pISSN 1976-913X
eISSN 2092-805X

J Inf Process Syst, Vol.8, No.4, December 2012 http://dx.doi.org/10.3745/JIPS.2012.8.4.575

An Active Co-Training Algorithm for Biomedical Named-Entity Recognition

Tsendsuren Munkhdalai*, Meijing Li*, Unil Yun**, Oyun-Erdene Namsrai*** and Keun Ho Ryu*

Abstract—Exploiting unlabeled text data with a relatively small labeled corpus has been an active and challenging research topic in text mining, due to the recent growth of the amount of biomedical literature. Biomedical named-entity recognition is an essential prerequisite task before effective text mining of biomedical literature can begin. This paper proposes an Active Co-Training (ACT) algorithm for biomedical named-entity recognition. ACT is a semi-supervised learning method in which two classifiers based on two different feature sets iteratively learn from informative examples that have been queried from the unlabeled data. We design a new classification problem to measure the informativeness of an example in unlabeled data. In this classification problem, the examples are classified based on a joint view of a feature set to be informative/non-informative to both classifiers. To form the training data for the classification problem, we adopt a query-by-committee method. Therefore, in the ACT, both classifiers are considered to be one committee, which is used on the labeled data to give the informativeness label to each example. The ACT method outperforms the traditional co-training algorithm in terms of f-measure as well as the number of training iterations performed to build a good classification model. The proposed method tends to efficiently exploit a large amount of unlabeled data by selecting a small number of examples having not only useful information but also a comprehensive pattern.

Keywords—Biomedical Named-Entity Recognition, Co-Training, Semi-Supervised Learning, Feature Processing, Text Mining

1. INTRODUCTION

As biomedical literature on servers grows exponentially in the form of semi-structured documents, biomedical text mining has been intensively investigated to find information in a more accurate and efficient manner. One essential task in developing such an information extraction system is the biomedical Named-Entity Recognition (NER) process, which basically defines the boundaries between typical words and biomedical terminology in a particular text, and assigns

※ This study was supported by Korea Biobank project (4851-307) of the Korea Centers for Disease Control and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (No. 2012-0000478)
Manuscript received February 13, 2012; accepted September 20, 2012.
Corresponding Author: Keun Ho Ryu
* Database/Bioinformatics Laboratory, Chungbuk National University, Cheongju, Korea ({tsendeemts, mjlee, khryu }@dblab.chungbuk.ac.kr)
** Dept. of Computer Science, Chungbuk National University, Cheongju, Korea (yunei@chungbuk.ac.kr)
*** Dept. of Information Technology, Mongolian National University, Ulaanbaatar, Mongolia (oyunerdene@num.edu.mn)

the terminology to specific categories based on domain knowledge.

The biomedical NER task is not trivial due to several factors. NER performance in the newswire domain is indistinguishable from human performance, as it has an accuracy that is above 90%. However, performance has not been the same in the biomedical domain. It has been hampered by problems such as the number of new gene names that are being created on a regular basis, the lack of standardization of technical terms between authors, and often the fact that technical terms such as gene names often occur with other terminology [1].

Proposed solutions include rule-based, dictionary based, and machine learning approaches. Recently, Semi-Supervised Learning (SSL) techniques have been applied to NER. SSL is a Machine Learning (ML) approach that typically uses a large amount of unlabeled and a small amount of labeled data to build a more accurate classification model than that which would be built using only labeled data. SSL has received significant attention for two reasons. First, preparing a large amount of data for training requires a lot of time and effort. Second, since SSL exploits are effectively unlabeled, the accuracy of classifiers is generally improved.

One of the most popular SSL approaches is the co-training method [2] in which two classifiers are ideally designed in independent feature sets and initially learn from the same training data. During the co-training process, each classifier labels unlabeled data and selects examples showing the highest confidence score from the view of its feature set to retrain so that the accuracy of the classifiers will be improved. The co-training algorithm can exploit the unlabeled data when the assumptions of view independence and view sufficiency hold.

However, the problem with co-training, which recent studies including our previous paper [3] have reported, is that since there is no informativeness control, the classifiers can supply each other with the same pattern of examples in different rounds and the improvement of classifier performance no longer exists. Active Learning (AL) can complement the co-training method by querying the informative examples from the unlabeled data [4]. AL is a technique that helps experts to label a small amount of the unlabeled data to learn strong classifiers.

