi-supervised clustering[25], there is an approach "Distance Metric Learning"[26),27], which attempts to obtain better clusters by learning distance measure between instances using the restriction among data. In the approach by Xing, et al.[26], the data set over the input space $\mathbb{R}^n$ is expressed as $\{x_i\}_{i=1}^m$. A distance metric $d(x, y)$ between data x and y is defined as follows,

$$d(x, y) = \|x - y\|_A = \sqrt{(x - y)^T A(x - y)}, \tag{1}$$

where A is the weighting diagonal matrix. The restrictions S and D between data x_i and x_j are given:

$S : (x_i, x_j) \in S$ *if* x_i *and* x_j *are similar*

$D : (x_i, x_j) \in D$ *if* x_i *and* x_j *are dissimilar*

Under these restrictions, A in (1) is calculated by solving the optimization problem described below.

$$\min_A \sum_{(x_i,x_j)\in S} \|x_i - x_j\|_A^2$$
$$s.t. \sum_{(x_i,x_j)\in D} \|x_i - x_j\|_A \geq 1,\ A \succeq O.$$

In this paper, each sentence is considered as one instance, labels of sentences are considered as the restriction among data, and the distance between sentences is redefined by Eq. (1). By redefining the distance, sentences with the same labels come close to each other, and sentences with different labels separate from each other. Therefore, from the distribution of sentences with true labels around a sentence with a temporary label in the reconstructed space, the degree that the temporary label is correct (reliability) can be calculated more properly.

In the reconstructed space, the reliability of a temporary label is calculated using distances to sentences with true labels in the following manner. It is consider that examples which are close to each other tends to be assigned the same labels. About a sentence with a temporary label, it is highly possible that the temporary label is correct if there are many sentences with true labels whose values are equal to the temporary label adjacent to the target sentence. On the other hand, if there are many sentences with labels whose values are different from the temporary label, it is highly possible that the temporary label is mistaken. From this viewpoint, the reliability of the temporary label based on the distances from sentences with true labels whose values are equal to the temporary label is defined as f (similarity reliability). The reliability based on the distances from sentences whose label values are different from the temporary label is defined as g (dissimilarity reliability).

Definition 1 (the similarity reliability and dissimilarity reliability)
Let s_x be a sentence with a temporary label l_x, and $s_1,\ s_2, \ldots, s_n$ be sentences with true labels $l_1,\ l_2, \ldots, l_n$. The distance between s_x and s_i is denoted by $d_i(s_x, s_i)$. $f(s_x)$, the reliability based on sentences with the true label which is equal to the temporary label (the similarity reliability), and $g(s_x)$, the reliability based on ones whose label is not equal to the temporary label (the dissimilarity reliability) are defined as follows:

$$f(s_x) = \sum_i \frac{1}{d_{(s_x,s_i)}}\ (l_x = l_i) \tag{2}$$

$$g(s_x) = \sum_i \frac{1}{d_{(s_x,s_i)}}\ (l_x \neq l_i) \tag{3}$$

■

The definitive reliability r which is the criterion for selecting sentences is defined based on similarity and dissimilarity reliability as follows.

Definition 2 (the definitive reliability)
The *definitive reliability* $r(s_x)$ is defined as follows:

$$r(s_x) = f(s_x) - g(s_x) \tag{4}$$

■

The procedure to extend the training data by selecting sentences with temporary labels using the definitive reliability is described in 3.2.2.

3.2.2 The Procedure of Training Data Extension

Based on the definitive reliability, the procedure to extend the training data by selecting sentences in the reference data is shown in **Fig. 2**, where T_t and T_b are thresholds for Extension I and Ⅱ respectively. T_t is the threshold of the number of sentences with high reliability temporary labels that should be added to the training data. T_b is the threshold of the number of sentences with low reliability temporary labels that are presented to users in order to check whether each of the temporary labels is correct. Because temporary labels with high reliability are likely to be correct, sentences with high reliability temporary labels are added to the training data (Extension I). On the other hand, sentences with low reliability temporary labels are assigned true labels obtained by users' feedback, and added to the training data (Extension Ⅱ). Since all reliabilities of temporary labels are modified by new feedback, sentences added to the training data by Extension I are re-selected at each step in the iterative process. Therefore, as a whole, sentences in the training data are increased by the number of sentences assigned true labels by feedback in Extension Ⅱ at each step.

```
Procedure : extending the training data with distance metric learning
About sentences in the reference data : $\{S_j\}_{j=1}^{m}$
for $j = 0 \ldots m$
    calculate $r(j)$ by (4)
end
From $\{r(j)_{j=1}^{m}\}$,
 the top $T_t \rightarrow$ add to the training data (Extension I)
 the bottom $T_b \rightarrow$ present to users (Extension Ⅱ)
```

Fig. 2 Procedure to extend the training data with redefining the distance metric between sentences in the reference data.

