IPSJ Transactions on Bioinformatics Vol. 2 93–100 (Sep. 2009)

Original Paper

Selection of Effective Sentences from a Corpus to Improve the Accuracy of Identification of Protein Names

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

As the number of documents about protein structural analysis increases, a method of automatically identifying protein names in them is required. However, the accuracy of identification is not high if the training data set is not large enough. We consider a method to extend a training data set based on machine learning using an available corpus. Such a corpus usually consists of documents about a certain kind of organism species, and documents about different kinds of organism species tend to have different vocabularies. Therefore, depending on the target document or corpus, it is not effective for the accurate identification to simply use a corpus as a training data set. In order to improve the accuracy, we propose a method to select sentences that have a positive effect on identification and to extend the training data set with the selected sentences. In the proposed method, a portion of a set of tagged sentences is used as a validation set. The process to select sentences is iterated using the result of the identification of protein names in a validation set as feedback. In the experiment, compared with the baseline, a method without a corpus, with a whole corpus, or with a part of a corpus chosen at random, the accuracy of the proposed method was higher than any baseline method. Thus, it was confirmed that the proposed method selected effective sentences.

1. Introduction

Protein function information is useful in various fields (for example, drug discovery and understanding life phenomena). The function information is stated in a number of documents about protein structure analysis. Thus, it is required to automatically extract the information from a number of documents [1),2)], and for that purpose it is important to identify protein names in those documents.

†1 Graduate School of Science and Technology, Kobe University
†2 Organization of Advanced Science and Technology, Kobe University
†3 Graduate School of Engineering, Kobe University

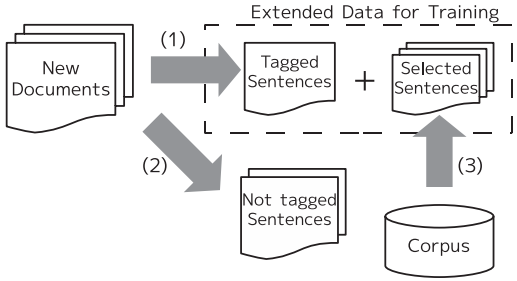

Fig. 1 Concept of training data extension.

Various studies about identification of protein names have been conducted [3)–5)], and almost all of these are about methods to identify names by using machine learning based on features of a word (part of speech, spelling and so on) and a context. These methods are based on the assumption that a sufficient training data set is given. Therefore, the accuracy of identification is not high when not enough training data are given. It is considerably difficult to collect enough training data, especially when targeting documents about an organism species on which a corpus has not been created. It is expected that the accuracy of identification would be improved if a training data set could be extended effectively by using an available corpus about different organism species from targeting documents. However, since the documents from which such a corpus has been created tend to have different vocabularies and styles, not all sentences in the corpus contribute to the identification. Thus, it is not effective to add a whole corpus to the tagged sentences in targeting documents. In addition, to simply add a corpus to the tagged sentences may have negative effects on the accuracy.

Motivated by the above background, we propose a method to extend a training dataset by selecting effective sentences for identification from an available corpus. The concept of training data extension is shown in **Fig. 1**. Some sentences from parts of the target documents are first marked with protein tags and the others are not, shown as (1) and (2). The sentences selected from a corpus are added to the tagged sentences as (3), and these sentences (tagged sentences in the target documents and selected ones from a corpus) are used as a training data set.

Since sentences in targeting documents and ones in a corpus have different

vocabularies, we consider that effective sentences can be selected based on the structures of sentences rather than based on certain keywords. Thus, we propose a selection method not using features of words themselves but using features of structures of sentences. The tagged sentences to which the sentences selected by the proposed method are added, are used as a training data set, and a model is generated by using the training data set, which tags each word in input sentences based on whether it is a protein name or not.

There is a similar research field addressing the problem called domain adaptation [6)–9)]. Domain adaptation is a branch of transfer learning [10)–13)], and is to fit the model learned based on the data set from one domain (called the source domain) to one from another domain (called the target domain). There are several studies in which domain adaptation is applied to named entity recognition, and are classified roughly into two categories. One assumes that texts in the target domain are tagged, but the other assumes that they are not tagged.