To solve the problem of the traditional co-training, this paper proposes an Active Co-Training (ACT) algorithm for NER adopting the idea of the AL querying. The ACT selects those that are relatively informative to both the classifiers to form an unlabeled data pool, where the classifiers are employed to retrain. We design a new classification problem to measure the informativeness of an example. Because the examples should be informative to both classifiers, the Informativeness Metric Classifier (IMC) classifies the examples based on both views of the feature sets. The Query-By-Committee (QBC) approach of the AL considers an example as informative when the classification models of the committee most disagree on its class decision [5]. In the ACT, both classifiers form one committee and an example is informative if the classifiers disagree on its class label. Thanks to the QBC approach, the ACT derives a training dataset where an instance is labeled as either informative or not for the IMC from the labeled data. Once the IMC is trained on the dataset, it is used on the unlabeled data to identify the informative examples. Finally, the informative examples are fed to the classifiers. The processes proceed iteratively until a given threshold is met. Therefore the decisions from both views are effectively reformed with an unlabeled data. The classifier performance is therefore cooperatively improved.

The rest of this paper is organized as follows. We review the previous studies for biomedical named-entity recognition in Section 2. Section 3 introduces the proposed method and the elements of the active co-training method. Section 4 includes a report of the environmental im-

provement in learning the classifiers and a discussion about the experiment. Finally, we summarize the main conclusions and present our future work direction in Section 5.

2. RELATED WORK

Many approaches have been proposed to recognize biomedical named entities from unstructured text. The ones that have been proposed are the rule-based, dictionary-based, and machine learning approaches. In the dictionary-based approach, a previously-prepared terminology list is matched though a given text to retrieve chunks containing the location of the terminology word. However, medical and biological text can contain a new terminology that might not be recognized by a dictionary-based approach.

The rule-based approach defines particular rules by observing the general features of biomedical entities in a biomedical text [6]. In order to identify any named entity in text data, a rule-generation process has to process a huge amount of text to collect accurate rules. On the other hand, domain experts collect the rules and this requires a lot of effort.

Since the machine learning approach was adopted, significant progress in biomedical NER has been achieved with methods like the Markov Model [7], the Support Vector Machine (SVM) [8-13] the Maximum Entropy Markov Model [14], and Conditional Random Fields (CRF) [15-19].

The authors of [10] showed that the dictionary-based method can be combined with machine learning by matching an example in the dataset through a dictionary to extract features, which comprise particular domain knowledge. Collier et al. [11] observed the effect of variations of character-level orthographics and Part-Of-Speech (POS) tag features and introduced a ranked table that contains features according to their probabilities of predicting a class. As shown in the table, the orthographic features indicate the class clearly, such as a feature named the GreekLetter, which has a 0.96 probability value of predicting a class.

A conditional random field, which is a classic sequence-labeling tool in text mining and natural language processing, has successfully been used in a large number of studies on biomedical NER, since it applied the advantage of sequence labeling. Hsu et al. [17] integrate two different directions of parsing CRF models, backward and forward, and selected a composite output score based on two results. The most recent study on using CRF is [18], which constructs a large dictionary from unlabeled data via a feature generalization method and combines the dictionary with a CRF tagger to improve the base CRF tagger performance.

All of these methods are based on a BIO tagging format in which each example in the dataset is classified either at the beginning, inside or outside of a named entity. Thus, the methods, especially SVM due to its kernel nature, may suffer from the multi-label classification problem. In our previous work [20], we solve the name boundary problem of named entities, which is discussed in [11], by introducing our BFSM algorithm that builds a Bayesian-Finite State Automaton Model. The BFSM model analyzes the sentence structure and selects the candidate of a bio named entity, and then a classifier is employed to classify a candidate word as “true” or “false”.

3. ACTIVE CO-TRAINING ALGORITHM FOR BIOMEDICAL NAMED-ENTITY RECOGNITION

This paper proposes a new semi-supervised learning algorithm—the active co-training algorithm (ACT). Both labeled and unlabeled data are used to train classifiers of the ACT. The training data is refreshed by a set of useful examples from the unlabeled data in each round. The flowchart of the ACT is shown in Fig. 1. The main advantage of this algorithm exists in its active querying component. In every iteration of the ACT, the active querying component forms a set of useful patterns, which are relatively informative to both classifiers (C_1 and C_2) from the unlabeled data. The flowchart of the active querying component is shown in Fig. 2. To select the informative examples, we designed a new classification problem. In this classification problem, an instance of the unlabeled data is classified as either informative or non-informative from both views (C_1 U C_2) and the informative one is added to the useful set to retrain the classifiers. Therefore, in each round of the ACT, the classifiers learn about useful patterns and their performance is improved cooperatively as the round is increased. The ACT terminates when a prede-