4. Evaluation

We evaluate the effectiveness of the proposed method by using documents shown in **Table 1**, each of which is referred by PDB. PDB-ID is the identifier of the protein registered in PDB, and the "correct sentence" means the sentence containing protein function information. Named entities in these documents are already tagged manually. In our experiment, seven documents are used for learning. Ten sentences in the documents are used as the initial training data with true labels, and the rest of the documents are used as the reference data. In addition, 16 documents are used for evaluation. We conduct eight experiments with different combinations of documents.

SVM (Support Vector Machine) is used as a classifier, and one sentence is regarded as one instance. Features shown in Section 2 are used for training the classifier. The number of keywords used as features is 45, and the number of patterns is 19.

The threshold T_t is 25, and T_b is between 6 and 200. That is to say, 25 sentences that are assigned temporary labels with high reliability are added to the training data. The transition of the accuracies as iterating feedback steps is shown in

Table 1 Documents used in the experiment.

PDB-ID	# of sentences	# of correct sentences	PDB-ID	# of sentences	# of correct sentences
1a0f	382	46	1a0h	359	26
1a0k	683	19	1a0o	148	12
1a0q	295	23	1a1s	285	24
1a23	528	5	1a26	243	13
1a3a	544	17	1a3h	275	8
1a3l	272	23	1a3r	299	21
1a3s	306	7	1a3y	209	3
1a4j	190	13	1a5a	113	10
1a5h	296	39	1a5i	324	73
1a5v	277	20	1a5y	291	33
1a5z	428	8	2a2g	365	13
2a39	312	4			

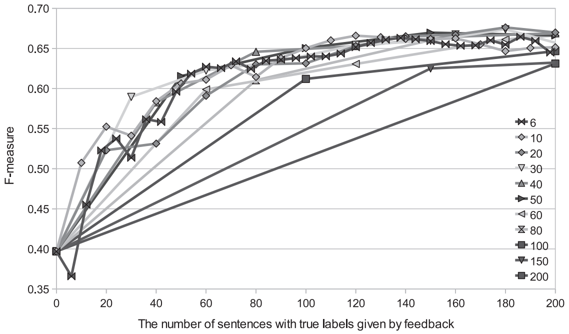


Fig. 3 The accuracies of the proposed method ($T_b = 6 \sim 200$).

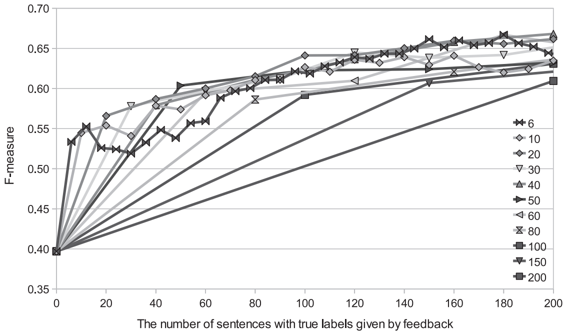
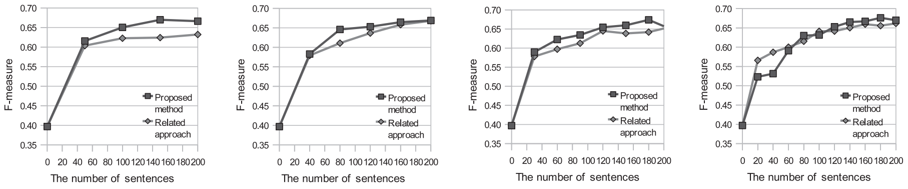


Fig. 4 The accuracies of the related approach ($T_b = 6 \sim 200$).