As a typical example of the former, Arnold, et al. [14)] proposed a method to generate robust features by exploiting tagged abstracts and non-tagged main texts of papers, and to try to identify protein names in non-tagged captions based on the features. Jiang, et al. [15)] proposed a method to train a classifier based on the most generalizable features across source domains and to apply the trained classifier to the non-tagged target domain. These studies use the non-tagged target domain, and use structural frequency features or generalizable features in order to adapt the model learned on the source domain to the target domain. However, in this paper, a part of documents from the target domain is tagged and used as a training data set, and we extend the training data set by selecting and adding useful sentences from the source domain.

On the other hand, as for the latter of the domain adaptation, Daumé III [16)] presented an approach to using a combination of common features in both domains and unique features in each domain. It seems that the accuracy is negatively affected based on the source domain, since a number of texts from the source domain are used. In addition, a training set from the target domain is large compared to a test set, and that is different from our assumption. Arnold, et al. [17)] proposed a method to adapt a model across domains or tasks by exploiting hierarchical features. While this method also generates a model using a whole source domain and tunes it using tagged data from the target domain, but our proposed method focuses on using just a useful part of the source domain. In this paper, we focus on the way to select the effective sentences from the source domain. Therefore, we evaluate the effectiveness by adding the selected sentences to a training data set, learning a model on the extended training data set, and applying it to non-tagged sentences from the target domain.

2. A Method to Select Sentences

A document tends to have a unique vocabulary (for example, names of proteins or related substances) depending on what organism species is a target in it. In fact, protein names vary between organism species. However, it would appear that sentences containing protein names have some common structural features. Excerpts of sentences containing protein names are shown in **Fig. 2**. The organism species targeted by the document (PMID:10381570 and PMID:10455134) is respectively fly and human. "*Ttk*" and "*AML1 and BSAD*" are protein names, and the combination of the subject, the predicate and the object is respectively "*Ttk* activates transcription" and "*AML1 and BSAD* activate transcription". Between the two sentences, the protein names themselves are different, but the predicate and the object related to the protein names are common to both sentences. Therefore, we propose a method to select effective sentences based on their structures.

2.1 Structures of Sentences Containing Protein Names

A modification relation and a co-occurrence relation in a sentence are considered as the structural features of the sentence. Characteristic structures of sentences containing protein names are captured from structures obtained as the

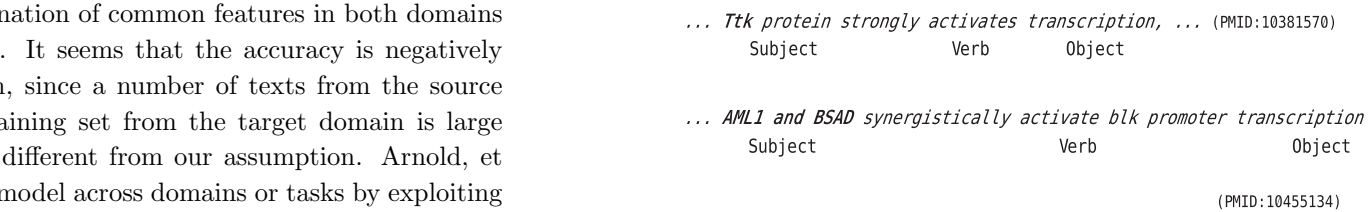

Fig. 2 Excerpts of sentences containing protein names.

result of syntax analysis by a parser. In this paper, we use the Stanford parser [18]. In Fig. 2, in the case where subjects, verbs (predicates), and objects are extracted as elements of sentences, the relation of subject–predicate, "[protein] activates", is obtained as a feature of the sentences.

Similarly, modification relations (subject–predicate, object–predicate, a pair of nouns connected by a preposition, and so on) and co-occurrence relations (words appearing with a protein name in a sentence) are assigned to each sentence as its features.