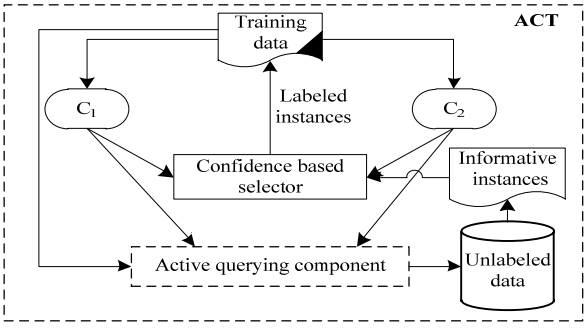

Fig. 1. Flowchart of the active co-training algorithm

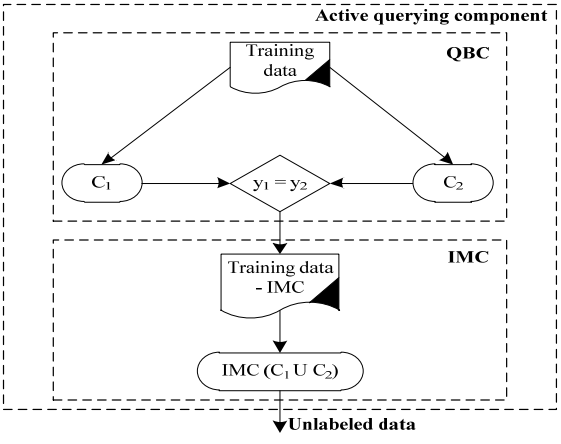

Fig. 2. Flowchart of the active querying component, where y_1 and y_2

Fig. 2. Flowchart of the active querying component, where y1 and y2 are predicted class

fined threshold, such as the number of rounds or a performance measurement, is met. The output of the algorithm will then be two accurate classification models, which can be combined to make a final decision on a test instance. All steps composing the algorithm will be described in the subsequent sections.

3.1 Data Preprocessing and Feature Processing for Biomedical NER Tasks

Preprocessing where text data is cleaned and processed via Natural Language Processing (NLP) is a preparatory task for feature processing, which extracts and converts different types of feature sets.

In preprocessing, the text data is cleaned by removing non-informative characters and replacing special characters with corresponding strings. The text data is then processed through the NLP task to have its sentences parsed and Parts-Of-Speech (POS) tagged. The sentence parser detects sentence boundaries in biomedical text data, and the POS tagger then annotates each token in a sentence with POS tags based on their context. Finally, we applied the BFSM model from our previous study [20] to the POS tagged data to assemble candidate tokens made up of

Table 1. The active co-training algorithm

```
Given:  A small set L of training examples;
        A large set U of unlabeled examples;
        Two sets F1 and F2 of features, redundantly sufficient;
        Co-training parameters N, n and p;

for i = 0 to N do // loop for N iterations
    for j = 1 to 2 do
        train_classifier(Cj, Fj, L);
            // training C1, C2 classifiers based on F1, F2 feature
               sets
    end for
        // active querying component starts
    for each example x ∈ L do
        y1 = classify(C1, x);
        y2 = classify(C2, x);
        if y1 = y2 then
            training_data_IMC.add(x, "Non-informative");
        else
            training_data_IMC.add(x, "Informative");
        end if
    end for
    train_classifier(IMC, F1 U F2, training_data_IMC);
    u = query_informative_examples(U, IMC);
        // active querying component ends
    for j = 1 to 2 do
        classify_all(Cj, u); // classify the informative dataset
        pick_top_confident(p, u, "TRUE");
            // pick most confident p number of positive examples
        pick_top_confident(n, u, "FALSE");
            // pick most confident n number of negative examples
    end for
    update_dataset(L);
        // refresh training data L with newly picked examples
end for
```

Table 2. The sample of the regular expressions for orthographic feature extraction

Feature description	Regular expressions
Roman number	[ivxdlcm]+|[IVXDLCM]+
Punctuation	[,\\.;:?!]
Start with dash	"-.*
Nucleotide sequence	[atgcu]+
Number	[0-9]+
Capitalized	[A-Z] [a-z]*
Quote	[\"`']

those parts that are most likely to form a named entity and to eliminate unnecessary tokens. This process enables us to deal with a candidate classification problem where a candidate example is classified as either "true" if it is a named entity or "false" if it is not a named entity.