Fig. 3 when the number of sentences returned at one feedback step is increased between 6 and 200. The X-axis is the number of sentences with true labels given by feedback, the number increases each time the feedback is returned. The accuracy is higher at an earlier step when the number of sentences returned at one feedback step is 30 (i.e., $T_b = 30$). For example, in the proposed method, the number of sentences with true labels given by feedback is 60, and the F-measure is over 62 percent after two steps of feedback with 30 sentences. By contrast, the F-measure reaches almost the same level after two steps of feedback with 60 sentences, but the number of sentences is 120. As a result, a certain accuracy is achieved when the number of sentences with true labels given by feedback with 30 sentences is 60 less than one by feedback with 60 sentences. Therefore, the accuracy in the case of returning feedback little by little is higher when the number of sentences with true label given by feedback is 30 or more. However, when T_b is smaller than 30, the accuracy is lower at early steps, because sentences with true labels are too few at early steps to learn correctly. Especially, when T_b is 6, the accuracy falls at the initial step, goes up and down after the second step, and becomes stable with around 60 sentences with true labels given by feedback.

There is a related approach in the field of active learning, which uses a distance from a hyperplane of SVM as a criterion for presenting to the users [28)–30)]. Because it is considered that the reliability of an instance which is close to the hyperplane is low, such instances are presented to the users, and their labels are corrected if they are mistaken. In our comparative experiment, this approach is applied, the reliability of the temporary label is calculated based on the distance from the hyperplane, and feedback steps are iterated. The transition of the accuracy by this related approach is shown in **Fig. 4**. Compared with the proposed method, the whole accuracy is low. Especially, the accuracy is lower at early steps of feedbacks.



(a) $T_b = 50$ (b) $T_b = 40$ (c) $T_b = 30$ (d) $T_b = 20$

Fig. 5 The comparison of accuracies.

In order to clarify the difference between the proposed method and the related method, the comparisons of accuracies are shown in **Fig. 5** (a)–(d) when T_b is 50, 40, 30 and 20 respectively. When T_b is 50, 40 and 30, the accuracies of the proposed method become high at early steps. In the case of $T_b = 20$, although the accuracy of the proposed method rise little compared with the related approach at early steps, it is almost equal to one of the related approach when the sum of sentences with true labels given by feedback is more than 60. The accuracy of the proposed method is relatively stable and high when T_b is more than a certain

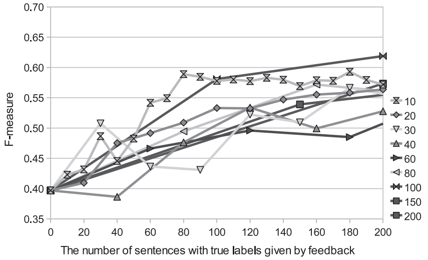


Fig. 6 The accuracy in the case of selection at random.

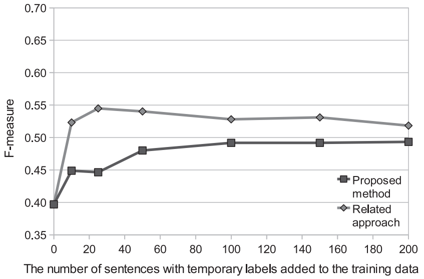


Fig. 7 The accuracies without feedback ($T_b = 0$).

value. In practice, most appropriate T_b cannot be decided strictly. Therefore, the proposed method whose accuracy is higher at a wide range of T_b, and converges early is better than the related approach.

The transition of accuracies are shown in **Fig. 6** when sentences for presenting to users are selected at random. In the iterative steps, the accuracies are affected more negatively, and show little sign of convergence. The proposed criterion is useful for selecting effective sentences which are presented to users, and returned with true labels. In order to compare the accuracies with just Extension I (without users' feedback), the accuracies without feedback process are shown in **Fig. 7** as the number of sentences added to the training data (T_t) is varied. The both accuracies of the proposed method and the related approach are converged at low values. Especially, in the proposed method, the accuracy is lower, which means that users' feedback contributes to improving the accuracy.

As above, in the proposed method, the accuracy is improved when the iteration count of feedbacks and the number of sentences returned by a feedback are small. Additionally, the accuracy increases monotonically as iterating feedbacks when T_b is more than a certain value. Thus, it is confirmed that the reliability based on the distance metric learning is effective.