2.2 Selection Based on Structures of Sentences

Because it is not obvious beforehand which features are effective, a portion of tagged sentences is used as a validation set, and a model that puts a protein tag on sentences is generated by the iterative process, which optimizes the accuracy of tagging protein names in the validation set. The details of a model used in this paper and the method of learning are described in Section 3. The outline of the iterative process to select effective sentences is shown in **Fig. 3**. First, tagged sentences in input documents are split into two sets, the training set and validation set(arrows① in Fig. 3. In the initial setting, sentences are selected from a corpus at random(arrow②). Next a model is generated from the training set and the selected sentences, and applied to the validation set(arrow③). Based on the tagged result, weights of syntactic features of sentences are updated(arrow④). Sentences are selected in the order of descending weights again, and a new model is generated. This process is iterated while the accuracy of tagging is rising.

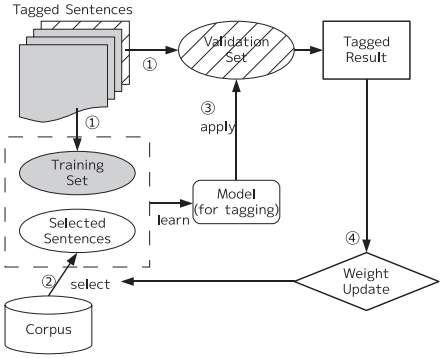

Fig. 3 Outline of sentence selection.

Figure 4 shows the procedure to update weights of features, where $F_j(1 \leq j \leq M)$ means features of sentences in a corpus, M is the number of all features, $f_i(1 \leq i \leq n)$ means features of sentences containing mistagged words, and n is the number of features. W_{F_j} means the weight of the feature F_j, and the initial value of W_{F_j} is the normalized number of times F_j appears in the training set. $U(> 0)$ is the increase ratio of the weight. By this procedure, just the weights of the sentences containing mistagged words are increased up to U times. **Figure 5** shows the procedure to update weights of sentences, where $S_i(1 \leq i \leq N)$ means sentences in a corpus, N is the number of all sentences, and W_{S_i} means the new weight of the sentence S_i. $\tilde{W}_{S_i}$ is the sum of the weights of the features of the sentence S_i. Finally, the new weight W_{S_i} is obtained by normalizing at line **7**. The procedure to select sentences based on updated weights is shown in **Fig. 6**, where W_{S_i} is the new weight of the sentence S_i and W'_{S_i} is the previous weight,

Procedure : update weights of features
1 for $(i = 1..n)$
2 **for**$(j = 1..M)$
3 **if**$(F_j == f_i)$
4 $W_{F_j} = W_{F_j} \times U$

Fig. 4 Procedure to update weights of features of sentences.

Procedure : update weights of sentences
1 initialize : $\tilde{W}_S = 0$
2 for$(i = 1..N)$
3 **for**$(j = 1..M)$
4 **if**$(S_i \; has \; the \; feature \; F_j)$
5 $\tilde{W}_{S_i} = \tilde{W}_{S_i} + W_{F_j}$
6 for$(i = 1..N)$
7 $W_{S_i} = \tilde{W}_{S_i} \Big/ \sum_{k=1}^{N} \tilde{W}_{S_k}$

Fig. 5 Procedure to update weights of sentences.

Procedure : select sentences
1 select = {}
2 **for**$(i = 1..N)$
3 **if**$(W_{S_i} > W'_{S_i} \parallel rank(W_{S_i})$ *is superior to* $T_R)$
4 select = select $\cup \{S_i\}$

Fig. 6 Procedure to select sentences.

word	stem	pos	chunk	tag
at	at	IN	B-PP	O
position	position	NN	B-NP	O
187	187	CD	I-NP	O
in	in	IN	B-PP	O
esterase	esterase	NN	B-NP	K
6	6	CD	I-NP	K
contributes	contribute	VBZ	B-VP	O
significantly	significantly	RB	B-ADVP	O

Fig. 7 A region of the word "esterase".

and the function $rank(W)$ returns the rank of the weight W in all weights of sentences. Sentences whose weights are increased by updating or are ranked in the top T_R in all sentences are selected.