As in the ACT algorithm shown in Table 1, two independent feature sets (F1, F2), which show one example from two different views as the co-training assumption, need to be designed and extracted. We selected one of the feature sets, F2, as an orthographic feature that was utilized as the best indicator in biomedical NER. There are many studies based on orthographic features [8-20] and we found the best to be [11]. Another set of features, F1, was extracted as a context-based feature set with lexicons, which provides domain knowledge in a consistent manner. Here, we use the lexicons prepared by the authors [21]. The regular expressions that reveal orthographic information are matched to the candidates to give orthographic information. A sample of the regular expressions is shown in Table 2. Analogous to the extraction of orthographic features, the context-based features are generated by matching the candidates through the lexicon. In order to design a scalable feature generation schema, the lexicons are stored in a prefix tree, which is an ordered tree data structure.

3.2 Querying Informative Examples from Unlabeled Data

The active querying component queries examples, which are relatively informative to the current classification models to retrain so that their performance is definitely improved. A Query-By-Committee (QBC) approach of AL considers an example as being informative when the classification models of the committee most disagree on its class decision [5]. In the ACT, both classifiers (C_1 and C_2) form one committee and an example is informative if the classifiers disagree on its class label. Hence, the training dataset (Training data - IMC in Fig. 2) for the Informativeness Metric Classifier (IMC) is constructed in the following way: an example is labeled as informative if the classifiers disagree on its class label or is labeled as non-informative if the classifiers agree on its class label.

However, the IMC is deployed on a feature space constructed by joining the feature spaces of the classifiers C_1 and C_2, since the set of useful instances should be relatively informative to both of the classifiers. Once the IMC learns from the training data, it is used on the unlabeled data to query the informative instances (as shown in Fig. 1).

Finally, the C_1 and C_2 classifiers label the informative instances and augment their training data by choosing a subset of the labeled instances showing the most confidence from their views. These procedures iteratively proceed in the ACT and the ACT exploits the unlabeled data with

the active querying component as its round goes. The proposed ACT algorithm is shown in Table 1.

3.3 Base Classifiers of ACT: The Naïve Bayes Classifier

The ACT can be used as a wrapper of the classification of algorithms that are capable of producing a predicted class with a confidence value. The ACT performs well with classification algorithms which spend a short amount of time training. In this paper, we use the Naïve Bayes classifier as the base classifier of the ACT.

The Naïve Bayes classifier estimates a class-conditional probability by assuming that the attributes are conditionally independent, given a class label. For categorical attributes, we only lead to a counter for each attribute value and for each class, and for continuous attributes we can assume a particular distribution for the values of the continuous attribute or discretize the attribute in pre-processing.

The ACT performs well with the Naïve Bayes classifier since it has a small number of parameters that must be manually tuned, and takes a short amount of time to train.

4. EXPERIMENTAL RESULTS

4.1 Labeled and Unlabeled Data

To conduct experiments showing the efficiency of the proposed integration approach, we used the GENIA v3.02 corpus [22], in which the biomedical named entities are annotated with their semantic labels. The GENIA corpus consists of 2,000 MEDLINE abstracts, 18,546 sentences, 400,000 words, and 528,113 tokens. The reason that we selected the GENIA corpus is that it provides the largest class set for a biomedical NER task. The 36 GENIA corpus classes are taken from the leaf nodes of a top-level taxonomy of 48 classes based on a chemical classification [11].

Exactly 254,139 PUBMED abstracts containing gene and protein keywords, which have been gathered since 2009, were collected as unlabeled data for semi-supervised learning. The abstracts are stored in one-line-abstract format so that the co-training algorithm randomly selects a number within 254,139, which is the line number corresponding to an abstract.

4.2 Experimental Setup

We ran the ACT until no improvement was observed in a round. Once the classifiers were finished learning from training dataset, the classification models were stored and evaluated on the test dataset. We reported average performance measurements of 5 runs in order to provide robust experimental results.

The parameter selection (L, n, p in Tab. 1) for the ACT algorithm is the one important consideration. We started with as little labeled data as possible for the initial seed of ACT to verify the effectiveness of the semi-supervised algorithm. Hence we randomly selected 500 abstracts from the GENIA corpus as the initial labeled seed (L), while the rest of the GENIA corpus was used as the evaluation dataset for the classifiers. We varied parameters n and p during different runs of the ACT.

4.3 Comparison with Traditional Co-Training Algorithm

We compared the performance of the proposed active co-training algorithm with the traditional co-training algorithm [2]. Fig. 3 shows the plots of classification F-measure versus the number of rounds for the three different runs, where the number of examples to be picked from the unlabeled data is given various values and shown on the top of each plot. C_1 and C_2 are classifiers based on context and orthographic features, respectively. Classifier C is a simple linear combination of the C_1 and C_2 classifiers. The linear combination benefits from an efficient co-training algorithm.