5. Conclusion

In this paper, we proposed the method to extract sentences containing protein function information with training data extension using user's feedback. In the proposed method, appropriate sentences for extending the training data are selected by conducting the distance metric learning and calculating the reliability of temporary labels assigned to sentences.

Compared with an approach in the field of active learning, the accuracy by the proposed method is higher when the iteration count of feedbacks and the number of sentences returned by a feedback are small. Therefore, it is confirmed that the training data is appropriately extended based on users' feedback by the proposed method. In addition, this result serves a purpose of reducing users' cost.

Our future work will focus on the accuracy at earlier steps of feedbacks. We will consider that the accuracy will be improved exploiting features of sentences returned as a feedback.

References

1) Berg, J.M., Tymoczko, J.L. and Stryer, L.S.: Biochemistry fifth edition, *WH Freeman and Company*, Vol.423, pp.436–437 (2002).

2) Wu, C.H., Yeh, L.S.L., Huang, H., Arminski, L., Castro-Alvear, J., Chen, Y., Hu, Z., Kourtesis, P., Ledley, R.S., Suzek, B.E., Vinayaka, C.R., Zhang, J. and Barker, W.C.: The protein information resource, *Nucleic Acids Research*, Vol.31, pp.345–347 (2003).

3) Berman, H.M., Battistuz, T., Bhat, T.N., Bluhm, W.F., Bourne, P.E., Burkhardt, K., Feng, Z., Gilliland, G.L., Iype, L., Jain, S., Fagan, P., Marvin, J., Padilla, D., Ravichandran, V., Schneider, B., Thanki, N., Weissig, H., Westbrook, J.D. and Zardecki, C.: The protein data bank, *Acta Crystallographica Section D: Biological Crystallography*, Vol.58, No.6, pp.899–907 (2002).

4) Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.C., Estreicher, A., Gasteiger, E., Martin, M.J. and Michoud, K.: The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003, *Nucleic Acids Research*, Vol.31, pp.365–370 (2003).

5) Tsai, R.T.H., Sung, C.L., Dai, H.J., Hung, H.C. and Sung, T.Y.: NERBio: Using selected word conjunctions, term normalization, and global patterns to improve biomedical named entity recognition, *BMC Bioinformatics*, Vol.7 (Suppl 5):S11 (2006).

6) Sun, C., Guan, Y., Wang, X. and Lin, L.: Biomedical named entities recognition using conditional random fields model, *Fuzzy Systems and Knowledge Discovery*, pp.1279–1288, Springer Berlin/Heidelberg (2006).

7) Lafferty, J., McCallum, A. and Pereira, F.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data, *Proc. International Conference on Machine Learning* (*ICML '01*), pp.282–289 (2001).