3. A Method for Learning a Model and Identifying Protein Names

In this paper, CRF (Conditional Random Fields) [19),20)] is used in order to put a protein tag on sentences. Here, each sentence is part-of-speech tagged by Brill's tagger [21)] and chunked by CRF++ [22)]. A stemming, a part-of-speech, and a chunk are used as features of each word in input sentences. In addition, two words around the current word in the sentence are considered for learning. An example of a region is shown in **Fig. 7**. In this paper, borders of protein names are not focused on, and the accuracy is evaluated by a partial match. That is to say, it is considered as correctly tagging that a model puts a protein tag on the words which are a part of protein name. Thus, each word is assigned one of just two types of tags ("K" or "O") depending on whether it is a protein name or not. A model is learned based on these features and contexts.

As described in Section 2.2, tagged sentences in input documents are split into the training set and the validation set. Therefore, several models are generated

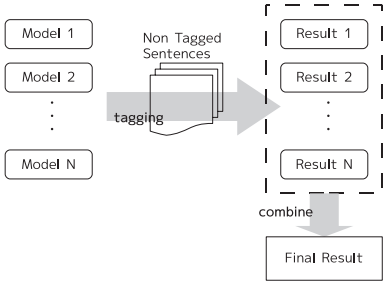

Fig. 8 Flow of combining tagged results.

depending on the manner of splitting. By applying each model to non-tagged sentences, the same number of tagged results are obtained. By combining these results, the final result is obtained. This flow is shown in **Fig. 8**. In the proposed method, given a threshold T_M, a word is regarded as a protein name, if more than or equals to T_M models agree with the result.

4. Evaluation

We evaluate the effectiveness of the proposed method by using an abstract set of the GENIA corpus [23)] and abstract sets about "mouse" and "fly" in the corpora of BioCreAtIvE1 Task 1B [24)]. Several hundreds of abstracts are extracted from a set of "GENIA" and they are treated as input documents, namely 'New Documents' in Fig. 1. Of those, between 10 and 100 abstracts are treated as a set of tagged sentences and 500 abstracts as a set of non-tagged sentences. Also 5,000 abstracts about "mouse" or "fly" are treated as a corpus to extend a training set. In addition, in a set of tagged sentences, 90 percent of all abstracts are a training set and the others are a validation set. Ten kinds of this combination of datasets are generated and 10 kinds of models are obtained. In the iteration of updating weights, 800 sentences are selected at random and added to the training set in the initial setting. In each step, the increase ratio of a weight of a feature U is 2, and the threshold T_R is 1,000. Moreover, the top 500 sentences are selected from ones whose weights are increased by updating to avoid selecting too many sentences. Since each model is not a weak learner, simple voting does not always result in the highest accuracy. Therefore, the threshold of a combination of models T_M is

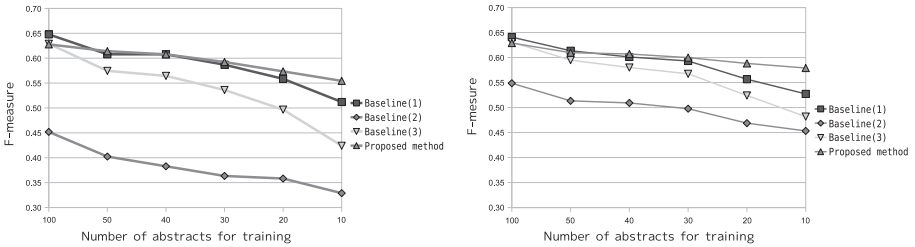

Fig. 9 Comparison of accuracies of "mouse" (left) and "fly" (right).

1 based on a preliminary experiment.