As shown in Fig. 3, the classifiers of the active co-training beat the classifiers of the traditional co-training algorithm. The F-measures of the active co-training classifiers are cooperatively

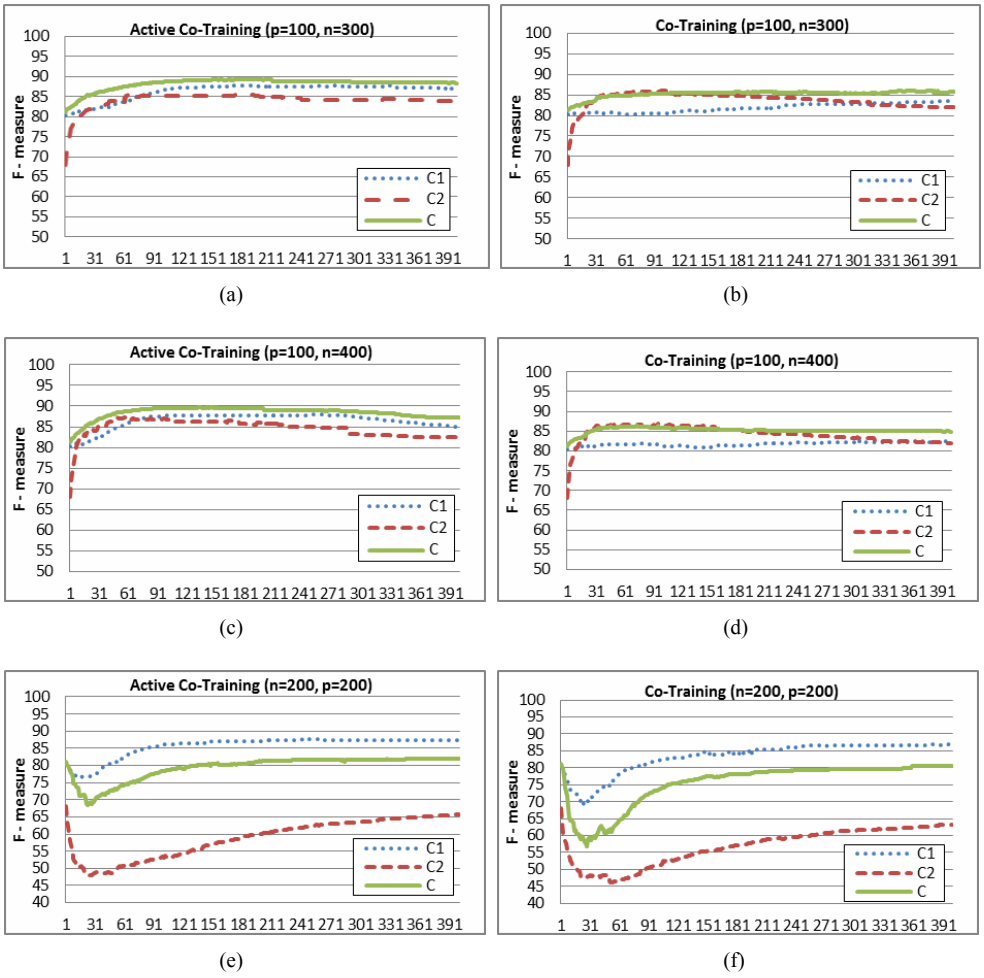

Fig. 3. Comparative results of the active co-training with the co-training [2], where the y axis represents the number of iterations performed. C_1 and C_2 are classifiers based on context and orthographic features, respectively, and form the combined classifier C

Table 3. The best F-measure values compared

Algorithm	Parameter value	Max. F-measure (%)	Round No.
ACT	p=100, n=300	89.28	185
Co-training		85.49	367
ACT	p=100, n=400	89.66	147
Co-training		85.67	170
ACT	p=200, n=200	82.3	328
Co-training		80.55	382