8) Bunescu, R., Ge, R., Kate, R.J., Mooney, R.J., Wong, Y.W., Marcotte, E.M. and Ramani, A.: Learning to extract proteins and their interactions from medline abstracts, *Proc. ICML-2003 Workshop on Machine Learning in Bioinformatics*, pp.46–53 (2003).
9) Rabiner, L.R.: A tutorial on hidden Markov models and selected applications inspeech recognition, *Proc. IEE*, Vol.77, No.2, pp.257–286 (1989).
10) Vapnik, V.N.: *The nature of statistical learning theory*, Springer (1995).
11) Fukuda, K., Tamura, A., Tsunoda, T. and Takagi, T.: Toward information extraction: Identifying protein names from biological papers, *Pacific Symposium on Biocomputing*, pp.707–718 (1998).
12) Tanabe, L. and Wilbur, W.J.: Tagging gene and protein names in biomedical text, *Bioinformatics*, Vol.18, No.8, pp.1124–1132 (2002).
13) Cooper, J.W. and Kershenbaum, A.: Discovery of protein-protein interactions using a combination of linguistic, statistical and graphical information, *BMC bioinformatics*, Vol.6, No.143 (2005).
14) Hao, Y., Zhu, X., Huang, M. and Li, M.: Discovering patterns to extract protein-protein interactions from the literature: Part II, *Bioinformatics*, Vol.21, No.15, pp.3294–3300 (2005).
15) Munna, M.A. and Ohkawa, T.: A method to extract sentences with protein functional information from literature by iterative learning of the corpus, *IPSJ Transactions on Bioinformatics*, Vol.47, No.SIG 17(TBIO 1), pp.22–30 (2006).
16) McCallumzy, A.K. and Nigamy, K.: Employing EM in pool-based active learning for text classification, *Proc. International Conference on Machine Learning* (*ICML*), pp.359–367 (1998).
17) Tong, S. and Koller, D.: Support vector machine active learning with applications to text classification, *Proc. International Conference on Machine Learning* (*ICML*), pp.999–1006 (2000).
18) Settles, B. and Craven, M.: An analysis of active learning strategies for sequence labeling tasks, *Proc. Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), pp.1069–1078 (2008).
19) Muslea, I., Minton, S. and Knoblock, C.A.: Selective sampling with redundant views, *Proc. National Conference on Artificial Intelligence* (*AAAI*), pp.621–626 (2000).
20) Zhou, Z.H., Chen, K.J. and Jiang, Y.: Exploiting unlabeled data in content-based image retrieval, *Proc. European Conference on Machine Learning* (*ECML*), pp.425–435 (2004).
21) Zhu, X., Lafferty, J. and Ghahramani, Z.: Combining active learning and semi-supervised learning using Gaussian fields and harmonic functions, *Proc. ICML workshop on The Continuum from Labeled to Unlabeled Data*, pp.58–65 (2003).
22) Tur, G., Hakkani-Tür, D. and Schapire, R.E.: Combining active and semi-supervised learning for spoken language understanding, *Speech Communication*, Vol.45, No.2, pp.171–186 (2005).
23) Yu, K., Bi, J. and Tresp, V.: Active learning via transductive experimental design, *Proc. International Conference on Machine Learning* (*ICML*), pp.1081–1087 (2006).
24) Tomanek, K. and Hahn, U.: Semi-supervised active learning for sequence labeling, *Proc. Association for Computational Linguistics* (*ACL*), pp.1039–1047 (2009).
25) Cohn, D., Caruana, R. and McCallum, A.: Semi-supervised clustering with user feedback, *Constrained Clustering: Advances in Algorithms, Theory, and Applications*, pp.17–31 (2008).
26) Xing, E.P., Ng, A.Y., Jordan, M.I. and Russell, S.: Distance metric learning with application to clustering with side-information, *Advances in neural information processing systems*, pp.521–528 (2003).
27) Bar-Hillel, A., Hertz, T., Shental, N. and Weinshall, D.: Learning distance functions using equivalence relations, *Proc. Twentieth International Conference on Machine Learning* (*ICML*), Vol.20, No.1, pp.11–18 (2003).
28) Sassano, M.: An empirical study of active learning with support vector machines for Japanese word segmentation, *Proc. 40th Annual Meeting on Association for Computational Linguistics* (*ACL*), pp.505–512 (2001).
29) Schohn, G. and Cohn, D.: Less is more: Active learning with support vector machines, *Proc. 17th International Conference on Machine Learning*, pp.839–846 (2000).
30) Tong, S. and Koller, D.: Support vector machine active learning with applications to text classification, *The Journal of Machine Learning Research*, Vol.2, pp.45–66 (2002).

(Received June 2, 2010)
(Accepted August 4, 2010)
(Released October 13, 2010)

(Communicated by *Tsuyoshi Kato*)

Kazunori Miyanishi received his B.E. and M.E. degrees from Kobe University in 2003 and 2005, respectively. He is a Ph.D. student in the Graduate School of Science and Technology, Kobe University. His current research interests include information extraction and bioinformatics.

Tomonobu Ozaki received his Ph.D. in Media and Governance from Keio University in 2002. Now he is a Specially Appointed Assistant Professor of Cybermedia Center, Osaka University. He is a member of the Japanese Society for Artificial Intelligence.

Takenao Ohkawa received his B.E, M.E., and Ph.D. degrees from Osaka University in 1986, 1988, and 1992, respectively. He is currently a Professor at the Development of Information Science, Graduate School of System Informatics, Kobe University. His research interests include intelligent software and bioinformatics. He is a member of the IEEE, the Institute of Electronics, Information, and Communication Engineers, the Institute of Electrical Engineers of Japan, and the Japanese Society for Artificial Intelligence.