We compare the proposed method with three baseline experiments. **Baseline (1)**: a model is trained based on only tagged sentences and applied to non-tagged sentences. **Baseline (2)**: a model is learned based on tagged sentences and a whole corpus, and applied to non-tagged sentences. **Baseline (3)**: a model is learned based on tagged sentences and 50 abstracts chosen at random from a corpus, and applied to non-tagged sentences. Each method is evaluated by recall, precision, and F-measure. The results of the comparison are shown in **Fig. 9**. From the result of baselines (1) and (2), when a whole corpus is added to a training set, the accuracy declines significantly. In addition, from the result of baseline (3), when abstracts chosen at random from a corpus are added, the accuracy is higher than the case of adding a whole corpus (baseline(2)), but is lower than the case of adding no abstract from the corpus (baseline(1)). In contrast, in the proposed method, when a training set is sufficient, the accuracy is nearly as high as that in the case of baseline (1). However, when a training set is small, the accuracy is higher than any other cases. In particular, when the number of abstracts in a training set is less than 30, the accuracies of all baselines are remarkably low, but the accuracy of the proposed method does not decline comparatively. The average numbers of sentences in an original training set and added sentences in the proposed method and baseline (3) are shown in **Figs. 10** and **11**. Here, the numbers of sentences in a whole corpus of "mouse" and "fly" are 41,345 and 38,510 respectively. When a whole corpus is added, the different vocabularies in the corpus have negative effects on the accuracy, because added sentences are much more than sentences in an original training set. Furthermore,

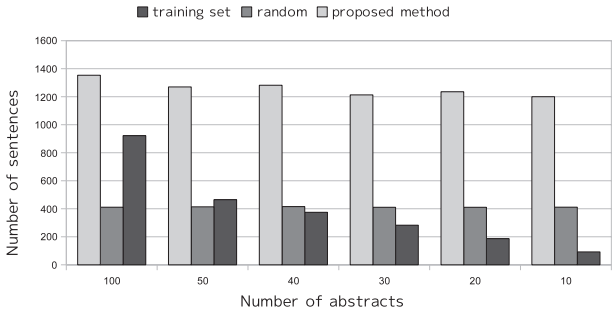

Fig. 10 Number of original training set and added sentences ("mouse").

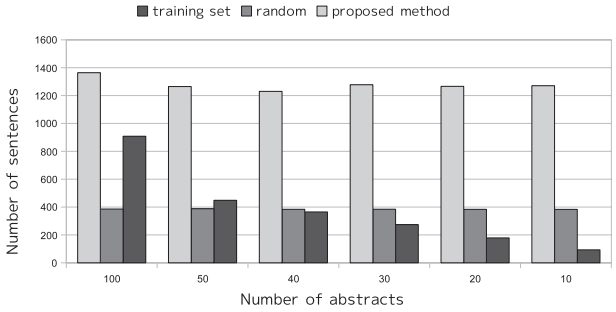

Fig. 11 Number of original training set and added sentences ("fly").

when abstracts chosen at random are added, the number of added sentences is only about 400, but the accuracy is negatively affected. On the other hand, in the proposed method, between 1,200 and 1,400 sentences are added, and the accuracy is improved especially when an original training set is small. Therefore, it is confirmed that sentences that have a positive effect on the identification are selected in the proposed method in order to complement a small training set.

Next, we discuss recall, precision, and F-measure of results of the baselines and the proposed method. Recall, precision, and F-measure of each result are shown in **Fig. 12**. Here, 10 abstracts are used as a set of tagged sentences. In baseline (2), recall is lowest but precision is highest among all methods. This is because the fact that most data (tagged sentences and a whole corpus) are used

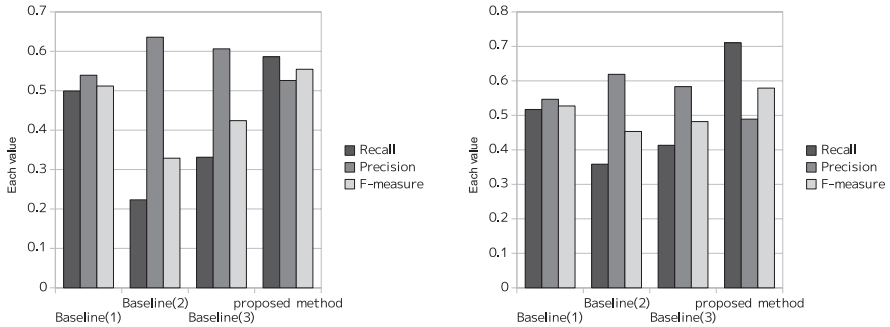

Fig. 12 Recall, precision, and F-measure of "mouse" (left) and "fly" (right).