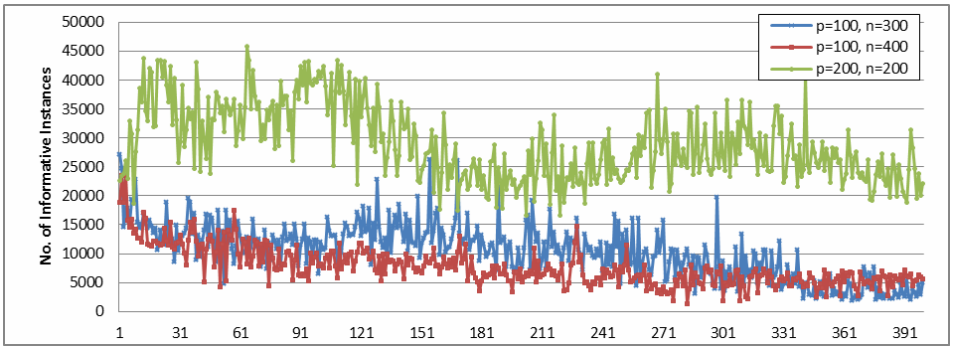

Fig. 4. The number of informative instances (selected by the active querying component of ACT) versus the training round, where the y—axis represents the number of iterations performed

improved, whereas the F-measures of the co-training classifiers are alternately improved. Therefore, the performance of the combined classifier is the highest. In addition, the performances of the active co-training classifiers significantly increased in the earlier rounds. The best value of the F-measure metric is compared in Table 3, where the last column contains the round where the value is achieved.

The plots shown in (e) and (f) illustrate a noisy case where the number of positive examples to be picked is doubled and thus it is possible to select noises or mislabeled examples for training.

The performances of the classifiers are even decreased in the early rounds, but the performances of the active co-training classifiers are not decreased as low as those of the co-training classifiers. This means that active co-training can prevent its classifiers from training on the noisy labeled data by not selecting such instances.

The number of examples selected by the active querying component is shown in Fig. 4, and it can be deduced that when the ACT is working efficiently, the number of informative examples decreases as training rounds increase, because the plot that shows the number of informative instances selected in ACT with parameters p=200 and n=200 is higher than others and there is no decrease as with the others.

4.4 Performance Evaluation

We evaluated the proposed active co-training algorithm with varying sizes of unlabeled data. To show the influence of unlabeled data sizes, we observed three different runs of ACT with the

same training data (500 of the GENIA abstracts), and with the unlabeled data differing in their size. 50,000, 100,000 and 200,000 MEDLINE abstracts were randomly selected for unlabeled data from the entire body of MEDLINE abstracts that we collected.

The F-measure is shown in Fig. 5, where the F-measure value is increased slowly when the amount of unlabeled data is as small as 50,000 abstracts and the F-measure value is increased efficiently, when the unlabeled data is larger than 200,000 abstracts. We were also interested in the number of informative examples selected in the runs, which is demonstrated in Fig. 6. The performance of the classifier is improved in earlier rounds of ACT with large amounts of unlabeled data, even if the number of informative examples selected in this run is identical to that of the other runs (as suggested in Fig. 6). Before performing this experiment, we assumed that the number of informative examples will be as large as the size of unlabeled data from which it is selected, yet it turns out that the number of informative examples is nearly the same and the performance improvement is different. Therefore, it can be inferred that the ACT method yields learning advancement by selecting a more comprehensive pattern from a large amount of unlabeled data.

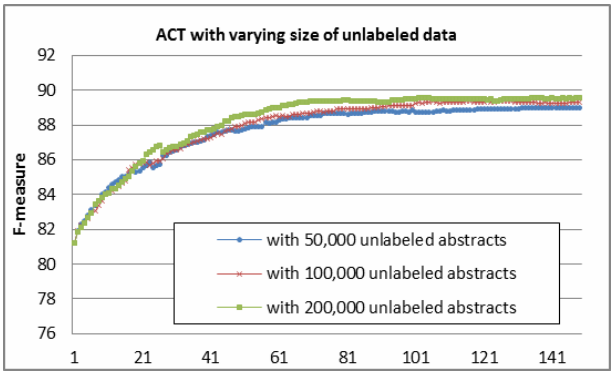

Fig. 5. The F-measure of combined classifiers with varying sizes of training data

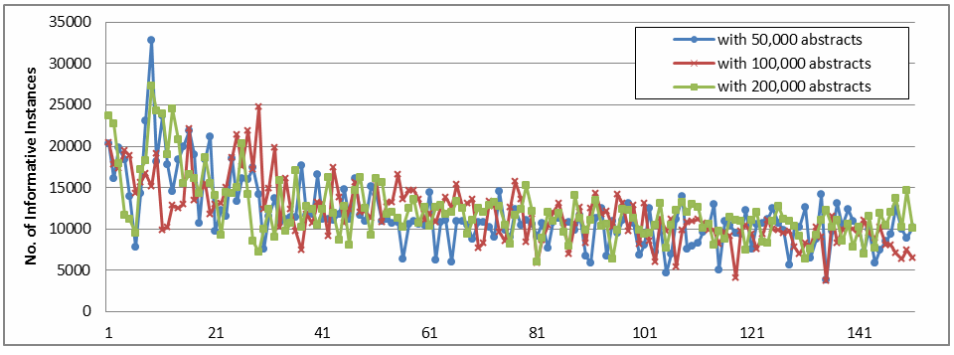

Fig. 6. The number of informative instances selected from varying sizes of unlabeled data, where the y axis represents the number of iterations performed

4.5 Discussion

From the experimental results, we observed a weakness of active co-training that is worth mentioning. As we saw in Fig. 3, learning advancement from one round to the next is different when the parameters n and p are different. The large number of examples to be picked improves the classifier F-measure in a few rounds, but the very large number of examples to be picked degrades the classifier performance. When the number of examples to be picked is small, the improvement of the classifier F-measure is also small, but a decrease performance is not observed. The large number of examples to be picked increases the probability under which noise corresponding to the examples is selected as the correct example, and this fact degrades the performance in future rounds. On the other hand, the small number of examples to be picked slowly increases performance. Therefore, selecting the right parameters is critical in the proposed ACT method.

5. CONCLUSION

Since the amount of biomedical literature sharply increases on open access servers, exploiting unlabeled text with a relatively small labeled corpus has been an active and challenging research topic. This paper proposes an active co-training method for biomedical NER. An active querying component is proposed and utilized to solve the main problem of the traditional co-training method, which is formulated as the informative example selection problem. The active querying component dynamically selects informative examples from unlabeled data so that classification performance is improved by learning from a labeled set of informative examples.

The experimental results show that the proposed active co-training algorithm outperforms the traditional co-training algorithm in terms of f-measure and the number of iterations required to build a better classification model. The one interesting conclusion that we derived is about how active co-training works with varying sizes of unlabeled data. The proposed method tends to efficiently exploit a large amount of unlabeled data by selecting a small number of examples that not only have useful information but also comprise a comprehensive pattern.

However, a parameter-tuning step is still needed in active co-training, and thus our future work should be to find a solution to perform an active semi-supervised training process without any input parameters. We expect that the active co-training algorithm can be applied to other real-world applications where unlabeled data is less expensive to collect, such as the identification of protein-protein interaction.

REFERENCES

[1] H. Dai, Y. Chang, R. T. Tsai, and W. Hsu, “New Challenges for Biological Text-Mining in the Next Decade,” Journal of computer science and technology, 2010, 25(1): 169.

[2] A. Blum, and T. Mitchell, “Combining Labeled Data with Co-Training,” 11th Annual Conference Computational Learning Theory, 1998.

[3] T. Munkhdalai, M. Li, T. Kim, O. Namsrai, S. Jeong, J. Shin, and K.H. Ryu, “Bio Named Entity Recognition based on Co-training Algorithm,” AINA 2012, 2012.

[4] B. Settles, Active learning literature survey, 2010. Univ. of Wisconsin-Madison, Madison, WI, Computer Sciences Tech., Rep.1648.

[5] H.S. Seung, M. Opper, and H. Sompolinsky, "Query by committee," The proc. of the ACM Workshop on Computational Learning Theory, 1992, pp.287-294.

[6] L.J. Gong, and X. Sun, "ATRMiner: A system for Automatic Biomedical Named Entities Recognition," ICNC 2010, 2010, pp.3842-3845.

[7] S. Zhao, "Named Entity Recognition in Biomedical Texts using an HMM Model," The Proc. of the COLING 2004 International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA), 2004.

[8] Z. GuoDong, S. Jian, N. Collier, P. Ruch, and A. Nazarenko, "Exploring Deep Knowledge Resources in Biomedical Name Recognition," COLING 2004 International Joint workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP), 2004, pp.99-102.

[9] K. M. Park, S. H. Kim, D. G. Lee and H. C. Rim, "Boosting Lexical Knowledge for Biomedical Named Entity Recognition," The Proc. of the Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA-2004), 2004, pp.7599.

[10] 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, 2005.

[11] N. Collier, and K. Takeuchi, "Comparison of character-level and part of speech features for name recognition in biomedical texts," Journal of Biomedical Informatics, 2004, pp.423-435.

[12] Z. Ju, J. Wang, and F. Zhu, "Named Entity Recognition From Biomedical Text Using SVM," Bioinformatics and Biomedical Engineering (iCBBE 2011), 2011.