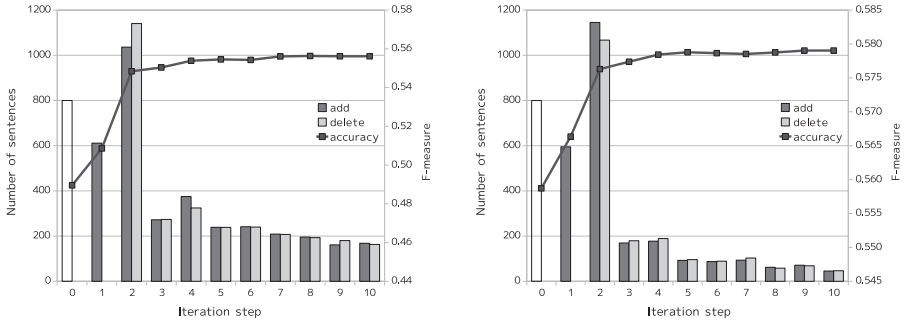

Fig. 13 Number of added or deleted sentences and accuracy at each iteration step of "mouse" (left) and "fly" (right).

for learning a model has a positive effect on precision. However, added sentences from a corpus have negative effects on recall. As a result, F-measure becomes lower than that in baseline (1) in which only tagged sentences are used for the learning. The result of baseline (3) shows the same tendency. On the other hand, in the proposed method, precision is a little lower than that in baseline (1), but recall is highest. This shows that the result of the proposed method is not affected negatively by a corpus and proper sentences are selected.

In detail, we analyze the transition of the number of added or deleted sentences and the accuracy of identification in the iteration of updating weight. **Figure 13** show the average number of added or deleted sentences and the accuracy in each step in the case of 10 abstracts for initial training sentences and "mouse" or "fly" for the expansions. The number of added sentences in the initial setting is 800 as previously described, and sentences are only added in the first iteration step, since the number of added sentences in the initial setting is small. In the second step, more than 1,000 sentences are added and deleted, and both added and deleted sentences are decreasing after the third step. The accuracy rises in the first two steps, and then converges. Therefore, it is confirmed that effective sentences for tagging protein names are certainly selected in the iterations.

5. Conclusion

In this paper, we proposed a method to select effective sentences from a corpus and extend the training set. In the proposed method, a set of tagged sentences is split into a training set and a validation set. The process to select sentences with syntactic features is iterated using the result of the identification of protein names in a validation set as feedback. The F-measure of the proposed method is higher than the method without using a corpus and also higher than the method with a whole corpus, or with a part of a corpus chosen at random. Therefore, it is confirmed that the method selects effective sentences.

Our future work will focus on improving the accuracy of a model. We will consider that a model will be trained using syntax information as a method to be adapted for the sentence selection based on syntax information.

References

1) Miyanishi, K., Takeuchi, M., Ozaki, T. and Ohkawa, T.: Iterative learning with feature update for extracting sentence containing protein function information, *Proc. 7th Atlantic Symposium on Computational Biology & Genome Informatics* (*CBGI2007*), pp.96–102, Salt Lake, USA (2007).
2) Munna, Md. 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, SIG 17(TBIO 1), pp.22–30 (2006).
3) Zhou, G.D., Shen, D., Zhang, J., Su, J. and Tan, S.H.: Recognition of protein/gene names from text using an ensemble of classifiers, *BMC Bioinformatics 2005*, Vol.6 (Suppl 1):S7 (2005).
4) Murata, M., Mitsumori, T. and Doi, K.: Overfitting in protein name recognition on biomedical literature and method of preventing it through use of transductive