[13] M. Li, T. Munkhdalai, T. Kim, P. Li, and K. H. Ryu, "A Bio-Textmining System for Protein-Protein Interaction Extraction," The proc. of 8th International Conference on Ubiquitous Healthcare, 2011.

[14] J. Finkel, S. Dingare, H. Nguyen, M. Nissim, C. Manning, and G. Sinclair, "Exploiting Context for Biomedical Entity Recognition: From Syntax to the Web," Joint Workshop on Natural Language Processing in Biomedicine and Its Applications at Coling 2004, 2004.

[15] B. Settles, "Biomedical Named Entity Recognition Using Conditional Random Fields and Rich Feature Sets," The proc. of the Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA-2004), 2004.

[16] S. Chan, and W. Lam, "Efficient Methods for Biomedical Named Entity Recognition," Bioinformatics and Bioengineering, 2007.

[17] C. Hsu, Y. Chang, C. Kuo, Y. Lin, H. Huang, and I. Chung, "Integrating high dimensional bidirectional parsing models for gene mention tagging," Bioinformatics, 2008.

[18] Y. Li, H Lin, and Z. Yang, "Integrating rich background knowledge for gene named entity classification and recognition," BMC Bioinformatics, 2009.

[19] L. Yang, and Y. Zhou, "Two-phase Biomedical Named Entity Recognition based on Semi-CRFs," Bio-inspired Computing: Theories and Applications (BIC-TA), 2010.

[20] T. Munkhdalai, M. Li, E. Namsrai, O. Namsrai, and K. H. Ruy, "BFSM: Finite State Machine Learned as Name Boundary Definer for Bio Named Entity Recognition," ICAST 2011, 2011.

[21] L. Tanable, and J. Wilbur, "Tagging Gene and Protein names in Full Text articles," Workshop on Natural language processing in the Biomedical Domain, 2002.

[22] J.D. Kim, T. Ohta, Y. Tateishi, and J. Tsujii, "GENIA corpus-a semantically annotated corpus for bio-text mining," Bioinformatics 2003, 2003, 19(Suppl. 1):18-2.

Tsendsuren Munkhdalai

He received the BS in Computer Science from Mongolian National University, Ulaanbaatar, Mongolia in 2010. In 2012, He received a MS degree at Database and Bioinformatics Laboratory of the Department of Computer Science, Chungbuk National University, Cheongju, South Korea and has joined Ph.D. program at the same laboratory. His research interests include database, bioinformatics and data mining, specially bio data, information extraction and text mining.

Meijing Li

She received a MS degree at Database and Bioinformatics Laboratory, Chungbuk National University, Cheongju, South Korea in 2010. She received BS degree in the School of Information and Computing Science from Dalian University, China, in 2007. Currently, she is a Ph.D. candidate at the same laboratory of the Department of Computer Science, Chungbuk National Univ., Rep. of Korea since 2010. Her major research interests include database, bioinformatics and data mining.

Unil Yun

He received a Ph.D. degree in Computer Science from Texas A&M University, Texas, USA, in 2005. He worked at Multimedia Laboratory, Korea Telecom, from 1997-2002. After receiving the Ph.D. degree he worked as a post-doctoral associate for almost 1year at the Computer Science Department of Texas A&M University. After then, he worked as a senior researcher in Electronics and Telecommunications Research Institute (ETRI). Currently, he is an assistant professor in School of Electrical & Computer Engineering, of Chungbuk National University. His research interests include data mining, database systems, information retrieval, artificial intelligence and digital libraries.

Oyun-Erdene Namsrai

She received a Ph.D. degree in Computer Science from Chungbuk National University, Cheongju, South Korea in 2008. She has been an assistant professor at Mongolian National University since 2010.

Her research interests are in the area of data mining, Bioinformatics, and advanced database systems.

Keun Ho Ryu

He received a Ph.D. degree from Yonsei University, Seoul, Korea. Currently he is a professor with Chungbuk National University and a Leader of Database and bioinformatics laboratory, Cheongju, Korea. He served the Korean Army as ROTC. He was not only a Postdoctoral Researcher and Research Scientist at the University of Arizona, Tucson, but also the Electronic and Telecommunications Research Institute, Daejeon, Korea. His research interests include temporal databases, the spatiotemporal database, stream data processing, knowledge-base information retrieval, database security, data mining, bioinformatics, and biomedical.

Dr. Ryu has served on numerous program committees, including Demonstration Co-Chair of the International Conference on Very large Databases, the PC committee member of APWeb, and the Advanced Information Networking and Applications. He has been a member of the ACM and IEEE since 1983.