SVM, *Proc. 10th International Conference on Information Technology*, pp.583–588, Rourkela, India (2007).
5) Koike, A. and Takagi, T.: Gene/protein/family name recognition in biomedical literature, *Proc. BioLINK 2004: Linking Biological Literature, Ontologies, and Databases*, pp.9–16, Boston, Massachusetts, USA (2004).
6) Daumé III, H. and Marcu, D.: Domain adaptation for statistical classifiers, *Journal of Artificial Intelligence Research*, Vol.26, pp.101–126 (2006).
7) Roark, B. and Bacchiani, M.: Supervised and unsupervised PCFG adaptation to novel domains, *Proc. Conference of the North American Chapter of the Association for Computational Linguistics and Human Language Technology* (*NAACL/HLT*), pp.126–133, Edmonton, Canada (2003).
8) Blitzer, J., McDonald, R. and Pereira, F.: Domain adaptation with structural correspondence learning, *Proc. Empirical Methods in Natural Language Processing* (*EMNLP*), pp.120–128, Sydney, Australia (2006).
9) Ben-David, S., Blitzer, J., Crammer, K. and Pereira, F.: Analysis of representations for domain adaptation, *Advances in Neural Information Processing Systems*, Vol.19, pp.137–144, Cambridge, MA, USA (2007).
10) Baxter, J.: A Bayesian/Information Theoretic Model of Learning to Learn via Multiple Task Sampling, *Machine Learning*, Vol.28, pp.7–39 (1997).
11) Thrun, S.: Is learning the n-th thing any easier than learning the first, *Advances in Neural Information Processing Systems*, Vol.8, pp.640–646 (1996).
12) Caruana, R.: Multitask Learning, *Machine Learning*, Vol.28, pp.41–75 (1997).
13) Arnold, A., Nallapati, R. and Cohen, W.W.: A comparative study of methods for transductive transfer learning, *Proc. Seventh IEEE International Conference on Data Mining Workshops*, pp.77–82, Omaha, Nebraska, USA (2007).
14) Arnold, A. and Cohen, W.W.: Intra-document structural frequency features for semi-supervised domain adaptation, *Proc. 17th ACM Conference on Information and Knowledge Management*, pp.1291–1300, Napa Valley, California, USA (2008).
15) Jiang, J. and Zhai, C.X.: Exploiting domain structure for named entity recognition, *Proc. Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pp.74–81, Morristown, NJ, USA (2006).
16) Daumé III, H.: Frustratingly easy domain adaptation, *Proc. 45th Annual Meeting of the Association for Computational Linguistics*, pp.256–263, Prague, Czech Republic (2007).
17) Arnold, A., Nallapati, R. and Cohen, W.W.: Exploiting feature hierarchy for transfer learning in named entity recognition, *Proc. ACL-08: HLT*, pp.245–253, Columbus, Ohio, USA (2008).
18) Klein, D. and Manning, C.D.: Accurate unlexicalized parsing, *Proc. 41st Meeting of the Association for Computational Linguistics*, pp.423–430, Sapporo, Japan (2003).
19) Lafferty, J., McCallum, A. and Pereira, F.: Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data, *Proc. Eighteenth International Conference on Machine Learning*, pp.282–289 (2001).
20) Sha, F. and Pereira, F.: Shallow parsing with conditional random fields, *Proc. 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pp.134–141, Edmonton, Canada (2003).
21) Brill, E.: Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging, *Computational Linguistics*, Vol.21, pp.543–565 (1995).
22) Kudo, T.: CRF++: Yet Another CRF toolkit (2005). Available at http://crfpp.sourceforge.net/
23) Collier, N., Park, H.S., Ogata, N., Tateishi, Y., Nobata, C., Ohta, T., Sekimizu, T., Imai, H., Ibushi, K. and Tsujii, J.: The genia project: corpus-based knowledge acquisition and information extraction from genome research papers, *Proc. 9th Conference on European chapter of the Association for Computational Linguistics*, pp.271–272, Association for Computational Linguistics Morristown, NJ, USA (1999).
24) Hirschman, L., Colosimo, M., Morgan, A. and Yeh, A.: Overview of biocreative task 1b: Normalized gene lists, *BMC Bioinformatics 2005*, Vol.6 (Suppl 1):S11 (2005).

(Received April 17, 2009)
(Accepted July 8, 2009)
(Released September 24, 2009)

(Communicated by *Kenji Satou*)

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 an assistant professor of Organization of Advance Science and Technology, Kobe University. He is a member of JSAI.

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 Computer Science and Systems Engineering, Graduate School of Engineering, 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 in Japan, and the Japanese Society for Artificial Intelligence.